

International Journal of Computer Science Issues

**Volume 8, Issue 1, January 2011
ISSN (Online): 1694-0814**

IJCSI proceedings are currently indexed by:



Cogprints Google scholar



IJCSI Publicity Board 2011

Dr. Borislav D Dimitrov

Department of General Practice, Royal College of Surgeons in Ireland
Dublin, Ireland

Dr. Vishal Goyal

Department of Computer Science, Punjabi University
Patiala, India

Mr. Nehinbe Joshua

University of Essex
Colchester, Essex, UK

Mr. Vassilis Papataxiaris

Department of Informatics and Telecommunications
National and Kapodistrian University of Athens, Athens, Greece

EDITORIAL

In this first edition of 2011, we bring forward issues from various dynamic computer science fields ranging from system performance, computer vision, artificial intelligence, software engineering, multimedia, pattern recognition, information retrieval, databases, security and networking among others.

Considering the growing interest of academics worldwide to publish in IJCSI, we invite universities and institutions to partner with us to further encourage open-access publications.

As always we thank all our reviewers for providing constructive comments on papers sent to them for review. This helps enormously in improving the quality of papers published in this issue.

Google Scholar reported a large amount of cited papers published in IJCSI. We will continue to encourage the readers, authors and reviewers and the computer science scientific community and interested authors to continue citing papers published by the journal.

Apart from availability of the full-texts from the journal website, all published papers are deposited in open-access repositories to make access easier and ensure continuous availability of its proceedings free of charge for all researchers.

We are pleased to present IJCSI Volume 8, Issue 1, January 2011 (IJCSI Vol. 8, Issue 1). Out of the 257 paper submissions received, 73 papers were retained for publication. The acceptance rate for this issue is 28.4%.

IJCSI Editorial Board
January 2011 Issue
ISSN (Online): 1694-0814
© IJCSI Publications
www.IJCSI.org

IJCSI Editorial Board 2011

Dr Tristan Vanrullen

Chief Editor

LPL, Laboratoire Parole et Langage - CNRS - Aix en Provence, France

LABRI, Laboratoire Bordelais de Recherche en Informatique - INRIA - Bordeaux, France

LEEE, Laboratoire d'Esthétique et Expérimentations de l'Espace - Université d'Auvergne, France

Dr Constantino Malagôn

Associate Professor

Nebrija University

Spain

Dr Lamia Fourati Chaari

Associate Professor

Multimedia and Informatics Higher Institute in SFAX

Tunisia

Dr Mokhtar Beldjehem

Professor

Sainte-Anne University

Halifax, NS, Canada

Dr Pascal Chatonnay

Assistant Professor

Maître de Conférences

Laboratoire d'Informatique de l'Université de Franche-Comté

Université de Franche-Comté

France

Dr Karim Mohammed Rezaul

Centre for Applied Internet Research (CAIR)

Glyndwr University

Wrexham, United Kingdom

Dr Yee-Ming Chen

Professor

Department of Industrial Engineering and Management

Yuan Ze University

Taiwan

Dr Vishal Goyal

Assistant Professor

Department of Computer Science

Punjabi University

Patiala, India

Dr Dalbir Singh

Faculty of Information Science And Technology
National University of Malaysia
Malaysia

Dr Natarajan Meghanathan

Assistant Professor
REU Program Director
Department of Computer Science
Jackson State University
Jackson, USA

Dr Deepak Laxmi Narasimha

Department of Software Engineering,
Faculty of Computer Science and Information Technology,
University of Malaya,
Kuala Lumpur, Malaysia

Dr. Prabhat K. Mahanti

Professor
Computer Science Department,
University of New Brunswick
Saint John, N.B., E2L 4L5, Canada

Dr Navneet Agrawal

Assistant Professor
Department of ECE,
College of Technology & Engineering,
MPUAT, Udaipur 313001 Rajasthan, India

Dr Panagiotis Michailidis

Division of Computer Science and Mathematics,
University of Western Macedonia,
53100 Florina, Greece

Dr T. V. Prasad

Professor
Department of Computer Science and Engineering,
Lingaya's University
Faridabad, Haryana, India

Dr Saqib Rasool Chaudhry

Wireless Networks and Communication Centre
261 Michael Sterling Building
Brunel University West London, UK, UB8 3PH

Dr Shishir Kumar

Department of Computer Science and Engineering,
Jaypee University of Engineering & Technology
Raghogarh, MP, India

Dr P. K. Suri
Professor
Department of Computer Science & Applications,
Kurukshetra University,
Kurukshetra, India

Dr Paramjeet Singh
Associate Professor
GZS College of Engineering & Technology,
India

Dr Shaveta Rani
Associate Professor
GZS College of Engineering & Technology,
India

Dr. Seema Verma
Associate Professor,
Department Of Electronics,
Banasthali University,
Rajasthan - 304022, India

Dr G. Ganesan
Professor
Department of Mathematics,
Adikavi Nannaya University,
Rajahmundry, A.P, India

Dr A. V. Senthil Kumar
Department of MCA,
Hindusthan College of Arts and Science,
Coimbatore, Tamilnadu, India

Dr Jyoteesh Malhotra
ECE Department,
Guru Nanak Dev University,
Jalandhar, Punjab, India

Dr R. Ponnusamy
Professor
Department of Computer Science & Engineering,
Aarupadai Veedu Institute of Technology,
Vinayaga Missions University, Chennai, Tamilnadu, India.

N. Jaisankar
Assistant Professor
School of Computing Sciences,
VIT University
Vellore, Tamilnadu, India

IJCSI Reviewers Committee 2011

- Mr. Markus Schatten, University of Zagreb, Faculty of Organization and Informatics, Croatia
- Mr. Vassilis Papataxiaris, Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece
- Dr Modestos Stavrakis, University of the Aegean, Greece
- Dr Fadi KHALIL, LAAS -- CNRS Laboratory, France
- Dr Dimitar Trajanov, Faculty of Electrical Engineering and Information technologies, ss. Cyril and Methodius University - Skopje, Macedonia
- Dr Jinping Yuan, College of Information System and Management, National Univ. of Defense Tech., China
- Dr Alexis Lazanas, Ministry of Education, Greece
- Dr Stavroula Mougiakakou, University of Bern, ARTORG Center for Biomedical Engineering Research, Switzerland
- Dr Cyril de Runz, CReSTIC-SIC, IUT de Reims, University of Reims, France
- Mr. Pramodkumar P. Gupta, Dept of Bioinformatics, Dr D Y Patil University, India
- Dr Alireza Fereidunian, School of ECE, University of Tehran, Iran
- Mr. Fred Viezens, Otto-Von-Guericke-University Magdeburg, Germany
- Dr. Richard G. Bush, Lawrence Technological University, United States
- Dr. Ola Osunkoya, Information Security Architect, USA
- Mr. Kotsokostas N. Antonios, TEI Piraeus, Hellas
- Prof Steven Totosy de Zepetnek, U of Halle-Wittenberg & Purdue U & National Sun Yat-sen U, Germany, USA, Taiwan
- Mr. M Arif Siddiqui, Najran University, Saudi Arabia
- Ms. Ilknur Icke, The Graduate Center, City University of New York, USA
- Prof Miroslav Baca, Faculty of Organization and Informatics, University of Zagreb, Croatia
- Dr. Elvia Ruiz Beltrán, Instituto Tecnológico de Aguascalientes, Mexico
- Mr. Moustafa Banbouk, Engineer du Telecom, UAE
- Mr. Kevin P. Monaghan, Wayne State University, Detroit, Michigan, USA
- Ms. Moira Stephens, University of Sydney, Australia
- Ms. Maryam Feily, National Advanced IPv6 Centre of Excellence (NAv6) , Universiti Sains Malaysia (USM), Malaysia
- Dr. Constantine YIALOURIS, Informatics Laboratory Agricultural University of Athens, Greece
- Mrs. Angeles Abella, U. de Montreal, Canada
- Dr. Patrizio Arrigo, CNR ISMAC, Italy
- Mr. Anirban Mukhopadhyay, B.P.Poddar Institute of Management & Technology, India
- Mr. Dinesh Kumar, DAV Institute of Engineering & Technology, India
- Mr. Jorge L. Hernandez-Ardieta, INDRA SISTEMAS / University Carlos III of Madrid, Spain
- Mr. AliReza Shahrestani, University of Malaya (UM), National Advanced IPv6 Centre of Excellence (NAv6), Malaysia
- Mr. Blagoj Ristevski, Faculty of Administration and Information Systems Management - Bitola, Republic of Macedonia
- Mr. Mauricio Egidio Cantão, Department of Computer Science / University of São Paulo, Brazil
- Mr. Jules Ruis, Fractal Consultancy, The Netherlands

- Mr. Mohammad Iftekhar Husain, University at Buffalo, USA
- Dr. Deepak Laxmi Narasimha, Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia
- Dr. Paola Di Maio, DMEM University of Strathclyde, UK
- Dr. Bhanu Pratap Singh, Institute of Instrumentation Engineering, Kurukshetra University Kurukshetra, India
- Mr. Sana Ullah, Inha University, South Korea
- Mr. Cornelis Pieter Pieters, Concast, The Netherlands
- Dr. Amogh Kavimandan, The MathWorks Inc., USA
- Dr. Zhinan Zhou, Samsung Telecommunications America, USA
- Mr. Alberto de Santos Sierra, Universidad Politécnica de Madrid, Spain
- Dr. Md. Atiqur Rahman Ahad, Department of Applied Physics, Electronics & Communication Engineering (APECE), University of Dhaka, Bangladesh
- Dr. Charalampos Bratsas, Lab of Medical Informatics, Medical Faculty, Aristotle University, Thessaloniki, Greece
- Ms. Alexia Dini Kounoudes, Cyprus University of Technology, Cyprus
- Dr. Jorge A. Ruiz-Vanoye, Universidad Juárez Autónoma de Tabasco, Mexico
- Dr. Alejandro Fuentes Penna, Universidad Popular Autónoma del Estado de Puebla, México
- Dr. Ocotlán Díaz-Parra, Universidad Juárez Autónoma de Tabasco, México
- Mrs. Nantia Iakovidou, Aristotle University of Thessaloniki, Greece
- Mr. Vinay Chopra, DAV Institute of Engineering & Technology, Jalandhar
- Ms. Carmen Lastres, Universidad Politécnica de Madrid - Centre for Smart Environments, Spain
- Dr. Sanja Lazarova-Molnar, United Arab Emirates University, UAE
- Mr. Srikrishna Nudurumati, Imaging & Printing Group R&D Hub, Hewlett-Packard, India
- Dr. Olivier Nocent, CReSTIC/SIC, University of Reims, France
- Mr. Burak Cizmeci, Isik University, Turkey
- Dr. Carlos Jaime Barrios Hernandez, LIG (Laboratory Of Informatics of Grenoble), France
- Mr. Md. Rabiul Islam, Rajshahi university of Engineering & Technology (RUET), Bangladesh
- Dr. LAKHOUA Mohamed Najeh, ISSAT - Laboratory of Analysis and Control of Systems, Tunisia
- Dr. Alessandro Lavacchi, Department of Chemistry - University of Firenze, Italy
- Mr. Mungwe, University of Oldenburg, Germany
- Mr. Somnath Tagore, Dr D Y Patil University, India
- Ms. Xueqin Wang, ATCS, USA
- Dr. Borislav D Dimitrov, Department of General Practice, Royal College of Surgeons in Ireland, Dublin, Ireland
- Dr. Fondjo Fotou Franklin, Langston University, USA
- Dr. Vishal Goyal, Department of Computer Science, Punjabi University, Patiala, India
- Mr. Thomas J. Clancy, ACM, United States
- Dr. Ahmed Nabih Zaki Rashed, Dr. in Electronic Engineering, Faculty of Electronic Engineering, menouf 32951, Electronics and Electrical Communication Engineering Department, Menoufia university, EGYPT, EGYPT
- Dr. Rushed Kanawati, LIPN, France
- Mr. Koteshwar Rao, K G Reddy College Of ENGG.&TECH,CHILKUR, RR DIST.,AP, India
- Mr. M. Nagesh Kumar, Department of Electronics and Communication, J.S.S. research foundation, Mysore University, Mysore-6, India

- Dr. Ibrahim Noha, Grenoble Informatics Laboratory, France
- Mr. Muhammad Yasir Qadri, University of Essex, UK
- Mr. Annadurai .P, KMCPGS, Lawspet, Pondicherry, India, (Aff. Pondicherry University, India)
- Mr. E Munivel , CEDTI (Govt. of India), India
- Dr. Chitra Ganesh Desai, University of Pune, India
- Mr. Syed, Analytical Services & Materials, Inc., USA
- Mrs. Payal N. Raj, Veer South Gujarat University, India
- Mrs. Priti Maheshwary, Maulana Azad National Institute of Technology, Bhopal, India
- Mr. Mahesh Goyani, S.P. University, India, India
- Mr. Vinay Verma, Defence Avionics Research Establishment, DRDO, India
- Dr. George A. Papakostas, Democritus University of Thrace, Greece
- Mr. Abhijit Sanjiv Kulkarni, DARE, DRDO, India
- Mr. Kavi Kumar Khedo, University of Mauritius, Mauritius
- Dr. B. Sivaselvan, Indian [Institute](#) of Information Technology, Design & Manufacturing, Kancheepuram, IIT Madras Campus, India
- Dr. Partha Pratim Bhattacharya, Greater Kolkata College of Engineering and Management, [West Bengal University of Technology](#), India
- Mr. Manish Maheshwari, Makhanlal C University of Journalism & Communication, India
- Dr. Siddhartha Kumar Khaitan, Iowa State University, USA
- Dr. Mandhapati Raju, General Motors Inc, USA
- Dr. M.Iqbal Saripan, Universiti Putra Malaysia, Malaysia
- Mr. Ahmad Shukri Mohd Noor, University Malaysia Terengganu, Malaysia
- Mr. Selvakuberan K, TATA Consultancy Services, India
- Dr. Smita Rajpal, Institute of Technology and Management, Gurgaon, India
- Mr. Rakesh Kachroo, Tata Consultancy Services, India
- Mr. Raman Kumar, National Institute of Technology, Jalandhar, Punjab., India
- Mr. Nitesh Sureja, S.P.University, India
- Dr. M. Emre Celebi, Louisiana State University, Shreveport, USA
- Dr. Aung Kyaw Oo, Defence Services Academy, Myanmar
- Mr. Sanjay P. Patel, Sankalchand Patel College of Engineering, Visnagar, Gujarat, India
- Dr. Pascal Fallavollita, Queens University, Canada
- Mr. Jitendra Agrawal, Rajiv Gandhi Technological University, Bhopal, MP, India
- Mr. Ismael Rafael Ponce Medellín, Cenidet (Centro Nacional de Investigación y Desarrollo Tecnológico), Mexico
- Mr. Supheakmungkol SARIN, Waseda University, Japan
- Mr. Shoukat Ullah, Govt. Post Graduate College Bannu, Pakistan
- Dr. Vivian Augustine, Telecom Zimbabwe, Zimbabwe
- Mrs. Mutalli Vatila, Offshore Business Philipines, Philipines
- Mr. Pankaj Kumar, SAMA, India
- Dr. Himanshu Aggarwal, Punjabi University, Patiala, India
- Dr. Vauvert Guillaume, Europages, France
- Prof Yee Ming Chen, Department of Industrial Engineering and Management, Yuan Ze University, Taiwan
- Dr. Constantino Malagón, Nebrija University, Spain
- Prof Kanwalvir Singh Dhindsa, B.B.S.B.Engg. College, Fatehgarh Sahib (Punjab), India

- Mr. Angkoon Phinyomark, Prince of Singkla University, Thailand
- Ms. Nital H. Mistry, Veer Narmad South Gujarat University, Surat, India
- Dr. M.R.Sumalatha, Anna University, India
- Mr. Somesh Kumar Dewangan, Disha Institute of Management and Technology, India
- Mr. Raman Maini, Punjabi University, Patiala(Punjab)-147002, India
- Dr. Abdelkader Outtagarts, Alcatel-Lucent Bell-Labs, France
- Prof Dr. Abdul Wahid, AKG Engg. College, Ghaziabad, India
- Mr. Prabu Mohandas, Anna University/Adhiyamaan College of Engineering, india
- Dr. Manish Kumar Jindal, Panjab University Regional Centre, Muktsar, India
- Prof Mydhili K Nair, M S Ramaiah Institute of Technnology, Bangalore, India
- Dr. C. Suresh Gnana Dhas, VelTech MultiTech Dr.Rangarajan Dr.Sagunthala Engineering College,Chennai,Tamilnadu, India
- Prof Akash Rajak, Krishna Institute of Engineering and Technology, Ghaziabad, India
- Mr. Ajay Kumar Shrivastava, Krishna Institute of Engineering & Technology, Ghaziabad, India
- Mr. Deo Prakash, SMVD University, Kakryal(J&K), India
- Dr. Vu Thanh Nguyen, University of Information Technology HoChiMinh City, VietNam
- Prof Deo Prakash, SMVD University (A **Technical** University open on I.I.T. Pattern) Kakryal (J&K), India
- Dr. Navneet Agrawal, Dept. of ECE, College of Technology & Engineering, MPUAT, Udaipur 313001 Rajasthan, India
- Mr. Sufal Das, Sikkim Manipal Institute of Technology, India
- Mr. Anil Kumar, Sikkim Manipal Institute of Technology, India
- Dr. B. Prasanalakshmi, King Saud University, Saudi Arabia.
- Dr. K D Verma, S.V. (P.G.) College, Aligarh, India
- Mr. Mohd Nazri Ismail, System and Networking Department, University of Kuala Lumpur (UniKL), Malaysia
- Dr. Nguyen Tuan Dang, University of Information Technology, Vietnam National University Ho Chi Minh city, Vietnam
- Dr. Abdul Aziz, University of Central Punjab, Pakistan
- Dr. P. Vasudeva Reddy, Andhra University, India
- Mrs. Savvas A. Chatzichristofis, Democritus University of Thrace, Greece
- Mr. Marcio Dorn, Federal University of Rio Grande do Sul - UFRGS Institute of Informatics, Brazil
- Mr. Luca Mazzola, University of Lugano, Switzerland
- Mr. Nadeem Mahmood, Department of Computer Science, University of Karachi, Pakistan
- Mr. Hafeez Ullah Amin, Kohat University of Science & Technology, Pakistan
- Dr. Professor Vikram Singh, Ch. Devi Lal University, Sirsa (Haryana), India
- Mr. M. Azath, Calicut/Mets School of Enginerring, India
- Dr. J. Hanumanthappa, DoS in CS, University of Mysore, India
- Dr. Shahanawaj Ahamad, Department of Computer Science, King Saud University, Saudi Arabia
- Dr. K. Duraiswamy, K. S. Rangasamy College of Technology, India
- Prof. Dr Mazlina Esa, Universiti Teknologi Malaysia, Malaysia
- Dr. P. Vasant, Power Control Optimization (Global), Malaysia
- Dr. Taner Tuncer, Firat University, Turkey
- Dr. Norrozila Sulaiman, University Malaysia Pahang, Malaysia
- Prof. S K Gupta, BCET, Guradspur, India

- Dr. Latha Parameswaran, Amrita Vishwa Vidyapeetham, India
- Mr. M. Azath, Anna University, India
- Dr. P. Suresh Varma, Adikavi Nannaya University, India
- Prof. V. N. Kamalesh, JSS Academy of Technical Education, India
- Dr. D Gunaseelan, Ibri College of Technology, Oman
- Mr. Sanjay Kumar Anand, CDAC, India
- Mr. Akshat Verma, CDAC, India
- Mrs. Fazeela Tunnisa, Najran University, Kingdom of Saudi Arabia
- Mr. Hasan Asil, Islamic Azad University Tabriz Branch (Azarshahr), Iran
- Prof. Dr Sajal Kabiraj, Fr. C Rodrigues Institute of Management **Studies** (Affiliated to University of Mumbai, India), India
- Mr. Syed Fawad Mustafa, GAC Center, Shandong University, China
- Dr. Natarajan Meghanathan, Jackson State University, Jackson, MS, USA
- Prof. Selvakani Kandeeban, Francis Xavier Engineering College, India
- Mr. Tohid Sedghi, Urmia University, Iran
- Dr. S. Sasikumar, PSNA College of Engg and Tech, Dindigul, India
- Dr. Anupam Shukla, Indian Institute of Information Technology and Management Gwalior, India
- Mr. Rahul Kala, Indian Institute of Inforamtion Technology and Management Gwalior, India
- Dr. A V Nikolov, National University of Lesotho, Lesotho
- Mr. Kamal Sarkar, Department of Computer Science and Engineering, Jadavpur University, India
- Dr. Mokhled S. AlTarawneh, Computer Engineering Dept., Faculty of Engineering, Mutah University, Jordan, Jordan
- Prof. Sattar J Aboud, Iraqi Council of Representatives, Iraq-Baghdad
- Dr. Prasant Kumar Pattnaik, Department of CSE, KIST, India
- Dr. Mohammed Amoon, King Saud University, Saudi Arabia
- Dr. Tsvetanka Georgieva, Department of Information Technologies, St. Cyril and St. Methodius University of Veliko Tarnovo, Bulgaria
- Dr. Eva Volna, University of Ostrava, Czech Republic
- Mr. Ujjal Marjit, University of Kalyani, West-Bengal, India
- Dr. Prasant Kumar Pattnaik, KIST,Bhubaneswar,India, India
- Dr. Guezouri Mustapha, Department of Electronics, Faculty of Electrical Engineering, University of Science and Technology (USTO), Oran, Algeria
- Mr. Maniyar Shiraz Ahmed, Najran University, Najran, Saudi Arabia
- Dr. Sreedhar Reddy, JNTU, SSIETW, Hyderabad, India
- Mr. Bala Dhandayuthapani Veerasamy, Mekelle University, Ethiopia
- Mr. Arash Habibi Lashkari, University of Malaya (UM), Malaysia
- Mr. Rajesh Prasad, LDC Institute of Technical Studies, Allahabad, India
- Ms. Habib Izadkhah, Tabriz University, Iran
- Dr. Lokesh Kumar Sharma, Chhattisgarh Swami Vivekanand Technical University Bhilai, India
- Mr. Kuldeep Yadav, IIIT Delhi, India
- Dr. Naoufel Kraiem, Institut Superieur d'Informatique, Tunisia
- Prof. Frank Ortmeier, Otto-von-Guericke-Universitaet Magdeburg, Germany
- Mr. Ashraf Aljammal, USM, Malaysia
- Mrs. Amandeep Kaur, Department of Computer Science, Punjabi University, Patiala, Punjab, India
- Mr. Babak Basharirad, University Technology of Malaysia, Malaysia

- Mr. Avinash singh, Kiet Ghaziabad, India
- Dr. Miguel Vargas-Lombardo, Technological University of Panama, Panama
- Dr. Tuncay Sevindik, Firat University, Turkey
- Ms. Pavai Kandavelu, Anna University Chennai, India
- Mr. Ravish Khichar, Global Institute of Technology, India
- Mr Aos Alaa Zaidan Ansaef, Multimedia University, Cyberjaya, Malaysia
- Dr. Awadhesh Kumar Sharma, Dept. of CSE, MMM Engg College, Gorakhpur-273010, UP, India
- Mr. Qasim Siddique, FUIEMS, Pakistan
- Dr. Le Hoang Thai, University of Science, Vietnam National University - Ho Chi Minh City, Vietnam
- Dr. Saravanan C, NIT, Durgapur, India
- Dr. Vijay Kumar Mago, DAV College, Jalandhar, India
- Dr. Do Van Nhon, University of Information Technology, Vietnam
- Mr. Georgios Kioumourtzis, University of Patras, Greece
- Mr. Amol D.Potgantwar, SITRC Nasik, India
- Mr. Lesedi Melton Masisi, Council for Scientific and Industrial Research, South Africa
- Dr. Karthik.S, Department of Computer Science & Engineering, SNS College of Technology, India
- Mr. Nafiz Imtiaz Bin Hamid, Department of Electrical and Electronic Engineering, Islamic University of Technology (IUT), Bangladesh
- Mr. Muhammad Imran Khan, Universiti Teknologi PETRONAS, Malaysia
- Dr. Abdul Kareem M. Radhi, Information Engineering - Nahrain University, Iraq
- Dr. Mohd Nazri Ismail, University of Kuala Lumpur, Malaysia
- Dr. Manuj Darbari, BBDNITM, Institute of Technology, A-649, Indira Nagar, Lucknow 226016, India
- Ms. Izerrouken, INP-IRIT, France
- Mr. Nitin Ashokrao Naik, Dept. of Computer Science, Yeshwant Mahavidyalaya, Nanded, India
- Mr. Nikhil Raj, National Institute of Technology, Kurukshetra, India
- Prof. Maher Ben Jemaa, National School of Engineers of Sfax, Tunisia
- Prof. Rajeshwar Singh, BRCM College of Engineering and Technology, Bahal Bhiwani, Haryana, India
- Mr. Gaurav Kumar, Department of Computer Applications, Chitkara Institute of Engineering and Technology, Rajpura, Punjab, India
- Mr. Ajeet Kumar Pandey, Indian Institute of Technology, Kharagpur, India
- Mr. Rajiv Phougat, IBM Corporation, USA
- Mrs. Aysha V, College of Applied Science Pattuvam affiliated with Kannur University, India
- Dr. Debotosh Bhattacharjee, Department of Computer Science and Engineering, Jadavpur University, Kolkata-700032, India
- Dr. Neelam Srivastava, Institute of engineering & Technology, Lucknow, India
- Prof. Sweta Verma, Galgotia's College of Engineering & Technology, Greater Noida, India
- Mr. Harminder Singh BIndra, MIMIT, INDIA
- Dr. Lokesh Kumar Sharma, Chhattisgarh Swami Vivekanand Technical University, Bhilai, India
- Mr. Tarun Kumar, U.P. Technical University/Radha Govinend Engg. College, India
- Mr. Tirthraj Rai, Jawahar Lal Nehru University, New Delhi, India
- Mr. Akhilesh Tiwari, Madhav Institute of Technology & Science, India
- Mr. Dakshina Ranjan Kisku, Dr. B. C. Roy Engineering College, WBUT, India
- Ms. Anu Suneja, Maharshi Markandeshwar University, Mullana, Haryana, India
- Mr. Munish Kumar Jindal, Punjabi University Regional Centre, Jaito (Faridkot), India

- Dr. Ashraf Bany Mohammed, Management Information Systems Department, Faculty of Administrative and Financial Sciences, Petra University, Jordan
- Mrs. Jyoti Jain, R.G.P.V. Bhopal, India
- Dr. Lamia Chaari, SFAX University, Tunisia
- Mr. Akhter Raza Syed, Department of Computer Science, University of Karachi, Pakistan
- Prof. Khubaib Ahmed Qureshi, Information Technology Department, HIMS, Hamdard University, Pakistan
- Prof. Boubker Sbihi, Ecole des Sciences de L'Information, Morocco
- Dr. S. M. Riazul Islam, Inha University, South Korea
- Prof. Lokhande S.N., S.R.T.M.University, Nanded (MH), India
- Dr. Vijay H Mankar, Dept. of Electronics, Govt. Polytechnic, Nagpur, India
- Dr. M. Sreedhar Reddy, JNTU, Hyderabad, SSIETW, India
- Mr. Ojesanmi Olusegun, Ajayi Crowther University, Oyo, Nigeria
- Ms. Mamta Juneja, RBIEBT, PTU, India
- Dr. Ekta Walia Bhullar, Maharishi Markandeshwar University, Mullana Ambala (Haryana), India
- Prof. Chandra Mohan, John Bosco Engineering College, India
- Mr. Nitin A. Naik, Yeshwant Mahavidyalaya, Nanded, India
- Mr. Sunil Kashibarao Nayak, Bahirji Smarak Mahavidyalaya, Basmathnagar Dist-Hingoli., India
- Prof. Rakesh.L, Vijetha Institute of Technology, Bangalore, India
- Mr B. M. Patil, Indian Institute of Technology, Roorkee, Uttarakhand, India
- Mr. Thipendra Pal Singh, Sharda University, K.P. III, Greater Noida, Uttar Pradesh, India
- Prof. Chandra Mohan, John Bosco Engg College, India
- Mr. Hadi Saboohi, University of Malaya - Faculty of Computer Science and Information Technology, Malaysia
- Dr. R. Baskaran, Anna University, India
- Dr. Wichian Sittiprapaporn, Mahasarakham University College of Music, Thailand
- Mr. Lai Khin Wee, Universiti Teknologi Malaysia, Malaysia
- Dr. Kamaljit I. Lakhtaria, Atmiya Institute of Technology, India
- Mrs. Inderpreet Kaur, PTU, Jalandhar, India
- Mr. Iqbaldeep Kaur, PTU / RBIEBT, India
- Mrs. Vasudha Bahl, Maharaja Agrasen Institute of Technology, Delhi, India
- Prof. Vinay Uttamrao Kale, P.R.M. Institute of Technology & Research, Badnera, Amravati, Maharashtra, India
- Mr. Suhas J Manangi, Microsoft, India
- Ms. Anna Kuzio, Adam Mickiewicz University, School of English, Poland
- Mr. Vikas Singla, Malout Institute of Management & Information Technology, Malout, Punjab, India, India
- Dr. Dalbir Singh, Faculty of Information Science And Technology, National University of Malaysia, Malaysia
- Dr. Saurabh Mukherjee, PIM, Jiwaji University, Gwalior, M.P, India
- Dr. Debojyoti Mitra, Sir Padampat Singhania University, India
- Prof. Rachit Garg, Department of Computer Science, L K College, India
- Dr. Arun Kumar Gupta, M.S. College, Saharanpur, India
- Dr. Todor Todorov, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria

- Mr. Akhter Raza Syed, University of Karachi, Pakistan
- Mrs. Manjula K A, Kannur University, India
- Prof. M. Saleem Babu, Department of Computer Science and Engineering, Vel Tech University, Chennai, India
- Dr. Rajesh Kumar Tiwari, GLA Institute of Technology, India
- Dr. V. Nagarajan, SMVEC, Pondicherry university, India
- Mr. Rakesh Kumar, Indian Institute of Technology Roorkee, India
- Prof. Amit Verma, PTU/RBIEBT, India
- Mr. Sohan Purohit, University of Massachusetts Lowell, USA
- Mr. Anand Kumar, AMC Engineering College, Bangalore, India
- Dr. Samir Abdelrahman, Computer Science Department, Cairo University, Egypt
- Dr. Rama Prasad V Vaddella, Sree Vidyanikethan Engineering College, India
- Prof. Jyoti Prakash Singh, Academy of Technology, India
- Mr. Peyman Taher, Oklahoma State University, USA
- Dr. S Srinivasan, PDM College of Engineering, India
- Mr. Muhammad Zakarya, CIIT, Pakistan
- Mr. Williamjeet Singh, Chitkara Institute of Engineering and Technology, India
- Mr. G.Jeyakumar, Amrita School of Engineering, India
- Mr. Harmunish Taneja, Maharishi Markandeshwar University, Mullana, Ambala, Haryana, India
- Dr. Sin-Ban Ho, Faculty of IT, Multimedia University, Malaysia
- Mrs. Doreen Hephzibah Miriam, Anna University, Chennai, India
- Mrs. Mitu Dhull, GNKITMS Yamuna Nagar Haryana, India
- Mr. Neetesh Gupta, Technocrats Inst. of Technology, Bhopal, India
- Ms. A. Lavanya, Manipal University, Karnataka, India
- Ms. D. Pravallika, Manipal University, Karnataka, India
- Prof. Ashutosh Kumar Dubey, Assistant Professor, India
- Mr. Ranjit Singh, Apeejay Institute of Management, Jalandhar, India
- Mr. Prasad S.Halgaonkar, MIT, Pune University, India
- Mr. Anand Sharma, MITS, Lakshmangarh, Sikar (Rajasthan), India
- Mr. Amit Kumar, Jaypee University of Engineering and Technology, India
- Prof. Vasavi Bande, Computer Science and Engneering, Hyderabad Institute of Technology and Management, India
- Dr. Jagdish Lal Raheja, Central Electronics Engineering Research Institute, India
- Mr G. Appasami, Dept. of CSE, Dr. Pauls Engineering College, Anna University - Chennai, India
- Mr Vimal Mishra, U.P. Technical Education, Allahabad, India
- Dr. Arti Arya, PES School of Engineering, Bangalore (under VTU, Belgaum, Karnataka), India
- Mr. Pawan Jindal, J.U.E.T. Guna, M.P., India
- Prof. Santhosh.P.Mathew, Saintgits College of Engineering, Kottayam, India
- Dr. P. K. Suri, Department of Computer Science & Applications, Kurukshetra University, Kurukshetra, India
- Dr. Syed Akhter Hossain, Daffodil International University, Bangladesh
- Mr. Nasim Qaisar, Federal Urdu Univetrsty of Arts , Science and Technology, Pakistan
- Mr. Mohit Jain, Maharaja Surajmal Institute of Technology (Affiliated to Guru Gobind Singh Indraprastha University, New Delhi), India
- Dr. Shaveta Rani, GZS College of Engineering & Technology, India

- Dr. Paramjeet Singh, GZS College of Engineering & Technology, India
- Prof. T Venkat Narayana Rao, Department of CSE, Hyderabad Institute of Technology and Management , India
- Mr. Vikas Gupta, CDLM Government Engineering College, Panniwala Mota, India
- Dr Juan José Martínez Castillo, University of Yacambu, Venezuela
- Mr Kunwar S. Vaisla, Department of Computer Science & Engineering, BCT Kumaon Engineering College, India
- Prof. Manpreet Singh, M. M. Engg. College, M. M. University, Haryana, India
- Mr. Syed Imran, University College Cork, Ireland
- Dr. Namfon Assawamekin, University of the Thai Chamber of Commerce, Thailand
- Dr. Shahaboddin Shamshirband, Islamic Azad University, Iran
- Dr. Mohamed Ali Mahjoub, University of Monastir, Tunisia
- Mr. Adis Medic, Infosys ltd, Bosnia and Herzegovina
- Mr Swarup Roy, Department of Information Technology, North Eastern Hill University, Umshing, Shillong 793022, Meghalaya, India
- Mr. Suresh Kallam, East China University of Technology, Nanchang, China
- Dr. Mohammed Ali Hussain, Sai Madhavi Institute of Science & Technology, Rajahmundry, India
- Mr. Vikas Gupta, Adesh Institute of Engineering & Technology, India
- Dr. Anuraag Awasthi, JV Womens University, Jaipur, India
- Dr. Dr. Mathura Prasad Thapliyal, Department of Computer Science, HNB Garhwal University (Central University), Srinagar (Garhwal), India
- Mr. Md. Rajibul Islam, Ibnu Sina Institute, University Technology Malaysia, Malaysia
- Mr. Adnan Qureshi, University of Jinan, Shandong, P.R.China, P.R.China
- Dr. Jatinderkumar R. Saini, Narmada College of Computer Application, India
- Mr. Mueen Uddin, Universiti Teknologi Malaysia, Malaysia
- Mr. S. Albert Alexander, Kongu Engineering College, India
- Dr. Shaidah Jusoh, Zarqa Private University, Jordan
- Dr. Dushmanta Mallick, KMBB College of Engineering and Technology, India
- Mr. Santhosh Krishna B.V, Hindustan University, India
- Dr. Tariq Ahamad Ahanger, Kausar College Of Computer Sciences, India
- Dr. Chi Lin, Dalian University of Technology, China
- Prof. VIJENDRA BABU.D, ECE Department, Aarupadai Veedu Institute of Technology, Vinayaka Missions University, India
- Mr. Raj Gaurang Tiwari, Gautam Budh Technical University, India
- Mrs. Jeysree J, SRM University, India
- Dr. C S Reddy, VIT University, India
- Mr. Amit Wason, Rayat-Bahra Institute of Engineering & Bio-Technology, Kharar, India
- Mr. Yousef Naeemi, Mehr Alborz University, Iran
- Mr. Muhammad Shuaib Qureshi, Iqra National University, Peshawar, Pakistan, Pakistan
- Dr Pranam Paul, Narula Institute of Technology Agarpara. Kolkata: 700109; West Bengal, India
- Dr. G. M. Nasira, Sasurie College of Engineering, (Affiliated to Anna University of Technology Coimbatore), India
- Dr. Manasawee Kaenampornpan, Mahasarakham University, Thailand
- Mrs. Iti Mathur, Banasthali University, India
- Mr. Avanish Kumar Singh, RRIMT, NH-24, B.K.T., Lucknow, U.P., India

- Dr. Panagiotis Michailidis, University of Western Macedonia, Greece
- Mr. Amir Seyed Danesh, University of Malaya, Malaysia
- Dr. Terry Walcott, E-Promag Consultancy Group, United Kingdom
- Mr. Farhat Amine, High Institute of Management of Tunis, Tunisia
- Mr. Ali Waqar Azim, COMSATS Institute of Information Technology, Pakistan
- Mr. Zeeshan Qamar, COMSATS Institute of Information Technology, Pakistan
- Dr. Samsudin Wahab, MARA University of Technology, Malaysia
- Mr. Ashikali M. Hasan, CelNet Security, India

TABLE OF CONTENTS

| | |
|---|-----------------|
| 1. ADaPPT: Enterprise Architecture Thinking for Information Systems Development Hanifa Shah and Paul Golder | Pg 1-7 |
| 2. Modular Design of Call Control Layer in Telephony Software Ilija Basicevic | Pg 8-12 |
| 3. Generating a Performance Stochastic Model from UML Specifications Ihab Sbeity, Leonardo Brenner and Mohamed Dbouk | Pg 13-21 |
| 4. Tetris Agent Optimization Using Harmony Search Algorithm Victor II M. Romero, Leonel L. Tomes and John Paul T. Yusiong | Pg 22-31 |
| 5. Scenario-Based Software Architecture for Designing Connectors Framework in Distributed System Hamid Mcheick, Yan Qi and Hafedh Mili | Pg 32-41 |
| 6. Improved FCM algorithm for Clustering on Web Usage Mining K. Suresh, R. Madana Mohana and A. Rama Mohan Reddy | Pg 42-45 |
| 7. Causally Ordered Delivery Protocol for Overlapping Multicast Groups in Broker-based Sensor Networks Chayoung Kim and Jinho Ahn | Pg 46-54 |
| 8. Mobile Agent PLM Architecture for extended enterprise Abdelhak Boulaalam, El Habib Nfaoui and Omar El Beqqali | Pg 55-61 |
| 9. Load Balancing using High Performance Computing Cluster Programming Sumit Srivastava, Pankaj Dadheeck and Mahender Kumar Beniwal | Pg 62-66 |
| 10. Creating Multiuser Web3D Applications Embedded in Web Pages Xandre Chourio, Francisco Luengo and Gerardo Pirela | Pg 67-73 |
| 11. A user-centric PKI based-protocol to manage FC 2 digital identities Samia Bouzefrane, Khaled Garri and Pascal Thoniel | Pg 74-80 |
| 12. Visual Attention Shift based on Image Segmentation Using Neurodynamic System Lijuan Duan, Chunpeng Wu, Faming Fang, Jun Miao, Yuanhua Qiao and Jian Li | Pg 81-86 |
| 13. Technology Marketing using PCA , SOM, and STP Strategy Modeling Sunghae Jun | Pg 87-92 |
| 14. Redesigning the user interface of handwriting recognition system for preschool children Mohd Nizam SAAD, Abd. Hadi ABD. RAZAK, Azman YASIN and Nur Sukinah AZIZ | Pg 93-98 |

| | |
|---|------------|
| 15. Texture Classification Using an Invariant Texture Representation and a Tree Matching Kernel Somkid Soottitantawat and Surapong Auwatanamongkol | Pg 99-106 |
| 16. Performance Measurement of Some Mobile Ad Hoc Network Routing Protocols Ahmed A. Radwan, Tarek M. Mahmoud and Essam H. Houssein | Pg 107-112 |
| 17. Evolution of Computer Virus Concealment and Anti-Virus Techniques: A Short Survey Babak Bashari Rad, Maslin Masrom and Suhaimi Ibrahim | Pg 113-121 |
| 18. Performance Evaluation of Two Node Tandem Communication Network with Dynamic Bandwidth Allocation having Two Stage Direct Bulk Arrivals K. Srinivasa Rao, Kuda Nageswara Rao and P. Srinivasa Rao | Pg 112-130 |
| 19. Fast Scalar Multiplication in ECC Using The Multi Base Number System G. N. Purohit and Asmita Singh Rawat | Pg 131-137 |
| 20. Accurate methods of calculating the Coronary Sinus Pressure plateau Loay Alzubaidi | Pg 138-140 |
| | |
| 22. QUESEM: Towards building a Meta Search Service utilizing Query Semantics Neelam Duhan and Ashok Kale Sharma | Pg 145-154 |
| 23. Multihop Routing In Self-Organizing Wireless Sensor Networks Rajashree V. Biradar, S. R. Sawant, R. R. Mudholkar and V. C. Patil | Pg 155-164 |
| 24. Service-Oriented Architecture and model for GIS Health Management: Case of cancer in Morocco Zouiten Mohammed, Harti Mostafa and Nejjari Chakib | Pg 165-170 |
| 25. Token Ring Algorithm To Achieve Mutual Exclusion In Distributed System - A Centralized Approach Sandipan Basu | Pg 171-175 |
| 26. University Grades System Application using Dynamic Data Structure Nael Hirzallah, Dead Al-Halabi and Baydaa Al-Hamadani | Pg 176-181 |
| 27. A Paper Presentation on Software Development Automation by Computer Aided Software Engineering (CASE) Nishant Dubey | Pg 182-184 |
| 28. Simulation and Performance Analysis of Adaptive Filtering Algorithms in Noise Cancellation Lilatul Ferdouse, Nasrin Akhter, Tamanna Haque Nipa and Fariha Tasmin Jaigirdar | Pg 185-192 |

| | |
|---|-------------------|
| 29. A Novel Architecture for Real Time Implementation of Edge Detectors on FPGA Sudeep K C and Jharna Majumdar | Pg 193-202 |
| 30. FARS: Fuzzy Ant based Recommender System for Web Users Shiva Nadi, Mohammad H. Saraee, Mohammad Davarpanah Jazi and Ayoub Bagheri | Pg 203-209 |
| 31. Fusion Of Facial Parts And Lip For Recognition Using Modular Neural Network Anupam Tarsauliya, Shoureya Kant, Saurabh Tripathi, Ritu Tiwari and Anupam Shukla | Pg 210-216 |
| 32. Health Care Implementation by Means of Smart Cards Magdy E. Elhennawy, Mohamed Amer and Ashraf Abd El-Hafeez | Pg 217-225 |
| 33. Enhancing Decision Making Using Intelligent System Solution Sushanta Kumar Panigrahi, Amaresh Sahu and Sabyasachi Pattnaik | Pg 226-231 |
| 34. Teaching Software Engineering: Problems and Suggestions Osama Shata | Pg 232-235 |
| 35. Software Vulnerabilities, Banking Threats, Botnets and Malware Self-Protection Technologies Wajeb Gharibi and Abdulrahman Mirza | Pg 236-241 |
| 36. A Remote Robotic Laboratory Experiment Platform with Error Management Chadi Rimani | Pg 242-248 |
| 37. Performance of Distributed System Abdellah Ezzati, Abderrahim Beni hssane and Moulay Lahcen Hasnaoui | Pg 249-252 |
| 38. A Generalized Framework for Energy Conservation in Wireless Sensor Network V. S. Anita Sofia and S. Arockiasamy | Pg 253-256 |
| 39. Platform for Assessing Strategic Alignment Using Enterprise Architecture: Application to E-Government Process Assessment Kaoutar Elhari and Bouchaib Bounabat | Pg 257-264 |
| 40. Towards an Adaptive competency-based learning System using assessment Noureddine El Faddouli, Brahim El Falaki, Mohammed Khalidi Idrissi and Samir Bennani | Pg 265-274 |
| 41. An Efficient Searching and an Optimized Cache Coherence handling Scheme on DSR Routing Protocol for MANETS Rajneesh Kumar Gujral and Anil Kapil | Pg 275-283 |
| 42. Fingerprint Matching Using Hierarchical Level Features D. Bennet and S. Arumuga Perumal | Pg 284-288 |
| 43. Computation of Fifth Degree of Spline Function Model by Using C++ Programming Faraidun K. HamaSalh, Alan Anwer Abdulla and Khanda M. Qadir | Pg 289-294 |

| | |
|---|-------------------|
| 44. A Method for Designing an Operating System for Plug and Play Bootstrap Loader USB Drive T. Jebarajan and K. Siva Sankar | Pg 295-301 |
| 45. A Survey of Connectionless Network Service Protocols for Mobile AdHoc Networks Maniyar Shiraz Ahmed, Syed Abdul Sattar and Fazeela Tunnisa | Pg 302-309 |
| 46. An authorization Framework for Grid Security using GT4 Debabrata Singh, Bhupendra Gupta, B. M. Acharya and Sarbeswar Hota | Pg 310-314 |
| 47. Anti-Trust Rank: Fighting Web Spam Jyoti Pruthi and Ela Kumar | Pg 315-319 |
| 48. Triangular Pyramid Framework For Enhanced Object Relational Dynamic Data Model for GIS Barkha Bahl, Navin Rajpal and Vandana Sharma | Pg 320-328 |
| 49. Simulation and performance Analysis of a Novel Model for Short Range Underwater Acoustic communication Channel Using Ray Tracing Method in Turbulent Shallow Water Regions of the Persian Gulf Mohammad Javad Dargahi, Abdollah Doosti Aref and Ahmad Khademzadeh | Pg 329-337 |
| 50. Classification of Electrocardiogram Signals With Extreme Learning Machine and Relevance Vector Machine S. Karpagachelvi, M. Arthanari and M.Sivakumar | Pg 338-345 |
| 51. Negotiation in Multi-Agent System using Partial-Order Schedule Ritu Sindhu, Abdul Wahid and G. N. Purohit | Pg 346-355 |
| 52. Hybrid CHAID a key for MUSTAS Framework in Educational Data Mining G.Paul Suthan and Santhosh Baboo | Pg 356-360 |
| 53. Feature-Level based Video Fusion for Object Detection Anjali Malviya and S. G. Bhirud | Pg 361-366 |
| 54. A Fuzzy Based Stable Routing Algorithm for MANET Arash Dana and Mohamad Hadi Babaei | Pg 367-371 |
| 55. Enhanced Digital Watermarking Algorithm for Directionally Selective and Shift Invariant Analysis B. Benita | Pg 372-375 |
| 56. Antenna selection for performance enhancement of MIMO Technology in Wireless LAN Rathnakar Achary, V. Vaityanathan, Pethur Raj Chellaih and S. Nagarajan | Pg 376-381 |
| 57. A Fuzzy-Ontology Based Information Retrieval System for Relevant Feedback Comfort T. Akinribido, Babajide S. Afolabi, Bernard I. Akhigbe and Ifiok J. Udo | Pg 382-389 |

| | |
|---|------------|
| 58. Distortion Analysis Of Tamil Language Characters Recognition Gowri N. and R. Bhaskaran | Pg 390-394 |
| 59. Implementation of Clustering Through Machine Learning Tool Sree Ram Nimmagadda, Phaneendra Kanakamedala and Vijay Bashkarreddy Yaramala | Pg 395-401 |
| 60. Improving Performance on WWW using Intelligent Predictive Caching for Web Proxy Servers J. B. Patil and B. V. Pawar | Pg 402-408 |
| 61. An Efficient Approach to Prune Mined Association Rules in Large Databases D. Narmadha, G. NaveenSundar and S. Geetha | Pg 409-415 |
| 62. Implementation of Reduced Power Open Core Protocol Compliant Memory System using VHDL Ramesh Bhakthavatchalu and Deepthy G R | Pg 416-421 |
| 63. A Novel Energy Efficient Mechanism for Secured Routing of Wireless Sensor Network Anupriya Sharma, Paramjeet Rawat, Suraj Malik and Sudhanshu Gupta | Pg 422-428 |
| 64. M-AODV: AODV variant to Improve Quality of Service in MANETs Maamar Sedrati, Bilami Azeddine and Mohamed Benmohamed | Pg 429-436 |
| 65. Optimization of LSE and LMMSE Channel Estimation Algorithms based on CIR Samples and Channel Taps Saqib Saleem | Pg 437-443 |
| 66. Classification rules for Indian Rice diseases A. Nithya and V. Sundaram | Pg 444-448 |
| 67. Predict Success or Failure of Remote Infrastructure Management Satinder Pal Ahuja and Sanjay P. Sood | Pg 449-454 |
| 68. MRI Mammogram Image Segmentation using NCut method and Genetic Algorithm with partial filters Pitchumani Angayarkanni | Pg 455-459 |
| 69. A Schematic Technique Using Data type Preserving Encryption to Boost Data Warehouse Security M. Sreedhar Reddy, M. Rajitha Reddy, R. Viswanath, G. V. Chalam, Rajya Laxmi and Md. Arif Rizwan | Pg 460-465 |
| 70. QEMPAR: QoS and Energy Aware Multi-Path Routing Algorithm for Real-Time Applications in Wireless Sensor Networks Saeed Rasouli Heikalabad, Hossein Rasouli, Farhad Nematy and Naeim Rahmani | Pg 466-471 |

- 71. DPCC: Dynamic Predictive Congestion Control in Wireless Sensor Networks** Pg 472-477
Saeed Rasouli Heikalabad, Ali Ghaffari, Mir Abolgasem Hadian and Hossein Rasouli
- 72. Off-Line Handwritten Signature Identification Using Rotated Complex Wavelet Filters** Pg 478-482
M. S. Shirdhonkar and Manesh Kokare
- 73. Ternary Tree and Memory-Efficient Huffman Decoding Algorithm** Pg 483-489
Pushpa R. Suri and Madhu Goel

ADaPPT: Enterprise Architecture Thinking for Information Systems Development

Hanifa Shah and Paul Golder

Birmingham City University,
TEE, Millennium Point, Birmingham, B4 7XG, UK

Abstract

Enterprises have architecture: whether it is visible or invisible is another matter. An enterprises' architecture determines the way in which it works to deliver its business objectives and the way in which it can change to continue to meet its evolving business objectives. Enterprise architectural thinking can facilitate effective strategic planning and information systems development. This paper reviews enterprise architecture (EA) and its concepts. It briefly considers EA frameworks. It describes the ADaPPT (Aligning Data, People, Processes and Technology) EA approach as a means to managing organisational complexity and change. Future research directions are discussed.

Keywords: Business Strategies, Enterprise Architecture, Organisational Processes, Enterprise Planning, Technologies Alignment.

1. Introduction

Much progress has been made in recent years in developing structures to describe the enterprise and to facilitate the development of information systems that appropriately complement the strategy of the enterprise. Despite the success of the enterprise architecture approach there are still major problems in achieving organisational change and in driving the re-alignment of IT systems. The complexity of modern organisations in terms of the business, legal and technological environment demands an architectural approach. Businesses are faced with ongoing and continual change to which they must respond in order to ensure success and even survival. The increasingly competitive environment demands a customer-focused approach. All of these factors contribute to the complexity and uncertainty faced by organisations resulting in an inability to be appropriately responsive to both internal and external events. Underlying this complexity and uncertainty is the gap between an organisation's business objectives and its underlying IT infrastructure. There is a need for information that is timely and understood in order to facilitate appropriate analysis and to appreciate the

relevant impacts of decisions made. Organisations are continually faced with the challenge of their IT delivering the business value demanded and responding speedily to the changing business needs.

2. Enterprise Architectural Thinking

An EA approach can help to provide a vehicle for organisational communication. Improving communication and discussion between business and IT staff enabling a shared understanding of the business and its supporting infrastructure that can facilitate improved decision making and more effective deployment of change. The approach provides a basis for standardisation and agreed notations and representations, processes and information become more transparent. Project costs can become more stable and better predicted, the time taken to bring about change either by enhancing current services or by offering new ones can be reduced.

Clearly identifying the key components through an enterprise architecture approach of business processes, information, technology applications and organisation and how these relate to each other facilitates focussing on the appropriate component as required in a particular situation. EA can be used to manage complexity and describe the interdependencies in a usable manner.

EA can facilitate a better return on an organisations investment by providing a means to identify cost saving opportunities, gaps and inconsistencies as well as facilitating the installed systems and applications being exploited. An enterprise architecture approach leads to improved scoping and coordination of programmes and projects.

3. The Challenge of Change

It is commonplace to identify the forces of change to which modern businesses are exposed. It is relevant to discriminate between forces for change that affect the business being carried out and those which affect only the way the business is delivered. A new computer system leaves organisational models unchanged but may change the data models and applications used to support them. A change in market or a merger will change the organisational model itself. The pressures for change on the organisation are such that a process of continual evolution even revolution is affecting all organisations. This means that the construction of an enterprise architecture is not a single event generating a static description of the organisation which thereafter impedes the process of change. On the contrary the continual evolution of the enterprise architecture is a process in parallel with the evolution of the business strategy. The question should be asked how do we architect the business to meet its evolving strategic needs and the answer should lie in the continual evolution of the architecture. The architecture is the interface between the strategic, what the enterprise wants to do, and the operational, what it does.

Strategic change in the organisation can lead to evolutionary changes in the enterprise architecture but may require more radical change. For example the merging of two organisations may require the integration of their existing enterprise architectures into a new common EA. This is a similar problem to the evolving enterprise architecture one but is likely to require more substantive change. For example we may not be able to assume that the concept 'Customer' is exactly the same in the two merging organisations so may need to examine this at some detail in order to achieve successful integration. However in an organisation that is evolving from concrete to virtual trading, the concept of customer may be undergoing equally significant change and the significance of this change may be overlooked in the assumption that it is evolutionary.

An organisation has a business strategy at a particular time; corresponding to this strategy it has (or is in the process of developing) the corresponding enterprise architecture for delivering this strategy. That existing enterprise architecture describes and specifies a number of business processes, data objects and applications which 'operationalises' the architecture. Next the business introduces a new strategy, corresponding to this we have desired enterprise architecture and its corresponding business processes, data objects and applications. The practical problem becomes how do we migrate from one EA to the next? Moreover we would want to know the series of architectures (or roadmap) that would take us

through the required transitional architectures. How do we identify the changes necessary in business processes, data objects, and applications required and how do we manage the transitions.

The following examples serve to highlight what is needed in the management of organisational change through evolving enterprise architectures.

- We need to be able to examine the ontology of concepts - what is a customer?
- We need to be able to identify the dynamics of the elements - how does a customer come into existence, what determines the life of a customer, how is it terminated?
- We need to be able to identify the agents responsible - who authorises the creation of a customer, who determines when a customer is no longer?
- We need to be able to specify the business rules related to the behaviour of customers and agents.

Existing methodologies and tools do not help use with these problems they are mainly focused on the storage and retrieval of data, and the specification of data manipulation processes. An enterprise architectural approach can facilitate this thinking.

Enterprise architecture has been widely adopted as a means to cope with the ever-increasing complexity of organizations and to ensure that the technical resources are appropriately employed and optimized [1]. EA is the fundamental organization of the system, embodied in its elements, their relationships to each other and to the environment and the principles guiding its design and evolution [2], [3]. Enterprise architecture is described as organizing logic for business processes and IT infrastructure, reflecting the integration and standardization requirements of the company's operating model in order to achieve business agility and profitable growth [4]. Currently, there exist a number of professional societies and organizations that are working on the definition and the management of enterprise architecture such as The Open Group, Microsoft, and IBM. Indeed, EA represents much more than IT architecture. It is an integrated and holistic vision of how the business processes across the enterprise, people, information, applications and technologies align to facilitate strategic objectives. EA frameworks identify the scope of EA and decompose various elements of the architecture onto structured layers/levels and elements. Several EA frameworks have been adopted for operational use in many private and governmental organizations.

EA emerged as an idea in 1980 and is embodied in the early EA framework developed by Zachman (1987) [5]. EA has re-emerged as a means to cope with the ever-increasing complexity of organizations. This re-emergence is closely related to the evolution of new business trends and to the evolution of IT, particularly to the advances in Internet technologies. These business trends comprise globalization, mergers and acquisitions, e-commerce, as well as the increasing importance of customer relationship management (CRM) and supply chain management. IT trends, on the other hand, comprise the advances in Internet technologies, hardware platform, application servers, and workflow servers. Due to the increasing importance of EA, certification opportunities in EA are being offered by several companies such as The Open Group and IBM in order to standardize an open method for IT architecture to solve business problems.

An EA approach is beneficial in aligning business and IT resources and in conforming to fundamental principles and common methodologies that govern the entire life cycle of the IS development process. In that sense, architectural frameworks are considered to be a convenient way to support such methodologies, and to separate roles that facilitate and implement these methodologies as needed. Still, there are many organizational and technical EA challenges.

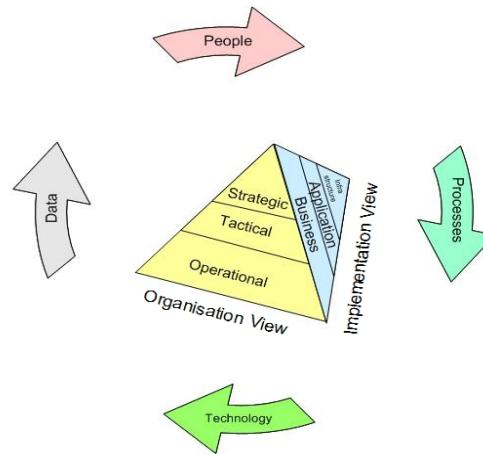
4. Enterprise Architecture Framework

EA frameworks identify the scope of the enterprise architecture and decompose various elements of the architecture onto structured levels and elements [6]. More formally, EA frameworks describe a method for designing IS in terms of a set of building blocks and how these blocks fit together. Several EA frameworks such as ARIS [7] and DODAF [3] have been adopted for operational use in many organizations. For example, the Federal EA [8], has been adopted by the US government as a business-driven framework in order to optimize some strategic areas. These areas include budget allocation, information sharing, performance measurement, and component based architecture. More specifically, EA frameworks contain a list of recommended standards and compliant products that are used to implement the building blocks for an IS. EA frameworks are useful in terms of simplifying architecture development and ensuring complete coverage of the designed solutions through a common terminology. In that sense, these frameworks are language independent by providing generic concepts and common terminology through which different EA stakeholders can communicate without making any assumptions about each others' language. Pragmatically, EA frameworks play a dual role. Firstly, they serve as implementation tools; secondly, they can serve as organisational planning tools.

5. The ADaPPT Approach

Fig. 1 ADaPPT Approach.

ADaPPT was developed in work with organisations using the ALTAR (Achieving Learning Through Action Research) methodology [1]. ADaPPT has four domains (elements¹): people, processes, technology and data.



ADaPPT : Aligning Data, People, Processes and Technology
There are two main views in ADaPPT: a organisational view and an implementation view. The model recognises that everything the enterprise does involves people, processes, technology and data and that these need to be aligned.

*'A process driven by people
consumes resources (technology)
and generates data'*

In the ADaPPT framework the term **people** represents not only individual people but groups within the organisation, departments, sections etc. and also roles such as marketing manager etc. *In as much as they can initiate actions and be responsible for the processes of the organisation. The term agent is also used.*

In ADaPPT **process** means all activity and actions within the organisation. This includes high level business processes – marketing, production etc.; middle level activities - launching new products etc; operational activities - checking an invoice – etc.

¹ We call these elements because just as the ancients believed that every thing was composed of the four basic elements, Fire, Water, Air, Earth. So ADaPPT believes that every business activity combines the basic elements of, People, Process, Data, Technology.

In ADaPPT **technology** includes all services, material and equipment used by the organisation. This includes computer and IT hardware and software, raw materials and processed product.

In ADaPPT **data** means all information both static and dynamic within the organisation such as management targets, performance measures and operational data : customer details etc.

The **organisational view** in ADaPPT recognises that the nature the enterprise varies throughout the organisational hierarchy:

- Strategic
- Tactical
- Operational

As we climb the hierarchy processes become less well defined, soft data becomes more important, the scope of responsibility becomes wider. ADaPPT does not attempt to be a complete framework for all enterprise needs it is focused on the business / technology issues.

The **implementation view** recognises that there is a spectrum from business through application to infrastructure. This can apply equally to each domain. For example data can be viewed at: the business level expressed as E-R and other data models; at the application level expressed as data structure diagrams; at the infrastructure level expressed through allocation of data on storage devices. If we consider the interactions between views we will see that the implementation view and the business views are independent of each other. Thus an operational level business problem will need to be addressed at the business level as a specification of the business problem, at the application level as design of the solution to the problem and at the infrastructure level by provision of software and other resources to implement the solution.

The ADaPPT framework thus has four domains each of which can be described with a three by three matrix of views.

'This is because enterprise planning is a complex process involving many thousands of elements. The ADaPPT approach aims to organise and simplify the process of thinking about and managing these thousands of elements'.

5.1 ADaPPT as Implementation Tool

ADaPPT in common with other EA frameworks provides a comprehensive representation of IS in terms of its building blocks. In this context, ADaPPT relates the necessary IS aspects/dimensions such as business

processes, data, and organization units to different perspectives at certain levels of abstraction. These perspectives rely mainly on the difference in EA stakeholders' views of the architecture that span different level of details. EA frameworks, as components specification tools, encompass the documentation of the architectural layers, architectural domains, architectural models, and architectural artefacts.

Typically, EA frameworks such as ADaPPT are decomposed to three architectural layers, which are business layer, application layer, and technology infrastructure layer [9]. The business layer describes the business entities such as business processes and relevant business information, and how these entities interact with each other to achieve enterprise wide objectives. The application layer determines the data elements and the software applications that support the business layer. The technology infrastructure layer comprises the hardware platforms and the communication infrastructure that supports the applications. Such layers are naturally characterized by information aspects, behavioural aspects, and structural aspects. As organizations consist of several units, the structural aspects determine the static decomposition of these units to several sub-units. The behavioural aspects show behaviour manifested in the sequence of activities and business processes performed to produce the needed services. These units exchange information in order to carry out business tasks. Each layer is naturally composed of several domains that reflect the information, behavioural, and structural aspects of the organizations. These domains specify the architectural aspects such as process architecture, product architecture, information architecture, technical architecture, and application architecture. Indeed, these domains are the means to separate the architectural concerns and reflect the view of different EA stakeholders to the architecture. For example, the process domain, which is a part of the business layer, describes business processes or business functions that offer the products or services to an organization. These architectural domains are typically described and documented by different architectural models such as business process models, value chain diagrams, and organization charts. Architectural models serve as a basis for documenting the different architectures by annotating the artefacts and their inter-relationship that are needed to model an organization from different perspectives. Architectural artefacts represent the necessary constructs and architectural elements such as data, business processes, resources, and events that represent the real world objects needed to design distinct model types.

5.2 ADaPPT as Implementation Tool

ADaPPT in common with other some other EA frameworks provide a holistic view of EA through the hierarchical layering, which implies the alignment between, business, application, and technology infrastructure layer. As such, business decisions and architecture planning can be made in the context of whole instead of standalone parts. In other words, EA frameworks such as ADaPPT make use of the abstractions in order to simplify and isolate simple IS aspects/dimensions without losing sense of the complexity of the enterprise as a whole. As an organisation planning tools, ADaPPT entail baseline architecture, future architecture, architectural roadmaps, and transition plans. Baseline architecture, which is also known as ‘as-is’ view, encompasses the documentation of different layers and the existing components (models, diagrams, documents etc). This architecture serves as a baseline for identifying the relationships between different components and the gaps that should be filled for better organizational performance. Target architecture, which is also referred to as the ‘to be’ view, specify the new EA components and the strategic initiatives that should be carried out for the sake of bridging the gaps and ensuring competitive advantage. This architecture should also identify the IT resources and technological infrastructure that are needed for supporting the new EA components in order to integrate the organization structure, business processes, data, and technical resources. Architectural roadmaps represent the intermediary EA alternatives of the baseline architecture generated in the process mitigating the risks and analyzing the existing gaps in order to shift to the target architecture. These roadmaps annotate the architectural milestones performed prior to reaching the target architecture. EA transition plans are merely specifications of an ‘as-is’ and ‘to-be’ view in terms of managing the feasibility of architectural transition such as risk assessment, gap analysis, and the supporting resources of the transition. More specifically, transition plans document the activities that need to be undertaken to shift from the baseline architecture to the target architecture. Such plans are means to determine the desired future state of the enterprise wide goals, business processes, technical resources, organization units, and data.

5.3 ADaPPT and other EA Frameworks and Tools

A range of tools can be used to model the architecture appropriate to the different views. Where appropriate familiar tools are used across several views so we do not need 36 different models as in the Zachman [5] approach eg E-R modelling is used for the Data Domain and UML can be used in the Process Domain Business and

Application Views. Figure 2 illustrates a mapping between the Zachman approach and ADaPPT.

| | Zachman | | | | | |
|----------------|---------|----------|---------|--------|------|------------|
| ADaPPT Domains | Data | Function | Network | People | Time | Motivation |
| People | | | X | X | X | X |
| Process | | X | X | | X | |
| Technology | | | X | | X | |
| Data | X | | X | | X | |

Fig. 2 ADaPPT and Zachman Framework.

One of the leading EA toolset is ARIS [7]. Whilst a toolset may support many frameworks it will also have an implicit ontology. ARIS is a complex tool with an underlying Process Model. ARIS manages complexity with four Views: Data View, Organization View, Function view, Product Service View. ARIS supports a detailed process oriented view of the organisation. However the basic units of ARIS fit easily within the ADaPPT framework. There is no inconsistency in using ADaPPT as an EA Framework and ARIS as the toolset to support the management and operation or the Enterprise's Architecture Repository. The four ARIS views are apparently consistent with the ADaPPT framework. However it is worth examining some of the lower level ARIS concepts to see if this apparent alignment in high level concepts is reflected in the detail (see Figure 3).

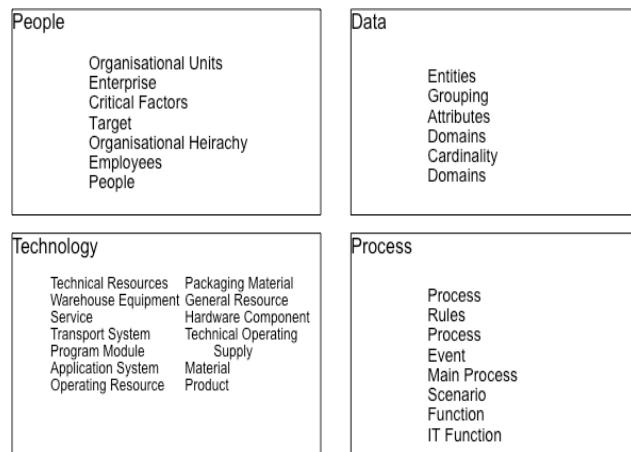


Fig. 3 ARIS concepts within ADaPPT Domain.

In ADaPPT it is possible to use conventional diagramming tools such as MS Visio or complex EA diagrammers or even full blown modellers such as IDS Sheer. It is the way the different elements combine which creates business value.

The main relationships in ADaPPT (see figure 4) are:

People: Initiate processes, Use data, Specify **technology**

Processes: Run on **technology**, Use **technology**, Generate data

Data: Stored on **technology**

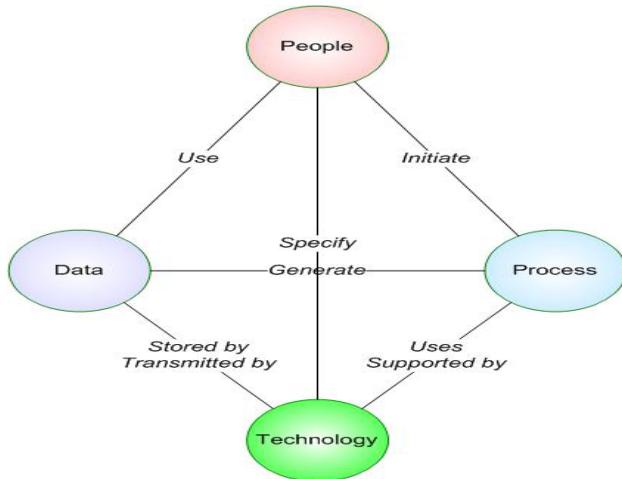


Fig. 4 The main relationships in ADaPPT.

These can be represented in the various EA tools. These main relationships are important. Representing, exploring and planning related to these elements is facilitated by, recognising and taking account of the main relationships and their content from organisational and implementation perspectives as appropriate for the particular context under consideration.

6. Conclusions

Our ongoing work is applying ADaPPT based enterprise architecture thinking for information systems development in public and private sector organisations. It provides a strong foundation for understanding the strategic, managerial and operational issues in aligning people, processes, data and technology and also in developing strategic, managerial and operational approaches while ensuring the alignment of people, processes, data and technology in an IS context. It is being used as the basis for understanding how knowledge management can be improved and new technologies exploited by organisations. Providing a basis for enabling the conceptualising of holistic, integrated and detailed consideration as appropriate to the development stage and stakeholder perspective.

References

- [1] Author, Eardley WA and Wood-Harper AT (2007) ALTAR: Achieving Learning Through Action Research, *European Journal of Information Systems*, Vol 16 No 6, Pp. 761-770.
- [2] IEEE (2006), "IEEE Standards Association, IEEE Std 1471-2000 IEEE Recommended Practice for Architectural Description of Software-Intensive Systems, http://standards.ieee.org/reading/ieee/std_public/description/s/e/1471-2000_desc.html," 2006.
- [3] DOD (2006), "DOD Architecture Framework, Systems and Software Consortium, <http://www.software.org/pub/architecture/dodaf.asp>," 2006.
- [4] J. W. Ross, P. Weill, and D. C. Robertson (2006), *Enterprise Architecture as Strategy: Creating a Foundation for Business Execution*. Boston, Massachusetts: Harvard Business School Press Book.
- [5] J. A. Zachman (1987) "A Framework for Information Systems Architecture," *IBM Systems Journal*, Vol. 26, No. 3.
- [6] Author and El Koudi M (2007), Enterprise Architecture Frameworks, *IEEE IT Professional*, Vol 9, No 5.
- [7] A.-W. Scheer (1999), *Business Process Engineering: Reference Models for Industrial Enterprises*, 2nd ed. Berlin: Springer.
- [8] FEA (2003) CIO, "Practical Guide to Federal Enterprise Architecture," Enterprise Architecture Program Management Office, <http://www.feapmo.gov>.
- [9] H. Jonkers R. v. Buuren, F. Arbab, F. d. Boer, M. Bonsangue, H. Bosma, H. t. Doest, L. Groenewegen, J. G. Scholten, S. J. B. A. Hoppenbrouwers, M. E. Iacob, W. Janssen, M. M. Lankhorst, D. v. Leeuwen, H. A. Proper, A. Stam, L. v. d. Torre, and G. E. V. v. Zanten, (2003), "Towards a Language for Coherent Enterprise Architecture Descriptions," in *7th IEEE International Enterprise Distributed Object Computing Conference (EDOC 2003)*. Brisbane, Australia, 2003.

Hanifa Shah is Associate Dean (Research) and Professor of Information Systems at Birmingham City University. She provides leadership and strategic direction to the research of the Faculty of Technology, Engineering and the Environment where staff are involved in research in three Centres. In the Centre for Low Carbon Research (CLCR) staff are investigating bio-energy generation, low carbon transportation, sustainable, intelligent buildings and retrofit and knowledge based engineering. In the Institute for Digital Experience and Applications (IDEAS) staff are contributing to research to help maximise economic and societal benefits of digital technologies. The Centre for Environment and Society Research (CESR) considers how planning policy and technology affect trends both in society and the environment in light of historical, contemporary and future contexts. Prof Shah was previously at Staffordshire University and at Aston University and is currently Visiting Professor at Manchester University. She has over 30 years in-depth knowledge and experience of both the higher education sector and business. Prof Shah has been at the forefront of the development of industrial links and partnerships for research, enterprise and teaching by Universities. She has supervised over twenty PhDs and has specialized in linking PhDs to industrial collaborations producing results with practical impact as well as academic significance. She has been invited to advise public and private sector organizations on strategic IT issues and also on the professional development of IT staff and university-academic collaboration opportunities. Prof Shah has helped to shape national and international thinking on work-based learning and university qualifications for IT Professionals through her work

on transformational development programmes for large corporations and public sector organizations. As a result of her work she has published over seventy papers, a number of book chapters and a book. Prof Shah's current research includes research methodologies for information systems research with collaborating corporations (ALTAR), organisational and professional development in IT (STEP) and the exploitation of mobile technologies in health informatics and the improvement of patient processes in hospitals. She is investigating enterprise architecture based approaches for information systems and knowledge management (ADaPPT) in areas such as healthcare and e-government.

Dr Paul Golder is a Visiting Research Fellow in the Faculty of Technology, Engineering and the Environment at Birmingham City University. His career has included time as an applied statistician with periods in various public sector organisations including the European Commission in Luxembourg. After that he concentrated on the data management aspects of decision making within the Computer Science Group of Aston University, and at the same time collaborated with colleagues within Aston Business School (ABS). Since retiring from Aston University in 2004 he has remained active in research continuing to collaborate with colleagues at Staffordshire University where he was appointed Visiting Research Fellow, during this time he undertook a part time employment as a Research Fellow to bring to completion an EPSRC funded research project. He has continued to assist in the supervision of PhD students and to carry out industrial placement visits for ABS. He has substantial experience in all areas of information management. He has expertise both in the development of information strategies and also skills in many of the areas central to the successful implementation of an information strategy. biography appears here. Degrees achieved followed by current employment are listed, plus any major academic achievements. Do not specify email address here.

Modular Design of Call Control Layer in Telephony Software

Ilija Basicevic

University of Novi Sad, Faculty of Technical Sciences
Novi Sad, 21000, Serbia

Abstract

An important property of a telephony system is the call control model on which it is based. It is noted that many call control models in the past, especially those in PSTN/ISDN networks follow centralized model. For such a model, typical is significant coupling of modules belonging to different services with the basic call control module which is aware of all active telephony features in the system. Although sometimes based on distributed model, VoIP call control models still manifest some of the listed problems of their predecessors. In this paper we present a fully distributed model which exhibits minimal coupling of modules belonging to different services and a simple basic call control module. The model is based on taxonomies of call control services which are presented in the paper. Also, the implementation of several typical services is described.

Keywords: Call Control, telephony services, Voice over IP, VoIP session transfer, VoIP session redirection

1. Introduction

Call control model can be described as a formal representation (and design) of a distributed software system for telephony communications. Typically, there is a network consisting of infrastructure and endpoints. This network can be represented as a graph. Infrastructure nodes are responsible for routing. Endpoint nodes are nodes that have only one adjacent node and are usually responsible for end users' access to services of telephony network. Call control model specifies the design and mutual interaction of software modules that are responsible for call processing.

The aim of this paper is twofold. One is to present the specific call control model that is developed here. An important issue with respect to that is the issue of modular development of telephone features. For rather long time, designers have strived to develop fully modular features, decoupled from the code of the so called "basic call control" and other features. The other aim of this paper is to focus on the feature management module, which is an important part of the call control layer.

Section 2 describes related work, section 3 presents two taxonomies of call processing features that are elaborated in this paper, and the manner in which some of the features

are implemented. Section 4 presents the implementation of a feature that belongs to the class of network based services. Section 5 contains concluding remarks.

2. Related Work

There has been a lot of research in the area of feature management and call control models. Influential call control models that are used in circuit switched telephony have been published by Telecommunication Standardization Sector of the International Telecommunications Union (ITU-T). We mention here Q.71 [1] model used in PSTN/ISDN networks and Intelligent Network (IN) model [2]. In Q.71 model, each new feature that is introduced to the call processing system leaves its "fingerprint" in the basic call control module. The IN is the first model in software industry that features systematic definition and operational adoption of service orientation [3]. The IN architectural stack clearly identifies several well defined layers of service with distinct responsibilities and roles. The next step in the telecommunications industry has been the development of object oriented application programming interfaces (API). We mention here Parlay, the 3GPP Open Service Architecture and Java APIs for Integrated Networks (JAIN). A simplified version of Parlay/OSA has been later extended with support for Web-services, and XML [4] resulting in Parlay X [5], which is a de-facto standard Web-services API today.

Distributed Feature Composition [6] has been an attempt to develop a highly distributed and decoupled model. The advent of Voice over IP (VoIP) telephony brought H.323[7] and Session Initiation Protocol (SIP) [8] models. There has been continuous work on improvement of those models as in WSIP[9] and in compositional control of IP[10]. SIP protocol is based on two-step and three-step transactions, while in compositional control an idempotent signaling protocol based on unilateral descriptions is proposed. WSIP is an integration of two concepts, SIP and Web Services. The idea behind WSIP is separation of service integration signaling. Thus we have a three tier stack of service integration, signaling and media

transmission. In more recent papers, call control is researched as a part of collaborative streaming applications [11]. IP Multimedia Subsystem (IMS), which is based on combination of IN concepts and application of Internet protocols, most importantly SIP and Diameter[12], appeared in 2004. As of today, IMS is considered a global standard for unified service control platform converging fixed, mobile, and cable IP networks [13].

Although telephony end points of today are way simpler than switches of PSTN/ISDN/IN networks, the most important aspects of feature management problem are present in both types of systems.

3. Two Taxonomies of Call Processing Features

The concept presented in this paper is implemented in the framework [14], but for convenience of readers, some details required for understanding the paper are repeated here. EndUser class models the end user of telephony endpoint, and contains the typical telephony endpoint interface. SignalingDevice is a class that interfaces the underlying telephony protocol stack (SIP, for example). Feature class is the parent of all classes that model telephony features (e.g. Session, CallWaiting, SessionRedirect etc.). Session class models the basic call feature with first party call control interface. P3Session also models the basic call feature but with the third party call control interface. The important methods of Session are: Invite, AcceptSession, Disconnect, ModifySession, and the callbacks for messages from the remote peer: OnAccepted, OnDisconnect, OnModifySession. The Invite message from the remote peer is handled in the EndUser object - at this moment the local Session object does not exist. FeatureMng is the feature manager class which dispatches received messages to feature modules. It is the responsibility of this class to determine which active features, and in which order will process the received message. EndPoint models the session terminal available to end users for utilizing network services. RoutingPoint is the infrastructure node responsible for routing of messages. RoutingPointExt is the infrastructure mode extended with the software modules of network side applications.

We have introduced simple taxonomy of features, based on analyses of standards and implementation results. Each feature is either primitive, derivative or composite. Basic service session has two classes, one with the interface for first and the other with the interface for third party call control. Basic session with the interface for first party call control is considered primitive feature and the root of this taxonomy. Derivative features are session transfer, session redirect, session waiting, etc. Composite features are basic session with interface for third party call control and conference.

The relationship between composite and primitive feature is similar to "has" relationship (aggregation) in Unified Modeling Language (UML).

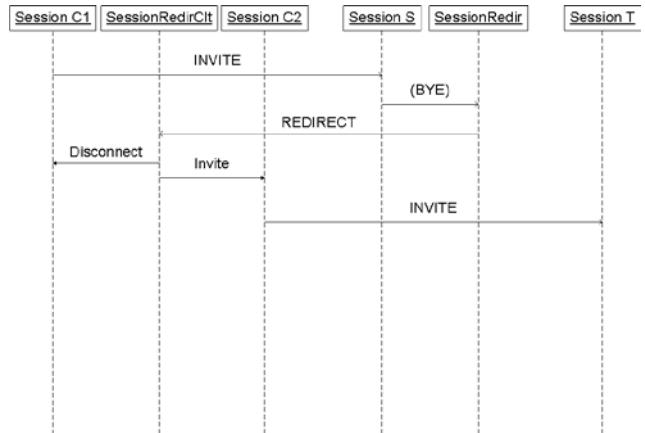


Figure 1: Session redirect MSC

Each feature's intelligence knows about itself and its primitive elements' features. Each feature registers at the feature manager for notification about call state changes. In notification, manager follows the rule that active primitive features are notified before derivative and composite ones. For example basic call (Session) is notified about session change before call waiting feature. For sessions, this rule reduces to the following: the basic call is the first feature to be notified about any session state changes.

Certain derivative features have read access to basic call state. This is used for checking preconditions of certain session control operations. No feature has full (read and write) access to another feature data. Feature manager fully recognizes only primitive features. All derivative and composite features are considered by feature manager only as instances of the generic class Feature. Regarding unwanted feature interaction, it is assumed that end user will not simultaneously activate features that result in unwanted feature interaction. Although this is a very simplifying hypothesis, we consider unwanted feature interaction to be out the scope of this paper.

Typical feature communicates the EndUser object (for receiving end user commands), the SignalingDevice object (for sending signaling messages to remote peers), the feature manager (for receiving commands from remote peers), but there is a category of derivative features that communicate the EndUser only for the reasons of feature activation and configuration. During operation phase there is no interaction with local end user in such features. Those features manipulate the call automatically, without the local user explicitly taking part in the call control. Example is call redirection.

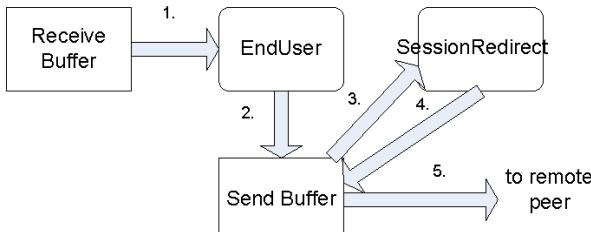


Figure 2: Chain processing of received messages (send buffer)

There is another taxonomy of telephony features, depending on the programming model. In the first class there are features that are realized as peer-to-peer (symmetrical) modules. Such a feature is basic call. Both sides in the session communication are the same. In the second class there are features that are realized using client-server model (asymmetrical). An example is session redirection feature, see Fig. 1. There is SessionRedirect class implementing server side that receives INVITE message (sent by Session object) and responds with REDIRECT message, and there is SessionRedirectClt class that receives REDIRECT message, closes the first session (Session object that sent the INVITE) and opens a new one by sending INVITE message to address referred to in REDIRECT message. The SessionRedirectClt class implements the client side of the feature. While SessionRedirect extends the Session class, SessionRedirectClt inherits the Feature class. An important aspect of operation of this feature is the chain processing of received messages. At the redirect server side, end user module is the first to process received INVITE. In case end user has already been engaged in a call, BYE message with a reason that the end user is already in call will be sent. The next in chain is session redirect server, which processes the BYE message, removes it from the send buffer and replaces it with REDIRECT message. The chain processing is presented in Fig 2. It can be presented with the following pseudo code.

```

while (sendbuffer not empty){
    msg = get_message(sendbuffer)
    //message is deleted from sendbuffer, iterator
    //moves to next
    if(relevant_to_operation_of_the_feature(msg))
        process (msg)
}
    
```

Processing of the message typically includes placing a different message in the send buffer. When the last feature in the chain finishes processing, messages in the send buffer are actually sent over the network to remote peer. The order of features in the chain is the result of the order in which they were activated. This order is very important for the feature interaction, but as noted earlier, we assume that it is the responsibility of the system's end user.

In order to sustain a relatively small number of basic message types, we have introduced the following message information field to messages: source feature type. Thus we can use the same message type (ACCEPT) for confirming session establishment and session transfer, or for example the same message type (BYE) for session tear down both in first party and third party call control. In those cases, ACCEPT will carry either Session' or SessionTransfer' identifier as source feature type and BYE will carry either Session' or P3Session' identifier in that message information field. Another reason for introducing this message information field is the chain processing of received messages, thus each feature knows which feature reacted before it in the chain, and placed a message in the send buffer. For example, if BYE is from SessionTransferClt, SessionRedirect feature will ignore it. But it will react to BYE which was placed into the send buffer by Session feature.

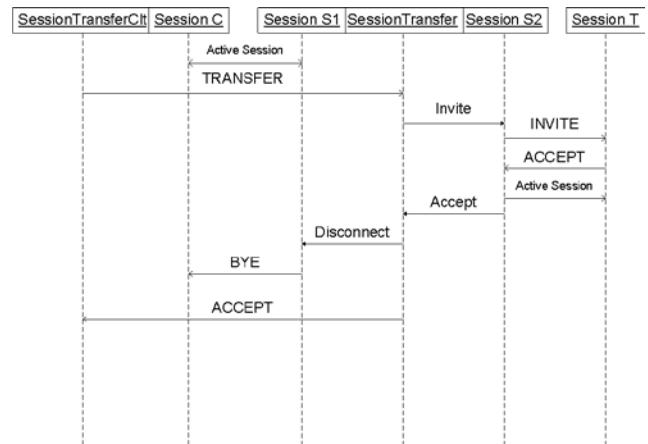


Figure 3: Session Transfer MSC

Another client-server based feature is transfer. There is SessionTransferClt class that implements the client side. This class contains the TransferSession method. When the user asks for transfer, EndUser object invokes the TransferSession method of SessionTransferClt object. In the next step, the SessionTransferClt's method sends TRANSFER message to SessionTransfer object at the remote peer. Two transfer feature objects are positioned at two sides of an active session. Upon receiving the message, SessionTransfer object instantiates another Session object, and invokes its Invite method. An INVITE is sent to the target point of transfer operation, see Fig. 3 We have assumed here that the target responds with ACCEPT. This message is processed by two feature modules: Session and SessionTransfer. As we already noted FeatureManager first dispatches the message to Session (as a primitive feature) and then to SessionTransfer (as a derivative feature).

SessionTransfer object disconnects the session to transfer client by sending BYE (it invokes Disconnect method of Session object after it gets the handle of the session object from feature manager). Immediately afterwards it confirms the successful transfer by sending ACCEPT to SessionTransferClt object at the remote side.

The example of session redirection feature (and session transfer) shows the value of modular approach. This feature can be implemented without the SessionRedirectClt class by placing its logic in Session class. (Also the transfer feature can be completely realized in Session class without introducing SessionTransfer and SessionTransferClt). In that way, the Session class grows more and more complex. It becomes the module that "knows everything" about call processing. The consequence is that call control layer becomes much more error prone. In contrast, modular development allows for gradual increase of functionality. The call control layer is easier to test and debug. The feature interactions are controlled in a more straight forward manner.

The third party call interface of basic session, realized in P3Session class, provides the following operations: Establish, and Terminate. The Establish method, that establishes basic session between two remote points is based on transfer and monitoring features. Monitoring is based on publish/subscribe event notification mechanism.

4. Network based services

The services mentioned in the Section 3 are end point based. There is no service specific processing in infrastructure points. In this section we will give an example of implementation of network based service. Line hunt is another ISDN supplementary service. It belongs to the group of network side services. In the framework such services are usually implemented by inheriting RoutingPointExt class. Thus a new class RoutingPointExt1 has been implemented. This class contains a linked list of hunt mappings. The mapping contains session layer address that is mapped, and network layer address and port it is mapped to. Since in line hunt one session layer address is mapped to a group of network layer addresses, the logical relationship of one line hunt mapping is realized as a group of list elements, all containing the same value of session layer address field.

The dynamics of the line hunt operation is realized in the following manner. The routing of the protocol message is intercepted. Typically, the router thread of SignallingDevice class passes the received message to the RouteProtMessage of RoutingPoint class, which then inspects the routing table and forwards the message according to the appropriate route in the table.

In case of the line hunt operation, router thread passes the received message to RoutingPointExt1 class. In case of INVITE and ACCEPT messages, the routing for line hunt is different from aforementioned general routing procedure. The processing of INVITE sets the flag of line hunt for that session layer address to active. A copy of INVITE message is sent to all addresses the target address of INVITE is mapped to. The first received ACCEPT from one of those addresses sets the flag to false, also a BYE is sent to all the other addresses that INVITE has been sent to.

5. Conclusions

During the decades long evolution of telephony software, there have been several approaches to design of call control layer. However, the majority of them featured a strongly centralized approach where basic call control module with inclusion of new features becomes a "know all" module. The legacy of PSTN/ISDN networks of fixed telephony has been the Q.71 model. With the appearance of the IN network, ITU-T invested in an approach where service plane and basic call control plane would be strongly separated. This paper presents strongly modular design of call control layer where inclusion of new features is possible without modification of existing ones, especially having in mind the basic call control module. Telephony features in this model are implemented as asymmetrical, client server modules, while basic call control module is implemented as symmetrical, peer-to-peer module. Implementations of session transfer and session redirect telephony features described in the paper show this separation between processing in the basic call control and other features. The implementation of line hunt service is given as an example of network based service.

References

- [1] ITU-T Q.71 - ISDN Circuit Mode Switched Bearer Services, (1993) International Telecommunication Union
- [2] ITU-T Recommendation Q.1200, Q-Series Intelligent Network Recommendation Structure, (1993) International Telecommunication Union
- [3] Tiziana Margaria, (2007), Service is in the Eyes of Beholder, Computer Magazine, vol 40, No 11, DOI:10.1109/MC.2007.398
- [4] Extensible Markup Language (XML) 1.0 (Fifth Edition), W3C Recommendation, 26 November 2008
- [5] ETSI OSA PArlay x 3.0 Specifications, European Telecommunications Standards Institute and The Parlay Group, 2007
- [6] Jackson M., Zave P., (1998), Distributed Feature Composition: A Virtual Architecture for Telecommunications Services, IEEE Transactions On Software Engineering, vol. 24, no. 10

- [7] Rosenberg J., Schulzrinne H., Camarillo G., Johnston A., Peterson J. Sparks R., Handley M. Schooler E., (2002), Session Initiation Protocol, RFC 3261, Internet Engineering Task Force
- [8] ITU-T H.323, Visual Telephone Systems and Equipment for Local Area Networks Which Provide A Non-Guaranteed Quality Of Service, (1996), International Telecommunication Union
- [9] Liu F., Chou W., Li L., Li J., (2004), WSIP - Web Service SIP Endpoint for Converged Multimedia/Multimodal communication over IP, IEEE International Conference on Web services (ICWS 04), pp.690
- [10] Zave P., Cheung E., (2006), Compositional Control of IP Media, International Conference On Emerging Networking Experiments And Technologies, Lisboa, Portugal
- [11] Kahmann V., Brandt J., Wolf L., (2006), Collaborative Streaming and Dynamic Scenarios, Communications of ACM, vol 49, no 11.
- [12] Calhoun P., Loughney J., Guttman E., Zorn G., Arkko J., (2003), Diameter Base Protocol, RFC 3588, Internet Engineering Task Force
- [13] Magedanz T., Blum N., Dutkowski S., (2007), Evolution of SOA concepts in Telecommunications, Computer Magazine, vol 40, No 11, 2007, DOI:10.1109/MC.2007.384
- [14] Basicevic I., (2009), Object-Oriented Framework for Development of Telephony Applications, Fourth International Conference on Digital Telecommunications, Colmar, France

Ilija Basicevic received his BSc, MSc, and PhD degrees from the University of Novi Sad in 1998, 2001 and 2009 respectively. Currently, he is assistant professor at the same university. His research interests are communication systems and computer security. He has published more than 30 papers. Ilija is member of ACM and IEEE.

Generating a Performance Stochastic Model from UML Specifications

Ihab Sbeity¹, Leonardo Brenner² and Mohamed Dbouk¹

¹ Lebanese University, Faculty of Sciences, Section I
Beirut - Lebanon

² LSIS Laboratory – Laboratoire des Sciences de l'information et des Systèmes
Marseille – France

Abstract

Since its initiation by Connie Smith, the process of Software Performance Engineering (SPE) is becoming a growing concern. The idea is to bring performance evaluation into the software design process. This suitable methodology allows software designers to determine the performance of software during design. Several approaches have been proposed to provide such techniques. Some of them propose to derive from a UML (Unified Modeling Language) model a performance model such as Stochastic Petri Net (SPN) or Stochastic process Algebra (SPA) models. Our work belongs to the same category. We propose to derive from a UML model a Stochastic Automata Network (SAN) in order to obtain performance predictions. Our approach is more flexible due to the SAN modularity and its high resemblance to UML's state-chart diagram.

Keywords: *Performance software engineering, UML, Stochastic Automata Network, Markovian analysis.*

1. Introduction

Quantitative analysis of software systems is being recognized as an important issue in the software development process. However, it is widely accepted that performance analysis techniques have suffered a lack of acceptance in the wider software design community. The most convincing reason is the reluctance of designers to learn the specialized formalism required by Markovian numerical solution techniques.

To encourage designers to incorporate performance analysis, a wave of explicit efforts got a serious attention in the last two decades [1, 2, 3, 4, 11, 12, 13, 16]. Since the initiation of the software performance engineering (SPE) methodology by Connie Smith [16], three large groups of researches in the area have been noticed. The first group aims at constructing performance based frameworks that are able to be used by designers. This is the case of formalisms such as Hit [2]. However, these formalisms had not received a good attention from designers because of

the wide evolution of other powerful and dominant specification and design techniques such as LOTOS, SDL, and more specifically UML. Thus, a second group of works [1, 9, 12] in SPE proposes to extend existing specification formalisms by introducing performance information such as time and probability distributions to these formalisms. In [1], a prototypic version of a program package is described, which takes a TSDL (Timed SDL) model as input and creates an internal representation of an equivalent Finite State Machine, so that validation and performance evaluation of TSDL models can be done automatically. In [12], Authors gave raise to stochastic LOTOS which modifies the semantics of LOTOS to allow performance information to be represented and performance results to be computed. Designers have to be careful with the new notations which mostly imply the formal power of the specification to be lost.

Extending the notation in order to support performance information in the Unified Modeling Language (UML), is a recent attempt to merge the most widely used object oriented design notations. It has been adopted by the industry body, the Object Management Group (OMG), as a draft standard [9]. A significant effort is underway to complete and refine this draft. Again, an additional effort is required from designers to deal with the proposed framework.

Recent efforts such as [3, 13] also show the eligible need to incorporate performance analysis in the UML specification process. However, these works treat specific applications and they do not give a general demarche that can be useful with other applications.

A significant approach does appear in SPE which consists of generating from specification formalisms, more specifically UML, a performance model. Up to date works in this area have explored the possibilities for simulation of UML models [7], generation of queuing network models

from UML deployment diagrams and collaboration diagrams. Moreover, mappings from UML to stochastic Petri net models (SPNs), more specifically Trivedi's SPNP [8] tool's variant of SPNs, and to stochastic process algebra models, particularly Hillston's PEPA [11], have been also developed. All of these show the potential of using UML's logical and behavioral notations to define the structure of performance models. Furthermore, the power of this approach results from the ability of the performance formalisms to represent models with large state space.

The purpose of this paper is to initiate a new methodology in the SPE area and it is similar to the approaches above in the sense that we propose to generate a performance model from UML specifications. Starting from a UML model, we suggest engendering a Stochastic Automata Network (SAN) model [10] which may be used to predicate performance for large systems. The SAN formalism is usually quite attractive when modeling a system with several parallel cooperative activities. An important advantage of the SAN formalism is that efficient numerical algorithms have been developed to compute stationary and transient measures [5, 15]. These algorithms take advantage of structured and modular definitions which allow the treatment of considerably large models. Another important advantage of the SAN formalism is the recent possibility of modeling and analyzing systems with (phase type) PH distributions [14] allowing to model deterministic activities. In addition, SAN permits to represent a system in modular way. A SAN model is a state-transition graph having a strong likeness with the UML state-chart diagram.

For these reasons, we believe that SAN is more than adequate to generate a performance model from a UML specification model.

This work opens the door to propose a more general demarche of generating a SAN model from a UML model. Here, we illustrate our approach in an informal way based on an example. The rest of the paper is structured as follows: Section 2 presents an informal definition of the SAN formalism. Section 3 considers how to exploit UML for performance analysis. This includes the case study of a chess game in order to show the direct use of UML. Section 4 explores how the UML model maps into SAN based on our case study. Some typical results obtained by solving the model are also presented. Section 5 concludes our paper and describes our ongoing works.

2. Stochastic Automata Network

Stochastic Automata Networks, SANs, were first proposed by Plateau in 1985 [10!]. The SAN formalism enables a complete system to be represented as a collection of

interacting subsystems. Each subsystem is represented by an automaton which is simply a directed and labeled graph whose states are referred to as local states, being local to that subsystem, and whose edges, relating local states to one another, are labeled with probabilistic and event information. The different subsystems apply this label information to enable them to interact with each other and to coordinate their behavior.

The states of a SAN are defined by the Cartesian product of the local states of the automata and are called the global states of the SAN. Thus, a global state may be described by a vector whose i th component denotes the local state occupied by the i th automaton. The global state of a SAN is altered by the occurrence (referred to as the firing) of an event. Each event has a unique identifier and a firing rate. At any moment, multiple events may be enabled to fire (we shall also use the word fireable to describe events that are enabled): the one which actually fires is determined in a Markovian fashion, i.e., from the relative firing rates of those which are enabled. The firing of an event changes a given global source state into a global destination state. An event may be one of two different types. A local event causes a change in one automaton only, so that the global source and destination states differ in one component (local state) only. A synchronizing event, on the other hand, can cause more than one automaton to simultaneously change its state with the result that the global source and destination states may differ in multiple components. Indeed, each synchronizing event is associated with multiple automata and the occurrence of a synchronizing event forces all automata associated with the event to simultaneously change state in accordance with the dictates of this synchronizing event on each implicated automata. Naturally, a synchronizing event must be enabled in all of the automata on which it is defined before it can fire.

Transitions from one local state to another within a given automaton are not necessarily in one-to-one correspondence with events: several different events may occasion the same local transition. Furthermore, the firing of a single event may give rise to several possible destinations on leaving a local source state. In this case, routing probabilities must be associated with the different possible destinations. Routing probabilities may be omitted only if the firing of an event gives rise to a transition having a single destination. Also, automata may interact with one another by means of functional rates: the firing rate of any event may be expressed, not only as a constant value (a positive real number), but also as a function of the state of other automata. Functional rates are defined within a single automaton, even though their parameters involve the states of other automata.

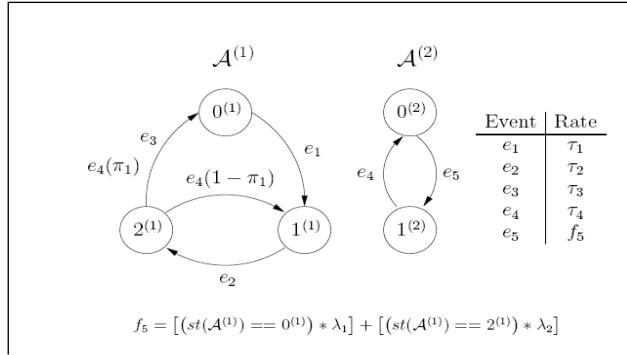


Fig 1: Example of a SAN model

As an example of the previous discussion, Figure 1 presents a SAN model with two automata, $A^{(1)}$ and $A^{(2)}$, the first with 3 local states, $0^{(1)}$, $1^{(1)}$ and $2^{(1)}$ and the second with two local states, $0^{(2)}$ and $1^{(2)}$. The model contains four local events, e_1 ; e_2 ; e_3 and e_5 and one synchronizing event, e_4 . When automaton $A^{(1)}$ is in local state $0^{(1)}$, and $A^{(2)}$ is in local state $0^{(2)}$, (global state $[0, 0]$), two events are eligible to fire, namely e_1 and e_5 . The event e_1 fires at rate τ_1 . This is taken to mean that the random variable which describes the time t from the moment that automaton $A^{(1)}$ moves into state $0^{(1)}$ until the event e_1 fires, taking it into state $1^{(1)}$, is exponentially distributed with a mean by $1/\tau_1$. Similar remarks hold for the firing rate of the other events. The firing of e_1 when the system is in global state $[0, 0]$ moves it to global state $[1, 0]$ in which e_5 is still eligible to fire, along now with event e_2 . The event e_1 cannot fire from this new state. The synchronizing event e_4 is enabled in global state $[2, 1]$ and when it fires it changes automaton $A^{(2)}$ from state $1^{(2)}$ to state $0^{(2)}$ while simultaneously changing automaton $A^{(1)}$ from state $2^{(1)}$ to either state $0^{(1)}$, with probability π_1 , or to state $1^{(1)}$ with probability $1-\pi_1$. Observe that two events are associated with the same edge in automaton $A^{(1)}$, namely e_3 and e_4 . If event e_3 fires, then the first automaton will change from state $2^{(1)}$ to state $0^{(1)}$; if event e_4 fires the first automaton to change from state $2^{(1)}$ to either state $0^{(1)}$ or state $1^{(1)}$ as previously described. There is one functional rate, f_5 , the rate at which event e_5 fires, defined as

$$f_5 = \begin{cases} \lambda_1 & \text{if } A^{(1)} \text{ is in state } 0^{(1)} \\ 0 & \text{if } A^{(1)} \text{ is in state } 1^{(1)} \\ \lambda_2 & \text{if } A^{(1)} \text{ is in state } 2^{(1)} \end{cases}$$

Thus event e_5 , which changes the state of automaton $A^{(2)}$ from $0^{(2)}$ to $1^{(2)}$, fires at rate λ_1 if the first automaton is in state $0^{(1)}$ or at rate λ_2 if the first automaton is in state $2^{(1)}$.

The event e_5 is prohibited from firing if the first automaton is in state $1^{(1)}$.

$$f_5 = [(st(A^{(1)}) == 0^{(1)}) * \lambda_1] + [(st(A^{(1)}) == 2^{(1)}) * \lambda_2]$$

Functional transitions are written more compactly, e.g., in which conditions such as $st(A^{(1)} == 2^{(1)})$ (which means 'the state of $A^{(1)}$ is $2^{(1)}$ ') have the value 1 if the condition is true and are equal to 0 otherwise. This is the notation used to describe functions in the PEPS software tool [4]. In this setting, the interpretation of a function may be viewed as the evaluation of an expression in the C programming language. The use of functional expressions in SANs is not limited to the rates at which events occur; indeed, probabilities also may be expressed as functions. Figure 2 shows the equivalent Markov chain transition rate diagram for this example.

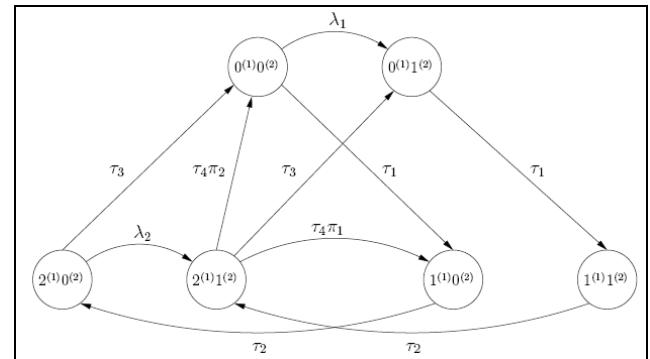


Fig 2: Transition rate diagram of corresponding Markov chain.

Furthermore, a new methodology has been recently incorporated into SANs: the use of phase-type distributions [14]. The exponential distribution has been the only distribution used to model the passage of time in the evolution of the different San components. In [14], it is shown how phase-type distributions may be incorporated into SANs thereby providing the wherewithal by which arbitrary distributions can be used, which in turn leads to an improved ability for more accurately modeling numerous real phenomena.

The real interest in developing stochastic automata networks lies, in addition to their specification, to the fact that the *transition matrix* of the underlying Markov chain of a SAN can be represented compactly by storing the elementary matrices corresponding to subsystems and never the transition matrix itself. The numerical analysis of the system is then done by using the elementary matrices, which extremely decreases the analysis cost [5].

3. A UML model for the chess game

The Unified Modeling Language (UML) [9] is a graphically based notation, which is being developed by the Object Management Group as a standard means of describing software oriented designs. It contains several different types of diagram, which allow different aspects and properties of a system design to be expressed. Diagrams must be supplemented by textual and other descriptions to produce complete models. For example, a use case is really the description of what lies inside the ovals of a use case diagram, rather than just the diagram itself. For a full account of UML see [6].

Here, we develop UML model which shows how a chess game might look if based on object oriented design. We do not develop in detail all the possible UML diagrams, concentrating on those that are essential to the objective of describing the structure and behavior of the system. In particular, we only introduce the class diagram (section 3.1), collaboration diagram (section 3.2) and more specifically the state-chart diagram (section 3.3).

We do not spend time in presenting the use case model, which is important in real software design projects, since it captures the requirements for the system. We take those very much as given.

3.1 The class diagram

UML is an object oriented design formalism. Thus the core of the language is the *class diagram*. A class model defines the essential types of object available to build a system; each class is described by a rectangle with a name. This can be refined by adding compartments below the name which list the attributes and operations contained in each instance of (object derived from) this class [11].

Classes are linked by lines known as *associations* which indicate that one of the classes knows about the other. The direction of this knowledge is known as the *navigability* of the association. In an implementation an association typically corresponds to one class having a reference variable of the type of the other class. Sometimes navigability has to be two ways, but it more often one way. This can be shown by adding arrow head to the end(s) of the association.

For the purpose of this paper, we assume that classes and objects exist as fundamental units of description within a design. In particular, classes encapsulate behavior, which can be described as state machine description.

The class model of our chess game is reduced to its essentials. We assume there are two kind of players X and

Y, the board and an umpire. A player spends time thinking before playing (achieving a movement on the board). Its movement may be valid or not. The umpire decides about the validity of the player movement.

Figure 3 shows classes for the chess game. We underline the existence of two classes (XPlayer and YPlayer) which inherit from the class Player. The inheritance relation is a form of generalization where one class is a super class; the other is a specification (sub class). The need to distinguish the two subclasses is related to the difference in behavior of the two players in term of state-transition. This is illustrated later in the state-chart diagram.

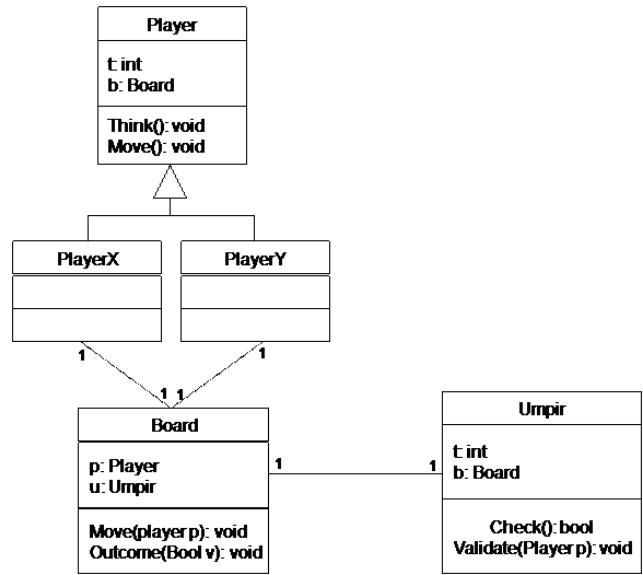


Fig 3: Chess game – The class diagram

3.2 The collaboration diagram

Collaborations are collections of objects, linked to show the relevant associations between their classes. Here “time” is not represented explicitly. Instead the emphasis is on showing which objects communicate with which others. Communications are represented by messages. Sometimes, these messages are numbered to show the order in which they should happen [8].

Figure 4 represents the collaboration diagram of our chess game. It acts of a communication between four objects: two players, x and y, aboard b and an umpire u. Also, we may for example imagine the existence of more than one player of type x and/or y that are playing on the same board. That implies the creation of new class instances which will be added to the collaboration diagram.

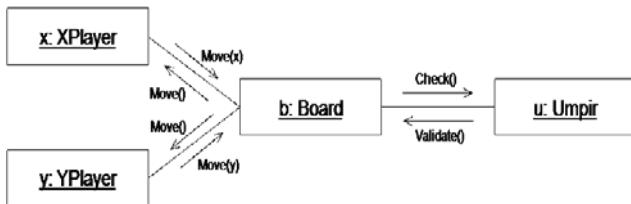


Fig 4: Chess game – the collaboration diagram

Collaboration (or communication) diagrams show a lot of the same information as sequence diagrams which are not represented here, but because of how the information is presented, some of it is easier to find in one diagram than the other. Communication diagrams show which elements each one interacts with better, but sequence diagrams show the order in which the interactions take place more clearly.

3.2 The state-chart diagram

UML defines state diagrams which allow a class to be defined in terms of states it can be in and the events (messages) which cause it to move between states.

Many software systems are event-driven, which means that they continuously wait for the occurrence of some external or internal event such as a mouse click, a button press, a time tick, or an arrival of a data packet. After recognizing the event, such systems react by performing the appropriate computation that may include manipulating the hardware or generating “soft” events that trigger other internal software components. (That’s why event-driven systems are alternatively called reactive systems.) Once the event handling is complete, the system goes back to waiting for the next event.

State-charts describe how instances of classes behave internally. In a complete design they provide a full description of how the system works. We insert the state-chart for each object into its box in the collaboration. At any point in the lifetime of this system, each object must be in one, and only one, of its internal states. Each time a message (event) is passed, it may cause a state change in the receiving object and this may cause a further message to be passed between that object and another with which it has an association.

The overall state of a system will be the combination of all the current internal state of its objects, plus the current values of any relevant attributes. Intuitively, readers may sense a strong similarity to the stochastic automata networks behavior. An automaton is a set of states, transitions and events. The global state of a SAN is a combination of the local states of its automata. The powerful point of mapping the state-chart diagram into a

SAN model is that the mapping process will not produce fundamental changes in the graph structure, only some information needed to represent relevant attributes and the time spent in a state is required. That may help designers to better understand the SAN performance model which is an emphasized advantage of our approach.

Figure 5 presents the state-chart diagram of our game model. For each object in the collaboration diagram, a chart is associated.

Each chart describes the internal behavior of the concerned object. Briefly, a chart is composed of states, transitions, triggers and actions. States are shown as lozenges; the initial state as a black filled circle. Transitions are the arrows between states, labeled with a trigger. Triggers represent the reason for an object to leave one state and follow the corresponding transition to another state; here, triggers are incoming messages or elapsing of time shown by the word *after* followed by the duration. Actions are resulting from a trigger carried out before entering the new state. They occurrence may follow a probability and/or they may involve sending messages to other objects, and these messages are prefixed by a caret. For example, in the automata corresponding to player X the label *after(t1)/^b.Move(x)* means that the action *Move(x)* of the Board *b* should occur when the trigger *after(t1)* is fired.

A player (X or Y) alternates between two states: the state where it is thinking about a move and the state where it is waiting its turn. Note that the initial state of X is thinking and that of Y is waiting (Player X starts the game). The two players cannot be simultaneously in the same state. When a player achieve a move, this move can be correct or not. The probability of a valid movement achieved by player X (respectively Y) is *p* (respectively *q*). The umpire alternates between two states: a state where it is checking a player’s move and the state where it is idle. Parameters *t1* and *t2* are respectively the time needed by players X and Y to think about a move. The parameter *T* represents the time need by the umpire in order to validate a player’s move.

4. Generating the SAN model

In this section, we present how a SAN model is directly generated for the UML chess game model presented in the previous section. The generation process is principally based on the state-chart diagram.

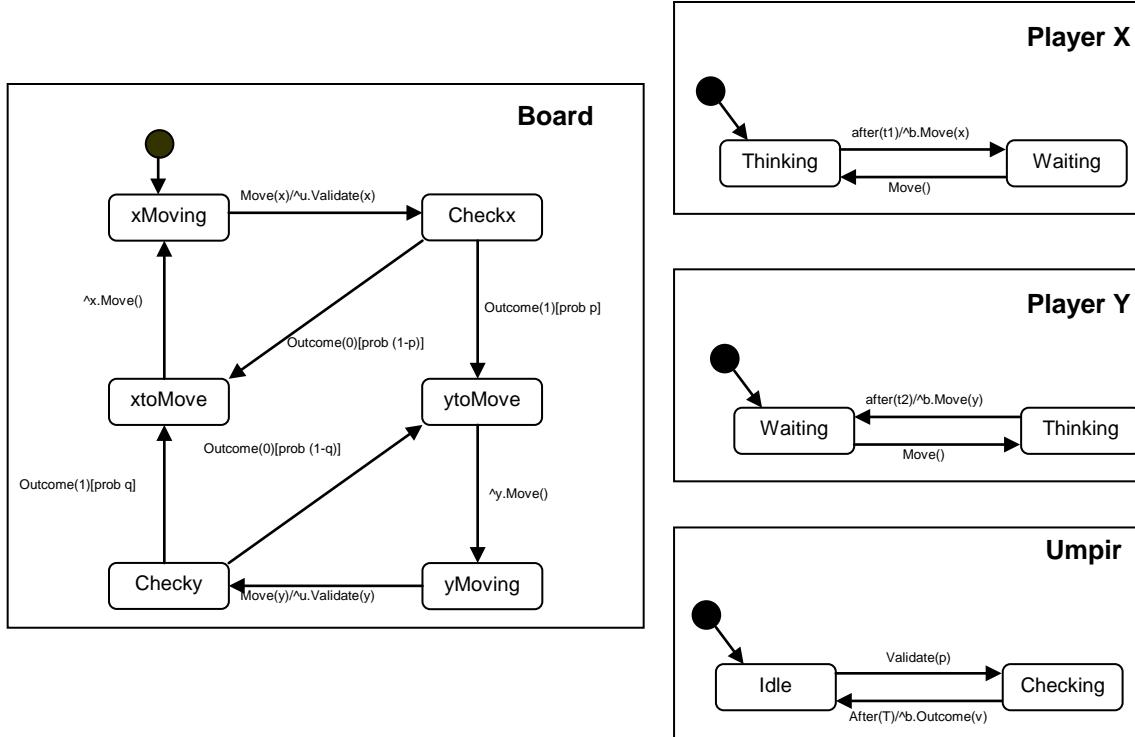


Fig 4: Chess game – the state-chart diagram

Even though the model of the chess game is not necessarily very significant from the performance point of view, but it is an excellent case study to show the simplicity of the generation process and consequently the significance of our methodology by using a SAN to generate a performance model from the UML specifications. Moreover, here, we focus only on the informal generation of the SAN model. The formal generation procedure is in the line of our future research activities.

The SAN model generated from the UML chess game model of section 3 is presented in figure 6.

As it can be noticed, the SAN model is robustly similar to the UML state-chart diagram of section 3.3. Each automaton of the state-chart diagram is mapped into a SAN automaton. But, in the general case, it is not necessarily that the number of the SAN automata is equal to that of the UML state-chart diagram. Sometimes, we need additional SAN automata in order to describe relevant attributes of the UML model. That is not the case of our chess game model. Moreover, each SAN automaton has a quasi-similar behavior to the state-chart, in term of states/transitions. SAN automaton X, Y, B and U correspond to UML's Player X, Player Y, Board and Umpire respectively (refer to figure 5). For all state-chart automata, the initial state

(black filled circle) is omitted in the SAN automata, as there is no time spent in this state. The same time criterion is applied to states xtoMove and ytoMove of the Board state-chart automata. Again, there is no time spent in these states (we will call such states *discrete states*). Their use in the state-chart is needed only to model the consequence of a valid move achieved by a player that implies the second player to go to the thinking state.

States 0 and 1 of automata X and Y represent respectively *Thinking* and *Waiting* states of players X and Y. States 0 and 1 of automata B represent respectively states *Idle* and *Checking* of the Board. States of automaton U (representing the Umpire) are obtained after the elimination of discrete states.

The generation of events is realized following a robust procedure: for each trigger corresponds an event. If the trigger does not fire an action in another automaton, the event is then a local event. Elsewhere, if the trigger fires an action in another automaton, the event is then a synchronizing event. If the fired action is a trigger for another action in another automaton, the same synchronizing event is used to synchronize all the affected automata. For example, looking to the state-chart of figure 5, the trigger *after(t1)* of Player X fires the action *Move(x)*

of the Board which, in turn, fires the action $Validate(p)$ of the Umpire. In the SAN model, this phenomena is represented by the synchronizing event “ a ” whose rate is equal to $1/t_1$. The SAN event “ b ” is generated in the same way as event “ a ”.

On the other hand, the trigger $after(T)$ of the state-chart Umpire fires the action $outcome(v)$ of the Board. This action may have different parameters’ value in the Board state-chart ($v = 0$ or 1). Thus, in principle, two events are needed to represent the firing of the trigger $after(T)$. However, as the occurrence of $outcome(0)$ and $outcome(1)$ is related to probabilities, a new event decomposition is achieved in order to carry out the probabilities. Thus, the generated SAN events are now $u1q$, $u2q$, $u1p$, $u2p$ that respectively correspond to $outcome(1)[prob q]$, $outcome(0)[prob (1-q)]$, $outcome(1)[prob p]$ and $outcome(0)[prob (1-p)]$. The rate of the each SAN events is equal to $1/T$ times the corresponding probability. In addition, as the occurrence of the action $outcome$ leads the state-chart Board to enter a discrete state, the action taken after leaving these discrete states ($x.move()$ and $y.move()$) is taken into consideration in the SAN model as a consequence of the occurrence of $outcome$. This is due to the fact that the discrete states are not represented in the SAN model. That is why, for example, synchronizing event $u1q$ appears also in the automata X (corresponding to the action $x.move()$ fired in the state-chart Board).

Thus, the SAN model is generated following a procedure that can be easily implemented. On more detail remains to specify in the SAN model concerns values of time parameters, i.e. t_1 , t_2 and T . In addition, the type of the time distribution should also be indicated. SANs are basically Continuous Time Markovian (CTM) formalism. In CTM models, the exponential distribution is usually used taking advantage from its memory-less property which respects Markov theory. However, other time distributions may be also used in the SAN model [14] based on Phase-Type distributions that can approximate any time distribution even deterministic.

As a typical result, we present one performance prediction resulting from the SAN model with exponential distribution. The analyzed performance metric is the average playing time of each players (X and Y) when the parameters values are: $p = q = 0.5$; $t_1 = 2$; $t_2 = 3$ and $T = 1$. The reflection time of player X is less than that of Y ($t_1 < t_2$), the obtained result indicated that the total thinking time of player Y (over the whole time of the game) is 10% more than the thinking time of player X.

The goal from this simple result is just to show the adequateness of our methodology from mapping the UML model to a SAN model. Of course, more sophisticated performance analysis may be applied to more significant and large applications. In summary, our methodology performs extremely well.

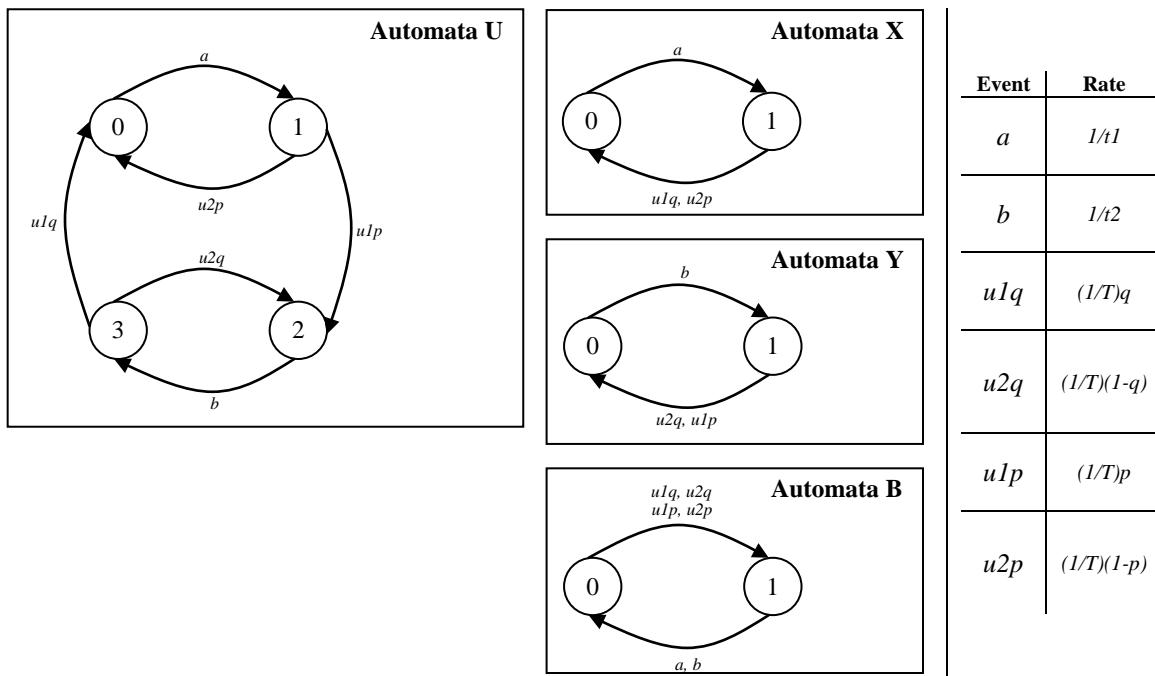


Fig 5: Generated SAN model

5. Conclusion

In this paper we have shown how a stochastic automata network may be directly generated from UML specifications. The key step is by using the state-chart diagram of UML which has a strong similarity to a SAN. Both, SAN and UML state-chart, are a set of automata. SAN's transitions are labeled by local events or/and by synchronizing events. UML's transitions are labeled by triggers that may fire actions in other automata. Mapping a state-chart automaton into SAN automaton is realized after eliminating UML discrete states. Mapping triggers and actions into events is achieved by following the path of actions carried out after the firing of a trigger. An informal description of the SAN generation procedure was presented based on a simple case study which goal is to show the simplicity and robustness of our demarche.

The methodology which we propose in this paper lays the groundwork for the development of a formal heuristic that permits to generate a SAN model from any UML state-chart model. In addition, the generation process is completely transparent to applications' designers. This opens the way for an efficient technique in the software performance engineering area by taking advantage of the power of SAN in analyzing large applications, SAN modularity, and its similarity to the UML state-chart diagram that may allow designers to easily interact with their performance model.

References

- [1] F. Bause and P. Buchholz, "Protocol analysis using a timed version of SDL", In Proceedings of the 3rd Int. Conf on Formal Description Techniques (FORTE' 90), Springer, 1991.
- [2] H. Beilner and F. J. Stewring, "Concepts and techniques of the performance modeling tool Hit", In Proceedings of the European Simulation Multiconference, SCS Europe, 1987.
- [3] Fateh Boutekkouk and Mohammed Benmohammed, "UML for Modelling and performance Estimation of Embedded Systems", Journal of Object Technology, vol. 8, no. 2, March-February 2009, pp. 95-118.
- [4] Leonardo Brenner, Paulo Fernandes, Brigitte Plateau, and Ihab Sbeity, "PEPS2007 - Stochastic Automata Networks Software Tool", QEST 2007: 163-164.
- [5] Paulo Fernandes, Brigitte Plateau and William J. Stewart, "Efficient Descriptor-Vector Multiplications in Stochastic Automata Networks", J. ACM 45(3): 381-414 (1998).
- [6] Martin Fowler and Kendall Scott, UML Distilled, Number ISBN 0-201-32563-2, Addison-Wesley, 1997.
- [7] Carina Kabajunga and Rob Pooley, "Simulating UML sequence diagrams", In Rob Pooley and Nigel Thomas, editors, UK PEW 1998, pages 198-207, UK Performance Engineering Workshop. July 1998.
- [8] P. King and R. Pooley, "Using UML to Derive Stochastic Petri Net Model", UKPEW' 99, Proceedings of the Fiftieth UK Performance Engineering Workshop, 1999.
- [9] Object Management Group, Response to the OMG RFP for Schedulability, Performance, and Time, OMG Document ad/2001-06-14, June 2001, <http://www.omg.org>.
- [10] B. Plateau, "On the stochastic structure of parallelism and synchronization models for distributed algorithms", Proc. of the SIGMETRICS Conference, Texas, 1985, 147-154.
- [11] R. Pooly and P. King, "Using UML to derive stochastic process algebra model", Proceedings of the Fiftieth UK Performance Engineering Workshop, 1999.
- [12] N. Rico and G. v. Bochman, "Performance description and analysis for distributed systems using a variant of LOTOS", in 10th int. IFIP Symposium on Protocol Specification, Testing and Validation, July 1990.
- [13] Vipin Saxena and Manish Shrivastava, "UML Modeling and Performance Evaluation of Multithreaded Programs on Dual Core Processor", International Journal of Hybrid Information Technology Vol.2, No.3, July, 2009.
- [14] Ihab Sbeity, Leonardo Brenner, Brigitte Plateau, and William J. Stewart, "Phase-type distributions in stochastic automata networks", European Journal of Operational Research 186(3): 1008-1028 (2008).
- [15] Ihab Sbeity, Mohamed Dbouk and Brigitte Plateau, "Stochastic Bounds for Microprocessor Systems Availability", IAJIT: International Arab Journal of Information Technology, vol 8, No 1, January 11.
- [16] Connie U. Smith, Performance Engineering of Software Systems. Addison-Wesley, Reading, Massachusetts, 1990.

Ihab Sbeity – received a Maîtrise in applied mathematics from the Lebanese university in 2002, a Master in computer science - systems and communications from the "Université Joseph Fourier" – France in 2003, and a PhD from "Institut National Polytechnique de Grenoble", France in 2006. His PhD work is related to Performance Evaluation and System Design. Currently, Dr. Sbeity occupies a full time position at the Lebanese university – Faculty of Sciences I – computer sciences department. His research interests include modeling and performance evaluation of parallel and distributed computer systems, numerical solution and simulation of large Markov models, UML modeling and Software Performance Engineering (SPE).

Leonardo Brenner - Leonardo Brenner is assistant professor of Computer Science at Université de Provence, France, and do his research at Laboratoire des Sciences d'Information et des Systèmes. He got his Ph.D. (2009) in Computer Science degree from Institut Polytechnique de Grenoble (Grenoble INP), France. His research interests include performance evaluation of computer and communication systems, as well as theory and numerical solution of stochastic modeling formalisms for large Markovian models. His current research topics include structured formalisms, in particular stochastic automata networks and Petri nets, and also practical applications of performance and reliability modeling.

Mohamed Dbouk - Bachelor's Honor" in Applied Mathematics; Computer Science, Lebanese university, Faculty of Sciences (I)-

Beirut, PhD from Paris-Sud 11 University (Orsay-France), 1997.
Specialty: Software Engineering and Information systems,
Performance modeling and optimization, Geographic Information
System & Hypermedia, Datawarehousing and Data mining.
He is a full time Associate Professor, Lebanese university -
Faculty of Sciences (I) - Department of Computer Science.
Academic Teaching: Database (including advanced topics),
Object Orientation, Software engineering, Information System &
Software Architectural Design, Distributed Architecture, Web
Application Architectural Design. Research and Publications in:
GIS, Cooperative and multi-agent systems, interoperability, 3D
Modeling, Groupware, Graphical user interface.

Tetris Agent Optimization Using Harmony Search Algorithm

Victor II M. Romero¹, Leonel L. Tomes² and John Paul T. Yusiong³

^{1,2,3}Division of Natural Sciences and Mathematics,
University of the Philippines Visayas Tacloban College
Tacloban City, Leyte, 6500, Philippines

Abstract

Harmony Search (HS) algorithm, a relatively recent meta-heuristic optimization algorithm based on the music improvisation process of musicians, is applied to one of today's most appealing problems in the field of Computer Science, Tetris. Harmony Search algorithm was used as the underlying optimization algorithm to facilitate the learning process of an intelligent agent whose objective is to play the game of Tetris in the most optimal way possible, that is, to clear as many rows as possible. The application of Harmony Search algorithm to Tetris is a good illustration of the involvement of optimization process to decision-making problems. Experiment results show that Harmony Search algorithm found the best possible solution for the problem at hand given a random sequence of Tetrominos.

Keywords: Harmony Search algorithm, Tetris, Intelligent Agent, Artificial Intelligence

1. Introduction

Problems and challenges have always been part of human life and of the human civilization itself. They define the difference between what is currently in existence and of what could be, after a goal has been achieved. In Computer Science, researchers are concerned with the search for solutions to computational problems and these problems may be categorized into two main classes: P-problems and NP-problems.

P-problems, otherwise known as Polynomial-time problems are problems whose solutions may easily be identified, that is, the procedure for finding the solution is already known. On the other hand, NP-problems are problems whose solutions have no proven optimal way of acquisition. NP-problems are also called "*I know it when I see it problems*" because of the fact that the validity of their solution may only be verified when tried and evaluated [4,5]. A good example of such problem is the creation of an intelligent agent for Tetris [5].

Tetris is a puzzle computer game originally created by Alexey Pajitnov [6]. An intelligent agent for Tetris is a program whose goal is to be able to play the game in the most optimal way possible. In such a case, we only know

the quality of the agent by assessing its performance when it has already played the game. To deal with such problems, where too little detail is known on the nature of a problem's solution, computer scientists use a different approach in the form of meta-heuristic algorithms.

Meta-heuristic algorithms are a primary sub-field of a larger class of algorithms and techniques called stochastic optimization [4]. They are called stochastic optimization because they employ some degree of randomness in searching for a solution. In other words, they are solving problems through a series of intelligent guesses. The primary ideas for achieving the series of intelligent guesses of existing meta-heuristic algorithms are driven by natural occurrences like biological processes and animal behaviors.

The popularity of Tetris has intrigued mathematicians and computer scientists to study its non-trivial nature and reveal its NP-Complete characteristics [5] triggering the motivation for the creation of an intelligent agent. A Tetris intelligent agent is an Artificial Intelligence (AI) program which plays or simulates the game with the goal of clearing as many rows as possible. In fact, in the past years scientists have successfully created intelligent agents using Evolutionary Algorithms [1,2] and Ant Colony Optimization [3].

Harmony Search (HS) algorithm is a meta-heuristic algorithm developed in 2001 by Geem *et al* [7]. It is modeled after the musical improvisation process, wherein a band of musicians continuously tries to create better harmony. This algorithm and its variants [8-9] have been applied to a wide array of real-life optimization problems such as structural design, ecological conservation, industrial operation and musical composition [7], [10-13]. The HS algorithm is a powerful optimization tool because of its ability to discover the high performance regions of the solution space in a reasonable amount of time. In addition, other characteristics enable the HS algorithm to increase its flexibility and produce better solutions, and these are [14]:

1. HS imposes fewer mathematical requirements.

2. HS uses stochastic random searches thus any derivative information is unnecessary.
3. HS creates a new solution vector after considering all of the existing solution vectors.

Since the Harmony Search algorithm has demonstrated its strength on various fields of discipline and has been successfully applied to many problem domains, this study explores the feasibility of using the Harmony Search algorithm in decision-making optimization problems, that is, to use the HS algorithm as the underlying optimization algorithm in facilitating the learning process of the Tetris intelligent agent.

The paper is organized as follows. A brief description about Tetris is presented in Section 2. Section 3 introduces the Harmony Search algorithm followed by a discussion on the proposed HS-based Tetris intelligent agent in Section 4. The experimental results are shown in Section 5 while Section 6 contains the conclusion.

2. The Tetris Game

Tetris is a game originally invented and programmed by Alexey Pajitnov in June 6, 1984 while working in Dorodnicyn Computing Center of the Academy of Science of the USSR in Moscow [6]. It is one of the most popular and most successful games to hit the market. In fact, Tetris' success as a computer game led to the creation of many other variants, sporting slightly different game play. The Tetris game and its variants are basically composed of two main components, the game pieces called Tetrominos and the game board.

The standard Tetris game board has a dimension of 10 x 20 and there are seven Tetrominos or game pieces, as shown in Figure 1 and these are O, J, L, I, S, T and Z. The game pieces differ significantly by the maximum number of rows that they are able to clear simultaneously. In fact, a simple analysis of the game pieces reveals that all are capable of clearing one and two rows. But only pieces "I", "J" and "L" are capable of clearing three rows and ultimately, only the "I" Tetromino is capable of clearing four rows (called "Tetris") [1].

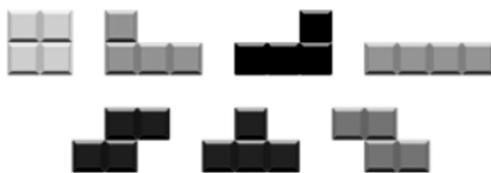


Fig. 1 Tetris Game Pieces-Tetrominos

In Tetris, Tetrominos fall from the top of the game board one at a time and aside from the current Tetromino being

manipulated, an advance view of the next piece is provided to the player. This is to enable the player to manipulate and position the current Tetromino in the game board such that one or more gapless row(s) of block is created. When such a scenario happens, the gapless row is cleared and all existing blocks above that row descend n units, where n is the number of rows cleared which is between 1 and 4. In addition, manipulation of a Tetromino can only be performed in two ways, either by doing a 90 degree rotation or a sideways movement while the Tetromino has not yet reached the bottom of the game board, at which case it fixes itself into position [1].

However, unlike most games, Tetris does not have a win condition, so the game continues until the stack of blocks in the game board disallows the entry of succeeding Tetrominos. This means that the only goal in the game is to be able to clear as many rows as possible for better and longer game play. As a result, staying in the game solely depends upon the number of rows cleared.

3. The Harmony Search (HS) Algorithm

Computer scientists have found a significant relationship between music and the process of looking for an optimal solution. This interesting connection led to the creation of the Harmony Search algorithm. It is a new kind of meta-heuristic algorithm mimicking a musicians' approach to finding harmony while playing music. When musicians try to create music, they may use one or a combination of the three possible methods for musical improvisation which are as follows: (1) playing the original piece, (2) playing in a way similar to the original piece, and (3) creating a piece through random notes.

In 2001, Geem *et al* [7] saw the similarities between the music improvisation processes and finding an optimal solution to hard problems and formalized the three methods as parts of the new optimization algorithm, the Harmony Search algorithm (HS); (1) harmony memory consideration (2) pitch adjustment and (3) randomization. These three methods are the main parameters of the algorithm and play a vital role in the optimization process [7], [10-13].

For musicians, one of the ways of producing good music is considering existing compositions and playing them as they are. In the Harmony Search algorithm, this is also the case, the use of harmony memory is vital as it ensures that potential solutions are considered as elements of the new solution vector. The second way of coming up with good music is by playing something similar relative to an existing composition, in HS this is called the pitch adjustment mechanism and may be referred to as the

exploitation mechanism in the Harmony Search algorithm; it is responsible for generating slightly varying solution from existing solutions. It is comparable to the mutation mechanism in genetic algorithms. Randomization, the last of the methods ensures that the search for solution is not isolated to the local optima. It makes the solution set more diverse by not limiting the search for solution in a confined area and is referred to as the exploration mechanism of the Harmony Search algorithm [7], [10-13].

So, in the Harmony Search algorithm, each musical instrument is represented as a decision variable. The value of each decision variable is set in a similar manner that a musician plays his instrument, contributing to the overall quality of the music created, thus the name Harmony Search. The pseudo-code of the Harmony Search algorithm as presented, shows that the optimization process is done on a per decision variable basis for each harmony (solution) in the harmony memory.

Furthermore, based on the pseudo-code and as shown in Figure 2, the optimization process of the Harmony Search algorithm may be described in three main steps:

1. **Initialization:** Program parameters are defined and the harmony memory is initialized by filling it up with random solutions; each harmony is evaluated using an evaluation or objective function.
2. **Harmony improvisation:** A new solution is created. The three methods of the Harmony Search algorithm are used to decide on the value that will be assigned to each decision variable in the solution.
 - a. **Creation of a new solution:** A new solution is created either (1) randomly with a probability of $1 - r_{accept}$ or alternatively (2) by copying an existing solution in the harmony memory, with a probability equal to r_{accept} .
 - b. **Adjustment:** With a probability of r_{pa} , the elements of the new harmony are then modified.
 - c. Using the objective function, the new harmony is evaluated.
3. **Selection:** When a terminating condition is met, the best harmony (solution) in the harmony memory is selected.

Also, there are several parameters that have to be defined before the start of the optimization process.

- (1) **Maximum number of cycles or iterations** – is the basis for terminating the optimization process.
- (2) **Harmony memory size** – refers to the number of harmonies that will be stored in the harmony memory.

- (3) **Number of decision variables** – each harmony is composed of several decision variables.
- (4) **Harmony Memory Consideration rate (r_{accept})** – determines the rate at which decision variables in the harmony are considered as elements of the new harmony that will be created.
- (5) **Pitch Adjustment rate (r_{pa})** - defines the probability for adjusting the values of decision variables copied from an existing harmony in the harmony memory by adding a certain value.

```

Begin
Define objective function  $f(x)$ ,  $x = (x_1, x_2 \dots x_d)^T$ 
Define harmony memory accepting rate ( $r_{accept}$ )
Define pitch adjusting rate ( $r_{pa}$ ) and other parameters
Generate Harmony Memory (HM) with random harmonies
While ( $t < max\ number\ of\ iterations$ )
    While ( $i <= number\ of\ variables$ )
        If(rand <  $r_{accept}$ )
            Choose a value from HM for the variable  $i$ 
        If(rand <  $r_{pa}$ )
            Adjust the value by adding certain amount
        End if
    Else
        Choose a Random Value
    End if
    End while
    Accept the new harmony (solution) if better
End while
Find current best solution
End

```

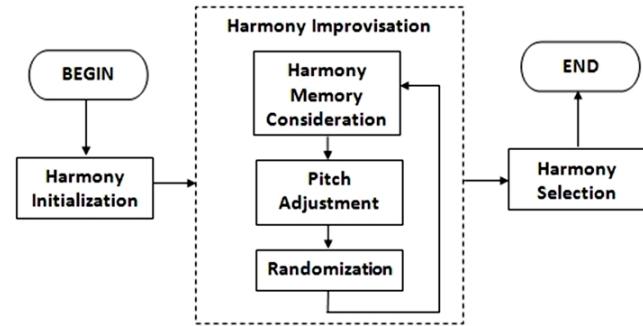


Fig. 2 The Harmony Search Optimization Process

4. HS-based Tetris Intelligent Agent

When a human first plays a Tetris game, he/she may play the game with no more than the goal of clearing as many rows as possible. However, as the game progresses or as he/she continues to play the game repeatedly, it becomes obvious that like every decision-making process in life, in choosing the best move for a game piece in Tetris, there are several factors to put into consideration for the maximization of the number of rows cleared.

Through the efforts of previous researches [1-3], [5], these factors were transformed into forms that can be implemented in computers and are called feature functions. Also, similar to the decision-making process in life, we assign weights to each of these factors to symbolize its significance or bearing to the final decision. Ultimately, the best move for a current piece is chosen using a linear (summation) combination of these feature functions with their corresponding weights, called the *state-evaluation function*, as shown in Eq. (1).

$$V(s) = \sum_{i=1}^{19} w_i f_i(s) \quad (1)$$

where s is the state (board configuration), w_i represents each of the weights of the i^{th} feature function $f_i(s)$ and $f_i(s)$ is a function that maps a state (a board configuration) to a real value. The goal of the optimization process is to be able to find an optimal set of weights that will result in the most number of cleared rows.

The 19 feature functions identified and used in this research are as follows: (1) pile height, (2) holes, (3) connected holes, (4) removed rows, (5) altitude difference, (6) maximum well depth, (7) sum of all wells, (8) landing height, (9) blocks, (10) weighted blocks, (11) row transitions, (12) column transitions, (13) highest hole, (14) block above highest hole, (15) potential rows, (16) smoothness, (17) eroded pieces, (18) row holes, and (19) hole depth.

Thus, the solutions generated by the program will come in the form of a vector of 19 weights, that is, w_1 to w_{19} . Feature functions 1-12 are from [2], 13-16 are taken from [3] and 17-19 are from [1]. A description of the feature functions is presented in Section 4.1.

As illustrated in Figure 3, the proposed HS-based Tetris intelligent agent called Harmonetris, is divided into two main parts;

- (1) the solution optimizer, and
- (2) the Tetris game simulator.

The solution optimizer comprises the Harmony Search algorithm part of the program while the Tetris game simulator is composed of our Tetris agent as well as the Tetris game itself.

The solution optimizer generates harmonies or solutions in the form of a set of weights; normally, w_1 to w_{19} . Each of this set of weights is passed to the Tetris game simulator, which plays one complete game of Tetris using the weights provided and a randomly generated sequence of Tetrominos. After playing the game, the game simulator returns the number of rows cleared by the agent based on the current assigned weights to the solution optimizer. The

number of cleared rows serves as the objective function, where more cleared rows means a better agent performance. When a terminating condition has been met, the solution optimizer then outputs the best solution created so far.

The Tetris game simulator uses the state-evaluation function in Eq. (1) to evaluate on a per Tetromino move basis for each feature function while the objective function rates an entire Tetris game in the form of maximum number of rows (lines) cleared which is the basis for the solution optimizer to determine the objective function value of each harmony.

The main idea behind this set up is to use the Harmony Search algorithm as the solution optimizer's underlying optimization algorithm.

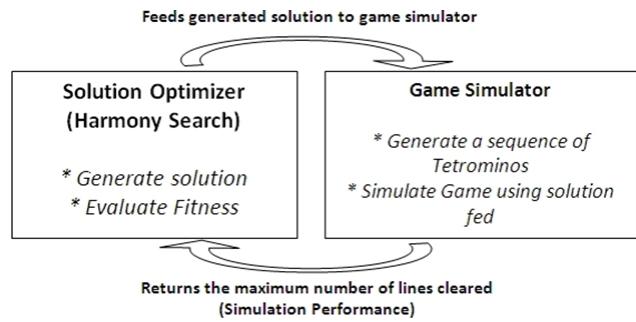


Fig. 3 The Harmonetris Overall Program Flow

4.1 The Feature Functions

A description of each of the feature functions, f_i , used in this research is presented in this section.

1. **Pile Height:** The row of the top most Tetromino in the board. Each of the filled cells reached directly from the top are compared and the row value of the topmost filled cell is the pile height. In Figure 4 the pile height is 13.
2. **Holes:** The number of all gaps with at least one occupied cell above them. In Figure 5(a), the holes in the board are marked with “1”. The number of holes in the board is 10.
3. **Connected Holes:** Similar to Holes, however, counts vertically connected gaps as one. Figure 5 shows the difference between Holes and Connected Holes. Connected Holes has a value of 7 with each connected hole in the board marked with “1”.

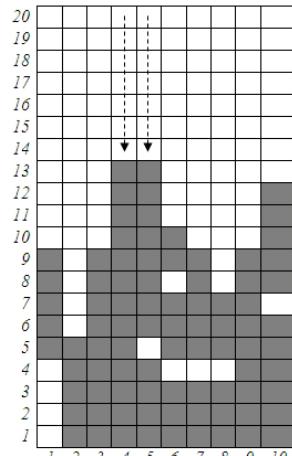


Fig. 4 The Feature Function: Pile Height

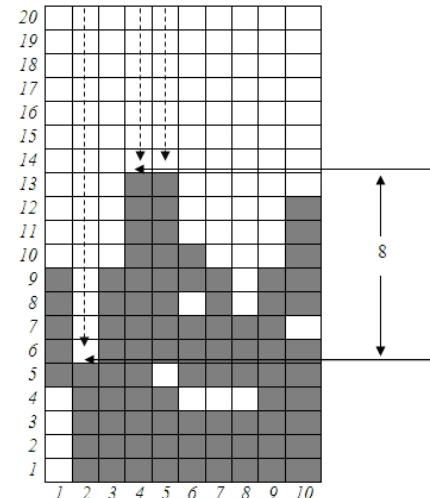


Fig. 6 The Feature Function: Altitude Difference

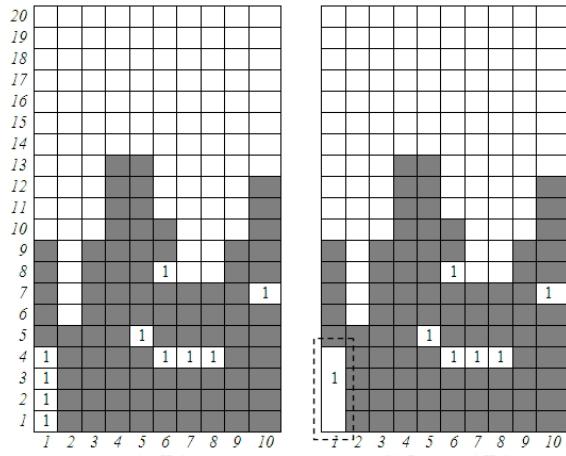


Fig. 5 Difference between Holes and Connected Holes

4. **Removed Rows:** The number of rows cleared by the last step. This is the number of rows that was cleared in order to arrive at the new board configuration.
5. **Altitude Difference:** The difference between the lowest gap directly reachable from the top and the highest occupied cell. Figure 6 has an altitude difference of 8.
6. **Maximum Well Depth:** The depth of the deepest well on the board. This is 4 in Figure 7.

7. **Sum of all Wells:** The sum of all wells on the game board. The board in Figure 7 has a value of 6.

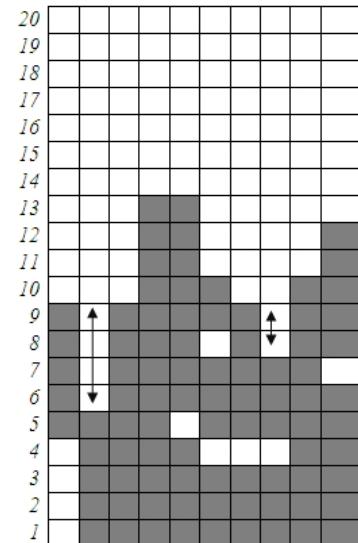


Fig. 7 The Feature Function: Sum of Wells

8. **Landing Height:** The height at which the last Tetromino has been placed. Figure 8 shows that the landing height of piece 'I' is 9.
9. **Blocks:** The number of cells that has been occupied in the board.

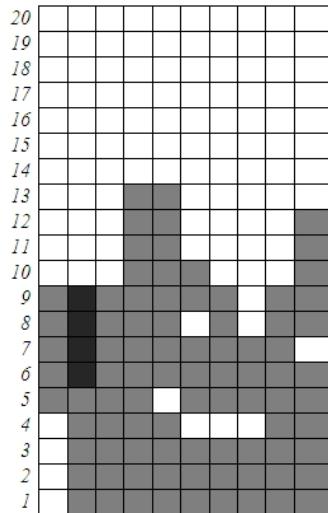


Fig. 8 The Feature Function: Landing Height

10. **Weighted blocks:** Same as Blocks, however counting from the bottom to the top, blocks located at row n count n -times as much as the blocks in row 1. Figure 9 illustrates the weight of each of the blocks.

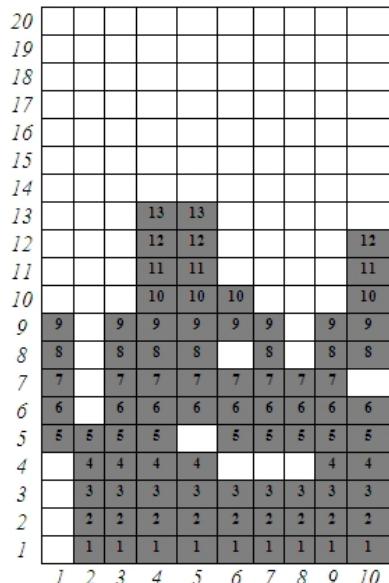


Fig. 9 The Feature Function: Weighted Blocks

11. **Row Transitions:** The sum of all occupied or unoccupied transitions. Each arrow in Figure 10 counts as one row transition.
12. **Column Transitions:** Same as Row Transitions, however, it only counts vertical transitions. Figure 11 provides a clear illustration of the column transitions in the board.

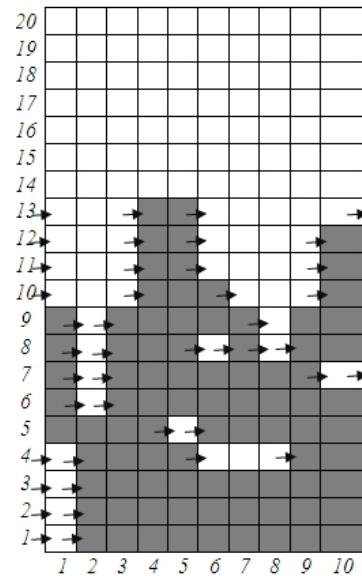


Fig. 10 The Feature Function: Row Transitions

13. **Highest Holes:** The height of the topmost hole on the game board. This is 8 in Figure 5(a).
14. **Blocks Above Highest Hole:** The number of blocks on top of the Highest Hole. In Figure 5(a) the highest hole in the board is at 8, so the number of blocks above it is 2.

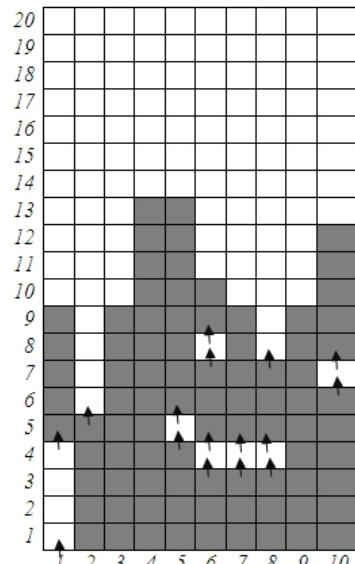


Fig. 11 The Feature Function: Column Transitions

15. **Potential Rows:** The number of rows located above the Highest Hole and in use by more than 8 cells. This is 0 in Figure 5(a).

16. **Smoothness:** The sum of all absolute differences of adjacent column height, as well as the difference of the first and last column. Using the board in Figure 11, Table 1 illustrates how the value of smoothness is computed.

Table 1: Sample Computation: Smoothness

| Columns | Column Heights | Value |
|-------------------|----------------|-----------|
| 1, 2 | $ 9-5 $ | 4 |
| 2, 3 | $ 5-9 $ | 4 |
| 3, 4 | $ 9-13 $ | 4 |
| 4, 5 | $ 13-13 $ | 0 |
| 5, 6 | $ 13-10 $ | 3 |
| 6, 7 | $ 10-9 $ | 1 |
| 7, 8 | $ 9-7 $ | 2 |
| 8, 9 | $ 7-9 $ | 2 |
| 9, 10 | $ 9-12 $ | 3 |
| 10, 1 | $ 12-9 $ | 4 |
| Smoothness | | 27 |

17. **Eroded Pieces:** The number of rows cleared in the last move multiplied with the number of cells of the last piece that were eliminated in the last move.
18. **Row Holes:** The number of rows with at least one hole. Figure 12 shows how the value for row holes is computed. In the said figure, the value is 7.

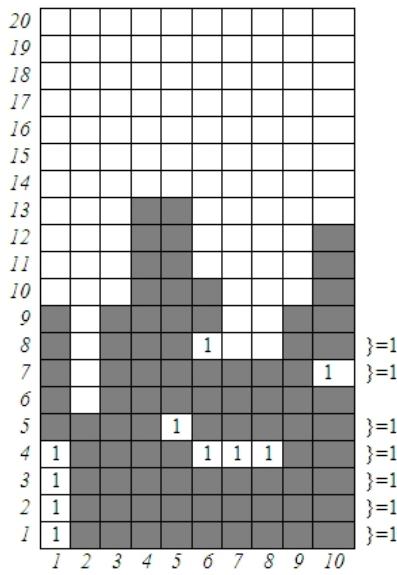


Fig. 12 The Feature Function: Row Holes

19. **Hole Depth:** The number of filled cells on top of each hole. Table 2 shows how to compute for the hole depth based on Figure 13.

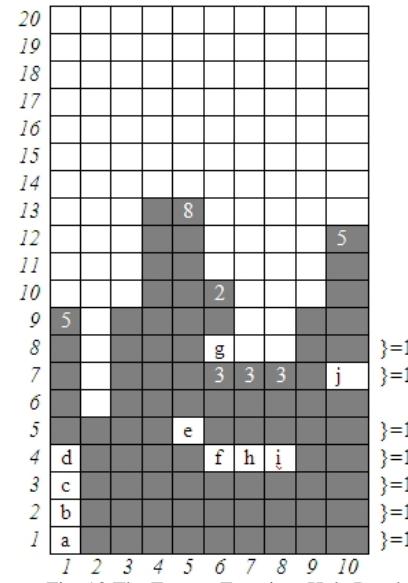


Fig. 13 The Feature Function: Hole Depth

Table 2: Sample Computation: Hole Depth

| Hole | Filled Cells |
|------|--------------|
| a | 5 |
| b | 5 |
| c | 5 |
| d | 5 |
| e | 8 |
| f | $3+2=5$ |
| g | 2 |
| h | 3 |
| i | 3 |
| j | 5 |

4.2 The Tetris Game Simulator

Every generated harmony of the solution optimizer in the form of a vector of weights is fed to the Tetris Game Simulator. The simulator performs a simulation of the game using the feature functions as basis for determining the best move, and then it returns the maximum number of cleared rows, which is the *objective function value* of the harmony.

The Tetris Game Simulator has two main components: (1) Tetris Game and (2) Tetris Agent. The Tetris Game defines the logical characteristics of the game and the rules on how it must be played. The Tetris Agent plays the Tetris Game until a termination condition is satisfied and then returns the required result. The agent always selects the move that will yield the best outcome, which in turn maximizes the result. The simulation will only terminate if the current state of the game satisfies a termination condition. There are two ways to end the simulation. One

is by limiting the number of pieces spawned and the other is waiting for the game to be over.

4.3 The Tetris Agent

In a Tetris Game, one piece is being played at a time. The player has to select from among the 10 possible translations (positions), that is, where to place the current piece. Also, the player may change the orientation (can be up to 4 possible orientations) of the game piece depending on how beneficial this move may be. Once the player has made the choice, it then moves the piece to the desired position and orientation. Afterwards, another piece will be spawned and the player has to repeat the same process. This will continue until the next piece can no longer be spawned, which means that the top row of the Tetris board is already occupied, or the maximum number of spawned pieces has been reached. In a Standard Tetris Game, only the current piece is known. Some other versions provide a preview of the next piece.

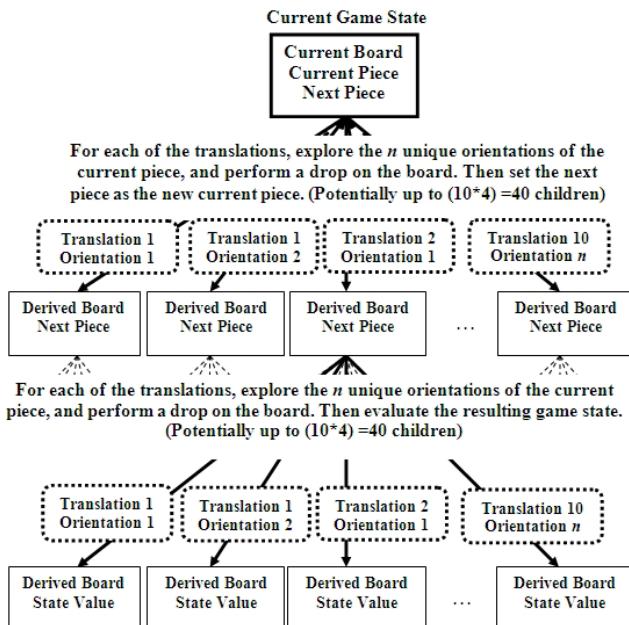


Fig. 14 The Tetris Agent Two-Piece Decision-making Process

In the case of knowing in advance the next piece, known as a *two-piece strategy*, the player may consider it in deciding for the next move. In this case, the player has to enumerate all the possible combinations of translations and orientations of the next piece for each of the derived game state of the current piece's possible moves. Then after that, the board is evaluated based on the preferences of the player. In this case, the state evaluation function is used. To ensure that the game will not end early, the player must select the move that has the highest state value. In this scenario, the player is performing a *greedy strategy*, in

which the player always selects the move with the highest reward (state value). Figure 14 provides a graphical view of the decision-making process.

5. Experiment Results and Discussion

The goal of the experiments is to determine the efficiency of the Harmony Search algorithm as the underlying optimization algorithm for the Tetris intelligent agent and to test the ability of the intelligent agent, *Harmonetris*, in finding the best possible solution with respect to the spawned game pieces. Each setup was subjected to 30 runs to make sure that the results are statistically acceptable. In this experimental setup, the following parameters are defined:

- Memory Improvisation (Number of Cycles) = 100
- Harmony Memory Size (Musical Pieces) = 5
- Harmony Consideration / Acceptance Rate = 0.95
- Pitch Adjustment Rate = 0.99

The results of our first experimental setup determined the performance of Harmony Search algorithm as the underlying optimization algorithm. After executing 120 runs in all, it has been observed that the maximum number of rows that the Tetris agent can clear is determined by Eq. (2).

$$\text{max rows} = (\# \text{ of spawned pieces} / 2.5) - 1 \quad (2)$$

According to C. Fahey [6], the theoretical best case for a Tetris game is to be able to clear one row using 2.5 numbers of spawned pieces. Such special case is illustrated in Figure 15 wherein a sequence of five "O" Tetrominos is able to clear two rows, thus the computed ratio of 5(spawned pieces) / 2(cleared rows) = 2.5.

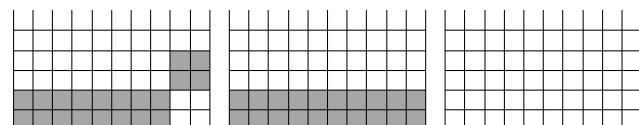


Fig. 15 The Optimal Tetris Case

This theory is taken from the fact that the Tetris board is 10 cells wide and a Tetromino has 4 cells each. This case however, is only applicable in instances wherein the number of spawned pieces is not random, which violates one of the basic specifications of Tetris.

Figure 16 to Figure 19 show the maximum number of cleared rows with respect to number of cycles over 30 runs with 100, 300, 500 and 1,000 spawned pieces, respectively.

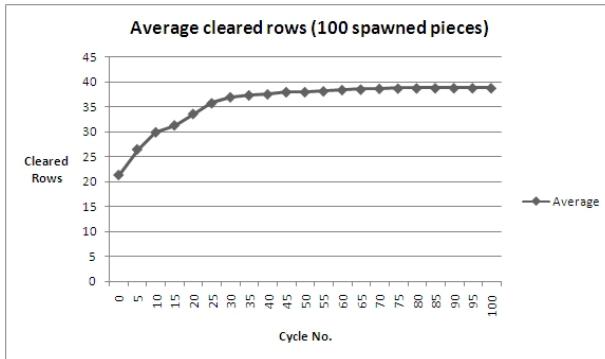


Fig. 16 Maximum number of cleared rows for 100 spawned pieces

The figures show the results of the experiments averaged over 30 runs. It is the average of the maximum number of cleared rows obtained by the best harmonies on all runs for a given cycle.

It can be observed from Figure 16 that the number of cleared rows approaches but does not reach 40. Thus, the maximum number of cleared rows in 30 runs for the 100 spawned pieces is 39.

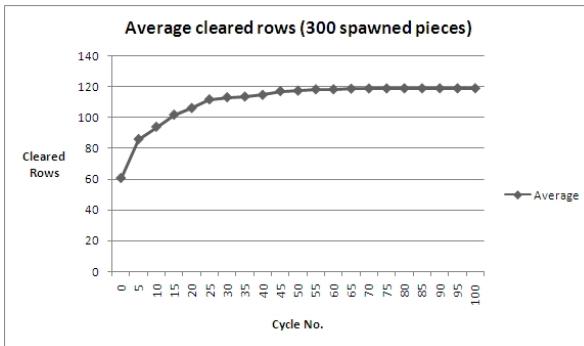


Fig. 17 Maximum number of cleared rows for 300 spawned pieces

Figures 17 to 19 also confirm the validity of Eq. (2), that is, the maximum number of rows that the Tetris agent can clear is determined by the said equation. Table 3 summarizes the experiment results obtained for the different setups.

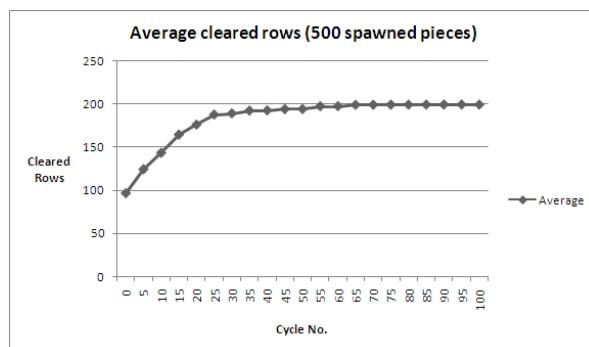


Fig. 18 Maximum number of cleared rows for 500 spawned pieces

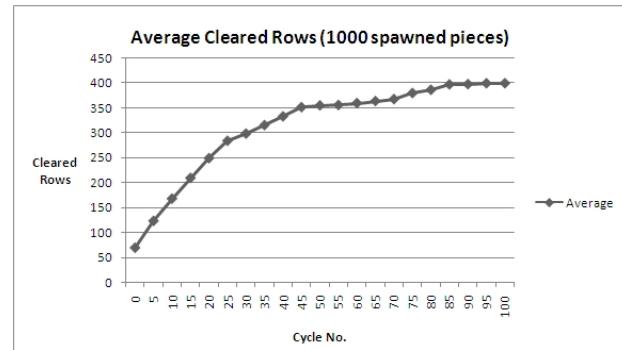


Fig. 19 Maximum number of cleared rows for 1,000 spawned pieces

Table 3: Spawned Pieces and Maximum number of Cleared Rows

| Total Number of Spawned Pieces (SP) | Maximum number of Cleared Rows (CR) | Eq. (2) (SP/2.5) - I | SP to CR Ratio |
|-------------------------------------|-------------------------------------|----------------------|----------------|
| 100 | 39 | 39 | 2.56410 |
| 300 | 119 | 119 | 2.52100 |
| 500 | 199 | 199 | 2.51256 |
| 1,000 | 399 | 399 | 2.50626 |

Another experiment was conducted to determine the best configuration for the feature functions, thus the terminating condition was set to “Game Over” only, with no restriction on the number of cycles and spawned pieces.

After allowing the program to run for two weeks straight, the Tetris agent was able to obtain the following harmonies, x_i , on its 304th cycle of harmony improvisation. Table 4 shows the results obtained by the 5 harmonies on the 304th cycle.

Table 4: The Performance of the Five Harmonies on the 304th cycle

| Harmony | Maximum number of Cleared Rows (CR) | Total Number of Spawned Pieces (SP) | SP to CR Ratio |
|---------|-------------------------------------|-------------------------------------|----------------|
| x_1 | 291,087 | 727,751 | 2.500115085867 |
| x_2 | 300,277 | 750,723 | 2.50010157288 |
| x_3 | 337,254 | 843,168 | 2.500097849098 |
| x_4 | 348,047 | 870,151 | 2.50009625136 |
| x_5 | 416,928 | 1,042,354 | 2.50008154885 |

As shown in Table 4, it can be observed that as solutions improve resulting to more cleared rows, the ratio of spawned pieces (SP) to cleared rows (CR) approaches the value of our theoretical best game play of 2.5.

Furthermore, analysis on the harmonies (weight configurations) reveals that the Tetris agent, in its attempt to come up with a better solution, gave emphasis on reducing the weight values of the number of holes, wells,

column transitions and row transitions, and at the same time increasing the weight value of potential rows.

6. Conclusions

In this paper, the researchers showed the efficiency of the Harmony Search algorithm as the underlying optimization algorithm for the Tetris intelligent agent and tested the ability of the intelligent agent in finding the best possible solution with respect to the random sequence of spawned game pieces. Experiment results reveal that the Harmony Search algorithm is an efficient optimization algorithm and the *Harmonetris*, the Tetris agent, is able to generate the best possible solution.

The best harmony (weight configuration) found in a span of two weeks was able to clear 416,928 rows in 1,042,354 spawned pieces yielding the Spawned Pieces to Cleared Rows ratio (SP/CR ratio) of 2.50008154885256. Thus, it can be observed that as the number of cleared rows increases, the SP/CR ratio approaches the optimum value of 2.5.

References

- [1] A. Boumaza, "On the Evolution of Artificial Tetris Players", In Proceedings of the 5th International Conference on Computational Intelligence and Games, 2009, pp. 381-393.
- [2] N. Böhm, G. Kókai and S. Mandl, "An Evolutionary Approach to Tetris", MIC 2005: The Sixth Metaheuristic International Conference, 2005.
- [3] X. Chen, H. Wang, W. Wang, Y. Shi and Y. Gao, "Apply Ant Colony Optimization for Tetris", In Proceeding of the 2009 Genetic and Evolutionary Computation Conference, 2009.
- [4] P. Collet and J.P. Rennard, Stochastic Optimization Algorithms, Handbook of Research on Nature Inspired Computing for Economics and Management, Hershey: IGR, 2007.
- [5] E.D. Demaine, S. Hohenberger, and D. Liben-Nowell, "Tetris is Hard Even to Approximate", In Proceedings of the 9th Annual International Conference on Computing and Combinatorics, Lecture Notes In Computer Science, 2003, pp. 351-363.
- [6] C.P. Fahey, Tetris, http://www.colinfahey.com/tetris/tetris_en.html.
- [7] Z.W. Geem, J.H. Jim and G.V. Loganathan, "A new heuristic optimization algorithm: Harmony Search", Simulation, 2001, 76(2), pp. 60-68.
- [8] X-Z. Gao, X. Wang and S.J. Ovaska, "Uni-modal and Multi-modal Optimization Using Modified Harmony Search Methods", International Journal of Innovative Computing, Information and Control, Volume 5, Number 10(A), 2009, pp. 2985-2996.
- [9] Z. Kong, L. Gao, L. Wang, Y. Ge and S. Li, "On an Adaptive Harmony Search Algorithm", International Journal of Innovative Computing, Information and Control, Volume 5, Number 9, 2009, pp. 2551-2560.
- [10] Z.W. Geem, "State-of-the-Art in the Structure of Harmony Search Algorithm", Studies in Computational Intelligence, vol. 270, 2010 pp. 1-10.
- [11] Z.W. Geem, M. Fesanghary, J-Y. Choi, M.P. Saka, J.C. Williams, M.T. Ayvaz, L. Li, S. Ryu and A. Vasebi, Recent Advances in Harmony Search (Studies in Computational Intelligence), 1st edition, 2010.
- [12] K.S. Lee and Z.W. Geem, "A New Meta-Heuristic Algorithm for Continuous Engineering Optimization: Harmony Search Theory and Practice", Computer Methods in Applied Mechanics and Engineering, 194(36-38), 2005, pp. 3902-3933.
- [13] S-B. Rhee, K-H. Kim and C-H. Kim, "Harmony Search Algorithm for Emission Considering Economic Load Dispatch Problems in Power Systems", International Journal of Innovative Computing, Information and Control-Express Letters, vol. 3, Issue 4(B), 2009, pp. 1269-1274.
- [14] X-S. Yang, "Harmony Search as a Metaheuristic Algorithm", Music-Inspired harmony Search Algorithm-Theory and Applications, vol. 191, 2009, pp. 1-14.

Scenario-Based Software Architecture for Designing Connectors Framework in Distributed System

Hamid Mccheick¹, Yan Qi² and Hafedh Mili³

¹ Computer Science Department, University Of Quebec At Chicoutimi,
Chicoutimi, Quebec G7H2B1, Canada

² Computer Science Department, University Of Quebec At Chicoutimi,
Chicoutimi, Quebec G7H2B1, Canada

³ Computer Science Department, University Of Quebec At Montreal,
Montreal, Quebec H3C3P8, Canada

Abstract

Software connectors is one of key word in enterprise information system. In recent years, software developers have facing more challenges of connectors which are used to connect distributed components. Design of connectors in an existing system encounters many issues such as choosing the connectors based on scenario quality, matching these connectors with design pattern, and implementing them. Especially, we concentrate on identifying the attributes that interest an observer, identifying the functions where these connectors could be applied, and keeping all applications clean after adding new connectors. Each problem is described by a scenario to design architecture, especially to design a connector based on architecture attributes. In this paper, we develop a software framework to design connectors between components and solution of these issues. A case study is done to maintain high level of independency between components and to illustrate this independency. This case study uses Aspect-Oriented Programming (AOP) and AspectJ, Design Pattern to and Program Slicing to solve main problems of design of connectors. A conclusion is given at the end of this paper.

Keywords: Design Connector, Attribute Driven Design, Software architecture, Scenario based system.

1. Introduction

1.1 General

In enterprise information system, each connector has a protocol specification that defines its properties [1]. Connectors can be also understood as some software elements which provide a conduit from one or many components to one or many components. The connector may also adapt the protocol and format of the message from one component to another [2]. Our design connectors process could be described as follows:

- choosing connectors based on scenario quality and stimulus described in software architecture [16],
- matching these connectors with design pattern and finally,
- implementing them using a programming language using aspect-oriented programming.

In software developing history, there are some classic connectors which are often used in desktop software system. These connectors are function call (method call), association class, class inheritance and share memory etc. These basic connectors work very well among components which are situated in a same application. Software developers always neglect them, because they do not need more code to implement them (even no code).

However, in the recent years distributed system is a growing trend. The distributed programs are making the software architecture and their connectors become more and more complicated. A distributed system is a system composed of several computers which communicate through network, hosting processes that use a common set of distributed protocols to assist the coherent execution of distributed activities [4]. The connectors situated in distributed system play a significant role in whole software architecture. Software engineers must face more challenges of connectors which are used to connect distributed components.

In this paper we mainly concentrate on the design of connectors in distributed system. Especially, we concern ourselves about the software maintenance in which new connectors are designed and added among existing components. Indeed, the idea is choosing the connectors, identifying the attributes that interest the observers, identifying the functions where these connectors could be

applied, and keeping all applications clean after adding new connectors.

We propose a set of methods to design and add the connectors. Among the methods, Aspect-Oriented Programming (AOP), Design Pattern and Program Slicing are three important technologies and they are described in detail.

1.2 Statement of the Problems

Software engineers must face more and more challenges of connectors which are used to connect distributed components, when they analyze, design, implement and maintain connectors. Naturally, this general problem raises many question: What is the reason to make connector design, implementation and maintenance become more difficult? How can we make this connector easy to reuse, replace and maintain? The answer described below to these questions leads to framing the problem area for our research:

The area of connector design and maintenance is seldom researched, especially, the maintenance of connector. Connector maintenance focuses on how to update or add connectors in an existing software system and how to keep the existing components clean instead of messing it after bringing new connector out. And this problem is quite different to design a connector for a new system.

1.3 Research Questions

The problems of our research can introduce the specific questions of research. In the following sections, we focus on two questions which software developers often meet.

- How to add the code related to new connector without modifying existing source code very much as far as possible? In other words, we should try our best to keep the existing components clean instead of messing it after bringing new connector out. And the component messed will become incompatible with the rest of your system [3].
- Where automatically to find the interesting points (or statements), which the other components want to know through connectors, in the existing source code? That means that we should determine the place of status or values in legacy component which should be provided to a new component. It is often difficult for the developer to grasp the basic architecture and relationships between code units, and this is made more difficult by the fact that the code may be poorly documented and poorly written [11].

In other words, the connectors should be added in distributed system without modifying more existing source code. And the point of source code which will be inserted a connector should be accurately found. For example, to maintain some large source code, if the software developers want to add a new connector between components, they must carefully analyze those codes and avoid messing the existing design and implementation. To implement these issues, we choose Aspect-Oriented Programming and Slicing tool techniques. A case study is done to proof the concept of maintaining the independency of components and then reduce the cost of the maintenance

1.4 Structure of Papers

The rest of the paper is organized as follows. We firstly present the overview of AOP, design pattern and programming slicing technique (section 2). Secondly, a mode for design of connectors is proposed in distributed system (section 3). Thirdly, an implementation based on AOP, design pattern and programming slicing are shown (section 4). Finally, a case study is given to illustrate that mode and those technologies (section 5) and a conclusion is given in section 6.

2. Background

Software maintenance in software engineering is the modification of a software product after delivery to correct faults, to improve performance or other attributes [5]. Designing a new connector for an existing distributed system belongs to software maintenance. And it will be a big job. Particularly, when the developers are not familiar with the system and they don't have enough documents about the project, they must spend much time analyzing and handling the existing source code. Software developers sometimes may put themselves in bad situation: they cannot find the right point (or statement) in large and complex source code files to provide the information for other components through connector. Or after adding a new connector, the source code becomes more and more difficult to read and understand. In this section we describe three important technologies: Aspect-Oriented Programming, Design Pattern and Program Slicing. They can be used to avoid the problems mentioned above

2.1 The classification of Connectors

In this section, we discuss the connector type, the interaction service between components and the usage of the existing connector.

There are eight types of software connectors [6]:

1. Procedure Call,
2. Event,

3. Data Access,
4. Linkage,
5. Stream,
6. Arbitrator,
7. Adaptor, and
8. Distributor.

The service categories provided by connector are described as below [6]:

- *Communication*: Communication connectors support transmission of data among components.
- *Coordination*: Coordination connectors support transfer of control among components.
- *Conversion*: These connectors convert the interaction required by one component to that provided by another.
- *Facilitation*: Facilitation connectors mediate and streamline component interaction.

Those four services provided by connector are requirements of component interactions. Different connector types can satisfy the requirements of the interactions between components. From Mehta et al.'s research [20], we can find that different connector types can and only can support specific service. In other words, one kind of type can only be used in some particular circumstances. For example, one connector which belongs to type Procedure Call is suitable to use for a situation in which two components need communication service. On the contrary, using a Procedure Call connector to provide a conversion service must be a bad idea.

However in Balek et al.'s research work [20], Balek et al., discuss that Mehta et al.'s taxonomy does not talk about the how to design connector types, and that those connector types are at different software levels. For example procedure calls are the assembly language of software interconnection [1]. It is not suitable to be put it together with event or data access.

2.2 Aspect-Oriented Programming (AOP)

Aspect-oriented programming is a paradigm that supports two fundamental goals: [7]

- Allow for the separation of concerns as appropriate for a host language.
- Provide a mechanism for the description of concerns that crosscut other components.

AOP isn't meant to replace OOP or other object-based methodologies. Instead, it supports the separation of components, typically using classes, and provides a way to separate aspects from the components [7].

We can use AOP to crosscut classes (components) to setup relationships between them without modifying functions in the code of original components. In other words, the code can be kept clean and independent by using Aspect-Oriented Programming.

2.3 Design Patterns

Design patterns are always used to describe relationships and interactions between components (classes or objects). The design patterns in Gamma et al.'s book [10] are descriptions of communication of objects and classes that are customized to deal with general design problems. And the communication of object and class should be loose coupling without becoming entangled in each other's data models and methods [8]. Gamma et al.'s design patterns particularly deal with problems at the level of software design, especially object-oriented software design. They can be classified by criterion scope which is used to specify whether the pattern is mainly used to classes or objects. So the design patterns have two kinds: class design pattern and object design pattern [10]. Object patterns are applied to object relationships, which can be modified at run-time and are more dynamic, such as Proxy Pattern, Observer Pattern etc.

In our research we primarily focus on the object design patterns and extend the area of the design patterns which are applied to distributed architecture (especially the messaging-oriented system). For example, the Observer pattern can be reinterpreted and redesigned in distributed architecture, which is sometimes called publish-subscribe style. In order to describe the relationship of components, we use Aspect-oriented programming (AOP).

2.4 Spring Framework

The Spring framework is a wide-ranging framework to develop enterprise Java applications. It provides a lightweight solution and a potential one-stop-shop for building enterprise-ready applications [18]. The Spring Framework is organized as modular. These modules are grouped into 6 types [18]: i) Core Container, ii) Data Access/Integration, iii) Web, iv) AOP (Aspect Oriented Programming) Instrumentation, and v) Test.

Core container is one of most important modules. Especially, in core container, Inversion of Control (IoC, it is also known as dependency injection) is key feature of Spring framework. Dependency Injection is based on Java language constructs, rather than the use of framework-specific interfaces. Instead of application code using framework APIs to resolve dependencies such as configuration parameters and collaborating objects, application classes expose their dependencies through methods or constructors that the framework can call with the appropriate values at runtime, based on configuration [19].

A connector situated in a distributed system must be based on certain transport mechanism (such as TCP/UDP socket, messaging system...) to connect components in network. According to some situation, the distributed connector needs to change the transport mechanism during the

runtime. For example, when the quality of network becomes more stable, the system has the intention of dynamically changing the TCP socket to UDP socket. For reuse of design (and code) and flexible loading mode of java class (or bean), we apply IoC of Spring framework to achieving this objective.

2.5 Program Slicing

Program slicing is a well-known program analysis and transformation technique that uses program statement dependence information to identify parts of a program that influence or are influenced by an initial set of program points of interest (called the slice criteria) [11]. A program slice constructed by identifying program points that affect a given program point is called a backward slice. A program slice composed of program points affected by a program point is called a forward slice [13].

Engineers apply program slicing technique in lots of programming area:

- Program Debugging;
- Program Comprehension;
- Program Testing.

When starting to design a connector, we firstly want to find a point (or statement) in one source code of component. The information included in that point will be notified to another component through connector. Fortunately, the finding process can be automatically done by using backwards program slicing starting with a variable as the slicing criteria.

2.6 Attribute Driven Design

The Attribute-Driven Design (ADD) method is an approach to defining a software architecture in which the design process is based on the quality attribute requirements the software must fulfill [16]. ADD is situated after requirements analysis and before high level design. So the input to ADD is a set of requirements which are regarded as a set of quality attribute scenarios; and the output is the input of high level design. There are 8 steps performed when designing an architecture using the ADD method [21]. These steps are described in Figure1.

The quality attribute scenarios mainly include availability, modifiability, performance, security, testability and usability. Among the steps of ADD, analyzing those quality attribute scenarios is the most important step of the design method.

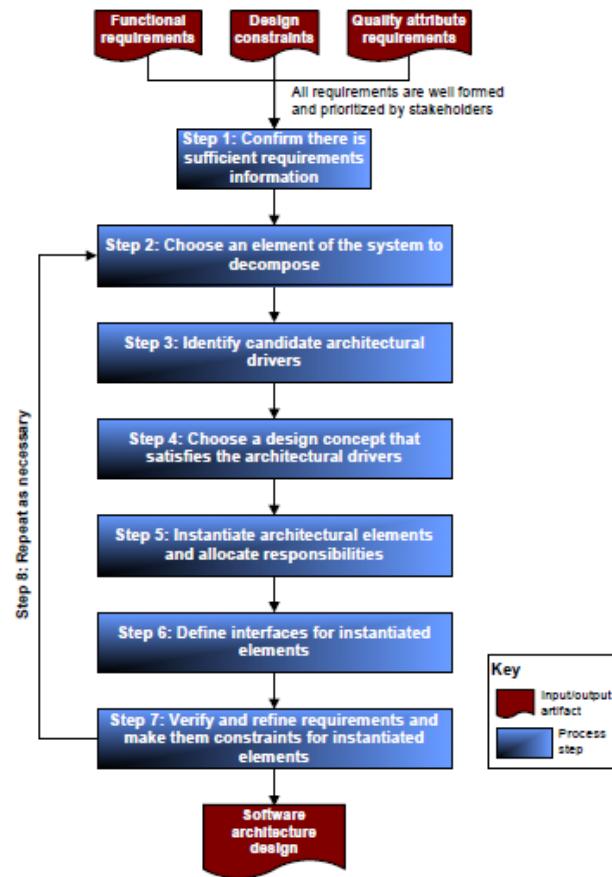


Fig. 1 Software architecture design process.

3. Model for Design of Connectors in Distributed System

In this section, we present a mode for design of connectors in distributed system. In distributed system, components are situated in separate node, which may be computer, PDA or server etc. Figure 1 shows distributed components (A, B, C and D) which are situated in different node (1, 2 and 3) in network.

In Figure 1, the lines are actually software connectors which are used for communication among components. In other words, connectors can be applied to describing the relationship between components (or distributed components).

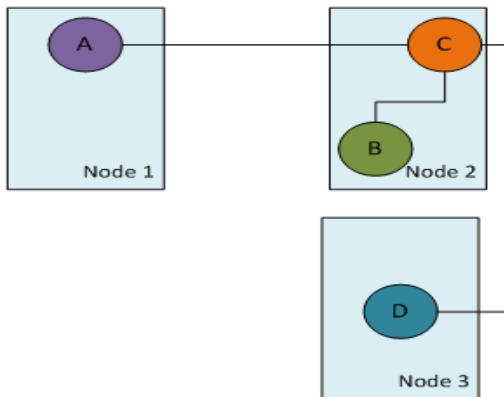


Fig. 2 Distributed components in different nodes.

We introduce a new model for design connectors based on the distributed architecture (Figure 2). According to some design pattern, component D wants to get some value or status of component A. So we must add a new connector to describe the relationship between components A and D.

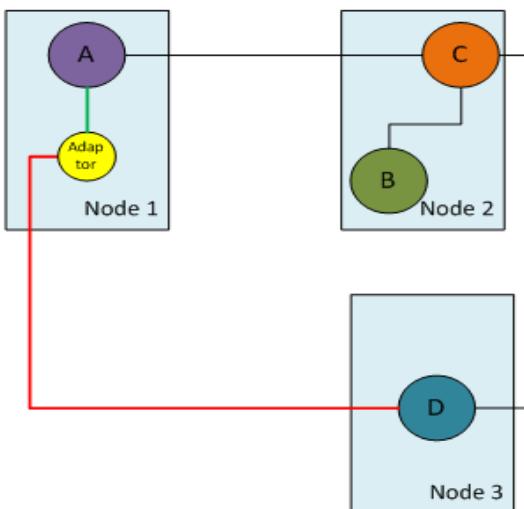


Fig. 3 New added connector between components.

In the figure 3, a new connector is described by two lines (a green and a red one) between component A and D. In other words, the new connector consists of two sub-connectors. The following diagram (figure 4) gives a description of the new connector. We add a new component which runs in Node 1 as an Adaptor. Aspect-oriented programming technique is applied to the sub-connector between component A and Adaptor according to design patterns. We can get the benefit from the AOP technique to deal with communication between A and Adaptor without any modification.

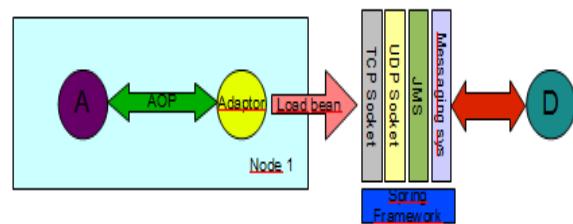


Fig. 4 The new connector.

The only thing we need to do about component A is to analyze the source code. That's mean we should find the interesting statements which are used to crosscut the component. Programming slicing technology is applied to analyze the source code. As for this kind of slicing, we can use backward slice to get data dependence among the source code. Another sub-connector between Adaptor and component D is usually implemented by using common network protocol or a messaging system, such as TCP (/UDP)/IP, JMS, etc. In order to support the switch dynamically from one transport mechanism to another, we use Spring framework to configure these java beans at run time.

At last we draw a conclusion about the mode – how to design a message-based connector in distributed system.

- i) Analyze the relationship between two components;
- ii) Choose a design pattern to describe the relationship;
- iii) Slice source code of component to find the place of the statement which another component is interested in;
- iv) Crosscut the component by using AOP;
- v) Add an adaptor to connect two different protocols;
- vi) Configure the different transport bean by using Spring.

This model can be not only applied to distributed system, but desktop application as well.

4. Implementation of Connector Model

The implementation of design of connector consists mainly of developing tools used by the mode. Details of implementation are given in case study (section 5.2). Note that the appendix shows the implementation of the protocol (connector) between two components.

4.1 AspectJ and Gang-of-Four Design Patterns

AspectJ is an implementation of AOP for the Java language built as an extension to the language. A compiler and a set of JAR files take common Java code and AspectJ

aspects and compile them into standard Java byte-code, which can be executed on any Java-compliant machine [7]. The Gang-of-Four (GoF) design patterns [10] offer flexible solutions to common software development problems. Each pattern is comprised of a number of parts, including purpose/intent, applicability, solution structure, and sample implementations [12]. GOF provides 23 well-known design patterns. Software developers can use one of them or combine some of them to describe the relationship between components. Design of connectors depends on it. Some results show that using AspectJ improves the implementation of many GoF patterns [12].

4.2 Indus Java Program Slicer and Kaveri

Indus Java program slicer is a part project of Indus [14]. The Indus slicer is the first and only publicly available Java slicing framework, and can handle almost all features of Java [11].

Kaveri is an eclipse plug-in front-end for the Indus Java slicer. It utilizes the Indus program slicer to calculate slices of Java programs and then displays the results visually in the editor. Kaveri is an effective tool for simplifying program understanding, program analysis, program debugging and testing [15] [17].

Using Kaveri plug-in can automatically slices the Java source code and provides the data and control dependence. According to the results of data and control dependency, developers can easily find the crosscutting points for design connectors.

5. Case Study

In this section, we present one case study to design connectors in distributed system. Firstly we do a design process by using ADD approach; then the actual design and implementation of the case are discussed by using the model described above and related tools.

5.1 ADD Design Process

In this section we adopt Attribute-Driven Design (ADD) approach to define a distributed software architecture in which the design process of the new connector is based on the quality attribute requirements the software must fulfill. The approach can follow a recursive process that decomposes a system or system elements by applying architectural tactics and patterns that satisfy its driving quality attribute requirements.

In the case study, we only focus on an important system quality attribute: modifiability scenarios. By following the ADD method, we discuss the steps of design in more detail.

Step1, choose the module to decompose. The existing system which runs on desktop is targeted, since it is the system's primary element.

Step2, choose the architectural drivers. As the input of ADD, one quality scenarios are chosen as the quality requirement: How to add the code related to new connector without modifying existing source code very much as far as possible? In other words, we should try our best to keep the existing components clean instead of messing it after bringing new connector out. And the component messed will become incompatible with the rest of your system. We present the possible value for each portion of modifiability scenario.

Table 1: Quality Attribute Scenario

| Modifiability Quality Attribute Scenario | |
|--|---|
| Portion of Scenario | Possible Value |
| Source | Developer and system administrator. |
| Stimulus | Hope to add a connector to setup the relationship between the two existing modules. |
| Artifact | Software structure and system environment. |
| Environment | Design time |
| Response | Makes modification without messing the existing source code |
| Response Measure | How much source code is modified |

Step 3, choose an architectural pattern. We choose Publish-Subscribe architectural pattern.

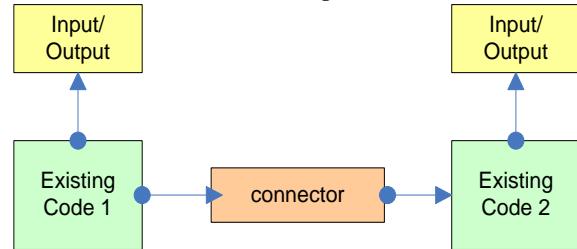


Fig. 5 Connector between components.

Step 4, instantiate modules and allocate functionality with module decomposition view.

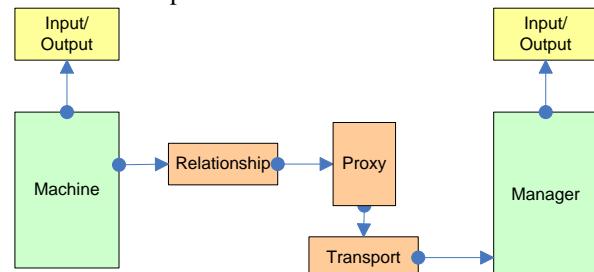


Fig. 6 Proxy between components.

Step 5, define interfaces of the child modules.

- Module Relationship provides the service to setup Observer design pattern;
- Module Proxy is the agent of Machine and it can produce the information;
- Module Transport carries the information produced by Proxy.
- Module Manager receives and consumes the information.

Step 6, Verify quality scenarios as constraint for the child modules. The modifiability quality attribute scenario can be satisfied by the multiple child modules. The Relationship module is designed to setup the interaction without modifying module Machine; Transport module is responsible for linking Manager and Proxy together.

5.2 Actual Design and Implementation

In this sub-section, following the design result of ADD, we present an example to design connectors by using the model and developing tools described above in order to reduce the cost of maintenance to compose software based on legacy components. In the example, we use Java and AspectJ as main developing language; use Observer design pattern [12] and UDP/IP network protocol to describe the relationship of components to implement a new connector; use Kaveri in Eclipse to slice a source code to get the “pointcut” for connector. Some key parts source code of the existing system is described as below:

Class AppSystem

```
public class AppSystem {
    public static void main(String[] args) {
        Machine OneMachine = new Machine();
        byte i = 0;
        double speed = 0.0;
        OneMachine.method4();
        OneMachine.methodF();
        OneMachine.method2();
        OneMachine.methodG();
        speed = OneMachine.getSpeed();
    }
}
```

In the simple application, there are two classes: Machine and AppSystem. Now we decide to upgrade simple application to a distributed system based on some requirements. Because another component named “Manager” which runs in another computer in network wants to be notified when the speed of the Machine has

been changed. Apparently, a new connector should be added in the system.

After analyzing the relationship between Machine and Manager, we decide to choose Observer design pattern to describe the relationship. In order to use Observer design pattern and AspectJ, we should firstly find the point around which the speed is changed. The point is also called pointcut in AOP.

Class Machine:

```
public class Machine {
    private double speed = 0.0;
    private double temperature = 0.0;
    public double getSpeed(){
        double ret = 0;
        ret = speed;
        return ret;
    }
    ...
    public void methodF(){
        method1();
        methodH();
        method2();
        method1();
    }
    public void methodG(){
        method2();
        methodH();
        method2();
    }
    public void methodH(){
        speed = speed + 1;
    }
}
```

We assume that the code of Machine is very large and complex. So if we manually scan the whole source code, we may be get the wrong statement or miss some points. So we do a program slicing for the system by using Kaveri. We pick statement “speed = OneMachine.getSpeed();” as a criteria. And I choose “value of the expression”. Then I start the process by using backward program slicing. The slice result is described as below:

Class Machine:

```
public class Machine {

    private double speed = 0.0;
    private double temperature = 0.0;
    public double getSpeed(){
        double ret = 0;
        ret = speed;
        return ret;
    }
    public void methodF(){

    }
}
```

```
        method1();
        methodH();
        method2();
        method1();
    }
    public void methodG(){
        method2();
        methodH();
        method2();
    }
    public void methodH(){
        speed = speed + 1;
    }
}
```

Class AppSystem

```
public class AppSystem {
    public static void main(String[]
args) {
    Machine OneMachine = new
Machine();
    byte i = 0;
    double speed = 0.0;
    OneMachine.method4();
    OneMachine.methodF();
    OneMachine.method2();
    OneMachine.methodG();
    speed =
OneMachine.getSpeed();
}
}
```

At last, we add a new class Adaptor in that simple application. The Adaptor works as a role of Observer and relays information about speed to component Manager. The Machine works as a role of Subject. So a new connector between Machine and Manager is designed and added in that distributed system. After adding all code about new connector, the existing component Machine is not changed. It is kept very clean in this situation. We just added some new files to the system which can be easily removed or changed without affecting the logic of original components.

6. Conclusion

This paper proposes a model and discusses the design of connectors in distributed system. At first, we present two problems when adding a new connector: 1) How to keep all source code clean after adding new connectors? and 2) Where automatically to find the place of status or values in legacy component which should be provided to a new component through connector? Then a connector

model is proposed to deal with the problems. In this model, AOP, design pattern and programming slicing are combined to resolve the problems of design of connector. Although, a case study is done to show the independency between components. It illustrates how to use the combination of these technologies to support high level of independency. For example, the example shows that it is easy to find a point in a complex source code and that the original code is not almost changed after being added a new connector.

In our future work, we are planning to analyze all kinds of connectors according to different design patterns. And we will provide different and generalized solutions of design connectors based on different results of analysis.

Appendix

This Appendix shows the implementation of the protocol (connector) between two components given in section 5. This protocol is implemented in two artifacts: ObserverProtocol and ObserverProtocolImpl.

```
package connector;

/*
 * This file implements the Observer
Protocol.
*/
import java.util.WeakHashMap;
import java.util.List;
import java.util.LinkedList;
import java.util.Iterator;

public abstract aspect ObserverProtocol
{
    /*
To set which class can be subject. */
    protected interface Subject { }

    /*To set which class can be
observers. */
    protected interface Observer { }

    /* Stores the mapping between
Subjects
    * and Observers. For each
subject, a
    * LinkedList is of its observers
is stored.*/
    private WeakHashMap
perSubjectObservers;

    /*

```

```

        * Returns a Collection of the
        observers of
        * a particular subject.
        * param s: the subject for which
        to return the observers
        * return a: Collection of
        observers of subject
        */
        protected List
getObservers(Subject s)
{
    if (perSubjectObservers == null)
    {
        perSubjectObservers = new
WeakHashMap();
    }
    List observers =
(List)perSubjectObservers.get(s);

    if (observers == null)
    {
        observers=new LinkedList();
        perSubjectObservers.put(s,
observers);
    }
    return observers;
}

/*
 * Adds an observer to a subject.
 * param s: the particular subject
to attach a new observer
 * param o: the new observer to
attach
*/
public void addObserver(Subject
s, Observer o)
{
    getObservers(s).add(o);
}
/*
 * Removes an observer from a
subject list.
 * param s: the particular subject
 * param o: the observer to remove
*/
public void
removeObserver(Subject s, Observer o)
{
    getObservers(s).remove(o);
}
/*
 * The join points after which to
do the update.
*/
protected abstract pointcut
subjectChange(Subject s);

/*

```

```

        * Call updateObserver after a
change of interest to
        * update each observer.
        * param s: the subject on which
the change occurred
        */
        after(Subject s):
subjectChange(s)
{
    Iterator iter =
getObservers(s).iterator();
    while (iter.hasNext())
    {
        updateObserver(s,
((Observer)iter.next()));
    }
}

/*
 * Defines how each Observer is to
be updated
 * when a change to a Subject
occurs.
 * param s: the subject on which a
change of interest occurred
 * param o: the observer to be
notified of the change
*/
protected abstract void
updateObserver(Subject s, Observer o);
}
```

```

package connector;

import main.*;

public aspect ObserverProtocolImpl
extends ObserverProtocol
{
    /*
     * Declare the Subjects and the
Observers
    */

    declare parents: Machine
implements Subject;

    declare parents: Adaptor
implements Observer;

    /*
     * Set the pointcut. Advise the
method methodH().
    */

```

```
protected pointcut
subjectChange(Subject subject):
    call(void
Machine.methodH()) && target(subject);

/*
 * Updating the observer, when
monitoring the change of the subject.
*/
protected void
updateObserver(Subject subject,
Observer observer)
{
    double speed = 0.0;
    Machine machine =
(Machine)subject;
    Adaptor adaptor =
(Adaptor)observer;
    speed = machine.getSpeed();
    adaptor.upDate(speed);
}
```

- [11] Venkatesh Prasad Ranganath, John Hatchiff, "Slicing Concurrent Java Programs using Indus and Kaveri", International "Journal on Software Tools for Technology" Transfer (STTT), Vol.9, Numbers 5-6, pp. 489-504, 2006.
- [12] Jan Hannemann and Gregor Kiczales, Design Pattern Implementation in Java and AspectJ, "ACM", USA, 2002.
- [13] Venkatesh Prasad Ranganath, Indus - Java Program Slicer, 2008.
- [14] Indus project, Available at <http://indus.projects.cis.ksu.edu/>
- [15] Ganeshan Jayaraman, Kansas State University, Indus-Kaveri, 2008.
- [16] Len Bass, Paul Clements, and Rick Kazman. Software Architecture in Practice, Addison-Wesley, 2003.
- [17] Venkatesh Prasad Ranganath , John Hatchiff, "Slicing concurrent Java programs using Indus and Kaveri", International Journal on Software Tools for Technology Transfer (STTT), v.9 n.5, p.489-504, October 2007
- [18] Reference Documentation of Spring Framework, Available at <http://www.springsource.org/>
- [19] Rod Johnson et al., Professional Java development with the Spring Framework, John Wiley & Sons, 2005
- [20] D. Balek and F. Plasil (2000). "Software connectors: A hierarchical model". Technical Report 2000/2, Charles University.
- [21] William G. Wood (2007). "A Practical Example of Applying Attribute-Driven Design (ADD)", Version 2.0. TECHNICAL REPORT CMU/SEI-2007-TR-005 ESC-TR-2007-005.

Acknowledgments

This work was sponsored by the Natural Sciences and Engineering Research Council (NSERC) of Canada, and by the University of Quebec at Chicoutimi (Canada).

References

- [1] Shown and Garlan, Software Architecture, Prentice-Hall 1996.
- [2] James McGovern, Scott W. Ambler, A practical guide to enterprise architecture, 2004.
- [3] Jeffrey Voas, Reliable Software Technologies, Maintaining Component-Based Systems, "IEEE Computer Society Press", USA, 1998.
- [4] Paulo Veríssimo, Luís Rodrigues, Distributed systems for system architects, Kluwer Academic Publishers, USA, 2001.
- [5] ISO/IEC 14764: Software Engineering — Software Life Cycle Processes — Maintenance, 2006.
- [6] Nikunj R. Mehta, Nenad Medvidovic, and Sandeep Phadke. "Towards a taxonomy of software connectors". In Proceedings of the 22nd International Conference on Software Engineering (ICSE 2000), pages 178-187, Limerick, Ireland, June 4-11, 2000.
- [7] Joseph D. Gradecki and Nicholas Lesiecki, Mastering AspectJ: Aspect-Oriented Programming in Java, Wiley, 2003.
- [8] James W. Cooper, Java design patterns: a tutorial, Addison-Wesley Professional, 2000.
- [9] Jan Hannemann, "Design Pattern Implementations using Aspect-Oriented Programming", available at <http://hannemann.pbworks.com/Design+Patterns>, 2008.
- [10] Erich Gamma, Ralph Johnson, John Vlissides, Richard Helm. Design Patterns: Elements of Reusable Object-Oriented Software, Addison-Wesley, 1995.

Hamid Mcleick professor of computer science at the University of Quebec in Chicoutimi, Chicoutimi, Canada. He has PhD in computer Science from Montreal University, Msc in Computer science from University of Quebec at Montreal, Canada, and Bachelor in Computer Science-Mathematic from Lebanese University. Professor Mcleick is a member of IEEE, IEEE Society and ACM. He is interested in software architecture, evolution and distributed object and service oriented computing.

Yan Qi Student Master degree at the University of Quebec at Chicoutimi. He has Bachelor in Computer science from China. He is interested in Software architecture area.

Hafedh Mili professor of computer science at the University of Quebec in Montreal, Montreal, Canada. Throughout his academic career, he worked on a variety of subjects, starting with knowledge representation, object-oriented software engineering, aspect-oriented development, service-oriented computing, and business process engineering.

Improved FCM algorithm for Clustering on Web Usage Mining

K.Suresh¹

R.Madana Mohana²

A.RamaMohanReddy³

¹ Department of Software Engineering, East China University of Technology, ECIT Nanchang Campus, Nanchang, Jiangxi-330013, P.R.China.

² Department of Information Technology, Vardhaman college of Engineering, Shamshabad, Hyderabad, A.P, India.

³ Department of Computer Science and Engineering , S.V.University College of Engineering, Tirupati, A.P, India.

Abstract

In this paper we present clustering method is very sensitive to the initial center values ,requirements on the data set too high, and cannot handle noisy data the proposal method is using information entropy to initialize the cluster centers and introduce weighting parameters to adjust the location of cluster centers and noise problems. The navigation datasets which are sequential in nature, Clustering web data is finding the groups which share common interests and behavior by analyzing the data collected in the web servers, this improves clustering on web data efficiently using improved fuzzy c-means(FCM) clustering. Web usage mining is the application of data mining techniques to web log data repositories. It is used in finding the user access patterns from web access log. Web data Clusters are formed using on MSNBC web navigation dataset.

Keywords: Datasets,cluterung,improved FCM clustering ,webusage mining.

1. Introduction

The World Wide Web has huge amount information [1,2]and large datasets are available in databases. Information retrieving on websites is one of possible ways how to extract information from these datasets is to find homogeneous clusters of similar units for the description of the data vector descriptions are usually used each its component corresponds to a variable which can be measured in different scales (nominal, ordinal, or numeric) most of the well known clustering methods are implanted

only for numeric data(k-means method) or are too complex for clustering large datasets(such as hierarchical methods based on dissimilarity matrices). Fuzzy clustering relevant for information retrieval as a document might be relevant to multiple queries, this document should be given in the corresponding response sets otherwise the user would not be aware of it ,Fuzzy clustering seems a natural technique for document categorization there are two basic methods of fuzzy clustering[4] ,one which is based on fuzzy c-partitions is called a fuzzy c-means clustering method and the other ,based on the fuzzy equivalence relations is called a fuzzy equivalence clustering method.

2. Clustering

Broadly speaking clustering algorithms[3] can be divided into two types partitioned and hierarchical. Partitioning algorithms construct a partition of a database D of n objects into a set of clusters where k is a input parameter.

Hierarchical algorithms create decomposition of the database D. they are a Agglomerative and divisive. Hierarchical clustering builds a tree of clusters, also known as a dendrogram. Every cluster node contains child cluster. An agglomerative clustering starts with one-point (singleton) Clusters and recursively merges two or more most appropriate clusters. A divisive clustering starts with one cluster of all data points and recursively splits into the most appropriate clusters. The process continues until a stopping criterion is achieved. There are two main issues in clustering techniques, Firstly finding the optimal number of clusters in a given dataset and secondly, given two sets of clusters, computing relative measure of goodness

between them. For both these purposes, a criterion function or a validation function is usually applied. The simplest and most widely used cluster optimization function is the sum of squared error [5]. Studies on the sum of squared error clustering were focused on the well-known k-Means algorithm [6] and its variants.

In conventional clustering objects that are similar are allocated to the same cluster while objects that differ are put in different clusters. These clusters are hard clusters. In soft clustering an object may be in more than two or more clusters. Clustering is a widely used technique in data mining application for discovering patterns in underlying data. Most traditional clustering algorithms are limited in handling datasets that contain categorical attributes. However, datasets with categorical types of attributes are common in real life data mining problem. For each pair of documents, a comparison vector is constructed that contains binary features that measure the overlap for highly informative but sparse features between the two documents and numeric features.

The aggregating the comparison vector into one value that belongs to interval. The aggregation step is performed by taking a weighted average the information gain has a tendency to favor features with many possible values over feature with fewer possible values ,we used a normalized version of information gain, called gain ration as weighting metric.

Clustering is of prime importance in data analysis, machine learning and statistics. It is defines as the process of grouping N item sets into distinct clusters based on similarity or distance function .A good clustering technique may yield clusters thus have high inter cluster and low intra cluster distance[7].The objective of clustering is to maximize the similarity of the data points within each cluster and maximize dissimilarity across clusters.

Information gain of feature is calculated as follows. Assume we have K, the set of class label (a binary set in our case: document pairs belonging to the same cluster or not) and M_i, the set of feature values for feature I, with this information; we can calculate the database information entropy; the probabilities are estimated from the relative frequencies in the training set.

$$H(k) = - \sum_{k \in K} P(k) \log_2 P(k)$$

The information gain of feature i is then measured by calculating the difference in entropy between the situations with and without the information about the values of the feature.

$$W_i = H(k) - \sum_{m \in M_i} P(m) \times H(k/m)$$

gain ration is a normalized version of information gain. it is information gain divided by split info li(i),the entropy of the feature values. This I just the entropy of the database restricted to a single feature.

$$W_i = \frac{H(k) - \sum_{m \in M_i} P(m) \times H(k/m)}{li(i)}$$

$$li(i) = - \sum_{m \in M_i} P(m) \log_2 P(m)$$

3. Improved FCM clustering algorithm

As the FCM algorithm is very sensitive to the number of cluster centers, cluster centers initialization often artificially get significant errors, and even get the actual opposite results .FCM algorithm[8] is hard on data sets too ,so the data sets must be quite regular, in order to solve problems, first of all we use information entropy to initialize the cluster centers to determine the number of cluster centers. it can be reduce some errors, and also can improve the algorithm introductions an weighting parameters after that combine with the merger of ideas and divide the large chumps into small clusters. Then merge various small clusters according to the merger of the conditions, so that you can solve the irregular datasets clustering. Document similarity measures as shown in below.

The algorithm as follows

```

Initialize number of clusters
Initialize Cj (cluster centers)
Initialize α (threshold value)
Repeat
For i=1 to n :update μj(Xi)
    For k=1 to p ;
        Sum=0
        Count=0
        For i=1 to n:
            If μj(Xi) is maximum in Ck then
                If μj(Xi)>= α
                    Sum=Sum+Xi
                    Count=Count+1
            Ck=sum/count
Until Cj estimate stabilize.

```

The clustering framing as follows
 A set clusters C={C₁, C₂, C₃,...,C_k}
 Maximum precision values:

$$\text{Purity} = \sum_{i=1}^k \left(\frac{|C_i|}{n} \right) \max_{j=1}^n \text{Precision}(C_i, L_j)$$

$$\text{Precision}(C_i, L_j) = \left(\frac{|C_i \cap L_j|}{|C_i|} \right)$$

$$\text{Inverse Purity} = \sum_{i=1}^m \left(\frac{|L_j|}{n} \right) \max_{j=1}^k \text{Recall}(C_i, L_j)$$

$$\text{Recall}(C_j, L_i) = \text{Precision}(L_i, C_j)$$

To calculate the harmonic mean, the F-means

$$\text{Purity-F} = \sum_{i=1}^m \left(\frac{|L_j|}{n} \right) \max_{j=1}^k \{F(C_j, L_i)\}$$

Where the maximum is taken over all cluster $F(C_j, L_i)$ is defined as

$$F(C_j, L_i) = \frac{2 \times \text{Recall}(C_j, L_i) \times \text{Precision}(C_j, L_i)}{\text{Recall}(C_j, L_i) + \text{Precision}(C_j, L_i)}$$

Web usage data set

In this section we describe the dataset used and the description of the dataset used for the experimental results.

Information about the dataset

Number of users : 989818
 Average number of visits per user : 5.7
 Number of URLs per category : 0 to 5000

Table 1:Clustering of web usage data

| sequence | Order of user page visits |
|----------|------------------------------|
| 1 | 1 1 |
| 2 | 2 |
| 3 | 3 2 2 4 2 2 2 3 3 |
| 4 | 5 |
| 5 | 1 |
| 6 | 6 |
| 7 | 1 1 |
| 8 | 6 |
| 9 | 6 7 7 7 6 6 8 8 8 8 |
| 10 | 6 9 4 4 4 10 3 10 5 10 4 4 4 |
| 11 | 1 1 1 1 1 1 1 1 |
| 12 | 12 12 |
| 13 | 1 1 |

Description of the Dataset we collected the data from the UCI dataset repository that consists of sever logs from msnbc.com for the month of September 1998. each sequence corresponds to page views of a

user during that 24 hour period. Each sequence in the dataset corresponds to the page views of a user during that twenty four hour period. Each event in the sequence corresponds to a users request for a page.

There are 17 page categories "FrontPage", "news", "tech", "local", "opinion", "on-air", "misc", "weather", "health", "living", "business", "sports", "summary", "bbs" (bulletin board service), "travel", "msn-news", and "msn-sports".

Each category is associated in order with an integer starting with "1". For example, "FrontPage" is associated with 1, "news" with 2, and "tech" with 3. Each row below "% Sequences:" describes the hits in order of a single user. For example, the first user hits "FrontPage" twice, and the second user hits "news" once.

The length of the user sessions ranges from 1 to 500 and the average length of session is 5.7 .

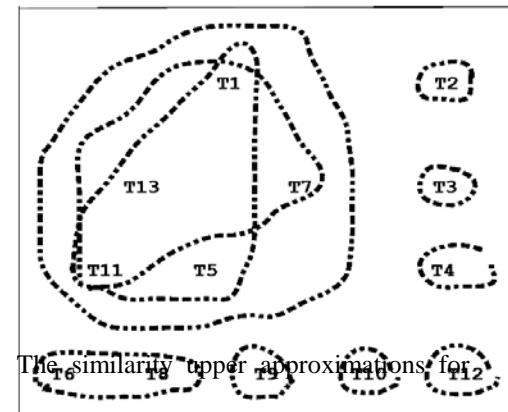
Similarity matrix with p=0.5

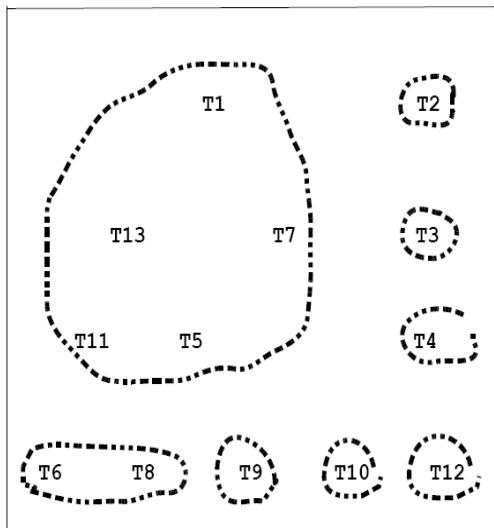
Table 1:Clustering of web usage data

The first similarity upper approximation at threshold value 0.5 is given by

$$\begin{aligned} R(T1) &= \{T1, T5, T7, T11, T13\} \\ R(T2) &= \{T2\} \\ R(T3) &= \{T3\} \\ R(T4) &= \{T4,\} \\ R(T5) &= \{T1, T5, T11, T13\} \\ R(T6) &= \{T6, T8\} \\ R(T7) &= \{T1, T7, T11, T13\} \\ R(T8) &= \{T6, T8\} \\ R(T9) &= \{T9\} \\ R(T10) &= \{T10\} \\ R(T11) &= \{T1, T5, T7, T11, T13\} \\ R(T12) &= \{T12\} \\ R(T13) &= \{T1, T5, T7, T11, T13\} \end{aligned}$$

Graphical representation of the clusters found after first upper approximations





$S_1 = \{T_1, T_5, T_7, T_{11}, T_{13}\}$

$S_2 = \{T_2\}$, $S_3 = \{T_3\}$, $S_4 = \{T_4\}$, $S_5 = \{T_1, T_5, T_7, T_{11}, T_{13}\}$, $S_6 = \{T_6, T_8\}$, $S_7 = \{T_1, T_5, T_7, T_{11}, T_{13}\}$, $S_8 = \{T_6, T_8\}$, $S_9 = \{T_9\}$, $S_{10} = \{T_{10}\}$, $S_{11} = \{T_1, T_5, T_7, T_{11}, T_{13}\}$, $S_{12} = \{T_{12}\}$, $S_{13} = \{T_1, T_5, T_7, T_{11}, T_{13}\}$.

This denotes that user visiting the hyper links in T_1 may visit the hyperlinks in T_5 ad then T_7 , T_{11} and hyper links in T_{13} .

One who visits the hyper links in T_6 may also visit the hyper links in T_8 .

4. Conclusions

In this paper we presented a clustering web usage data ,from msnbc.com which is useful in finding the user access patterns and the order of visits of the hyperlinks of the each user. The suggested approach was used for efficacy contained a hard clustering of the msbn.com data set and as the analysis indicated each of the clusters seems to contain observations with specific common charterstics and improve the algorithm efficiency with help of improved FCM algorithm .Experiments prove the improved algorithm has able to identify the initial cluster centers.

References

- [1] J. Srivastava, R. Cooley, M. Deshpande, PN. Tan, Web usage mining: discovery and applications of usage patterns from web data. SIGKDD Explorations, Vol. 1, No. 2, 2000, pp.12–23.
- [2] M. N. Garofalakis, R. Rastogi, S. Seshadri, K. Shim Data minino and the web: past, present and future, In Proc. of the second international workshop on webinformation and data management, ACM, 1999.
- [3] A. K. Jain and R. C. Dubes, "Data clustering: A review.," ACM Computing Surveys, vol. 31, 1999.

- [4] U. Maulik and S. Bandyopadhyay, "Genetic algorithm based clustering technique," Pattern Recognition, vol. 33, pp. 1455–1465, 2000.
- [5] P. Zhang, X. Wang, and P. X. Song, "Clustering categorical data based on distance vectors," The Journal of the American Statistical Association, vol. 101,no. 473, pp. 355–367, 2006.
- [6] A. Vakali, J. Pokorný and T. Dalamagas, An Overview of Web Data Clustering Practices, EDBT Workshops, 2004, pp. 597-606.
- [7] Lin Zhu, Fu-Lai Chung, Shitong Wang.Generalized Fuzzy C-Means Clustering Algorithm With Improved Fuzzy Partitions[J].IEEE Transactions on Systems,2009:39-3.
- [8] Cheul Hwang, Frank Chung-Hoon Rhee.Uncertain Fuzzy Clustering:Interval Type-2 Fuzzy Approach to C-Means[J]. IEEE Transcations on Fuzzy Systems,2007:15-1.



K.Suresh received his Bachelor degree and Master Degree in Information Technology from JNT University Hyderabad. He is currently working as Foreign Faculty in East China University of Technology, P.R.China and Visiting Faculty for Jiangxi Normal University, he worked in AITS, Rajampet, A.P, India .he has published more than 20 national and International conference papers and journals. He received best paper award in 2008 in National wide paper presentation. His field of interest is data mining, database technologies, information retrieval system.



R. Madana Mohana is an Associate Professor in the Information Technology department at Vardhaman College of Engineering, Hyderabad, Andhra Pradesh, India since 2007. He has 8 years of teaching experience at both UG and PG levels. He received his B. Tech in Computer Science and Information Technology from Jawaharlal Nehru Technological University, Hyderabad in 2003 and M. E in Computer Science and Engineering from Sathyabama University ,Chennai in 2006. He is doing Ph. D in Computer Science and Engineering at Sri Venkateswara University Tirupathi, Andhra Pradesh, India. He is a life member of ISTE Technical Association. His areas of interest include Data Mining, Automata Theory, Compiler Design and Database Systems.



Dr. A. Rama Mohan Reddy received the B. Tech. from JNT University, Hyderabad in 1986, M. Tech degree in Computer Science from National Institute of Technology in 2000 Warangal and Ph. D in Computer Science and Engineering in 2008 from Sri Venkateswara University, Tirupathi, Andhra Pradesh, India. He worked as Assistant Professor, Associate Professor of Computer Science and Engineering, Sri Venkateswara University College of Engineering during the period 1992 and 2005. Presently working as Professor of Computer Science and Engineering, Sri Venkateswara University College of Engineering. He has 28 years of Industry and Teaching experience. Currently guiding twelve Ph. D scholars. He is life member of ISTE and IE. Research interests include Software Architecture, Software Engineering and Data Mining. He has 10 international publications and 14 international conference Publications at International and National level.

Causally Ordered Delivery Protocol for Overlapping Multicast Groups in Broker-based Sensor Networks

Chayoung Kim¹ and Jinho Ahn^{2*}

¹ Contents Convergence Software Research Center, Kyonggi University
Suwon, Gyeonggi-do 443-760, Republic of Korea

² Dept. of Computer Science, College of Natural Science, Kyonggi University
Suwon, Gyeonggi-do 443-760, Republic of Korea

Abstract

In sensor networks, there is a lot of overlapping multicast groups because of many subscribers, associated with their potentially varying specific interests, querying every event to sensors/publishers. Also gossip-based communication protocols are promising as one of potential solutions providing scalability in P(Publish)/ S(Subscribe) paradigm in sensor networks. Moreover, despite the importance of both guaranteeing message delivery order and supporting overlapping multicast groups in sensor or P2P networks, there exist little research works on development of gossip-based protocols to satisfy all these requirements. In this paper, we present a causally ordered delivery guaranteeing protocol for overlapping multicast groups, based on sensor-brokers as delegates. In the protocol based on sensor-broker, sensor-broker might lead to make overlapping multicast groups organized by subscriber's interests. The message delivery order has been guaranteed consistently and all multicast messages are delivered to overlapping groups using gossip-based protocols by the sensor-broker. Therefore, these features of the protocol based on sensor-broker might be significantly scalable rather than those of the protocols by coordinated groups like traditional committee protocols.

Keywords: *Sensor Network, Group Communication, Overlapping Multicast Groups, Scalability, Reliability.*

1. Introduction

A wireless sensor network(WSN) has important applications such as remote environmental monitoring, target tracking, natural disaster relief, biomedical health monitoring, hazardous environment exploration and seismic sensing and virtual worlds such as massive multiplayer games [18, 25]. The design of WSN depends significantly on the application, and it must consider factors such as the environment, the application's design objects and system constraints [1, 5, 25]. The environment plays a key role in determining the size of the network and the network topology. For indoor environment, fewer sensor nodes are required to form a network in a limited

space whereas outdoor environments may require more sensor nodes to cover a larger area [1, 5, 25]. There are two types of WSNs: structured and unstructured. An unstructured WSN is one that constrains a dense collection of sensor nodes, deployed in an ad hoc manner into the field. In an unstructured WSN, network maintenance such as managing connectivity and detecting failures is difficult since there are so many nodes [25]. In a structured WSN, all or some of the sensor nodes are deployed in a pre-planned manner. The advantage of a structured network is that fewer sensor nodes can be deployed with lower network maintenance and management cost [25]. For reliable communication for this sensor networks, services such as congestion control, acknowledgements, and packet-loss recovery are necessary to guarantee reliable packet delivery [25]. Especially, these above applications need a variety of collaboration features, such as chat windows, white boards, p2p video and other media streams, and coordination mechanisms [11]. So, the use of p2p overlapping groups is expected to generate new models of interactive communication and cooperation to support these applications [25] in sensor networks. A new data dissemination paradigm for such sensor networks is different from mobile ad-hoc networks in a method of designing data propagation and aggregation generated by various and lots of sensor nodes [1, 5, 21]. There are several researches based on the P (publish)/S (subscribe) paradigm [8] in the area of sensor network communications to address the problem of querying sensors from mobile nodes in order to minimize the number of sent result packets [14, 20]. In P/S paradigm systems, a query node periodically runs an algorithm to identify the sensors it wishes to track and "subscribe" to these sensors of their interest, and the sensors periodically "publish" [14, 20]. So, the intermediate sensors in the networks, along the reverse path of interest propagation might aggregate the query results by combining reports from several sensors [14]. An important feature of that

interest and data propagation and aggregation are determined by localized sensors interactions [14]. And performing local computations to reduce the amount of data before transmission can obtain orders of magnitude energy savings [14, 20]. Recently, gossip-based protocols seem more appealing in many P/S systems because they are more scalable than traditional reliable broadcast [3] and network-level protocol deriving from IP Multicast for many of the various applications requiring reliable dissemination of events [7, 15]. In gossip-based protocols, when a process sends a multicast message, it randomly selects a small subset of members, called gossip targets. The number of gossip targets is called fan-out, which relates reliability of gossip-based protocols. Usually the time necessary to reach all processes in a group is $\log N$, where N is the size of the group, the maximum number of gossip rounds. This approach often relies on the assumption that every process knows every other process [4]. Gossip-based protocols have turned out to be adequate for large scale settings by achieving a "high degree of reliability" and strong message delivery ordering guarantees offered by deterministic approaches [6, 7, 10]. The seminal probabilistic broadcast (pbcast) algorithm of Birman et. al. [4] is originally described as a broadcast presented in the system based on global view and Eugster's algorithm(lpbcast) [7] is implemented for P/S systems as a broadcast. These previously developed gossip-based protocols implicitly assume that all processes in a group are interested in all events [4]. Such a flooding technique is not adequate when many events are only of interest for lower than half the processes in the overall groups [6]. PMCAST [6] deals with the case of multicasting events only to subsets of the processes in a large group by relying on a specific orchestration of process as a superimposition of spanning trees. The delegates of this protocol [6] yet are themselves not interested in the same topics as subscribers have. But, when multicasting an event, PMCAST [6] follows the underlying tree, by gossiping depth-wise, starting at the root and the interested subscribers of overlapping multicast groups receive their event messages [6]. An atomic broadcast on gossip-based protocols is implemented in Birman et. al. [4] and Eugster et. al. [6] for P/S systems. But, these protocols are performed by hierarchical membership protocols [6] for each delegate group or the totally ordered delivery properties are maintained by global member views [4]. These features are likely to be highly overloaded on each member and not scalable. Also, there is no causal order guaranteeing multicast protocol supporting overlapping multicast groups, useful for many distributed applications with a variety of collaboration features, such as chat windows, white boards, p2p video and other media streams, based on publisher(sensor-broker) in the previously developed protocols. A causal ordering protocol ensures that if two messages are causally

related and have the same destination, they are delivered to the application in their sending order [3]. Consider a distributed application that uses a sensor, a controller and a monitor for machine monitoring. The controller aggregates sensor notifications and controls the machine. In some cases, the controller decides to stop the machine due to a notification from the sensor, in which case the sensor also sends a reading to the monitor. In this case, causal order would ensure that the monitor would receive the sensor reading before the stop notification. The wrong order would falsely indicate a malfunction in the controller, that is, the delayed sensor reading could indicate that the machine was still operating [22]. In [16], Kim et al. suggested an efficient and scalable causal order guaranteeing multicast protocol to use only local views supporting overlapping multi-groups. In the proposed protocol, there is no sensor-broker. So, overlapping multi-groups are defined by only subscribers' interests. The messages of join/leave are disseminated by gossip communication based on its local views. Messages including causal context graphs [19] based on group identification are delivered to the application layer without any sensor-brokers [6, 9]. In this paper, we present a causal order guaranteeing multicast protocol based on sensor-brokers as delegates that aggregate the information of results in sensor networks, periodically gossip about the messages of them and guarantee causally ordered delivery of the messages in the face of transient member population. This protocol is appropriate for sensor networks in a pre-planned manner. Fewer sensor nodes can be deployed since nodes are placed at specific locations to provide small coverage. In this structured network, there are lower network maintenance and management costs because fewer sensor nodes cannot be changed frequently.

2. Background

2.1 System Model

In the distributed system, a group consists of a set of processes. Processes join and leave the system dynamically and have ordered distinct identifiers. The process maintains a local membership list called a "local view". It can send unicast messages to another process through the communication network. A finite set of processes communicate only by exchanging messages over a fully connected, point-to-point network. Processes communicate using the primitives *send(m)* and *receive(m)*. Communication links are fair-lossy, but correct processes can construct reliable communication links on top of fair-lossy links by periodically retransmitting messages. Each member performs operations according to a local clock. Clock rates at all members are the same. Runs of the system proceed in a sequence of rounds. Members may

undergo two types of failures, both probabilistic in nature. The first is process failure. There is an independent, per-process probability of at most Υ that a process has a crash failure during the finite duration of a protocol. Such processes are called faulty. Processes that survive despite the failures are correct. The second type of failures is message omission failure. There is an independent, per-message probability of at most δ that a message between non-faulty processes experiences a send omission failure. The union of all message omission failure events and process failure events are mutually independent. For simplicity, we do not include process recovery in the model. Also, we expect that both Υ and δ are small probabilities. There are no malicious faults, spurious messages, or corruption of message i.e. we do not consider Byzantine failures.

In proposed protocols, a group of processes is defined through two primitives PMCAST and PDELIVER, which use gossip protocols to provide probabilistic reliability in networks. Processes communicate with these two pairs of primitives, PMCAST and PDELIVER, which model unreliable communication associated with probability α of successful message transmission. We refer to probability α as the expected reliability degree. These primitives are as follows: **(Integrity)** For any message m , every correct process PDELIVER m at most once, and only if m was previously PMCAST by $sender(m)$. **(Validity)** If a correct process p PMCASTs a message m then p eventually PDELIVERS m . **(Probabilistic Agreement)** Let p and q be two correct processes. If p PDELIVERS a message m , then with probability α , q PDELIVERS m . In other terms, the only probabilistic property is Agreement. This probabilistic notion of agreement also captures a weakly consistent membership of local view, typical for large scale settings.

2.2 Related Work

In P. Eugster et. al [6], the protocol deals with the case of multicasting events only to subsets of the processes in a large group by relying on a specific orchestration of process as a superimposition of spanning trees. But, PMCAST [6] is not also a genuine multicast[13] because of considering delegates. Birman et al. [4] proposed a gossip-style protocol called bimodal multicast thanks to its two phases: a "classic" best-effort multicast such as IP multicast is used for the first rough dissemination of messages. The second phase assures reliability with a certain probability by using a gossip-based retransmission. But gossip-based broadcast protocols based on Lpbcast [7] proposes gossip-based broadcast membership mechanisms based on a partial view without a global view. Each process has a randomly chosen local view of the system. Lpbcast [7] is a completely as a decentralized membership protocol because of no dedicated messages for

membership management based on gossips. In Eugster's algorithm [9], atomic probabilistic broadcast (apbcast) implemented for publish/subscribe[8] programming is a hybrid approach. Its deterministic ordering of messages ensures the consistency of the delivery order of broadcast messages and its probabilistic propagation of broadcast messages and order information provides a high level of reliability in the face of an increasing number of process failures because of more heroic efforts by making use of the membership of delegates. However, building such the membership of delegates requires the global knowledge of membership, and it may be very difficult to maintain such the structure in the present of joins/leaves of processes. Probabilistic Atomic Broadcast (pabcast) [9] is fully probabilistic by mixing message atomic ordering and propagation, basing these on gossips without a membership of delegates. But, a promising approach for increasing scalability is to weaken the deterministic ordering guarantees to make the properties of dependencies between broadcast messages probabilistic. Also, it does not give the guarantees achieved for the consistency of the delivery order of overlapping groups. As a fundamental problem in distributed computing, much effort has been invested in solving atomic broadcast [3]. Early work such as [3] mostly focuses on stronger notions of Agreement and also membership than the proposed protocols discussed in this paper. In [19], an inter-process communication mechanism, called Psync, explicitly encodes partial ordering with each message. Psync is based on a conversation abstraction that provides a shared message space through which a collection of processes exchange messages. The general form of this message space is defined by a directed acyclic graph that preserves that partial order of the exchanged messages. And there are researches based on the P (publish)/S (subscribe) paradigm [8] in the area of sensor network communications to approach the problem of querying sensors from mobile nodes [14, 20]. Directed Diffusion [14] can be seen as publish-subscribe mechanism, which is implemented using the tree-based architecture rooted at the publisher. SENSTRACT [20] is mapping from queries to topics and the corresponding underlying sensor network structure. SENSTRACT [20] is a tree-based P/S system structured by service providers as roots, representing one of the data-centric routing protocols for data dissemination of sensor networks.

Recently, there is a gossip-based technique that GO [24] is a new platform support for gossip applications targeted to large-scale deployments. Adaptive Peer-Sampling [23] focuses on Newscast, a robust implementation of peer sampling service. It is an adaptive self-control mechanism for its parameters, namely-without involving failure detectors-nodes passively monitor local protocol events using them as feedback for a local control loop for self-tuning the protocol parameters. The research work [2]

presents P3Q, a fully, decentralized gossip-based protocol to personalize query processing in social tagging systems. This P3Q does not rely on any central server: users periodically maintain their networks of social acquaintances by gossiping among each other and computing the proximity between tagging profiles.

3. The Proposed Protocol

3.1 Basic Idea

In sensor-broker based protocol, some sensors that are designed as brokers might lead to make overlay networks and query nodes subscribe to all the topics that match their interest. The mapping of subscribers and brokers is entirely driven by the query application. Recently, much research has been devoted to designing broker selection methods that best suits application needs [15, 20]. Each sensor can provide information periodically with some of its brokers by peer-sampling services [15], assumed to be implemented in the systems and the brokers might aggregate the query results by combining reports from several sensors. The aim of peer-sampling services is to provide every node with peers to exchange information with[15]. This assumption has led to rigorously establish many desirable features of gossip-based broadcast protocols like scalability, reliability, and efficiency [15] and a wide range of higher funtions, which include information dissemination, aggregation, and network management [15]. We also consider the reliability of the service by examining its self-healing capacity and robustness to failure. If all brokers representing a sensor grid may be stale, all members of brokers are not changed because of periodical peer-sampling services [15], i.e., the failure of brokers having their interests is tolerant by self-healing functions. But, if all subscribers in a grid lose their interests in information published on a particular topic, the brokers representing the grid send leave messages to the corresponding overlay multicast groups and then all of their group members are updated by leaving brokers. Query nodes subscribe to the corresponding sensor-broker networks of their interest and receive the results of the queries. The sensor-broker periodically gossip about the messages of the results [20] and guarantee causally ordered delivery of the messages with aggregating the information of the results in overlapping networks. In this protocol, because every broker knows every other brokers, a VT(vector time) for each broker, π_i is a vector of length n , where $n =$ the number of broker members. And this protocol leads us to extend single VT to multiple VTs because a broker belongs to several interest overlapping groups [3]. Every sensor-broker maintains a VT for each group and attaches all the VTs to every message that they multicast. Also, epoch protocol [1] is assumed to be

implemented for the member leaving overlapping multicast groups and each group membership list updated by the member. The digest information of vector and membership list is sent or received periodically using gossip-based protocols [4]. Therefore, this protocol might make up transient faults of sensor-broker with the peer-sampling services [15] and deal with brokers' leave by membership management. That is, the processing of temporal faults is different from that of brokers' leave. So, these features of this protocol might result in its very low membership management cost compared with the cost incurred by maintaining member list for traditional committee in the previous protocols [4].

3.2 Algorithm Description

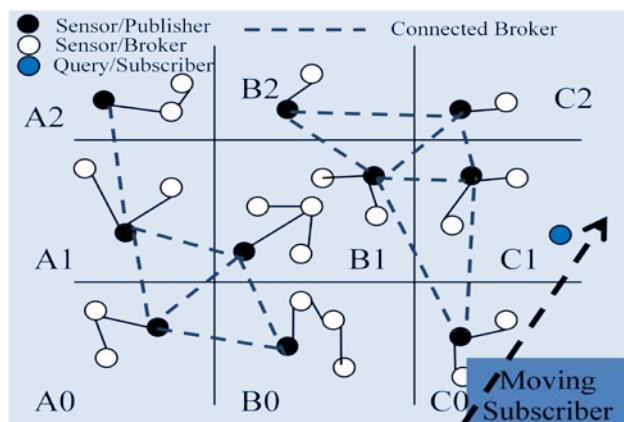


Fig. 1 Publishers(Sensors) vs. Subscribers

In figure 1, there is a two-dimensional area of interest(AoI), which the sensor-broker publishes messages to a particular topic, while query nodes subscribe to all the topics that match their interests. In this protocol based on sensor-broker, we present a causal order guaranteeing multicast protocol supporting overlapping subscriber groups and useful for these applications with a variety of collaboration features, such as chat windows, white boards, p2p video and other media streams, and coordination mechanisms requiring causally ordered delivery of messages.

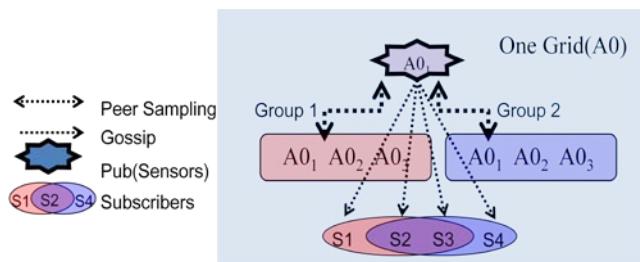


Fig. 2 Sensor-Broker A_{01} covering a Grid

In figure 2, we can see that a sensor-broker $A0_1$ in a grid $A0$ of a sensor network like one of figure 1 publishes desired messages of query results to all query nodes, $\{s1, s2, s3\}$ and $\{s2, s3, s4\}$ subscribing to their topics, Group1 and Group2 respectively using gossip-style disseminations. There are overlapped members= $\{s2, s3\}$ subscribing to the topics of Group1 and Group2.

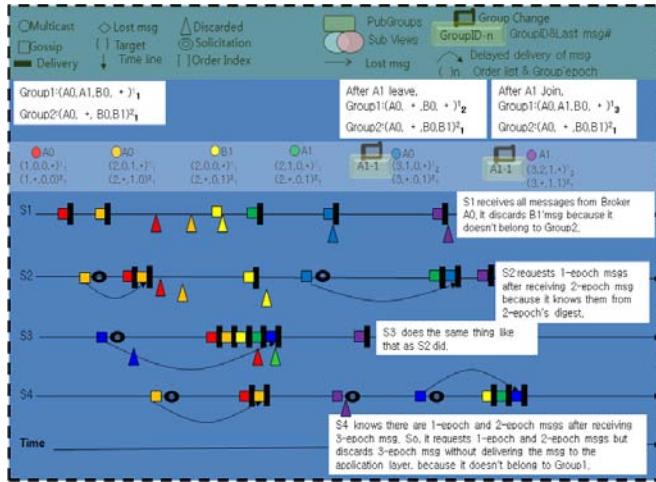


Fig. 3 Example of message deliveries from sensor-broker to subscribers

Figure 3 shows that sensor-broker gossip about multicast messages piggybacked with all VT clocks for all of interesting groups to guarantee causally ordered delivery of messages in a sensor network like one of figure 2. This example in figure 3 illustrates Group1={A0,A1,B0}, Group2={A0,B0,B1}, Subscribers={S1,S2,S3,S4} and maximum number of Gossip Rounds = 2. The overlapping subscriber members={S2,S3} receive all messages from Group1 and Group2, S1 receives messages only from Group1 and S4 receives messages only from Group2. In the figure 3, there are VT clocks $((0,0,0,*)^1_1, (0,*,0,0)^2_1)$ for each group, Group1 and Group2, to depict each VT_e^g as a vector of length n , with a subscript epoch variable, e for covering the cases of a process leave and join and a special entry * for each process that is not a member of Group $_e^g$. For each message generated by a member, each $VT_e^g(p_i)[i]$ is incremented by 1. So, if a member A0 generates a multicast message, then VT^1_1 and VT^2_1 is $((1,0,0,*)^1_1$ and $(1,*,0,0)^2_1$) respectively. And when a member A1 leaves and joins Group1 again, VT^1_2 for Group1 is from $((*,*,0,*)^1_2$ to $((*,0,0,*)^1_3$ because subscript epoch is changed from 2 to 3. Figures 2 and 3 show an example of the protocol based on sensor-broker with causal ordering VT clocks in what order is $A0 \rightarrow A0 \rightarrow B1 \rightarrow A1 \rightarrow A0 \rightarrow A1$. In this case that subscribers know what messages should be delivered according to causal ordering VT clocks piggybacked by multicast messages and delay some messages after comparing their causal ordering VT

clocks and validating their receipt of predecessor. Also, there are undesired messages are sent to a subscriber, forcing it to discard them. Subscriber S1 receives all messages from Broker A0. It discards B1's message without delivering it to the application layer because S1 doesn't belong to Group2. Subscriber S2 requests some of 1-epoch messages to the latest gossip-sender after receiving 2-epoch messages because it knows that some of 1-epoch messages are not received from piggybacked 2-epoch's digest. Subscriber S3 does the exactly same thing as S2 did because it knows that all of 1-epoch messages are not received. Subscriber S4 knows that 1-epoch and 2-epoch messages are not received after receiving 3-epoch message. So, it requests 1-epoch and 2-epoch messages but discards 3-epoch message because S4 does not belong to Group1 and 3-epoch message is generated by A1, the member of Group1. Subscriber S4 can check 3-epoch's digest for validating causal ordering VT clocks and discard the 3-epoch message without delivering it to the application layer. And S4 solicits the retransmission of 1-epoch messages and 2-epoch messages to the latest gossip-sender.

The data structures and procedures for sensors and subscribers in our protocol are formally given in figures 4 and 5.

```

Procedure RECEIVE_JOIN //Join MSG is received
Merge member=q with p.Local_View
Adjust p.Local_View by Fixed_Size //Randomly Selected Local View
Update member=q and epoch in Overlapped_Multicast_Vector

Procedure RECEIVE_LEAVE //Leave MSG is received
if q.Last_m.Seq not received then //Check the LAST MSG
    send Solicit_Retransmission(m) to Latest_Gossip_Sender
Merge q.Local_View with p.Local_View
Adjust p.Local_View by Fixed_Size //Randomly Selected Local View
Update don't care member=q and epoch in Overlapped_Multicast_Vector

Procedure RECEIVE_DIGEST
If m in Digest is not in Pending_List or not delivered then
    send solicit_Retransmission(r) to Latest_Gossip_Sender
    //Check the VTsummary whether or not interest MSG is not received

Procedure RECEIVE_SOLICITATION //Request MSG is received
if m in Pending_List or delivered then
    Digest(Overlapped_Multicast_Vector)
    send Gossip(m) to REQUESTER //Send the MSG piggybacked with VT

```

Fig. 4 Algorithm Description (Cont'd)

```

Procedure INITIALIZE // Initializing Vector and Local View
Overlapped_Multicast_Vector=(Group1(0,0,0,*),  

    Group2(*,0,0,0),...):*:not a member, subscript1: epoch1
Local_View={pid}

Procedure SEND_MULTICAST
Gossip_Count=ZERO
for All_Interest_Groups do //Sending Seq.=Sending Seq.+1
    Overlapped_Multicast_Vector(GroupID(pid)) =
        Overlapped_Multicast_Vector(GroupID(pid)+1)
m = (pid, All_Interest_Groups, Gossip_Count,
    Overlapped_Multicast_Vector) //Msg piggybacked with Vector
Unreliable_Multicast(m) //Using unreliable MCAST

Procedure SEND_JOIN
m = (pid, All_Interest_Groups, Gossip_Count)
Unreliable_Multicast(m)

Procedure SEND_LEAVE
m = (pid, All_Interest_Groups, Gossip_Count, Last_m.Seq, Local_View)
Unreliable_Multicast(m)

Procedure SEND_DIGEST //Select gossip-target from Local View
for All_Interest_Groups and All_Subscribers in
    Overlapped_Multicast_Vector do //Periodically gossip summary
        Digest(Overlapped_Multicast_Vector)
call Procedure SEND_GOSSIP

Procedure SEND_GOSSIP//Select gossip-target from Local View
for each p in Local_View and All_Subscribers s
    with probability rate //Randomly Selecting Small Set
        m.Gossip_Count=m.Gossip_Count+1
        send m including Local_View to p //Gossip about MSG with VT
        send m to s
do Garbage_Collection

Procedure RECEIVE_MULTICAST
if m is not in their interest then //MSG is received but not in interest
    check m.Overlapped_Multicast_Vector and send
Solicit_Retransmission(m)
    //Check the VT whether or not interest MSG is not received
else if m not in Pending_List then //MSG is received and in interest
    put m into Pending_List
call Procedure SEND_GOSSIP
delivery = TRUE
for All_Interest_Groups do //Check VT of MSG for causal order
    if(m.Overlapped_Multicast_Vector(GroupID(id)) >
        p.Overlapped_Multicast_Vector(GroupID(id)) and id=p) then
        delivery = FALSE, BREAK
    if(delivery = TRUE) then //If causal order is satisfied
        remove m from Pending_List
        deliver m to APPLICATION //Deliver the MSG to app.

```

4. Performance Evaluation

In this section, we compare average throughput of our protocol based on sensor-broker with that of a previous protocol based on traditional reliable committee [3]. In this comparison, we rely on a set of parameters referred to Bimodal Multicast [4] and LPBCast [7] for gossiping parameters. And we assume that processes gossip in synchronous rounds, gossip period is constant and identical for each process and maximum gossip round is $\log N$. The probability of network message loss is a predefined 0.1% and the probability of process crash during a run is a predefined 0.1% using UDP/IP. The group size of each sub-figure is 32(2), 64(4), 128(8) and 256(16).

Figure 6 shows the average throughput as a function of perturb rate for various group sizes. The x-axis is the group size (the number of overlapping groups) and the y-axis is the number of messages processed in the perturb rate, (a)20%, (b)30%, (c)40% and (d)50%. In the four sub-figures from 6(a) to 6(d), the average throughput of causally ordered delivery protocol based on sensor-broker is not a rapid change than that of the protocol based on traditional reliable committee. Especially, the two protocols are compared to each other in terms of scalability by showing how the number of messages required for maintaining membership list in perturbed networks with processes join and leave. The proposed sensor-broker protocol is more scalable because the brokers are selected by peer-sampling services [15, 20] and all messages including join and leave are gossiped by them.

And then you compare the message overhead of our protocol based on sensor-broker with that of the previous protocol based on local views [16]. We consider sensor nodes and query nodes. While sensor nodes are stationary, query nodes are mobile. The query node periodically, every 60s, sends its query and the sensors publish their current value every 50s(this value is able to be varied). We do the cell size could be set to 300m and the default AoI contains roughly 40 to 50 sensors with a total number of 600 sensors in coordinates $600 \leq x, y \leq 900$. In the four sub-figures from 7(a) to 7(d), the message overhead of our protocol based on sensor-broker is more slightly lower than that of the previous protocol based on local views [16]. We can evaluate the effects of the query node mobility and scalability with respect to the number of sensors and query nodes. We can see message overhead of sensor-broker protocol increases fast for low numbers of query nodes but then the increase diminishes with increasing number of query nodes. It shows that the sensor coverage fluctuates to some degree. But with a high number of query nodes, it becomes more likely that a broker to which a query node sends a subscription already has an active subscription. Therefore the increase in the

Fig. 5 Algorithm Description

message overhead eventually diminishes once all the subscriptions and update messages are no longer sent to the query nodes. In contrast, the message overhead of the protocol based on local views [16] sub-linearly increases with the number of query nodes. This is an attractive indication for the scalability and mobility of sensor-broker protocol because the message overhead increases with increasing number of query nodes and the sensor coverage is nearly the same for the numbers of sensor nodes.

However it is not always so, approached from a study in A.-M. Kermarrec [12]. Especially, online social networks are still growing regularly by the day. These networks constitute huge live platforms that are exploited in many ways. And it is clearly appealing to perform large-scale general purpose computations on such platforms and one might be tempted to use a central authority for that, namely one provided by the company orchestrating, likely the broker coordinating. Yet, this poses several privacy problems. In [12], they argue that a decentralized approach where the participants in the social network keep their own data and perform computations in a distributed fashion without any central authority. So it depends on the user applications which approach between sensor-broker and local views is more preferable.

5. Conclusions

In this paper, we present a causal order guaranteeing multicast protocol. The protocol based on sensor-brokers(publisher-brokers) as delegates periodically gossips about the messages in overlapping multicast groups and guarantees causally ordered delivery of the messages. In the protocol based on sensor-broker, some sensors might lead to make overlay multicast groups and query nodes subscribe to the corresponding sensor-brokers networks of their interest and receive the results of the queries, that is, aggregated messages of the information. The vector information of each interesting group piggybacked with every multicast message for causally ordered delivery are sent or received periodically using gossip-based protocols by sensor-brokers. So, these features of this protocol might be that causally ordered delivery properties by sensor-brokers are the same as those properties by traditional committee, but result in its very low communication cost compared with the cost incurred by the traditional committee because of gossip-style disseminations. And for future works, we clearly show the pros and cons according to applications by comparing two different protocols, the sensor-broker based one and the fully decentralized one based on local views.

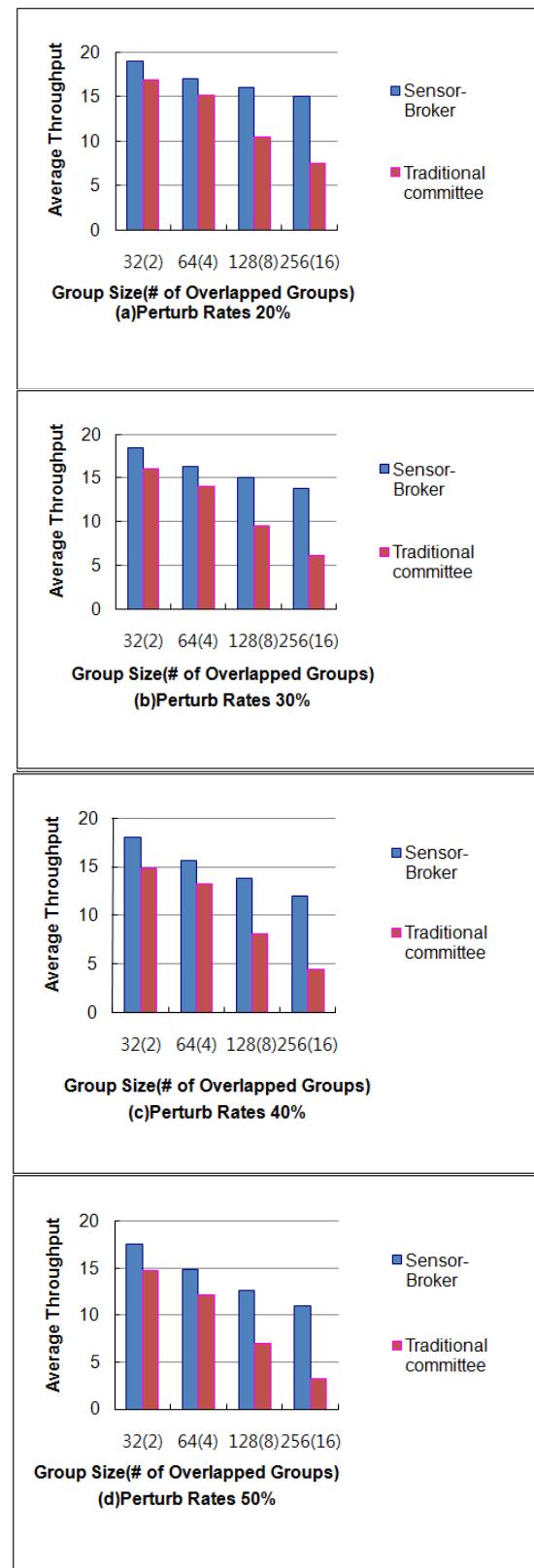


Fig. 6 Average Throughput by Perturb Rates

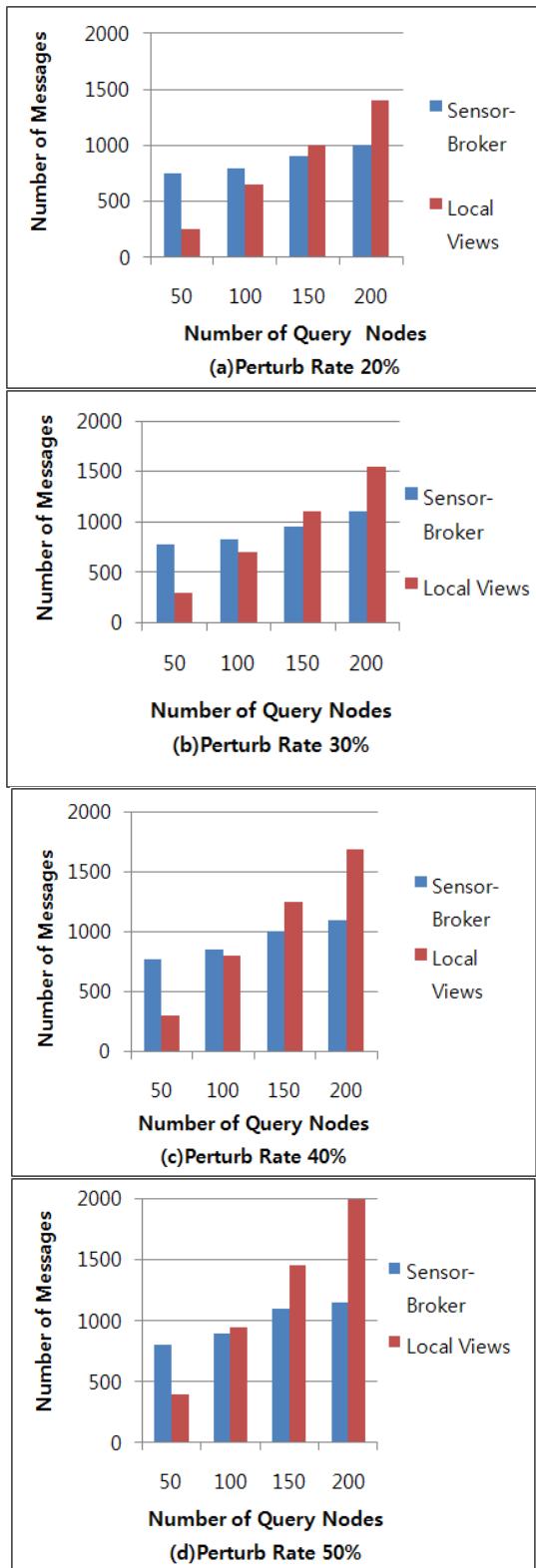


Fig. 7 Message Overhead by Query Nodes

References

- [1] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on Sensor Networks", IEEE Communications Magazine. Vol. 40, No. 8, 2002, pp. 102-114.
- [2] X. Bai, M. Bertier, R. Guerraoui, A.-M. Kermarrec, and V. Leroy, "Gossiping Personalized Queries", in Proceedings of 13th International Conference on Extending Database Technology, Lausanne, Switzerland, Mar. 2010, pp. 87-98.
- [3] K. Birman, A. Schiper, and P. Stephenson, "Lightweight Causal and Atomic Group Multicast", ACM Transactions on Computer Systems. Vol. 9, No. 3, 1991, pp. 272-314.
- [4] K. Birman, M. Hayden, O. Ozkasap, Z. Xiao, M. Budiu, and Y. Minsky, "Bimodal Multicast", ACM Transactions on Computer Systems. Vol. 17, No. 2, 1999, pp. 41-88.
- [5] D. Culler, D. Estrin, and M. Srivastava, "Overview of Sensor Networks", IEEE Computer, Vol. 37, No. 8, 2004, pp. 41-49.
- [6] P. Eugster, and R. Guerraoui, "Probabilistic Multicast", in Proceedings of the 2002 International Conference on Dependable Systems and Networks, Vienna, Austria, Jun. 2002, pp. 313-324.
- [7] P. Eugster, R. Guerraoui, S. Handurukande, P. Kouznetsov, and A.-M. Kermarrec, "Lightweight probabilistic broadcast", ACM Transactions on Computer Systems, Vol. 21, No. 4, 2003, pp. 341-374.
- [8] P. Eugster, P. Felber, R. Guerraoui, and A.-M. Kermarrec, "The many faces of Publish/Subscribe", ACM Computing Surveys, Vol. 35, No. 2, 2003, pp. 114-131.
- [9] P. Eugster, "Atomic Probabilistic Broadcast", EPFL, IC_TECH_REPORT_200303.
- [10] P. Felber, and F. Pedone, "Probabilistic Atomic Broadcast", in Proceedings of 21st IEEE Symposium on Reliable Distributed Systems (SRDS'02), Osaka, Japan, Oct. 2002, pp. 170-179.
- [11] D. Freedman, Ken. Birman, K. Ostrowski, M. Linderman, R. Hillman, and A. Frantz, "Enabling Tactical Edge Mashups with Live Objects", in Proceedings of the 15th International Command and Control Research and Technology Symposium(ICCRTS '10), Information Sharing and Collaboration Processes and Behaviors Track. Santa Monica, CA, USA, Jun. 2010, (Best Paper in Track; Best Paper in Conference.)
- [12] A. Giurgiu, R. Guerraoui, K. Huguenin, and A.-M. Kermarrec, "Computing in Social Networks", in Proceedings of 12th International Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS), New York, USA, Sep. 2010, LNCS Vol. 6366, pp.332-346.
- [13] R. Guerraoui, and A. Schiper, "Genuine Atomic Multicast in Asynchronous Distributed Systems", Theoretical Computer Science, Vol. 254, Issue 1-2, 2001, pp. 297-316.
- [14] C. Intanagonwiwat, R. Govindan, and D. Estrin, "Directed diffusion: A scalable and robust communication paradigm for sensor networks", in Proceedings of the Sixth Annual International Conference on Mobile Computing and Networking (MobiCOM '00), Boston, MA, Aug. 2000, pp. 56-67.
- [15] M. Jelasity, S. Voulgaris, R. Guerraoui, and A.-M. Kermarrec, and M. Steen, "Gossip-based Peer Sampling", ACM Transactions on Computer Systems, Vol. 25, No. 3, 2007, pp. 1-36.
- [16] C. Kim, and J. Ahn, "Decentralized Multi-Group Communication Protocol Supporting Causal Order", in

- Proceedings of the Second International Conference on Communication Software and Networks, 2010(ICCSN'10), Singapore, Feb. 2010, pp. 444-448
- [17] J. Lifton, M. Laibowitz, D. Harry, N.-W. Gong, M. Mittal, J. and A. Paradiso, "Metaphor and Manifestation—Cross-Reality with Ubiquitous Sensor/Actuator Networks", IEEE Pervasive Computing, Vol. 8, No. 3, 2009, pp. 24-33.
- [18] C. Meesookho, S. Narayanan, and C. S. Raghavendra, "Collaborative Classification Applications in Sensor Networks", in Proceedings of Sensor Array and Multichannel Signal Processing Workshop, Rosslyn, USA, Aug. 2002, pp. 370-374.
- [19] L. Peterson, N. Buchholzand, and R. Schlichting, "Preserving and using context information interprocess communication", ACM Transaction Computer Systems, Vol. 7, No. 3, 1989, pp. 217-246.
- [20] S. Pleisch, and K. Birman, "SENSTRAC: Scalable Querying of SENsor Networks from Mobile Platforms Using TRACKing-Style Queries", International Journal of Sensor Networks. Vol. 3, Issue 4, 2008, pp. 266-280.
- [21] G. Pottie, and W. Kaiser, "Wireless Integrated Network Sensors", Communications of the ACM, Vol. 43, No. 5, 2000, pp. 51-58.
- [22] L. Rodrigues, R. Baldoni, E. Anceaume, and M. Raynal, "Deadline-Constrained Causal Order," in Proceedings of Third IEEE International Symposium on Object-Oriented Real-Time Distributed Computing, 2000, pp. 234-243.
- [23] N. Tolgyesi, and M. Jelasity, "Adaptive Peer Sampling with Newscast", in Proceedings of Euro-Par 2009, LNCS, vol. 5704, Delft, the Netherlands, Aug. 2009, pp. 523-534.
- [24] Y. Vigfusson, K. Birman, Q. Huang, and D. Nataraj, "GO:Platform Support For Gossip Applications", in Proceedings of IEEE Ninth International Conference on Peer-to-Peer Computing(P2P 2009), Las Vegas, USA, Jan. 2009, pp.222-231.
- [25] J. Yick, B. Mukherjee, and D. Ghosal, "Wireless sensor network survey", Computer Networks, Vol. 52, Issue 22, Aug. 2008. pp. 2292-2330.

* Corresponding author: Tel.: +82 31 249-9674

CHAYOUNG KIM B.S. and M.S. degrees from the Sookmyung Women's University, Seoul, Korea, in 1996 and 1998, respectively and Ph.D. degree from the Korea University in 2006. From 2005 to 2008, she was a senior researcher in Korea Institute of Science and Technology Information, Korea, where she has been engaged in National e-Science of Supercomputing Center. Since 2009, she has been a researcher at Contents Convergence Software Research Center in Kyonggi University, Korea. Her research interests include distributed computing, group communications and peer-to-peer computing.

JINHO AHN(Corresponding author) received his B.S., M.S. and Ph.D. degrees in Computer Science and Engineering from Korea University, Korea, in 1997, 1999 and 2003, respectively. He has been an associate professor in Department of Computer Science, Kyonggi University. He has published more than 70 papers in refereed journals and conference proceedings and served as program or organizing committee member or session chair in several domestic/international conferences and editor-in-chief of journal of Korean Institute of Information Technology and editorial board member of journal of Korean Society for Internet Information. His research interests include distributed computing, fault-tolerance, sensor networks and mobile agent systems.

Mobile Agent PLM Architecture for extended enterprise

Abdelhak Boualaam*, El Habib Nfaoui*, Omar EL Beqqali*

* Sidi Mohamed Ben Abdellah University
GRMS2I FSDM B.P 1796 Fez-Atlas, Morocco

Abstract

Nowadays manufacturers are under increased pressure to have an add value for their products to struggle the low-cost production in emerging countries. Distributed control and Intelligent Product are a new and exciting opportunity to build more effective process networks for a wide range of applications in logistics and product development. Radio Frequency Identification is applied increasingly; this technology applied in conjunction with the Mobile Agent system can bring more values in managing and control the lifecycle of products by optimizing the three essential factors: cost, quality and deadline for the survival of a company in the competitive manufacturing world. In this paper we propose Mobile Agent PLM Architecture for extended enterprise, based on Mobile Agent and RFID or, more generally, Product Embedded Information Devices (PEID), for tracking and managing the information of the whole product lifecycle in the extended enterprise, and to satisfy new requirements for increased integrability, traceability, adaptability, extensibility, and closed-loop PLM. Mobile Agents are suitable for tracking information in distributing environment and the mobility aspect, at any time and any place. This paper proposes a first architecture based on these technologies.

Keywords: Intelligent product, PLM, closed-loop PLM, traceability, Mobile Agent, RFID.

1. Introduction

The decentralized information context, the distributed decision-making authority, the integration of physical and informational aspects, and the cooperative relationship among product lifecycle management, make the Intelligent Product a new and interesting paradigm, with great potential for meeting today's agile manufacturing challenges. Critical issues to be investigated include how to define intelligent product for a given problem within a specific context, what should be the appropriate system architecture, how to design effective cooperation mechanisms for good system performance, inspired by what happen in nature with us as human beings and the way we develop intelligence and knowledge [1]. In this paper, a new approach based on Mobile Agent which manages and controls the Intelligent Product is proposed and developed for extended enterprise.

However, equally or more importantly, Auto ID systems provide the basic infrastructure for reconsideration and possible alterations of the product and agent product. This is based on the observation that a physical product connected to a network can potentially access and affect its own functions. That is, through this network connection, a product (or a set of products) can interact indirectly with those operations that they come in contact with. In order to convey any status changes in real time, the PLM mobile agent uses radio frequency identification (RFID) technology to transform the physical processes and statuses into information flow. Receiving data from RFID middleware, the system conveys all changes to corresponding mobile agents automatically. We refer in this document to such products being 'intelligent' in a loose sense; we also introduce the concept of an Intelligent Product and we consider its potential impact on the entire product lifecycle.

This paper is organized as follows; section 2 will analyze different proposals for defining Intelligent Products; a classification of intelligence will be also presented. Section 3 present related technologies (RFID and Mobile Agent). In Section 4, we present our proposed approach for Intelligent Product. Finally, Section 5 provides some concluding remarks and future work.

2. Intelligent Product: Background review

In this section, we give an overview on the concept of Intelligent Product by presenting the recent principal definitions in the literature. All these definitions focus on certain aspects of Intelligent Products and on certain application areas or parts of the product lifecycle.

According to the Oxford English Dictionary an intelligent device or machine is one which is "able to vary its behaviors in response to varying situations, requirements and post experience". McFarlane [2] defines an Intelligent Product- as a product whose information content is permanently bound to its material content and which is able to influence decisions made about it; McFarlane

formalizes the concept of an intelligent product with the following working definition:

An intelligent product is a physical and information-based representation of a product which:

- Possesses a unique identification;
- Is capable of communicating effectively with its environment;
- Can retain or store data about itself;
- Deploys an adequate language to display its features, production requirements, etc.;
- Is able to participate in decisions relevant to its own destiny.

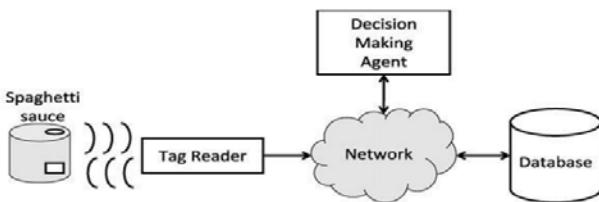


Fig. 1 Intelligent jar of spaghetti sauce [3].

Based on this definition, Wong and al. [3] have defined a two levels classification of intelligence. When the Intelligent Product only covers points 1–3, it is information oriented, and is called a product with level 1 product intelligence. A product with level 2 product intelligence covers all points, and is called decision oriented. Even though this Intelligent Product classification is quite generic concerning the level of intelligence of an Intelligent Product, it is based on a separation between the actual product and its information-based counterpart (as seen in Fig. 1). Therefore, it is mainly intended for describing the use of RFID technology in for example manufacturing and supply chain purposes, without covering for instance products with embedded processing

and communication capabilities.

We also found in the literature another definition complementary to previous ones, given by Karkkainen and al.; the fundamental idea behind an Intelligent Product according to Karkkainen and al. [4] is the inside-out control of the supply chain deliverables and of products during their lifecycle. In other words, the individual products in the supply chain themselves are in control of where they are going, and how they should be handled. To move to inside-out control of products, the products should possess the following properties:

- Globally unique identification code;
- Links to information sources about the product across organizational borders, either included in the identification code itself or accessible by some look-up mechanism;
- Can communicate what needs to be done with them to information systems and users when needed (even proactively).

Another definition of Intelligent Products is given by Venta in [5]. Venta refers by intelligence to products and systems that:

- Continuously monitor their status and environment.
- React and adapt to environmental and operational conditions.
- Maintain optimal performance in variable circumstances, also in exception cases.
- Actively communicate with the user, environment or with other products and systems.

Based on these definitions, a novel three-dimensional classification of Intelligent Products is introduced [6], which covers all the main aspects of the field. This classification model is shown in Fig. 2.

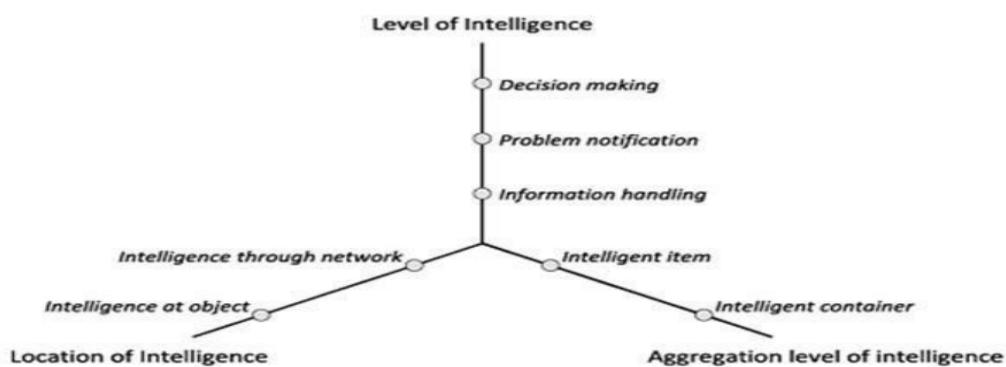


Fig. 2 Classification model of Intelligent Products [6]

3. Technologies enabling intelligent product

In this section we analyze the technologies behind Intelligent Products from two main points of view, the Mobile Agents technology and the Radio Frequency Identification.

3.1 Mobile Agent technology

Agents can be loosely defined as 'software that assist people on their behalf...and are delegated to perform task(s), and given constraints under which they can operate'.

In other term, in [7] Mobile agent is a kind of software. It is autonomous, and it migrates from one host to another in diverse network environments, and can be used in distributed architecture for the decision-making process [8]. It can transmit messages, distribute resources, and interact with other mobile agents or distributed resource systems. The mobile agent accepts tasks assigned by its owner, and then move onto the Internet to platforms that provide related services to carry out its task. When the work is complete, the mobile agent reports back to its owner. A mobile agent must have the following properties [9]:

- (1) It should be able to achieve one or more goals automatically.
- (2) It should be able to clone and propagate itself.
- (3) It should be able to collaborate and communicate

However, agents come in a myriad of different types, usually depending on the nature of their environment. What has been needed in a classification scheme to distinguish between different types of agents?

This paper does not propose a further definition of what an agent is, but will adopt the Franklin and Graesser [10] classification scheme and categorize the agents dealt with here as goal-oriented, communicative, mobile agents.

Franklin and Graesser definitions are as follows:

- goal oriented agents - agents that do no simply act in response to the environment;
- communicative agents - those able to communicate with other agents;
- Mobile agents - those able to transport themselves from one host to another.

Our interest in mobile agents should not be motivated by the technology per se, but rather by the benefits they provide for the creation of distributed systems, such as extended enterprise in paper. So there are seven reasons to use mobile agents [12]:

- (1) reduce network load;
- (2) overcoming network latency;
- (3) encapsulate protocols;
- (4) asynchronously execution and autonomously;
- (5) dynamic adaptation;

Table 1 List of commonly used mobile agents [11]

| <i>name</i> | <i>developer</i> | <i>Language</i> | <i>Application</i> |
|--------------------|-----------------------|---------------------|---|
| Agent Tcl | R. Gray, U Dart. | Tcl Tk | Information management |
| AgentSpace | Ichiro Sato, O. U. | Java | General purpose |
| AgentSpace | Alberto Sylva | Java | Support for dynamic and dist. Appl. |
| Aglet | IBM, Tokyo | Java | Internet |
| Ajanta | Minoseta U. | Java | General purpose |
| Ara | U Kaiserslautern | C/C++, Tcl, Java | Partially connected c. D.D.B. |
| Concordia | Mitsubishi E.I.T. | Java | Mobile computing, Data base |
| JATlite | Standford U. | Java | Information retrieval, Interface agent |
| Kafka | Fujitsu Lab. Japan | Java, UNIX-based | General purpose |
| Kali Scheme | NEC Research I. | Scheme | Distributed data mining, load balancing |
| Knowbots | CNRI | Python | Distributed systems/Internet |
| Messengers | UCI | C (Messenger-C) | General purpose |
| MOA | OpenGroup, UK | Java | General purpose |
| Mole | Stuttgart U. Germany | Java, UNIX-based | General purpose |
| OAA | SRI International, AI | C, C-Lisp, Java, VB | General purpose |
| Odyssey | General Magic | Telescript | Electronic commerce |
| Plangent | Toshiba Corporation | Java | Intelligent tasks |
| Tacoma | Norway & Cornell | C, UNIX-based, | Client/Server model issues/OS support |
| The Tube | David Halls, UK | Scheme | Remote execution of scheme |
| Voyager | ObjectSpace | Java | Support for agent systems |

- with other software and agents.
- (4) It must have a scope of competence.
 - (5) It should have some evolution states to record the computation status.
 - (6) naturally heterogeneous;
 - (7) robust and fault-tolerant.

Various projects applied the Mobile Agent paradigm to add the values in manufacturing intelligent product:

Holonic Manufacturing Systems (HMS) [13], in the manufacturing context, a Holonic Manufacturing System is seen as an autonomous and co-operative building block of a system for transforming, transporting, storing and/or validating information and physical objects.

Ubiquitous manufacturing, agent system paradigm to collaborative negotiation in a global manufacturing supply chain network [14].

Timon C. Du and al. in [15] propose a framework for using mobile agents to demonstrate autonomous behavior in the electronic marketplace.

Mobility and autonomy are keys characteristics when considering Mobile Agent technology as a component of intelligent product. This implies that agents can move around a network according to some itinerary. Various mobile agents systems (frameworks) support both mobility and itinerary, such as Aglet software development kit and framework. The Aglets software development kit conforms to the MASIF (Mobile Agent System Interoperability Facility) standard [16]. MASIF is a standard for mobile agent systems which has been adopted as an OMG technology.

Once created, an aglet object can be dispatched to and/or retracted from a remote host, deactivated and placed in secondary storage, then activated later (fig. 3).

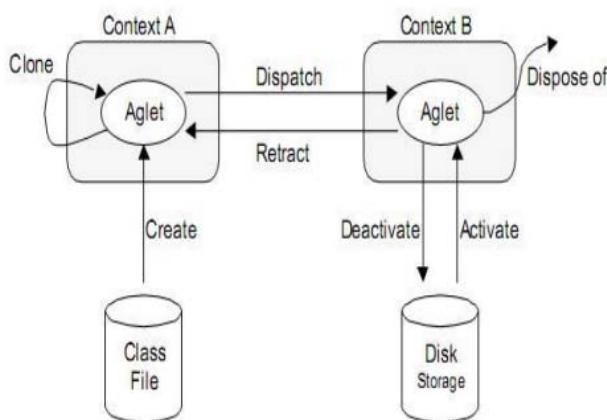


Fig. 3 Aglet Life-Cycle Model [15].

In the literature there are many available mobile agent systems (frameworks) for developing mobile agent. Table 1 summarizes the commonly used mobile agent systems for research and applications. Java Agent Development framework (JADE Framework) [17] is also used to develop agent applications in compliance with the

Foundation for Intelligent Physical Agents (FIPA) specifications.

3.2 RFID technology

In this part we present Radio Frequency Identification (RFID) technology and its potential applications in industry. RFID technology offers wireless communication between RFID tags and readers with non line-of-sight readability. These fundamental properties eliminate manual data entry and introduce the potential for automated processes to increase project productivity, construction safety, and project cost efficiency.

RFID can increase the service and performance of the construction industry with applications in materials management, tracking of tools and equipment, automated equipment control, maintenance and service, document control, failure prevention, quality control, field operations, and construction safety.

The number of radio frequency identification (RFID) applications in different industries increases continuously. Cumulative sales of RFID tags up to the beginning of 2006 reached 2.4 billion. In 2005 alone, 600 million tags were sold, which presents the trend in RFID allocation [18]. While the initial RFID application areas were generally industrial, applications like retail sector, supply chain management, warehouse management, logistics, manufacturing, military applications, and service sector also present potential applications for RFID technologies [18]. Radio-Frequency Identification (RFID) technology is a wireless sensor technology which is based on the detection of electromagnetic signals. They operate with up to 1MB of memory and have longer reading ranges because of the internal power supply. Their operating power is obtained from the transceiver device. However, they have shorter reading ranges and require a higher-powered reader than active tags.

The RFID is an automatic identification technology, relying on storing and remotely retrieving data by using devices called RFID tags or transponders. RFID provides a contact-less data accessing solution. If a product is attached with RFID tag, then the tag information will be read by the RFID reader and feedback to the backend PLM mobile agent system. Thus, the manual data input can be eliminated. When a RFID reader detects a tag, the mobile agent will receive the tag content and the reader ID. Then, the agent will determine an event type and initiate a RFID event message.

An RFID system is constituted by:

- an RFID device (tag);
- a tag reader with an antenna and transceiver;

- and a host system or connection to an enterprise system (Fig. 4).

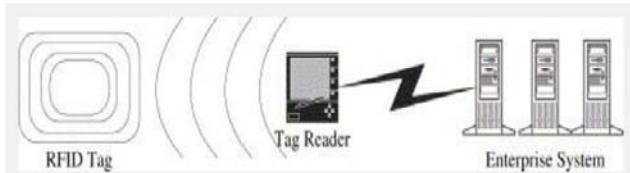


Fig. 4 A typical RFID system

Currently there is a considerable work being undertaken in the rationalization of frequency spectrum allocation between countries, development of standards and the introduction of many commercial applications. There are now over 350 patents registered with the US Patent Office related to RFID and its applications (Table 2).

Table 2 The decades of RFID [19]

| Decade | Event |
|-----------|--|
| 1940/1950 | Radar refined and used major World War II development effort. RFID invented in 1948. |
| 1950/1960 | Early explorations of RFID technology, laboratory experiments. |
| 1960/1970 | Development of the theory of RFID. Start of application field trials. |
| 1970/1980 | Explosion of RFID development. Tests of RFID accelerate. Very early adopter implementations of RFID. |
| 1980/1990 | Commercial applications of RFID mainstream |
| 1990/2000 | Emergence of standards. RFID widely deployed. RFID becomes a part of everyday life |

In contrast to the typical use of RFID technology today in warehouse management and supply chain applications, the paper proposes a Mobile AGENT PLM Architecture for extended enterprise, it combines Mobile Agent and RFID or, more generally, Product Embedded Information Devices (PEID), for tracking and managing the information of the whole product lifecycle in the extended enterprise, and to satisfy new requirements for increased integrability, traceability, adaptability, extensibility, and closed-loop PLM. Multi Mobile Agent and Radio Frequency Identification systems have been commonly recognized as enabling technologies for designing and implementing next generation of industrial control systems featuring.

After discussing the concepts of Intelligent Product and the technologies behind, the next section on our

proposed approach, and its contribution in product lifecycle management.

4. Proposed approach

Currently, the manufacturing industry is moving more and more from a supplier-driven to a customer-driven market. Due to the growing industrial capacity, customers are provided with a greater choice, and competition between suppliers is increased. As a result, companies must shorten product life cycles, reduce time-to-market, increase product variety and instantly satisfy demand, while maintaining quality and reducing investment costs. This is a great challenge to the manufacturing process itself; it must be more flexible and robust as well as demonstrate enhanced scalability. Therefore, the ends for introducing the Intelligent Product concept in manufacturing are to improve production planning and control, to enable customized products and to make change-over between product variants more effective.

This section is focused on our proposed approach, which is based on the next postulate “the product is an actor who manages its evolution in cooperation with the various actors of the life cycle, equipped with Intelligent Data Unit using Radio Frequency Identification (RFID), and the Mobile Agent”. This approach combines the Mobile Agent and RFID technologies for overcoming the challenges described above.

The system architecture is shown in Fig. 5. In this architecture, the physical product is identified and characterized by the intelligent agent, and the product is tracked and traced throughout its life cycle. Our approach involves the mutation of the simple physical product to an ambient product or actor capable to communicate with other players throughout its life cycle with the property of having the ‘inside-out’ communication. The main advantage of this architecture is that it has certain capability to perform tasks, communicate with other agents in a given environment, and to cooperate. We use RFID-based product information management to enabling item-level “track and trace”, and to complete, accurate, and timely information to support the product lifecycle management.

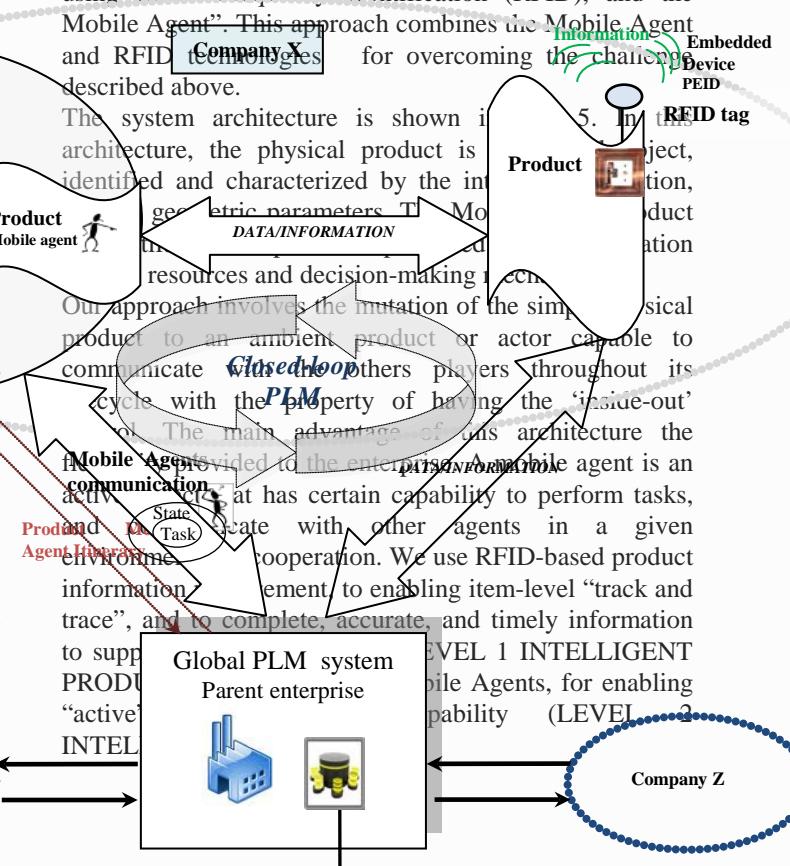


Fig. 5 Proposed architecture

PLM: Product Lifecycle Management

PEID: Product Embedded Information Devices

RFID: Radio Frequency Identification

PDKM/DB: Product Data Knowledge Management/Database

The complete architecture includes a Mobile Agent for each product. In our scenario, the main actor is the Product. It will be obviously consider as an active actor (RFID property, Mobile Agent party). But other Mobile Agents are acting in the same environment for cooperation. In the first second of our scenario, each Mobile Agent involved will have its unique identification (the same identification

PDKM/DB These agents could easily act on behalf of the owner (e.g. responsible designer). When the product is delivered for company X to the parent enterprise, the corresponding Mobile Agent migrates to accompanying her counterparty physical Product, with the possible feedback to its environment in the real-time, when there is a change in the product information.

Therefore the proposed architecture makes possible the tracking and managing the information of the whole product lifecycle in extended enterprise, with an automatic feedback of information about all product lifecycle phases, to build the knowledge database, manage and re-use the information.

5. Conclusion and future works

In this paper we reviewed the concept, the definitions, the classification of intelligence, and practical ends of Intelligent Products. As discussed above, the Intelligent Products combines many disciplines and could be used in many ways. We proposed a new architecture based on the Mobile Agents and the Radio Frequency Identification (RFID), for tracking and managing the information of the whole product lifecycle in the extended enterprise, and to satisfy new requirements for increased integrability, traceability, adaptability, extendibility, and closed-loop Product Lifecycle Management.

The architecture proposed in this research is the first step for a generic, « intelligent », layer interface based on mobile agents and RFID technologies with existing Product Lifecycle Management (PLM) systems. In the future work, we will model and implement the mobile agent architecture by using a framework. It must be able to communicate in real-time, and also able to interact with existing PLM systems and other information systems such as Enterprise Resource Planning (ERP), Advanced Planning and Scheduling (APS), Manufacturing Execution System (MES).

References

- [1] D. Kiritsis, Closed-loop PLM for intelligent products in the era of the internet of things, Computer-Aided Design, 2010.
- [2] D. McFarlane, S. Sarmab, J.L. Chirna, C.Y. Wonga, K. Ashtonb, Auto ID systems andintelligent manufacturing control, Engineering Applications of Artificial Intelligence 16 (2003) 365–376.

- [3] C.Y.Wong, D.McFarlane, A. Zaharudin, V. Agarwal, The intelligent product driven supply chain, in: Proceedings of SMC'02, 2002.
- [4] M. Kaarkainen, J. Holmstrom, K. Framling, K. Artto, Intelligent products—a step towards a more effective project delivery chain, Computers in Industry 50 (2003) 141–151.
- [5] O. Venta, Intelligent products and systems, Technical Report, VTT, 2007.
- [6] G.G. Meyer, K. Framling, J. Holmstrom, Intelligent Products: A survey, Computers in Industry 60 (2009) 137–148.
- [7] Y.F Chung, T.S Chen, M.W Lai, Eficient migration access control for mobile agents, Computer Standards & Interfaces 19) 1061–1068.
- [8] PDKM/DB rout, A. Bouras, E.H. Nfaoui, and O. El Beqqali, A Collaborative Decision-making Approach for Supply Chain Based on a Multi-agent System, Artificial Intelligence Techniques for Networked Manufacturing Enterprises Management, ISSN 1860-5168, ISBN 978-1-84996-118-9, Springer London Dordrecht Heidelberg New York, 2010, pp. 125–145.
- [9] T.K. Shih, Mobile agent evolution computing, Information Sciences, 137 No.1, 2001, pp. 53–73.
- [10] S. Franklin, A. Graesser, Is it an Agent, or just a Program? : A Taxonomy for Autonomous Agents, Proceedings of the 3rd Int. Workshop on Agent Theories, Architectures, and Languages, Berlin, Springer-Verlag.
- [11] W. Shen, Q. Hao, H.J. Yoon, D.H. Norrie, Applications of agent-based systems in intelligent manufacturing: An updated review, Advanced Engineering Informatics 20 (2006) 415–431.
- [12] D.B. Lange, M. Oshima, Mobile Agents with Java: The Aglet API.
- [13] S. Bain, H. Panetto, G. Morel, New paradigms for a product oriented modeling: Case study for traceability, Computers in Industry 60 (2009) 172–183.
- [14] Y. Zhang, G.Q. Huang, T. Qu, O. Ho, S. Sun, Agent-based smart objects management system for real-time ubiquitous manufacturing, Robotics and Computer-Integrated Manufacturing (2010).
- [15] T.C. Dua, E.Y. Lib, E. Weic, Mobile agents for a brokering service in the electronic marketplace, Decision Support Systems 39 (2005) 371– 383.
- [16] MASIF (Mobile Agent System Interoperability Facility) specification, OMG TC Document orbos/98-03-09. Available from <ftp://ftp.omg.org/pub/docs/orbos/98-03-09.pdf>, 1998.
- [17] F. Bellifemine, G. Cire , Greenwood D. Developing multi-agent systems with JADE. USA: John Wiley & Sons, Ltd; 2006.
- [18] B. Oztays, S. Baysan, F. Akpinar, Radio frequency identification (RFID) in hospitality, Technovation 29 (2009) 618–624.
- [19] The history of RFID, Association for Automatic Identification and Mobility, 2010, <http://www.aimglobal.org>.

Abdelhak Boulaalam is a PhD student at Sidi Med Ben AbdEllah University (GRMS2I group) in Morocco. He received his Master in Computer Science from the University of Sidi Md Ben AbdEllah in

2008. His current research interests are Multi-Agent Systems applied to Product Lifecycle Management.

El Habib Nfaoui is currently an associate Professor at the University of Sidi Med Ben AbdEllah, he obtained his PhD in Computer Science from University of Sidi Med Ben AbdEllah (GRMS2I group) in Morocco and University of Lyon (LIESP Laboratory) in France under a COTUTELLE (co-advising) agreement. His current research interests are Multi-Agent Systems and distributed simulation applied to supply chain context, decision-making and modeling.

Omar El Beqqali is currently Professor at Sidi Med Ben AbdEllah University. He is holding a Master in Computer Sciences and a PhD respectively from INSA-Lyon and Claude Bernard University in France. He is leading the 'GRMS2I' research group since 2005 (Information Systems engineering and modeling) of USMBA and the Research-Training PhD Unit 'SM3I'. His main interests include Supply Chain field, distributed databases and Pervasive information Systems. He also participated to MED-IST project meetings. O. El Beqqali was visiting professor at UCB-Lyon1 University, INSA-Lyon, Lyon2 University and UIC (University of Illinois of Chicago). He is also an editorial board member of the International Journal of Product Lifecycle Management (IJPLM).

Load Balancing using High Performance Computing Cluster Programming

Dr. Sumit Srivastava¹, Mr. Pankaj Dadheech² and Mr. Mahender Kumar Beniwal³

¹ Information Technology Department, Rajasthan Technical University, Poornima College of Engineering, Jaipur, Rajasthan 302025/Sitapura, India

² Computer Science & Engineering Department, Rajasthan Technical University, Swami Keshvanand Institute of Technology, Management & Gramothan, Jaipur, Rajasthan 302025/Jagatpura, India

³ Computer Science & Engineering Department, Rajasthan Technical University, Swami Keshvanand Institute of Technology, Management & Gramothan, Jaipur, Rajasthan 302025/Jagatpura, India

Abstract

High-performance computing has created a new approach to science. Modeling is now a viable and respected alternative to the more traditional experiential and theoretical approaches. High performance is a key issue in data mining or in image rendering. Traditional high performance clusters have proved their worth in a variety of uses from predicting the weather to industrial design, from molecular dynamics to astronomical modeling. A multicomputer configuration, or cluster, is a group of computers that work together. A cluster has three basic elements—a collection of individual computers, a network connecting those computers, and software that enables a computer to share work among the other computers via the network. Clusters are also playing a greater role in business. Advances in clustering technology have led to high-availability and load-balancing clusters. Clustering is now used for mission-critical applications such as web and FTP servers. For permanent clusters there are, for lack of a better name, cluster kits, software packages that automate the installation process. A cluster kit provides all the software you are likely to need in a single distribution. Cluster kits tend to be very complete. For example, the OSCAR distribution. Open Source Cluster Application Resources is a software package that is designed to simplify cluster installation. A collection of open source cluster software, OSCAR includes everything that you are likely to need for a dedicated, high-performance cluster. OSCAR takes you completely through the installation of your cluster. In this Paper with the help of Open Source Cluster Application Resource (OSCAR) cluster kit, attempt to setup a high performance computational cluster with special concern to applications like Integration and Sorting. The ease use of cluster is possible globally and transparently managing cluster resources. Cluster computing approach nowadays is an ordinary configuration found in many organizations to target requirements of high performance computing.

Keywords: Clustering, Performance analysis, Web clustering, Workload characterization, High Performance.

1. Introduction

Computing speed isn't just a convenience. Faster computers allow us to solve larger problems, and to find solutions more quickly, with greater accuracy, and at a lower cost. All this adds up to a competitive advantage. In the sciences, this may mean the difference between being the first to publish and not publishing. In industry, it may determine who's first to the patent office.

Traditional high-performance clusters have proved their worth in a variety of uses—from predicting the weather to industrial design, from molecular dynamics to astronomical modeling. High-performance computing (HPC) has created a new approach to science—modeling is now a viable and respected alternative to the more traditional experiential and theoretical approaches. High performance is a key issue in data mining or in image rendering. Advances in clustering technology have led to high-availability and load-balancing clusters. Develop the algorithms for the Sorting that are faster and more accurate than they were ever before. The problem with the huge data sorting is that it takes a lot of time and as a result a large amount of money is required. This huge money is just wasted in the sorting of data and not in any useful work so, it is highly required that this time be as less as possible. So we came up with the some new methods to lower that Complexity. The approach we followed was High Performance Computing.

2. Background

The field of High Performance Computing is the most popular for getting faster results for any application. The applications designed using High Performance Computing concepts are Numerical Integration, Quick Sorting. Application for Numerical Integration is designed for calculating the integration of function $\cos(x)$ under the limits 0 to $\pi/2$. Such integration application can be easily executed over several nodes simultaneously so as to reduce the work load on single computer for calculating the integration of function completely.

2.1 The main Applications we took up as the Research were:

- Numerical Integration of function $\cos(x)$ for limits 0, $\pi/2$
- Quick Sorting

The strength and weakness of each algorithm and methodology is to be analyzed and then new methodologies and algorithms are to suggested for doing the above stated work. This methodology could be evaluated in certain measures like:

- Sorting any amount of Data
- Complexity be as less as possible
- Fast in response and output
- Reliable

The basic unit of a cluster is a single computer, also called a “node”. Clusters can grow in size - they “scale” - by adding more machines. The power of the cluster as a whole will be based on the speed of individual computers and their connection speeds are. In addition, the operating system of the cluster must make the best use of the available hardware in response to changing conditions. This becomes more of a challenge if the cluster is composed of different hardware types (a “heterogeneous” cluster), if the configuration of the cluster changes unpredictably (machines joining and leaving the cluster), and the loads cannot be predicted ahead of time.

High performance is a key issue in data mining or in image rendering. Advances in clustering technology have led to high-availability and load-balancing clusters. OSCAR is a package of RPM's, Perl-scripts, libraries, tools, and whatever else is needed to build and use a modest-sized Linux cluster. OSCAR is an Open Source project. Every component within OSCAR is available under one of the well known Open-Source licenses (e.g. GPL). The goal of OSCAR is making clusters easy to build, easy to maintain and easy to use. In other words,

OSCAR contains the resources need to apply cluster computing to High Performance Computing problems.

3. Implementation and Applications

The numerical Integration algorithm can divide a big problem of integration into several smaller problems of integration. For decomposing the problem, the area under the curve of the function has to be divided into rectangles. These smaller segments of problem can be distributed over the cluster (i.e. the nodes) so that they can be executed simultaneously and hence result can be produced in a shorter time than computing the whole problem on a single machine. Using the concept of Message Passing Interface, we can also select the number of processes in which we want our problem to be decomposed and executed. On increasing the number of processes, the result can be generated by the processors more accurately.

Another application which can be designed using MPI libraries is Quick sort algorithm in which sorting of the numbers can be done using the Divide and Conquer Rule. In Quick sort, the array of numbers is divided into partitions according to the position of the pivot element. The elements lesser than pivot come before it and elements greater than pivot comes after it. Thus the partitions can be distributed over cluster and this quick sort mechanism can be called recursively so as to sort the partitions. Using HPC, these problems can be executed in a shorter time and more accurately.

The major difficulty in parallel programming is subdividing problems so that different parts can be executed simultaneously on different machines. MPI is used to determine how a problem can be broken into pieces so that it can run on different machines.

With most parallelizable problems, programs running on multiple computers do the bulk of the work and then communicate their individual results to a single computer that collects these intermediate results, combines them, and report the final results. As the program executes on each machine, it will first determine which computer it is running on and, based on that information, tackle the appropriate part of the original problem. When the computation is complete, one machine will act as a receiver and all the other machines will send their results to it. For this approach to work, each executing program or process must be able to differentiate itself from other processes.

To calculate the area under a curve, we use numerical integration method. This problem can be solved by parallel calculations because it can be easily decomposed into parts

that can be shared among the computers in a cluster. The numerical integration method used in this problem is data decomposition. Each process has a different set of bounds, so the area that each calculated is different but the procedure is the same.

To multiply two $m \times n$ matrices, we use decomposing the problem. It is essential to realize that there are a number of trade offs that must be balanced when dividing a problem. The idea is to break the algorithm into pieces of code or tasks based on the data required by that task. The best solution usually lies somewhere between maximizing concurrency and minimizing communication.

3. Results and Discussions

The process of parallel algorithm design can be broken into several steps. First, we must identify the portions of the code that can, at least potentially, be executed safely in parallel. Next, we must devise a plan for mapping those parallel portions into individual processes or onto individual processors. After that, we need to address the distribution of data as well as the collection and consolidation of results. OSCAR is designed with high-performance computing in mind. Basically, it is designed to be used with an asymmetric cluster. Generally decomposition strategies fall into two different categories—data decomposition or data partitioning, and control decomposition or task partitioning. With data decomposition, the data is broken into individual chunks and distributed among processes that are essentially similar. With control decomposition, the problem is divided in such a way that each process is doing a different calculation.

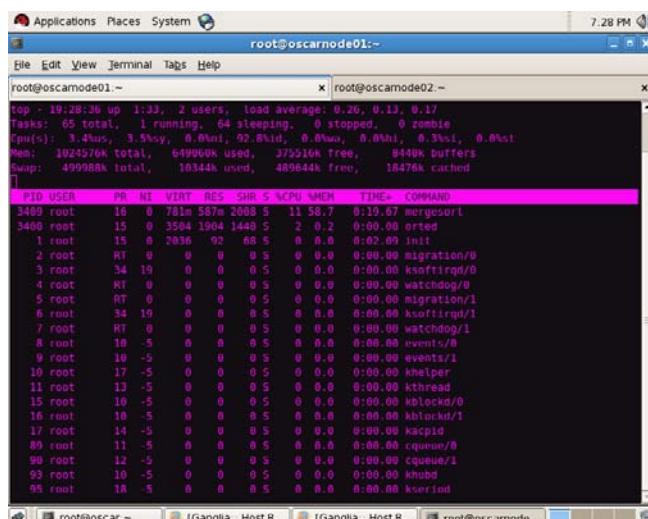


Fig. 1 Result on Node 1.

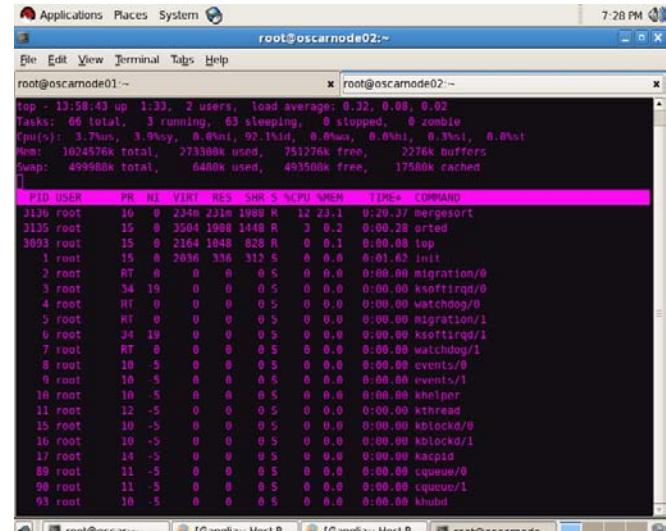


Fig. 1 Result on Node 2.

4. Conclusions

The basic objective of this paper is to setup a high performance computational cluster with special concern to applications like Integration and Sorting. By using a cluster kit such as OSCAR, the setting up a high performance cluster, could be performed and with the help of message passing libraries like LAM/MPI, the designing numerical integration and quick sort applications, is achieved. Finally, by designing test cases and checking the performance on two client nodes. A High Performance computational cluster using Open Source Cluster Application Resource Programming has achieved with better CPU utilization.

5. Future Work

Future work includes the formal analysis of existing code for automated conversion to a parallel version for cluster implementation. The research presented here is a work-in-progress. Our goal continues to be to meet the needs of performance tuning using high performance cluster programming with better CPU utilization.

Acknowledgments

Authors would like to thank the Department of CS/IT, of Poornima College of Engineering, Jaipur and Swami Keshvanand Institute of Technology, Management & Gramothan, Jaipur for providing computational facilities and Prof. C. M. Choudhary & Prof. Anil Chaudhary for their guidance.

References

- [1] Michael D. Kane, John A. Springer, "Integrating bioinformatics, distributed data management, and distributed computing for applied training in high performance computing", Proceedings of the 8th ACM SIGITE conference on Information technology education, Destin, Florida, USA, Pages: 33-36 Year of Publication: 2007 ISBN:978-1-59593-920-3
- [2] Ian Gorton, Daniel Chavarria-Miranda, Manoj kumar Krishnan, Jarek Nieplocha, "A High-Performance Event Service for HPC Applications", Proceedings of the 3rd International Workshop on Software Engineering for High Performance Computing Applications, Washington, DC, USA, Page: 1, Year of Publication: 2007, ISBN:0-7695-2969-0
- [3] Sadaf R. Alam, Jeffrey S. Vetter, Pratul K. Agarwal, Al Geist, "Performance characterization of molecular dynamics techniques for biomolecular simulations", Proceedings of the eleventh ACM SIGPLAN symposium on Principles and practice of parallel programming, New York, USA, Pages: 59 – 68, Year of Publication: 2006, ISBN:1-59593-189-9
- [4] Volodymyr Kindratenko, George K. Thiruvathukal, Steven Gottlieb, "High-Performance Computing Applications on Novel Architectures", Computing in Science & Engineering, Pages: 13 – 15, Volume : 10 , Issue:6, Nov.-Dec. 2008, ISSN : 1521-9615
- [5] James C Phillips, John E. Stone, "Probing Biomolecular Machines with Graphics Processors", Queue Bioscience, New York, NY, USA, Pages: 30, Volume 7 , Issue 9, October 2009, ISSN:1542-7730
- [6] Chrisila Pettey, Ralph Butler, Brad Rudnick, Thomas Naughton, "Simple maintenance of Beowulf clusters in an academic environment", Journal of Computing Sciences in Colleges, Pages: 208 – 214, Volume 18 , Issue 2, December 2002
- [7] Adit Ranadive, Mukil Kesavan, Ada Gavrilovska, Karsten Schwan, "Performance implications of virtualizing multicore cluster machines", Proceedings of the 2nd workshop on System-level virtualization for high performance computing, Glasgow, Scotland, Pages: 1-8, 2008, ISBN:978-1-60558-120-0
- [8] Rubing Duan, Radu Prodan, Thomas Fahringer, "Performance and cost optimization for multiple large-scale grid workflow applications", Proceedings of the 2007 ACM/IEEE conference on Supercomputing, Reno, Nevada, Article No.: 12, 2007, ISBN:978-1-59593-764-3
- [9] Michael A. Bauer, "High performance computing: the software challenges", Proceedings of the 2007 international workshop on Parallel symbolic computation, London, Ontario, Canada, Pages: 11 – 12, 2007, ISBN:978-1-59593-741-4
- [10] Marsha Meredith, Teresa Carrigan, James Brockman, Timothy Cloninger, Jaroslav Privoznik, Jeffery Williams, "Exploring Beowulf clusters", Journal of Computing Sciences in Colleges, Pages: 268 – 284, Volume 18, Issue 4 April 2003, ISSN:1937-4771
- [11] Lamia Youseff, Rich Wolski, Brent Gorda, Chandra Krintz, "Evaluating the Performance Impact of Xen on MPI and Process Execution For HPC Systems", Proceedings of the 2nd International Workshop on Virtualization Technology in Distributed Computing, Page: 1, 2006, ISBN:0-7695-2873-1
- [12] Xiao Qin, Hong Jiang, Adam Manzanares, Xiaojun Ruan, Shu Yin, "Dynamic load balancing for I/O-intensive applications on clusters", ACM Transactions on Storage (TOS), Article No.: 9, Volume 5 , Issue 3, November 2009, ISSN:1553-3077
- [13] Kathryn Mohror, Karen L. Karavanic, "A study of tracing overhead on a high-performance linux cluster", Proceedings of the 12th ACM SIGPLAN symposium on Principles and practice of parallel programming, San Jose, California, USA, Pages: 158 – 159, 2007, ISBN:978-1-59593-602-8
- [14] Becker, Donald J., Thomas Sterling, Daniel Savarese, John E. Darband, Udaya A. Ranawak, and Charles V. Packer. "BEOWULF: A parallel workstation for scientific computation." Proceedings of the 24th International Conference on Parallel Processing, Volume I. Boca Raton, Fla.: CRC Press, 1995.
- [15] Fang, Yung-Chin; Tau Leng, Ph.D.; Victor Mashayekhi, Ph.D.; and Reza Rooholamini, Ph.D. "OSCAR 1.1: A Cluster Computing Update." Dell Power Solutions, Issue 4, 2001.
- [16] Hsieh, Jenwei, Tau Leng, and Yung-Chin Fang. "OSCAR: A Turnkey Solution for Cluster Computing." Dell Power Solutions, Issue 1, 2001.
- [17] Vallee G, Scott S.L, Morin C, Berthou J.-Y, Prisker H, "SSI-OSCAR: a cluster distribution for high performance computing using a single system image", Cluster Computing and the Grid, 2001. Proceedings. First IEEE/ACM International Symposium on Issue Date: 2001 On page(s): 22 - 25 ISSN : 1550-5243 Print ISBN: 0-7695-2343-9 INSPEC Accession Number: 8598350
- [18] Luethke, Brian, Thomas Naughton, and Stephen L. Scott. "C3 Power Tools: The Next Generation." Paper to be presented at the Austrian-Hungarian Workshop on Distributed and Parallel Systems (DAPSYS 2002), Linz, Austria, September-October 2002.
- [19] Naughton, Thomas, Stephen L. Scott, Brian Barrett, Jeff Squyres, Andrew Lumsdaine, and Yung-Chin Fang. "The Penguin in the Pail-OSCAR Cluster Installation Tool." Paper presented at the World Multiconference on Systemics, Cybernetics and Informatics (SCI 2002), Orlando, Fla., July 2002.
- [20] Leangsukun, C. ; Shen, L. ; Tong Liu ; Hertong Song ; Scott, S.L. ; "Availability prediction and modeling of high mobility OSCAR cluster" Cluster Computing, 2003. Proceedings. 2003 IEEE International Conference on Issue Date: 1-4Dec.2003 On page(s): 380 - 386Print ISBN: 0-7695-2066-9, INSPEC Accession Number: 7947717
- [21] Limaye, K. ; Leangsukun, B. ; Munganuru, V.K. ; Greenwood, Z. ; Scott, S.L. ; Libby, R. ; Chanchio, K. ; "Grid aware HA-OSCAR" High Performance Computing Systems and Applications, 2005. HPCS 2005. 19th International Symposium on Issue Date : 15-18 May 2005 On page(s): 333 - 339 ISSN: 1550-5243 Print ISBN: 0-7695-2343-9INSPEC Accession Number: 8598352

- [22] Greidanus, P. ; Klok, G.W. ; "Using OSCAR to Win the Cluster Challenge" High Performance Computing Systems and Applications, 2008. HPCS 2008. 22nd International Symposium on Issue Date : 9-11 June 2008 On page(s): 47 -51 ISSN : 1550-5243 Print ISBN: 978-0-7695-3250-9 INSPEC Accession Number: 10067160
- [23] Mattson, T.G. ; "High performance computing at Intel: the OSCAR software solution stack for cluster computing" Cluster Computing and the Grid, 2001. Proceedings. First IEEE/ACM International Symposium on Issue Date : 2001 On page(s): 22 – 25 Print ISBN: 0-7695-1010-8 References Cited: 7 INSPEC Accession Number: 6957990
- [24] Laxman Nathawat ; "Tools and techniques for building High Performance Linux clusters" Source Journal of Computing Sciences in Colleges archive Volume 20, Issue 6 (June 2005)Pages: 55 - 56 Year of Publication: 2005 ISSN:1937- 4771
- [25] Kimura, K. ; Kodaka, T. ; Obata, M. ; Kasahara, H. ; "Multigrain parallel processing on OSCAR CMP" This paper appears in: Innovative Architecture for Future Generation High-Performance Processors and Systems, 2003 Issue Date : 17July2003 On page(s): 56 - 65 Print ISBN: 0-7695-2019-7 INSPEC Accession Number: 7979833



Mr. Mahender Kumar Beniwal received his M.Tech in Computer Science from JRN, Rajasthan Vidyapeeth University, Udaipur, India. He is pursuing his PhD at Suresh Gyan Vihar University, Jaipur. . He is currently working as a Reader in the Department of Computer Science & Engineering, Swami Keshvanand Institute of Technology, Management & Gramothan, Jaipur. He is a member of the IEEE and the IEEE Computer Society. His area of interest include Internet Technology, Internet Security Applications, OOPs, Network Programming, Software Engineering, Web Technology.



Dr. Sumit Srivastava received his PhD degree from University of Rajasthan, Jaipur, India. And he has received his M.Tech from Rajasthan Vidyapeeth, Jaipur and M.C.A. from Birla Institute of Technology, Mesra, Ranchi. He has more than 10 years of experience in teaching. He is currently working as an Associate Professor in the Department of Information Technology, Poornima College of Engineering, Jaipur. To his credit, he has more than 15 publishing in the proceedings of the reputed National & International conferences. He has 8 publications in various International Journals. He is also guiding thesis of many students of M.tech affiliate to Rajathan Technical University, Kota. His research interest include Statistical method in Data Mining, Grid & Cluster Computing and Network Security.



Mr. Pankaj Dadheech received his B.E. in Computer Science & Engineering from University of Rajasthan, Jaipur, India. He is currently working as a Senior Lecturer in the Department of Computer Science & Engineering, Swami Keshvanand Institute of Technology, Management & Gramothan, Jaipur. He has presented 7 papers in various National & International conferences. He is a member of many Professional Organizations like the IEEE and the IEEE Computer Society, CSI, & ACM. His area of interest include High Performance Computing, Security Issues in Grid Computing and Information Security.

Creating Multiuser Web3D Applications Embedded in Web Pages

Xandre Chourio, Francisco Luengo and Gerardo Pirela

Scientific Modeling Center (CMC), University of Zulia
Maracaibo, Zulia 4004, Venezuela

Abstract

It is not common to find web pages that show interactive multi-user 3D virtual environments as part of their contents without requiring special plug-ins for the web browser to be able to execute such an application. This paper presents a new architecture based on Java technology to create web portals which include 3D virtual scenarios that many users may share whose actions in that environment can incur in automatic updating of the rest of the portal's contents, without requiring any special additional software to be displayed by or installed unto any browser.

Keywords: *Web3D, java, web portal, client-server architecture.*

1. Introduction

Despite the increase of 3D applications on internet, the web is still being substantially a 2D world. It is uncommon to find web pages which include a 3D interactive environment among its elements. The usual mode of displaying information is still text, image, or video rendering on the webpage, and the predominant browsing mode remains via hyperlinks associated to those elements. Despite the great advancements in current technology, there still exists a considerable gap between the Web as we know it today and the Web3D as we imagine it.

Web applications use web browser as user interfaces. Their way of displaying information has evolved significantly from their beginnings. As Internet's presence and power, as well as diversity of contents, increase so does the user community's demand for more intuitive web applications with regards to information retrieval and handling. In fact, 3D web environments have existed for decades and languages such as VRML have tried to become the equivalent of HTML for 3D interactive environment visualization on the web.

A technology does not currently exist that offers the kind of information interaction and rendering capability found in desktop applications. The main reason for this is the heterogeneous network environment on which Internet

operates, as well as the security mechanisms needed to execute applications safely.

Current research and development trends allow us to visualize a new generation of web browsers with powerful GUIs which are able to handle and render realistic, interactive 3D content. But, in the meantime, many other research efforts point towards determining how to choose and combine current technology to achieve more effective and efficient development of web applications able to implement such interfaces as such technology combinations dictate the needs, uses, and capabilities of future web browsers.

In this article we propose a platform based on Java technology for publishing virtual 3D interactive, multiuser environments capable of interacting also with the rest of the web page's content and which can be viewed on standard web browsers without having to install any plugins. We additionally present a web portal to show the platform's functionality.

In section II we comment on Web3D and the attributes any Web3D application must possess. Section III presents an evaluation and selection of tools. Sections IV and V describe the platform we propose. Finally, conclusions are shown in section VI.

2. Web3D

Web3D groups any programming language, protocol, file format or any other technology that allows to integrate interactive 3D virtual environments as part of web page contents [19].

3D graphic applications for the web formally started with the virtual reality modeling language (VRML) which appeared in 1994 and has become since a standard for creating, transmitting, and representing 3D objects and 3D environments on the web. Later, in 2001 the Web3D Consortium [20] announced the release of X3D as the new standard succeeding VRML. X3D offered complete back-compatibility with its predecessor and included significant

improvement such as NURBS extension, humanoid animation, morphing, and the use of XML-like syntax, making it extensible and facilitating its use with other applications.

However, a web portal's vistosity is based not only on offering the user attractive elements, but also on how such elements behave under different events [1]. This functionality is commonly achieved by using programming languages, and hence companies such as Sun Microsystem Inc. have developed APIs for 3D graphics like Java3D [7], which is not oriented for web use but which can be combined with Java-Applets as an alternative for developing 3D web applications. Similarly, other Java APIs have emerged which try to take advantage of graphic hardware through OpenGL, such are the cases of JOGL [5] and LWJGL [9] among other products which do not require installing any additional software component, like jGL [2].

Java is one of the strongest languages for web application development due to its versatility, robustness, and multi-platform support. Its usability in the case of Web3D can be appreciated in applications such as Cortona Jet [3] and Shout3D [17] which combine proprietary technology with java-based interfaces to provide any web browser with capability for 3D interactive graphic rendering without requiring any special software installation. Another prominent example is the Project Wonderland [15] which was entirely developed in Java and uses Java3D and other technologies [8,14] to provide a set of tools which allow creation of virtual collaborative worlds in which the users can communicate and share applications in real time.

Currently, big companies and consortiums are renewing their interest in this technology, promoting the development of new tools for Web3D. In this particular, Google has recently launched O3D, an API written in JavaScript which allows construction of virtual, interactive 3D environment for the web [11]; however, this API seems to be oriented towards developing simple (low complexity) environments and this may become a limitation in the future. On the other hand, the Kronos Group is currently developing WebGL, an API similar to O3D but with native OpenGL support [21]. Another example of such new 3D graphics technology for the web is Ajax3D [13], which also uses JavaScript for drawing and interaction.

Even though the term Web3D refers to technology that allows to take 3D virtual environment to the Web, recently the term has also been used to encompass applications developed from such a technology, that is applications that deliver 3D content into Web [10,18].

A 3D application for the web exhibits specific properties according to its nature and which determine its functionality. Such properties and characteristics are listed below.

Such properties and characteristics are listed below.

- a. Hardware Independence – Due to the heterogeneous environment that surrounds Internet, it becomes necessary that applications exhibited therein may run into any platform.
- b. No software installation required – To protect users and ensure their safety, no installation should be required on the side of the client to run such applications.
- c. Graphic capability to create 3D virtual environments – Tools, such as OpenGL, add realism to the environment with the use of techniques such as shadowing, texturing, NURBS, etc., in order to create a web experience similar to that which desktop applications offer.
- d. Multiuser – Enabling sharing a common environment and allowing communication and interaction among several users.
- e. Highly interactive – The application should be versatile and offer the user easy control via a rich, friendly interface.
- f. Coupled with Web page contents – This implies that the application not only be embedded in a web page but that it also be able to complement itself interacting with other contents and resources in the same webpage, dynamically.

General web applications show many of the characteristics described above, except for graphic capability to create 3D virtual environments (point c in the list), which is specific to Web3D applications. However, access to this characteristic is precisely what drives developers to sacrifice some other attributes in the list. For example, in order to take advantage of the client's graphic hardware power, many Web3D applications require for its visualization that additional software be installed or that special applications be used. On the other hand, these applications tend to be independent from the rest of the hosting webpage's content; that is, the interaction in it does not affect the rest of the elements on the page (point f in the list) and hence, the application is not truly integrated to the web portal or it simply runs on a separate window.

We describe in the following sections implementation details of a Web3D application which show all the attributes listed above.

3. Technology Selection

Many researchers opt to combine several different tools to develop their Web3D applications because no single tool provides all the functionality they expect. Instances of this are products which integrate XHTML, scripting languages (JavaScript, Python), 3D markup languages (VRML, X3D),

web services, and AJAX-based technology to develop multiuser environments [4,12,23].

On the other hand, the Java platform offers developers the possibility to create embedded applications in web pages with high-level programming language and support for multiplatform development, security, communication, database connection, interaction, access to DOM (Document Object Model), among other capabilities [6]. Additionally, the Web3D Consortium promotes a library for X3D support written entirely in Java [22].

Both approaches allow the development of applications which satisfy the characteristic described in the previous section. However, since Java is a compiled programming language, its running time is faster than that of scripting languages such as JavaScript; moreover, Java development environments make it easier to create versatile, wide-range applications, which hence makes Java the ideal tool to develop Web3D applications as defined in section II.

However, Java's greatest drawback is that it does not offer native support for 3D graphic generation. One way of overcoming this drawback is using jGL [2] – a library for 3D drawing for Java which emulates OpenGL.

To illustrate jGL's scope for drawing 3D environments, we show below a performance comparative analysis against JOGL [5], a Java API which implements OpenGL and takes advantage of the client's graphic hardware power (for which it needs to be installed unto the computer where the application will run). Figure 1 shows the time the API takes to process a polygon grid at different resolution. Figure 2 shows the effect illumination and texturing techniques have upon both API's performance.

As expected, when taking advantage of the client's graphic hardware power, JOGL shows significantly superior performance as compared to jGL. However, these results allow us to set a polygon bound on 3D modeling and representation to achieve acceptable running time with jGL.

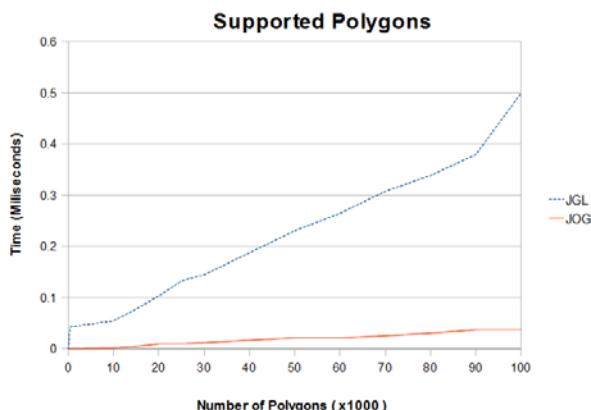


Fig. 1 jGL and JOGL's performances

Graphic Lighting and Texturing vs. Time

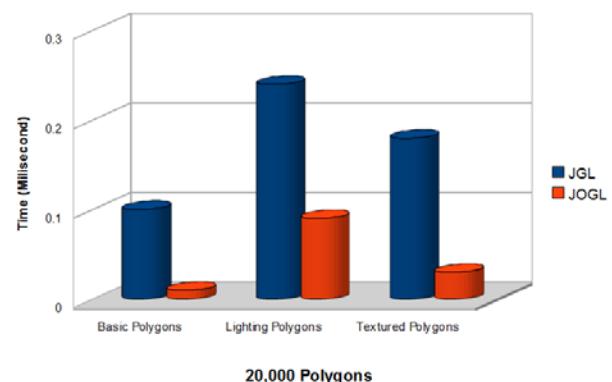


Fig. 2 Drawing, lighting and texturing polygons with jGL and JOGL.

4. Architecture

We describe below implementation details of a Web3D portal which meets all the requirements mentioned in Section II and which we call henceforth SMW3D (Multiuser Web3D System)

SMW3D relies upon three fundamental pillars, rendering, communication, and information. These pillars in turn orbit around the Java Virtual Machine – a robust, secure platform which current support allows it to be used in almost any operating system and web browser.

The structure of the portal consists of three java applets and other DHTML elements embedded in a web page (see Figure 3A). The main applet (Figure 3B) renders the virtual 3D environment, letting the user navigate through this environment, visualize other users who may be connected to the portal, and interact with them. This applet can also communicate with the rest of the web page elements to respond to the interaction being carried out. Two such elements are the other two applets (Figure 3C); one of them shows a 2D map of the environment with every connected user's location in it, and the other allows users' communication via chat. Figure 3D shows the rest of the elements of SMW3D – a frame to show a webpage and multimedia contents.

Additionally, SMW3D is built under a client-server architecture which deploys four server applications that provide data to make the portal fully functional (see Figure 4). These server applications are a Web Server (which attends to web page requests), a User Server (which keeps track of users connected to the portal), a Data Server (which provides data for 3D environment rendering and avatars), and a Communications Server (which allows communication among users). All those servers (except Web Server) were made using the Java Platform's main features (see [6]).

The SMW3D's architecture and communication schema among their components are described next.

4.1 Web Server

It is a standard web server (Apache, IIS, or any other) which hosts the portal and which answers the web browser's requests (when the portal is needed) as well as the main applet (to update the portal's DHTML elements with regards to the user interaction with the 3D environment).

4.2 User Server

This server is in charge of keeping track of all connected users, their position and orientation in the virtual environment, as well as their IP address. The server receives this information from every connected client's main applet every time a user's position within the environment is updated.

Every time the server receives an update of a client's position, orientation, or a new client's connection, it retransmits it to the rest of the connected clients. The application then is able to correctly draw all the avatars within all open portals' scenery. Additionally, the server sends the updated list of connected users to all other servers for them to correctly run their respective functions. Nevertheless, establishing peer-to-peer communication may overload the server due to constant traffic when many users are connected and in continual motion. An alternative could be a mixed communication scheme that includes multicast technology to avoid server overload; however, for security reasons, applets do not support multicast and so we opted for a broadcast communication scheme which also allows for more manageable data package size.

4.3 Data Server

This server is in charge of providing all open portals main applets the graphic information they need for the correct rendering of the 3D virtual environment (vertices, polygons, colors, lighting, textures, area identification, information associated with each area, delimitation of each area in the environment, etc.) In order to reduce the amount of data to be delivered for graphic rendering as users move through the environment, a database was created using MySQL DBMS and implementing a quadtree structure [16]. This data management technique is commonly used in videogames.

4.4 Communications Server

This server allows communication among users via chat, taking advantage of the connected user list provided by the user server. It manages connection better by listening for

open portals' chat applet requests through a different port than that used by the User server. It is important to mention that even though the User Server keeps and provides the list of connected users, the Communications Server also keeps its own list of chatting users (or users available for chat) and it is this latter list the one displayed in the chat applet as every user is autonomous to use this module or not without this decision affecting the user's status in the main environment

5. Multiuser Web3D System

Figure 4 illustrates this Multiuser Web3D System (SMW3D) which works as follows. The Web Server delivers to each connected user's portal all the resources associated with it. Once the portal has been loaded up onto the client's web browser, the main applet communicates with the User Server requesting use of the communication channel via an "alias" for the user (who must input such an alias along with gender and chat usage status). The User Server also checks the alias to avoid duplication. Once connection has been established, the main applet communicates with the Data Server requesting information about area in the virtual environment where the user is located so it can be rendered on the scene, along with information to be published in other components of the portal. Simultaneously, the User Server shares the alias with the Communication Server to establish a communication channel for the user with this resource at the user's disposal. At the same time, the applet which shows the 2D map (Figure 3C) is constantly listening to the User Server channel to update all users' positions in its map, differentiating its own user from the rest. The main applet, in turn, is also listening to the User Server for other users' position/orientation updates in the environment. When such updates occur, it compares the new information against its current information to render the corresponding avatars updates, generating corresponding animations (walking, turning).

Communication between the main applet and the User Server is carried out every time the user moves through the virtual environment. Similarly, when the user accesses an area for which the main applet has no graphic information to render, it talks to the Communication Server requesting information for the new area to be accessed.

Furthermore, when the user accesses certain areas of the virtual environment, said areas may have associated, informative content, maybe even a URL, which the main applet distributes to the other elements of the portal in charge of publishing it. For example, when the user accesses a certain office within the virtual environment which has a web page associated to it, the main applet establishes a communication with the HTML element

destined for that exact purpose via a DOM. There are many ways to communicate with a DOM in Java (JsObject, DomAPI, and applet-specific functions). In our case, we use applet-specific functionality, as shown below.

```
try {  
    applet-maestro.getAppletContext().  
    showDocument(newURL("javascript:  
        carga-dinamica(\""+url+"\")"));  
}  
catch (MalformedURLException me) { }  
}
```

Access to DOM allows the main applet to update any portal element according to user interaction within the virtual environment.

6. Conclusions

The main objective of this research was developing multiuser Web3D portals – web pages with embedded applications which render 3D virtual, secure, multi-user, interactive environments. The authors reviewed different Web3D technology alternatives and other contributions to the fields which ended up justifying the use of the Java platform as a viable, ideal technology to build Web3D portals according to the criteria presented in Section II.

As a result, we created a web portal for a University dependency which shows several elements related to such dependency such as the start page, a 2D map of its facilities, and a 3D environment (among other information). The application assigns an avatar to each connected user for his or her representation in the 3D virtual environment and identification therein. Users may navigate through the virtual environment which will automatically update all the other elements of the portal (contextual web page frame, banners, etc.) relative to user's location. Moreover, users which are connected to the portal can also communicate among one another via chat.

Even though many developers combine different technologies to take advantage of each one of them, the Java platform offers a wide variety of tools to create complex applications which can be used to build colorful, dynamic, multi-user 3D virtual environments, emphasizing speed at runtime (or application response time), interactivity, interconnectivity, and security. Furthermore, it is faster than those scripted languages. However, lack of native support for 3D rendering is still Java's only drawback.

The current tendency points towards web browsers with powerful GUIs which are able to take advantage of graphic hardware capabilities, where web contents may be interpreted, and which may deploy realistic, interactive 3D content without having to download any complementary applications of dubious provenance.

References

- [1] BOUSSINOT F., SUSINI J-F., DANG TRAN F., and HAZARD L. A reactive behavior framework for dynamic virtual worlds. Proceedings of the sixth international conference on 3D Web technology. pp 69-75. 2001.
- [2] CHEN B.Y. and NISHITA T. jGL and its Applications as a Web3D Platform. In ProceedingsWeb3D 2001. pp 85-91. 2001.
- [3] Cortona Jet, Parallel Graphics, Inc. [online] on <<http://www.parallelgraphics.com/products/jet>>
- [4] FRANCIS B., and STONE R. ebScylla: a 3D web application to visualise the colonisation of an artificial reef. Proceedings of the 14th International Conference on 3D Web Technology. pp 167-175. 2009.
- [5] Java Binding for OpenGL [online] on <<http://kenai.com/projects/jogl/pages/Home>>
- [6] Java SE Technologies at a Glance [online] on <<http://java.sun.com/javase/technologies/index.jsp>>
- [7] Java3D [online] on <<https://java3d.dev.java.net/>>
- [8] jME (jMonkeyEngine) [online] on <<http://www.jmonkeyengine.com/>>
- [9] Lightweight [online] on <<http://lwjgl.org/>>
- [10] LUENGO F., CONTRERAS M., LEAL A., and IGLESIAS A. Interactive 3D Applications Embedded in Web Pages. 4th Internacional ConferenceComputer Graphics, Imaging and Visualization. pp 14-17. 2007.
- [11] O3D API [online] on <<http://code.google.com/intl/in-IN/apis/o3d/>>
- [12] OSTROWSKI D. A Web-Based Interactive 3D Visualization Architecture. IEEE Visualization-Poster, Baltimore, MD. (USA). November 2006.
- [13] PARISI T. Ajax3D: The Open Platform for Rich 3D Web Applications. Media Machines, Inc. Whitepaper. 2006.
- [14] Project Darkstar [online] on <<http://projectdarkstar.com/>>
- [15] Project Wonderland [online] on <<https://lg3d-wonderland.dev.java.net/>>
- [16] SAMET H. Foundations of Multidimensional and Metric Data Structures. Morgan-Kaufmann, San Francisco. 2006
- [17] Shout3D v2.8 Shout Interactive, Inc. [online] on <<http://www.shout3d.com>>
- [18] STEFANO TORNINCASA. Web3D Technology applications for distance training and learning: the Leonardo project "WEBD". XII ADM International Conference. 2001.
- [19] WALSH A.E. and BOURGES-SEVENIER M.Core Web3D. Prentice Hall PTR. 2001.
- [20] Web3D Consortium [online] on <<http://www.web3d.org>>
- [21] WebGL – OpenGL ES 2.0 for the Web [online] on <<http://www.khronos.org/webgl/>>
- [22] Xj3D - Java based X3D Toolkit and X3D Browser [online] on <<http://www.web3d.org/x3d/xj3d/>>
- [23] YIN S. A Web-Based 3D Gaming Style Multi-user Simulation Architecture. 1st International Workshop on WEB3D GAMES 2007. pp 19-22. 2007.

Xandre Chourio works as Research Assistant assigned to University of Zulia's Scientific Modeling Center (Venezuela). He

holds a B.Sc. degree in Computer Science at the University of Zulia (2010). His fields of interest are client-server applications, distributed computing and grid computing.

Francisco Luengo is Titular Professor at the Department of Computer Science of the University of Zulia (Venezuela). He holds a B.Sc. degree in Computer Science at the University of Zulia (1992) and a M.Sc. in Computer Science at the Central University of Venezuela (1996) and a Ph.D. in Applied Mathematics and Computational Sciences at the University of Cantabria (Spain,

2005). His fields of interest are computer graphics and animation, parallel computing, artificial intelligence and numerical analysis.

Gerardo Pirela is Associate Professor at the Department of Computer Science of the University of Zulia (Venezuela). He holds a B.Sc. degree in Computer Science at the University of Zulia (1997) and a M.Sc. in Science and Computer Engineering at Oregon Graduate Institute of Science & Technology (USA, 1999). His fields of interest are Artificial Intelligence and Bioinformatic.

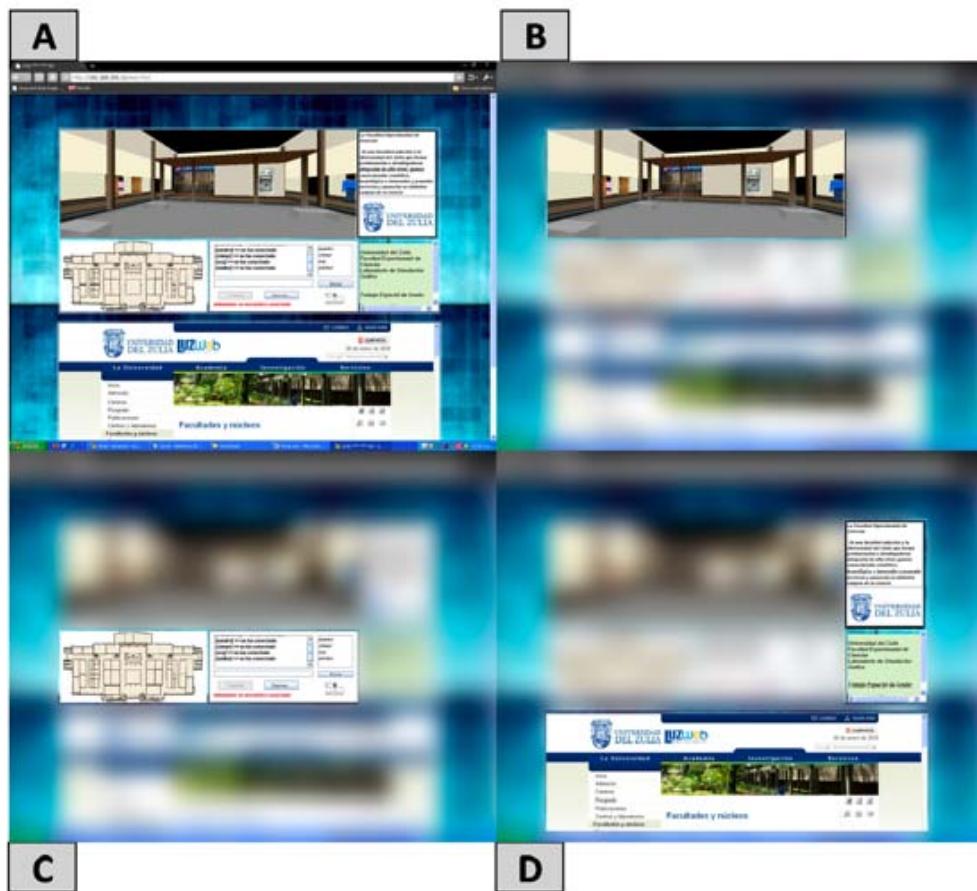


Fig. 3 Multiuser Web3D page.

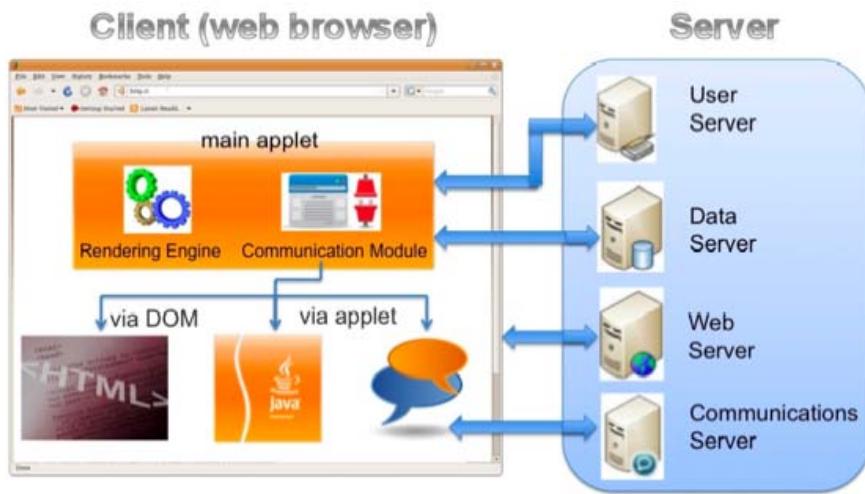


Fig. 4 Multiuser Web3D System Architecture

A user-centric PKI based-protocol to manage FC² digital identities

Samia Bouzefrane¹, Khaled Garri¹ and Pascal Thoniel²

¹ CEDRIC Laboratory, Conservatoire National des Arts et Métiers
292 rue Saint Martin, 75141, Paris Cedex 03, France

² NTX Research
111 avenue Victor Hugo, 75116 Paris - France

Abstract

The proliferation of e-services (e.g. e-commerce, e-health, e-government) within the emerging digital Identity Management Systems make Internet an undeniable convenient and powerful tool for users. However in this environment, users are required to manage several digital identities and a great number of personal data. As such, simplification of users' involvement is highly needed while increasing the users' confidence, and guaranteeing security. This paper proposes a low-cost authentication solution which leads to a reduction of users' identities, even across several circles of trust, while maintaining high-level security. This solution is suitable for FC², a platform dedicated to manage digital identities within circles of trust.

Key words: Identity management system, public key infrastructure, federated identity, circle of trust, digital identity, security.

1. Introduction

With a boom in online services generally accessed through a lot of login-password couples, Internet users are having an ever increasing number of digital identities.

Indeed, Internet was not originally designed with the digital identity idea and, some solutions have been proposed to deploy digital-identity management architectures using existing standards and protocols such as InfoCard standard that is a user-centric approach or Liberty Alliance standard that is based on the notion of identity federation.

A typical identity-management architecture requires basic components like an identity provider (IDP) that authenticates the user in a secure manner allowing him to access to a service provider (SP) and an attribute provider

(AP) to supply the user attributes to any authorized agent while not compromising privacy.

FC² (Federation of Circles of Trust) [1] is a French project initiated by several companies jointly with government and academic actors. It takes into account the following points: the user must have control on his personal data, a great number of certificates must be provided at low cost, multiple services may belong to distinct groups and accessible via various material supports like usb key, smart card or a mobile equipment. FC² tries to bring a solution to these requirements by implementing a comprehensive platform that allows new secure electronic services based on a transparent and interoperable federated identity management.

In this context, a new PKI¹-based protocol, called "2.0", has been proposed to guarantee secure access to electronic services at low cost.

Based on three levels, FC² project integrates:

- An international PKI that delivers and manages server certificates for identity providers, service providers and attribute providers.
- An internal PKI deployed by each registration authority (associated to each circle of trust) for all its agencies.
- A "user" PKI that addresses final users.

Our contribution, in this paper, concerns the "user" PKI that integrates an entity called "electronic notary" used instead of a certification authority, allowing the registration of new users (citizen/consumer/professional) within a registration authority that may be a proximity agency (telecom agency, banking agency) viewed as a trust third party. The proposed crypto-system is based on the same principle whatever asymmetric algorithm is used. The local registration authority delivers a "public key certificate" to the user along with a private key using his usb key, his smart card or his cell phone. The local

¹ Public Key Infrastructure

registration authority uploads user's "public key ownership certificate" to its central electronic notary server through a secure channel. Thus, anyone, any IDP, any application and any process can request this electronic notary server to authenticate the digital identity of the user. This trusted user is now able to access, at any time services belonging to distinct circles of trust, federated by FC² system in a transparent manner.

The topic of trust and PKI management has been addressed during the last few years like in [7], [8], [9], [10]. For example, John Linn in [7] presents and compares several trust models and applied for use with public-key certificate infrastructures based on the X.509 specification, including subordinated hierarchies, cross-certified CA, hybrid CA, bridge CAs, and trust lists. More recent works deal with identity management systems and focus on the use of PKI within a federated architecture like in Liberty Alliance [11]. Another work on Liberty Alliance targeted a pan-European multi Circle of Trust environment [12].

On the other hand, Windows CardSpace delivered with recent versions of .NET Framework manages identities according to a user-centric approach [13]. A more sophisticated work introduces a formal semantics based calculus of trust that explicitly represents trust and quantifies the risk associated with trust in PKI and identity management [14]. However, all these research works targeted a particular identity management system. Since CardSpace and Liberty Alliance are not interoperable, Jorstad et al. in [15] tried to integrate the current SIM authentication used in GSM with both Liberty Alliance and CardSpace such that it can be used for Internet services. Unfortunately this work is limited to mobile equipments and is not used with other physical supports.

The PKI-based approach proposed in this paper allows managing different circles of trust defined in FC² project while each circle may adopt any identity management system independently from the system used by other circles of trust. Moreover, with our solution the user may access to FC² services after registration and authentication phase thanks to any physical support (smart card, USB key, cell phone, etc.) he has.

In the rest of this paper, Section 2 recalls the objective of FC² platform. Section 3 describes the user-centric PKI-based approach that is proposed to access electronic services of different circles of trust while guaranteeing security at low cost. Section 4 details the way keys and certificates are generated and used. Section 5 explains the role of each involved actor through distinct use cases. Section 6 outlines the implemented software architecture, before concluding in Section 7.

2. The FC² project

The FC² project (see Figure 1) is a French R&D cross-sector initiative including companies, government and academic actors.

The project started on July 2007 and ended on June 2010. It aimed to:

- Define and implement interoperable identity federation architecture schemes, fully agnostic versus underlying technologies (Liberty Alliance [2], Microsoft/Cardspace [3], Higgins [4], Open-ID [5], etc.),
- Implement a dedicated infrastructure for service providers enrolment and support,
- Provide strong authentication and privacy management services,
- Provide a high level of protection against digital identity attacks,
- Provide a simple and convenient user experience, targeting user empowerment and trust as well as universality of use across a large variety of end-user devices,
- Create innovative business models, acceptable and/or adoptable by all players in the value chain,
- Provide the needed technologies and processes to master identity management technologies on a large scale basis, from national/government level to corporate/enterprise level.

The R&D developments within the project targeted especially the problems of Federated Identities in a cross-sector environment, encompassing the e-government, local administrations, telecommunications and financial domains.

Regarding the FC² issues, the interoperability of various identity-management technologies and the user privacy have been addressed through research works undertaken by FC² partners, like in [5] and [6].

In this paper, we focus on the access to on-line services within a federated identity platform while allowing secure, low-cost, and user-centric properties.

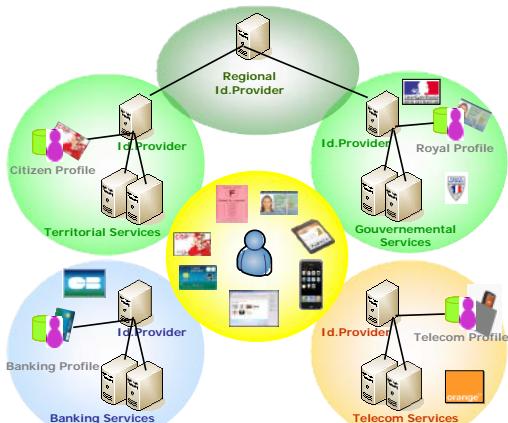


Fig.1. FC² architecture

3. PKI 2.0

3.1 The FC² context

As for any PKI, the main objective is naturally to ensure data exchange confidentiality, integrity and authenticity along with strong authentication of the actors. The non-functional objectives are the fact of allowing the materialization of trust within FC² framework, between distinct circles of trust and the user. It was decided in FC² project, not to impose higher certification authority nor hierarchy between the circles of trust, but to provide a security solution that may be collaborative between the actors, and that may be evolutionary, flexible for the integration of new partners (circles). The PKI-based solution that is deployed uses three levels:

- *A bridge PKI*: to which is assigned a certificate that is auto-signed or delivered by a known CA which signs for a given circle of trust the certificate of the internal CA.
- *An international PKI*: to deliver the public certificates of identity providers and attribute providers when communicating with users or electronic notaries.
- *An internal PKI*: with a root certificate auto-signed and a certificate signed by the bridge CA certificate; this PKI signs the certificates of the identity and attribute providers for their inter-circle communication, and those of each registration authority and each electronic notary when communicating each other.

3.2 PKI 2.0 principle

Our contribution concerns a protocol called PKI 2.0 to register new users within the FC² platform. It consists of two parts. The enrolment phase involves the registration of new users and the generation of keys and certificates. The verification phase that assumes the publication of new public-key ownership certificates on an electronic server

acting as a notary that checks the validity of the user certificates.

The first step concerns the Registration Authority (*RA*) that allows registering new users. The *RA* must have multiple proximity agencies distributed over the territory. In the real world, the *RA* and the proximity agencies are the following according to the circle of trust considered: In the governmental circle, the *RA* corresponds to "les Notaires de France" and the proximity agencies are notaries. In the banking circle, the *RA* is the bank and the proximity agencies are the banking agencies. In the telecom circle, the *RA* is the telecom operator and the proximity agencies are the local telecom agencies.

RA has the role of checking the identity of the user, and of executing a dedicated procedure to deliver finally an auto-signed public-key certificate and a "public-key ownership certificate". In such model, the cost is low since that certification authorities are not necessary.

In our solution, the user holds his key pair and his certificates (certificate of public-key and certificate of public-key ownership) generated by the proximity agency called the Local Registration Agency (*LRA*). The certificate of public-key ownership published by a *LRA* on the Electronic Notary (*EN*) server is used to check the validity of the auto-signed certificate of public key. To distinguish them from the certificates generated by a server, they are called "customer" certificates. The user can use his certificates for encryption, authentication or signature.

This protocol avoids the use of a certification authority to the benefit of the *EN* server. No certificate from registration authority is necessary since the user certificate is auto-signed. However, the *LRA* checks the certificate generation and insures the secured publication of the corresponding ownership certificate on the *EN* server. The auto-signature of a certificate does not bring any guarantee on its validity, it only insures to be in compliance with the standard X509v3 so as to be used by existing applications. In fact, the certificate validity is obtained on-line by requesting the *EN* server.

4. PKI 2.0 keys and certificates

In this section, we explain how keys and certificates are generated, and how the enrolment is performed.

4.1 Key-pair generation

To have a PKI 2.0 certificate, each user must have a pair of keys. The key generation can be done according to different ways:

- personal way: the user runs locally a software tool provided by FC². This software corresponds to an identity selector installed on his machine or his cell phone.
 - decentralized way: the user goes physically to a Local Registration Agency that generates keys for him.
 - centralized way (not recommended): a RA - one by circle of trust - generates the keys for each person.
- At this stage, the user has two keys: a public key K_{pub} and a private key K_{priv}.

4.2 Certificates Generation

The PKI 2.0 recommends for each user two pairs of keys, one pair is dedicated for self authentication and electronic signature and the other pair for encryption. These pairs of keys have to be stored in secure way, especially for the private keys.

The public-key certificates X509v3 are auto-signed and stored in plain text. The first one is an authentication/signature certificate whose legal value rises from European directive 1999/93. The second one is dedicated to encryption.

The generation of certificates can be also carried out according to a personal, centralized or decentralized way. However, the decentralized procedure is the optimal way since it allows the generation of a public-key certificate and a public-key ownership certificate that guarantees the authenticity of the latter. These certificates are added to the information system for the first time during the enrolment phase.

4.3 Enrolment operation

The user enrolment process is done directly within a Local Registration Agency (*LRA*) according to three steps:

- First step: checking user identity

The user presents one or more identity documents to the registration agent that physically authenticates user. This face-to-face step is easy to realize within FC² project thanks to the LRA that are distributed over the territory.

- Second step: Generating public-key certificate

Once the user identity is checked, the registration agent launches the public-key certificate generation after having generated the corresponding key pair. This certificate contains: third party nationality (FR), third party type (registration agency), third party circle of trust, timestamping (validity period of the certificate), user identity, public key, auto-signature (with the user private key). Once the certificate is generated, the registration agent must register it on the user physical device (USB key, smart card or cell phone).

- Third step: Generating certificate of public-key ownership

PKI 2.0 principle is that the consumer/citizen is responsible for certifying the ownership of his public key without involving any certification authority. For this purpose, PKI 2.0 adds a new certificate, called "public-key ownership certificate". This certificate does not contain the user public key. Instead, it contains the hash value of the public key (by using a hash function like MD5, SHA-1 or RIPE-MD). Hence, a certificate of public-key ownership contains: nationality of the third party (FR), type of the third party (registration agent), third party circle of trust, time-stamping, user identity, public key hash value. Once generated, this certificate is encrypted with the user private key.

The following section describes the different entities involved in the infrastructure.

5. Principal actors and use cases

PKI 2.0 gathers the following actors: *LRA*, *EN* servers, service providers, and finally users. The local agencies and servers of each circle of trust have an internal PKI which delivers the necessary certificates to establish a SSL communication. The electronic-notary servers have also certificates issued by a known certification authority for communication between them and users or service providers. Figure 2 illustrates the general architecture of the PKI 2.0 including the main actors and the interactions between them.

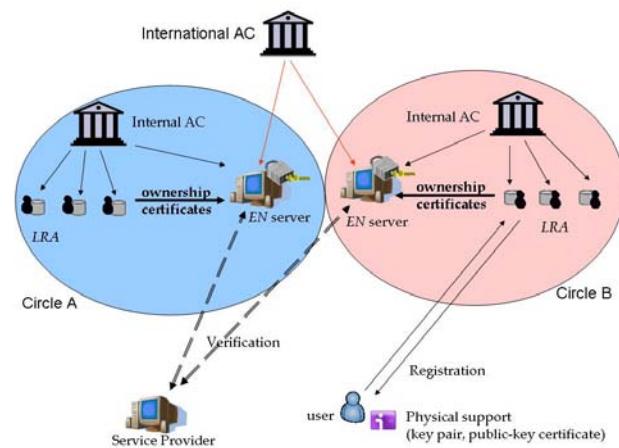


Fig.2. PKI 2.0 Architecture overview

4.4 Main actors

User entity

The user, as consumer or citizen, is linked to a circle of trust. First of all, he has to register himself within a *LRA* to access the system, it's the enrolment phase. Afterwards, he will be able to communicate with the system to manage his

own account (update, revoke or renew). Of course, he is also able to check any public-key certificates on any EN servers delivered by any RA of any circles of trust.

Local Registration Agency

This entity manages clients' enrolment by generating certificates and publishing and deleting of public-key ownership certificates onto the *EN* server.

Electronic Notary server

The *EN* server contains a register of ownership certificates. It is requested by other actors to authenticate the public-key certificates. Indeed, it stores the public-key ownership certificates published by *LRA*. The communication by a *EN* and a *LRA* is secured thanks to SSL certificates delivered by an internal PKI. *EN* server is requested to verify the users public-key certificates delivered by its own *LRAs*.

Service Providers

The service providers correspond to the web sites where users are identified using their public-key certificate. These providers ask *EN* to check the public-key certificates used by users to access the provided services.

Identity providers

The identity provider, one per circle of trust, stores public-key certificates of PKI 2.0 generated by *LRA*. The publication of these certificates is done automatically thanks to a software module integrated within the user identity selector.

5.1 Use Cases

New users enrolment

As stated before, this step is the first one allowing the user to access to FC² platform. It is performed by the user within a *LRA*. First of all, the user is authenticated by presenting an identity document. Once authenticated, several steps may be carried out in order to issue the public-key certificate, that is, to:

- register the user in a database which contains the list of users. The information stored in the data base are: Common Name, gender, date of birth, postal address and e-mail. These data items are inserted automatically into the certificates generated as in the following steps.
- generate the public-key certificate, that is auto-signed with the private key of the user.
- register the public-key certificate (in .P12 and .cer formats) on the physical device of the user.
- generate the public-key ownership certificate under .P12 format, and finally encrypt this certificate using the private key of the user.

Publication of public-key ownership certificate

The public-key ownership certificate allows checking the authenticity of the public key through the hash value of this latter that is inserted into the certificate. This certificate must be published onto the *EN* server. This is done thanks to SSL communication with the *NE* server. The SSL authentication is done mutually by using SSL certificates generated by the internal PKI. Once the mutual authentication is achieved successfully, a message is sent to *EN* with the following information:

$$\{Id_{LRA_i}, M, \{H(M)\}_{K_{priv\ LRA_i}}\}_{K_{pub\ EN}}$$

Where Id_{LRAi} : is the identity of the *LRA i*,

M : the message contains the ownership certificate encrypted with the private key of the user, and other information like: serial number, version, signature algorithm.

$H(M)$: the hash value of M . All these information are encrypted with the public key of *EN*.

The use of the signature enables us to guarantee the message integrity. The encryption with the public key of *EN* guarantees confidentiality since that only the appropriate *EN* will decrypt the message. Upon receiving the message, the *EN* begins decrypting the message using its private key, and then checks the signature. For this purpose, the *EN* reads the identity of the sending *LRA* because it has a register of all the public-key certificates of his local agencies, indexed with their serial number. Then, *EN* begins extracting the certificate that contains the public key of the concerned *LRA*, in order to verify the signature. *EN* computes the hash value of M and compares it to the one received. If they are equal, *EN* stores the certificate of ownership in its database, otherwise an error message is notified to the *LRA*.

User account management

This task concerns the update, delete and read operations carried out on the user account. The update may concern the modification of some personal information or even the generation of a public-key certificate with the new information. The account deletion implies the deletion of the public-key ownership certificate within *EN*. In this case, the *LRA* sends the following message to inform *EN* about the revocation:

$$\{Id_{LRA_i}, M, \{H(M)\}_{K_{priv\ LRA_i}}\}_{K_{pub\ EN}}$$

Where $M=\{\text{serial number of public-key certificate, removal}\}$

The renewal of the public-key certificate is necessary after removal.

Public-key certificate verification

This operation allows user, identity provider, or service provider to check a public-key certificate in real time, in order to access an Internet service or to exchange data with another actor. The verification process is done as in the following:

- request a *EN* server whose the address is in the certificate using a SSL communication and authentication with the concerned *EN*
- the public-key certificate is sent to the already authenticated *EN*
- then *EN* extracts the serial number that is also the serial number of the public-key ownership certificate,
- *EN* looks at the public-key ownership certificate in its data base,
- if the public-key ownership certificate is found in the data base, *EN* tries to decrypt the public-key ownership certificate with the public-key extracted from the given/received public-key certificate,
- if the public-key ownership certificate has been successfully opened, *EN* extracts the hash value from the public-key ownership certificate, and compares it with the one computed with the public key contained in the received certificate. If they are equal, the verification is successful.

As described in the following section, an implementation of PKI 2.0 has been performed.

6. Software architecture

We implemented the PKI 2.0 by developing three modules: the first module proposes a tool for the *LRA*. It allows to manage user accounts and to control the publication of ownership certificates on the *EN* server. As an example, Figure 4 illustrates the case of adding a new user into a circle of trust. The second module is devoted to the *EN*, it is composed of two programs: one to answer the publishing requests from his Local Registration Agencies, and the other to answer the check requests from any service providers and any users. The third module is devoted to external customers (users, IDP and service providers) that want to check a public-key certificate of PKI 2.0 (see Figure 3).

7. Conclusion

Our PKI-based solution seems to be an answer to the FC² needs in terms of security. This proposal is a user-centric system that fulfills the requirements of a large number of users at low costs. Moreover, the PKI 2.0 solves the problem of managing multiple digital identities by allowing the use of only one or few identities within several circles of trust, by replacing Certification

Authorities by Registration Authorities that have proximity agencies easily accessible to citizens/consumers.

As a perspective to our work, we aim to develop the module of checking as a plugin to Web browsers.

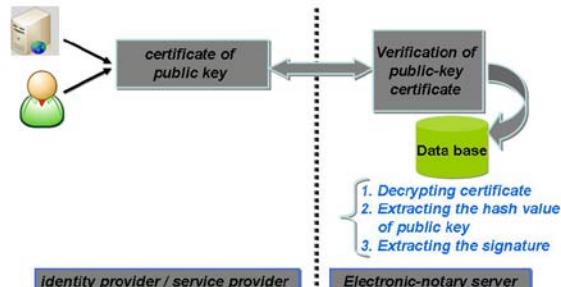


Fig.3. Checking a public-key certificate

Acknowledgments

This research is part of the project called FC2 (Federation of Circles of Trust -www.fc2-consortium.org). Khaled Garri has been partly financially supported by the DGCIS (Direction Générale de la Compétitivité de l'Industrie et des Services). We are thankful to students Nesrine Jlidi, and Mohamed Mammeri, who contribute in the implementation of the software tools.

References

- [1] Consortium, FC². Fédération de Cercles de Confiance et usages sécurisés de l'identité. <http://www.fc2-consortium.org/>.
- [2] Liberty Alliance Project, Available: <http://www.projectliberty.org/>
- [3] <http://www.eclipse.org/higgins/>
- [4] <http://openid.net/>
- [5] H.-B. Le et S. Bouzefrane "Identity management systems and interoperability in a heterogeneous environment", in **Int. Conf. on Advanced Technologies for Communications**, Hanoi, oct., pp. 243-246, IEEE, 2008.
- [6] A. Davoux, J.-C. Defline, L. Francesconi, M. Laurent-Maknavicius, K. Bekara, R. Gola, J.-B. Lezoray, V. Echebarne, "Federation of Circles of Trust and Secure Usage of Digital Identity", in **eChallenges e-2008**, Stockholm, Sweden, October 2008.
- [7] J. Linn, "Trust Models and Management in Public-Key Infrastructures," RSA Laboratories, Tech. Rep., 2000. <ftp://ftp.rsasecurity.com/pub/pdfs/PKIPaper.pdf>
- [8] M. Blaze, J. Feigenbaum, and J. Lacy, "Decentralized Trust Management," in Proceedings of the **1996 IEEE Symposium on Security and Privacy**, May 1996, pp. 164–173.
- [9] U. Maurer, "Modeling a public-key infrastructure," in Proceedings of the **4th European Symposium on Research in Computer Security** (ESORICS 96), ser. Lecture Notes in Computer Science, vol. 1146, September 1996, pp. 325–350.

- [10] Radia Perlman "An Overview of PKI Trust Models", **Journal of IEEE Networks**, Nov/Dec. 1999, pp. 38-43.
- [11] Liberty Alliance Project, "Liberty Alliance Trust Models", draft version 1.0-14, 13 April 2003.
- [12] Dao Van Tran, Pal Lokstad, Do Van Thanh, "Identity Federation in a Multi Circle-of-Trust Constellation", Report of Telektronikk 3/4.2007, pp. 103-118.
- [13] Windows CardSpace "Geneva",
<http://connect.microsoft.com/site642/content/content.aspx?ContentID=10104>
- [14] Jingwei Huang, David Nicol, "A calculus of trust and its application to PKI and identity management", Proceedings of the **8th Symposium on Identity and Trust on the Internet**, 2009, Pages: 23-37.
- [15] Ivar Jorstad, Do Van Thuan, Tore Jonvik & Do Van Thanh, "Bridging CardSpace and Liberty Alliance with SIM authentication", Pages: 12-25 , Proceedings of the **9th Symposium on Identity and Trust on the Internet**, 2010.

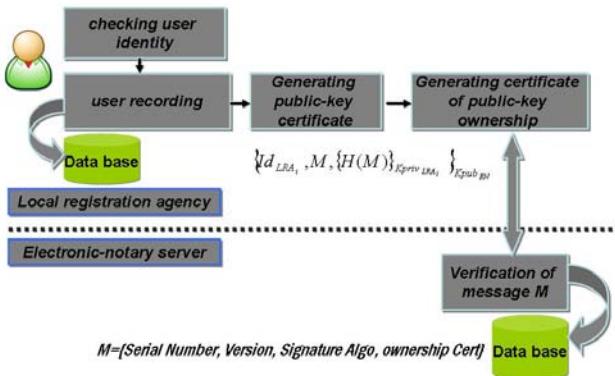


Fig.4. Adding a user in a circle of trust

Samia Bouzefrane is an associate professor at the CNAM (Conservatoire National des Arts et Métiers) in Paris. She received her Ph. D. in Computer Science in 1998 at the University of Poitiers (France). She joined the CEDRIC Laboratory of CNAM on September 2002 after 4 years at the University of Le Havre. After many research works on real-time systems, she is interested in smart objects. She took part in the MESURE project to evaluate the performance of Java Card platforms, which has received on September 2007 the Isabelle Attali Award from INRIA during "e-Smart" Conference. Furthermore, she is the author of two books: a French/English/Berber dictionary (1996) and a book on operating systems (2003). Currently, she is a member of the ACM-SIGOPS, France Chapter.

Pascal Thoniel holds a Master of Finance from IEP Paris (Sciences Po). After 10 years of experience in Business IT, Pascal has created NTX Research in 1997, a company specialized in Information Systems security. Pascal has 15 years of experience in IT Security : IT Security Audit and Policy designer, inventor XC Technology (strong authentication and confidentiality - patented), inventor of a new user-centric approach for PKI. NTX Research plays also an active role in the FC² project.

Khaled Garri is a PhD student from the SEMPIA team (embedded and mobile systems towards ambient intelligence)

of the CNAM in Paris. He has a Master's degree. In addition to FC² project, he is working currently on embedded systems security.

Visual Attention Shift based on Image Segmentation Using Neurodynamic System

Lijuan Duan¹, Chunpeng Wu¹, Faming Fang¹, Jun Miao², Yuanhua Qiao³ and Jian Li¹

¹ College of Computer Science and Technology, Beijing University of Technology
Beijing, 100124, China

² Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences
Beijing, 100190, China

³ College of Applied Science, Beijing University of Technology
Beijing, 100124, China

Abstract

A method of predicting visual attention shift is proposed based on image segmentation using neurodynamic system in this paper. The input image is mapped to a neural oscillator network. Each oscillator corresponding to a pixel is modeled by means of simplified Wilson-Cowan equations, and is coupled with its 8-nearest neighbors. Then the image is segmented by classifying the oscillation curves of the excitatory groups of all the oscillators. The classifier is constructed based on features of frequency, offset, phase and amplitude of the curves. The visual attention shift between the regions on the image is predicted according to the saliency strength of each region. Referring to the mechanism of winner-take-all competition, the saliency of a region is the aggregation of the dissimilarities between this region and all the other ones. Experimental results on images show the effectiveness of our method.

Keywords: Neural oscillation, Coupling method, Image segmentation, Saliency, Visual attention shift.

1. Introduction

Human vision system is able to select salient information among mass visual input to focus on. This selective attention mechanism enables us to efficiently understand the visual scenes without forming a complete, detailed representation of our surroundings [1]. When performing a visual task, according the mechanism of winner-take-all competition [2], the most salient region is attended first, and then our attention will shift to the less salient regions successively due to the adaptability of the visual system, while the attended regions may also be attended again after a few seconds. Computationally modeling such mechanism has become a popular research topic in recent years [3-5].

In this paper, a method of predicting visual attention shift is proposed based on image segmentation using neurodynamic system. In previous studies [2, 6], visual attention is often modeled to shift between pixels according the saliency strength of these pixels, i.e., these methods do not consider the semantic information in the image. We think commonly that visual attention should shift between meaningful regions on an image, and these regions should correspond to an object or at least part of an object. Therefore, in order to label the input image with meaningful regions, we apply our simplified Wilson-Cowan equations which we proposed in [7] to image segmentation. Wilson-Cowan equations [8] are based on the assumption that the features of an object are grouped based on the temporal correlation of neural activities [9]. Thus neurons that fire in synchronization would signal features of the same object, and groups desynchronized from each other represent different objects. Experimental observations [10] of the visual cortex of animals show that synchronization indeed exists in spatially remote columns and phase-locking can also occur between the striate cortex and extrastriate cortex, between the two striate cortices of the two brain hemisphere, and across the sensorimotor cortex. These findings have concentrated the attention of many researchers on the use of neural oscillators such as Wilson-Cowan oscillators [11].

The remainder of the paper is organized as follows: In Section 2, we firstly stated the framework of our visual attention shift method in details. Then we demonstrate our experimental results in Section 3. The summary is given in Section 4.

2. Proposed Method

2.1 Neural Oscillation and Synchronization

To segment a gray image by using neurodynamic system, we let the input image correspond to a neural oscillator network in our method. Therefore, each pixel in the image is mapped to a neural oscillator in the network, and the intensity of each pixel is considered to be the external input to the corresponding oscillator.

We describe an oscillator by means of simplified Wilson-Cowan equations. Such a model consists of two non-linear ordinary differential equations representing the interactions between two populations of neurons that are distinguished by the fact that their synapses are either excitatory or inhibitory. Thus, each oscillator consists of a feedback loop between an excitatory groups x_i and inhibitory groups y_i that obey the Equation (1):

$$\begin{aligned} \frac{dx_i}{dt} &= -r_1 x_i + r_1 H(ax_i - cy_i + I_i - \phi_x) \\ \frac{dy_i}{dt} &= -r_2 y_i + r_2 H(bx_i - dy_i - \phi_y) \end{aligned} . \quad (1)$$

Both x_i and y_i are interpreted as the proportion of active excitatory and inhibitory neurons respectively, which are supposed to be continuous variables and their values may code the information processed by these populations. Especially, the state $x_i = 0$ and $y_i = 0$ represents a background activity which correspond to the background in an image. The parameters in Equation (1) are as follows: a is the strength of the self-excitatory connection, d is the strength of the self-inhibitory connection, b is the strength of the coupling from x to y , and c is the strength of the coupling from y to x . Both ϕ_x and ϕ_y are thresholds, r_1 and r_2 modify the rate of change of the x and y group respectively. Figure 1(a) shows the model of an oscillator. All the above parameters are non-negative, and I_i is external input to the oscillator in position i which corresponds to a pixel in the image. $H()$ is a sigmoid activation function defined as in Equation (2):

$$H(z) = \frac{1}{1 + e^{-\frac{z}{T}}} . \quad (2)$$

T is a parameter that sets the central slope of the sigmoid relationship.

We locally couple the above simplified Wilson-Cowan oscillators as in Equation (3):

$$\begin{aligned} \frac{dx_i}{dt} &= -r_1 x_i + r_1 H(ax_i - cy_i + I_i - \phi_x) + \alpha \Delta x_i \\ \frac{dy_i}{dt} &= -r_2 y_i + r_2 H(bx_i - dy_i - \phi_y) + \beta \Delta y_i \end{aligned} . \quad (3)$$

α and β represent the strength of the connection between neurons. Δx_i and Δy_i represent coupling terms from all other adjacent oscillators in the neural network. An open chain of coupled oscillators is shown in Figure 1(b) which is preferable for 1-D neural network. Since an image corresponds to a 2-D network in our method, we couple each oscillator with its 8-nearest neighbors. The coupling method in our model is illustrated in Figure 1(c). In Figure 1(c), each ellipse represents an oscillator and all the oscillators enclosed by a rectangle represent the neural oscillator network corresponding to the input image. In Figure 1(c), for the red oscillator at t_2 , its coupling strength is related to five purple oscillators (including itself) at last moment t_1 . And the coupling strength of these five oscillators at t_3 are related to the red oscillator at t_2 . Note that for better illustration, each oscillator only connects with 4 other ones in Figure 1(c). Thus Δx_i and

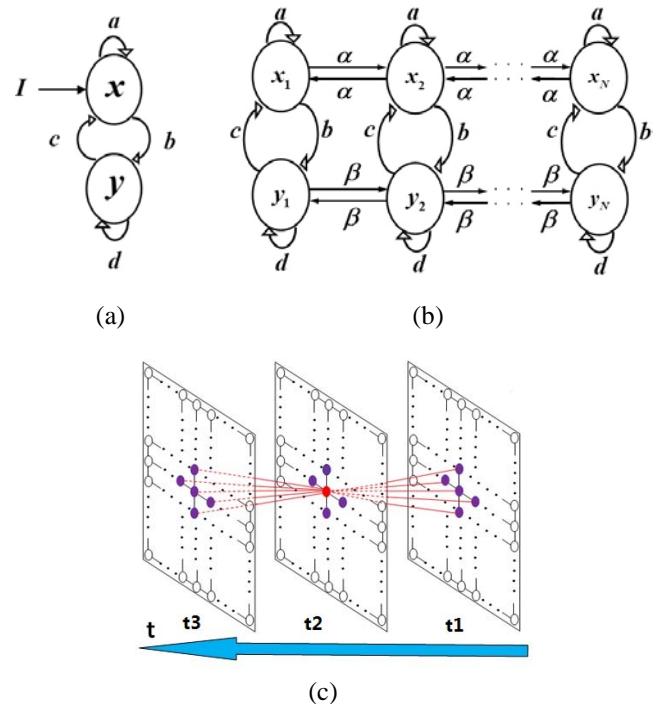


Figure 1. (a) A single oscillator. (b) An open chain of coupled oscillators. (c) The coupling method in our model. For the red oscillator at t_2 , its coupling strength is related to five purple oscillators (including itself) at last moment t_1 . And the coupling strength of these five purple oscillators at t_3 are related to the red oscillator at t_2 .

Δy_i in Equation (3) are computed as in Equation (4):

$$\begin{aligned}\Delta x_i &= \sum_{j \in N(x_i)} r_j (x_j - x_i) \\ \Delta y_i &= \sum_{k \in N(y_i)} r_k (y_k - y_i)\end{aligned}\quad . \quad (4)$$

$N(x_i)$ and $N(y_i)$ are the 8-nearest neighbors of x_i and y_i respectively. The weight r_j and r_k is determined as in Equation (5):

$$r_{ij} = \begin{cases} 1 & |I_i - I_j| < \phi \\ 0 & \text{otherwise} \end{cases} \quad . \quad (5)$$

ϕ is a threshold to decide whether two adjacent oscillators couple with each other.

All the parameters above are determined to make the differential equation in Equation (3) reach synchronization asymptotically, and we use 4th order Runge-Kutta method to find the iterative solution of this differential equation.

Figure 2 shows an input image and the corresponding oscillation curves of the excitatory groups of all the oscillators in the network. In Figure 2(b), except that the oscillators corresponding to background pixels in the input image become silent over time, the oscillation curves of all the other oscillators can be obviously clustered into 4 classes: the red arrow points to one class, the green arrow points to the other three classes. Consequently, the actually 4 objects in the input image can be reasonably segmented by classifying the neural oscillation curves. We will explain how to classify the oscillation curves automatically using the features extracted from the curves in the next subsection.

2.2 Image Segmentation by Classifying the Oscillation Curves

After observing and analyzing many oscillation curves, we assume that each oscillation curve generated by Equation (3) is the superposition of sine and cosine waves. Therefore, we fit Fourier curves to the data of oscillation curves as in Equation (6):

$$\hat{x} = m_0 + m_1 \cos(\omega \cdot t) + n_1 \sin(\omega \cdot t) \quad . \quad (6)$$

\hat{x} is an approximation of the x in Equation (3), i.e., the strength of x corresponding to each moment t in Figure 2(b). And m_1 , n_1 , ω are parameters. Equation (6) can also be written as Equation (7):

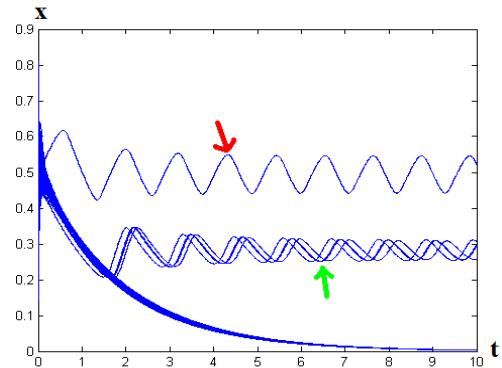
$$\hat{x} = m_0 + \sqrt{m_1^2 + n_1^2} \sin(\omega t + \theta) \quad . \quad (7)$$

where $\theta = \arctg(m_1 / n_1)$. In Equation (7), ω represents frequency, θ represents phase, $\sqrt{m_1^2 + n_1^2}$ represents

amplitude, and m_0 represents offset from t -axis as shown in Figure 2.



(a)



(b)

Figure 2. Input image and neural oscillation. (a) An input image. (b) Oscillation curves of the excitatory groups of all the oscillators.

As stated in last subsection, in order to segment the input image, we classify all the corresponding oscillation curves by combining the above four features (frequency, phase, amplitude and offset) extracted from each curve. A four-layer classifier is constructed as shown in Figure 3, and each layer classifies the oscillators using one feature respectively. We define the concept of an *oscillator slice* as a group of oscillators with same class label who connect with each other on the neural network, so each oscillator slice can also be labeled a class which is the same as any oscillator of this slice. By setting a threshold according to the feature, one oscillator slice from last layer can be classified into two or more slices on current layer. Moreover, at each layer, maybe there are several oscillator slices with same class label. We argue that if these slices with same class label connect with each other, the combination of these slices corresponds to a meaningful region in the input image, therefore these slices do not need to be further classified. Figure 4 shows results of image segmentation by classifying the oscillation curves.

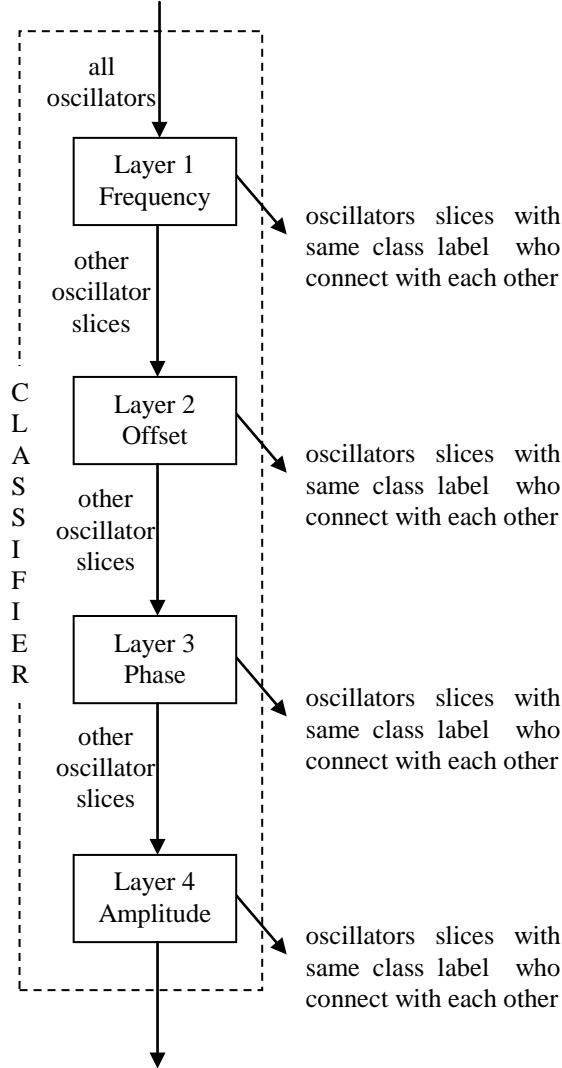


Figure 3. A four-layer classifier used for analyzing the neural oscillation curves.

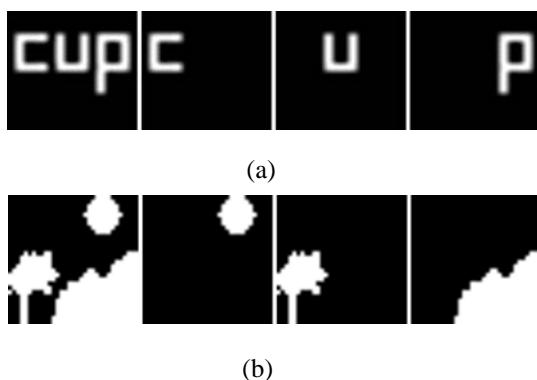


Figure 4. Results of image segmentation by classifying neural oscillation curves. (a) An input image and the image segmentation results. (b): Another Input Image and the corresponding segmentation results.

2.3 Visual Attention Shift between Regions

According to the mechanism of winner-take-all (WTA) competition, visual attention shifts from the most salient region to the least one. In our method, given an image, all the regions are labeled by classifying the neural oscillation curves based on simplified Wilson-Cowan equations. The saliency of each region is calculated in a local-global manner as follows: the dissimilarities between the current region and all the other regions of the image are calculated, and the aggregation of these dissimilarities is the saliency strength of the current region. If one region is more “irregular” measured by the above mentioned local-global method than other regions, this region is more salient. Given a gray image I with P regions labeled by the image segmentation method as stated above, we compute the average of intensity of each region as in Equation (8):

$$\text{AveIntensity}_i = \frac{1}{\text{Num}(i)} \sum_{(u,v) \in R(i)} I_{u,v} \quad (i = 1, 2, \dots, P). \quad (8)$$

$\text{Num}(i)$ is the total number of pixels in region i , $R(i)$ represents all the coordinates of pixels in region i , and $I_{u,v}$ is the intensity of pixel (u,v) on image I . Then the saliency strength of each region is calculated as in Equation (9):

$$\text{Saliency}_i = \sum_{j=1, j \neq i}^P |\text{AveIntensity}_i - \text{AveIntensity}_j|. \quad (9)$$

Then $\text{Saliency}_i (i = 1, 2, \dots, P)$ are sorted by descending order, and visual attention are supposed to shift from the most salient region to the least one. Figure 5(a) shows an input image, and Figure 5(b) – Figure 5(e) shows the whole process of visual attention shift predicted by our method. Note that the white region is currently attended in Figure 5(b) – Figure 5(e). The tree is attended firstly, and then the sun, the background, the hill.

3. Experimental Validation

To guarantee that the neurodynamic system described in Equation (3) reach synchronization asymptotically, the parameters in Equation (1) – Equation (5) are set as follows: $\alpha = 20$, $\beta = 14$, $r_1 = 1$, $r_2 = 1$, $\phi = 0.1$, $a = 1$, $b = 1$, $c = 2$, $d = 0.5$, $\phi_x = 0.2$, $\phi_y = 0.15$, $T = 0.025$.

Note that all pixel values of the input image are normalized to $[0, 1]$ before neural oscillation. We use 4th order Runge-Kutta method to find the iterative solution of the differential equations in Equation (3).

Figure 6 demonstrates the whole process of our method. Given a gray image as shown in Figure 6(a), we map this image to a neural oscillator network. Then Figure 6(c) illustrates the oscillation curves of the excitatory of all the

oscillators. The six color arrows point to six clusters of curves corresponding to six regions in the input image. Note that the blue and the green arrows actually point to two different clusters of curves although these two clusters are similar to each other. Figure 6(b) shows the course of visual attention shift predicted by our method. And the order is from upper left to lower right. The background is predicted to be attended firstly, and the jeep last.

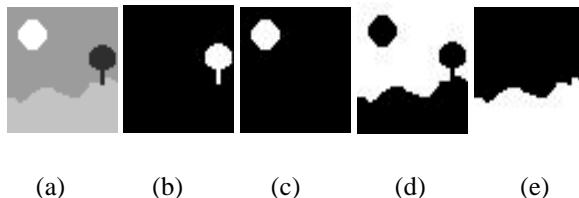


Figure 5. Prediction of visual attention shift. (a): An input image. (b) – (e): the whole process of visual attention shift predicted by our method. Note that the white region is currently attended. The tree is attended firstly, and then the sun, the background, and the hill.

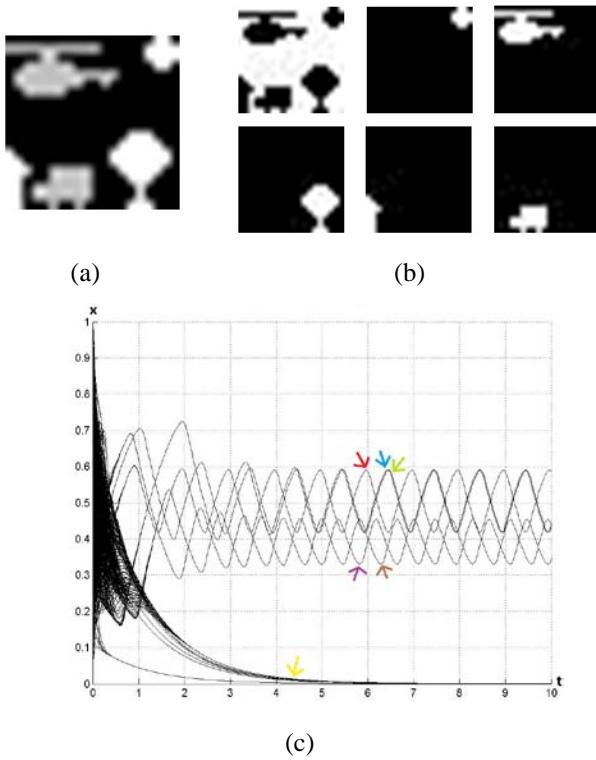


Figure 6. (a) Input image. (b) The whole course of visual attention shift predicted by our method, and the order is from upper left to lower right. Note that the white region in each image is currently attended. (c) Neural oscillation curves of the excitatory groups of all the oscillators.

4. Conclusions

In this paper, a method of predicting visual attention shift has been proposed based on image segmentation using neurodynamic system. We think commonly that visual attention should shift between meaningful regions on an image, and these regions should correspond to an object or at least part of an object. Therefore, we apply our simplified Wilson-Cowan equations to image segmentation. The input image is mapped to a neural oscillator network. Then by analyzing the features of frequency, offset, phase and amplitude of the oscillation curves, the image is labeled with different regions. To determine the order of visual attention shift, we define the saliency strength of each region as the aggregation of dissimilarities between this region and all the other ones, and more salient region is predicted to be attended earlier.

Acknowledgments

This research is partially sponsored by Natural Science Foundation of China (Nos.60702031, 60970087, 61070116 and 61070149), Beijing Natural Science Foundation(Nos.4072023 and 4102013), Beijing Municipal Education Committee (No.KM200610005012) and Beijing Municipal Foundation for Excellent Talents(No.20061D0501500211) and National Basic Research Program of China (Nos. 2007CB311100 and 2009CB320902).

References

- [1] R. Rensink, K. O'Regan, and J. Clark. To see or not to see: the need for attention to perceive changes in scenes. *Psychological Sciences*, 1997.
- [2] L. Itti, C. Koch and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1998.
- [3] T. Judd, K. Ehinger, F. Durand and A. Torralba. Learning to predict where humans look. *IEEE International Conference on Computer Vision (ICCV)*, 2009.
- [4] W. Wang, Y. Wang, Q. Huang, and W. Gao. Measuring visual saliency by site entropy rate. *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.
- [5] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-Aware saliency detection. *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.
- [6] Raghu Raj, W. S. Geisler, Robert A. Frazor, and A. C. Bovik. Contrast statistics for foveated visual systems: fixation selection by minimizing contrast entropy. *J. Opt. Soc. Am. A*, 2005.
- [7] Y. Meng, Y. Qiao, J. Miao, L. Duan, and F. Fang. Qualitative analysis in locally coupled neural oscillator network. *International Conference on Neural Networks (ICNN)*, 2009.

- [8] H. R. Wilson and J.D. Cowan. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical Journal*, 1972.
- [9] D. Wang. The time dimension for scene analysis. *IEEE Transactions on Neural Networks (TNN)*, 2005.
- [10] A. K. Engel, A. K. Kreiter, P. König, and W. Singer. Synchronization of oscillatory neuronal responses between striate and extrastriate visual cortical areas of the cat. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 1991.
- [11] S. Campbell and D. Wang. Synchronization and desynchronization in a network of locally coupled Wilson-Cowan oscillators. *IEEE Transactions on Neural Networks (TNN)*, 1996.

Lijuan Duan received her Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences in 2003. She is currently an Associated Professor at the College of Computer Science and Technology, Beijing University of Technology, China. Her research interests include artificial intelligence, neural networks, neural information processing, image understanding and biological vision. She has published more than 30 research articles in refereed journals and proceedings on image retrieval, visual neural information coding, image segmentation, visual perception and cognition.

Chunpeng Wu received the B.S. degree in computer science from the College of Computer Science and Technology, Beijing University of Technology, Beijing, in 2008. He is currently a candidate for Master of Computer Science and Technology, Beijing University of Technology, Beijing. His research interest is visual saliency detection.

Faming Fang received the B.S. degree in computer science from the College of Computer Science and Technology, YanTai University, Shandong, in 2005. He is currently a candidate for Master of Computer Science and Technology, Beijing University of Technology, Beijing. His research interest is image segmentation based on neurodynamics.

Jun Miao received the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in 2005. He is currently an Associated Professor at the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. His research interests include artificial intelligence, neural networks, neural information processing, image understanding, and biological vision. He has published more than 30 research articles in refereed journals and proceedings on face detection, visual neural networks, visual neural information coding, neural oscillation, image segmentation, visual perception and cognition.

Yuanhua Qiao received the Ph.D. degree of hydromechanics from the College of Life Science and Bioengineering in Beijing University of Technology, in 2005. She is currently an Associated Professor at The College of Applied Sciences, Beijing University of Technology, China. Her research interests include differential dynamics, neuron dynamics, pulse differential equations, mathematical model construction, biomathematics, neural networks and image understanding. She has published more than 15 research articles in refereed journals and proceedings in the field of biomathematics, Bioengineering, Image segmentation and visual perception.

Jian Li received the B.S. degree in 1984 from Beijing University of Technology. He is currently a Professor at the College of Computer Science and Technology, Beijing University of

Technology, China. His research interest is signal processing and information security.

Technology Marketing using PCA, SOM, and STP Strategy Modeling

Sunghae Jun

Department of Bioinformatics and Statistics, Cheongju University
Cheongju, Chungbuk 360-764, Korea

Abstract

Technology marketing is a total processing about identifying and meeting the technological needs of human society. Most technology results exist in intellectual properties like patents. In our research, we consider patent document as a technology. So patent data are analyzed by Principal Component Analysis (PCA) and Self Organizing Map (SOM) for STP(Segmentation, Targeting, and Positioning) strategy modeling. STP is a popular approach for developing marketing strategies. We use STP strategy modeling for technology marketing. Also PCA and SOM are used to analyze patent data in STP modeling. To verify improved performance of our study, we make experiments using patent data from USPTO.

Keywords: Technology Marketing, Principal Component Analysis, Self Organizing Map, Segmentation, Targeting, Positioning.

1. Introduction

This paper proposes a technology marketing method using machine learning algorithms and marketing engineering tools. We use principal component analysis (PCA) and self organizing map (SOM) for machine learning algorithms [6],[10]. STP (Segmentation, Targeting, and Positioning) strategy modeling is considered for marketing engineering tool in our research [9],[12]. Technology marketing is a total processing about identifying and meeting the technological needs of human society. Most technology results exist in intellectual properties like patents [17]. Recently, the researches for patent analysis have been published [1],[8],[11],[16],[18]. On our research, we consider patent document as a technology. So, patent documents are analyzed by text mining [2], PCA, SOM, and STP strategy modeling. Original patent documents are transformed into document-term matrix as structured data by keyword extraction of text mining [3]. And then, using PCA, the document-term matrix is reduced suitable dimension size to be analyzed [4]. STP is a popular approach for developing marketing strategies [9],[12]. We use STP strategy modeling for technology marketing. Also PCA and SOM are used to analyze patent data in STP

modeling. To verify improved performance of our study, we make experiments using patent data from United States Patent and Trademark Office (USPTO) [15].

2. Related Works

2.1 Technology Marketing

Technology marketing (TM) is at variance with general business marketing. TM is defined as total behavior of technology transfer for planning, selling, and sales promotion from technology developer to technology demander. According to development of industrial structure, the importance of technology marketing has been increased. We have to consider the following issues for technology marketing. They are organization of TM, selection of selling techniques, technology packaging, planning of TM, technology contract, post evaluation, and feedback.

2.2 SOM, PCA, and Text Mining

SOM: Self organizing map (SOM) was introduced by T. Kohonen [10]. SOM is an unsupervised learning algorithm for clustering. Also SOM is called as a neural networks model based on competitive learning. It has two layers which are input and feature layers. We can cluster all elements by feature map with two dimensions. Firstly SOM performs clustering with input vector x and weight matrix M . The data point x_i is treated one at a time. Also the closest m_j to x_i is found by Euclidean distance, and then m_j is updated as the following [7].

$$m_k = m_j + \alpha(x_i - m_j) \quad (1)$$

where m_j and m_k are current and new weights. So m_k moves to x_i . This learning is repeated until given conditions such as change rate of weights and the number of repeat.

PCA: Principal component analysis (PCA) is a dimension – reduction method [6],[7]. The aim of PCA is to transform input vector with high dimension (p) into feature vector with low dimension (k). Though the k is smaller than p, PCA has the minimum information loss.

Text mining: Text mining is a data mining technique for finding hidden and useful patterns from a large text database [3]. General text mining methodology consists of document preprocessing and indexing [2]. There are collection creation, document parsing, document segmentation, and text summarization in the document preprocessing [14]. Also keyword extraction, phase extraction, morphological analysis, stop-word filtering, term association, and term clustering are the methods for indexing. In addition, we consider topic clustering and mapping for our research. The methods of topic clustering are term selection, document categorization, and cluster title generation. Trend, query, aggregation, and zooming maps are used in topic mapping.

2.2 STP Strategy Modeling

Segmentation, targeting, and positioning (STP) is a popular strategy model in marketing [9],[13]. Segmentation clusters customers to similar groups by their wants and needs. In targeting, marketers determine one or two groups for their marketing approaches. Positioning explains the competition power of company's product to target segments.

Segmentation: Segmentation has five phases which are segmenting markets, describing market segments, evaluating segments attractiveness, selecting target segments and allocating resources to segments, and finding targeted customers [13]. Using factor analysis to reduce the data and forming segments by cluster analysis are popular methods for segmentation. Also a perceptual map is a good approach for segmentation. This map shows a visual representation of competitive alternatives and customers' preference. In this paper, we use recency – frequency (RF) chart as a perceptual map.

Targeting: Targeting is the approach how to assess the attractiveness of segments, determine one or two segments to serve, and identify target customers [13].

Positioning: Positioning is the behavior of planning company's offering and image for occupying a distinctive place in target customers [9]. The goal of targeting is to posit the product or brand in customers' mind and maximize the potential benefit of the company.

3. Technology Marketing using PCA, SOM, and STP

We propose a technology marketing approach by analysis of patent data. Patent data has complete information about technology. To discover knowledge in patent documents, we use text mining, PCA, SOM, and STP. Our proposed approach has six steps as the following.

Step1: Searching patent data

Using keywords equation in title, abstract, ...
: Retrieved patent documents

Step2: Document – term matrix

Using text mining
: Document(n) – term(k) matrix, ($n < k$)

Step3: Document – PC matrix

Using SVD – Principal Component Analysis
: Document(n) – PC(p) matrix, ($n > p$) and ($k > p$)

Step4: Segmentation

Using Self Organizing Map (SOM)
: Patents clustering

Step5: Targeting

Using top 5 keywords
: Defining one or two clusters for target market

Step6: Positioning

Using target result
: Increasing benefit and contribution of company

In the step 1, we retrieve patent documents from USPTO. A searching formula is needed for patent retrieval. Our study uses the searching formula from abstract in patent document. The number of terms (k) is very larger than the number of documents (n) in step 2. k and n are the numbers of variables and observations respectively. But in general PCA model, the number of variables has to be extremely smaller than the number of observations. So we are not able to do PCA to this document – term matrix directly. To settle this problem, we consider singular value decomposition (SVD). We can construct SVD of document – term matrix X [5],[7].

$$X = UDV^t \quad (2)$$

Many computational algorithms for this standard decomposition have existed. Where U is an orthogonal matrix and its columns u are called left singular vectors. V is also an orthogonal matrix with right singular vectors v . D is a diagonal matrix and its diagonal elements are $d_1 \geq d_2 \geq \dots \geq d_p \geq 0$ called singular values. According to SVD

– PCA in step 2, we get document – PC matrix in step 3. The number of PCs (p) is very smaller than n. Also we select the optimal number of PCs by screeplot with 10% or more variance. To get efficient segmentation, we use SOM as a clustering method. SOM has provided many results in diverse data types such as text, document [15]. After SOM clustering, each group is represented its detailed technology by extracted top five keywords. We can define the technology fields of all segments and determine target market. Finally we try positioning approach from determined target market. The following figure shows the process of our proposed method for technology marketing.

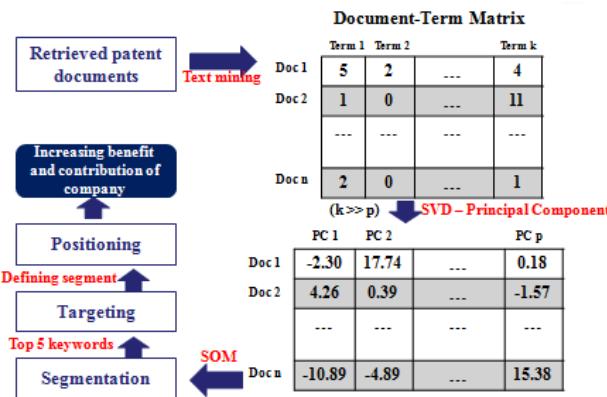


Fig. 1 Proposed technology marketing process.

Increasing company's benefit and contribution are effective results of our study. Also we can expect the diverse contributions for marketing approach.

4. Experimental Results

We apply our research to marketing engineering technologies. A company, developing business models for marketing engineering, wants to know which technologies are needed for 'marketing engineering' field. In this case, we analyze patent data of marketing engineering for solving above problem. The real data for our experiments are patent documents about marketing engineering. So, we retrieved patent data from United State Patent and Trademark Office USPTO). The keywords equation for searching patents is the following.

Abstract = Marketing * Engineering

We selected patents with marketing and engineering as keywords in their abstracts from 1948 to 2010. Also the time point of patent retrieval was November 9, 2010. Total number of selected patent documents was 80. The following table shows all searched patents.

Table 1: Total number of selected patent documents

| Years | # of Patents |
|-------------|--------------|
| 1948 – 1970 | 0 |
| 1971 – 1980 | 4 |
| 1981 – 1990 | 1 |
| 1991 – 2000 | 25 |
| 2001 – 2010 | 50 |
| Total | 80 |

Most patents of the technology of marketing engineering have been shown after 1991. So we are able to think this technology has enormous potential for research and development. Also we know the total number of registered patents through 1948 – 2010 is 80.

Using keywords extraction of text mining, we got document – term matrix. This matrix has 80 documents and 2018 terms. The number of columns (2018) is very larger than the rows (80). So we are not able to do PCA directly. To solve this problem, we use SVD. And then we have another problem of PCA. This is how many principal components should be retained for analysis. There is no clear answer but a couple of popular rules. One rule is to consider only those with variance over 10% of each principal component.

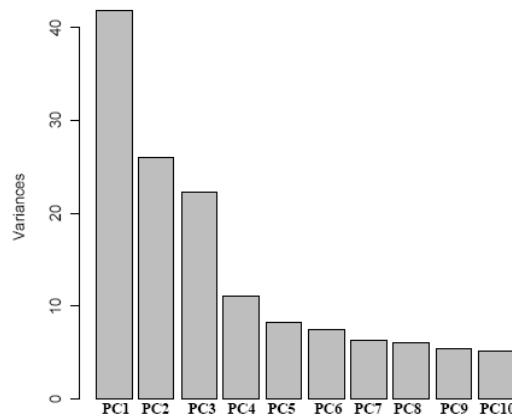


Fig. 2 Screeplot of Top 10 principal components (PCs).

From above figure, we determined 4 as the number of principal components (PCs). Then we got PCs for 80 documents using SVD-PCA. The following figure shows the result of SVD-PCA.

| Doc # | PC1 | PC2 | PC3 | PC4 |
|-------|-----------|------------|------------|-------------|
| 1 | 2.591274 | -3.0827508 | 0.4813304 | -0.7171706 |
| 2 | 2.514716 | -4.3660141 | -0.5926086 | 1.6769411 |
| 3 | 2.651611 | -6.2916608 | -2.0429489 | 2.4285974 |
| 4 | 2.651611 | -6.2916608 | -2.0429489 | 2.4285974 |
| 5 | 2.943499 | -3.6309257 | 2.4725174 | 3.3832317 |
| ... | | | | |
| 76 | 3.159518 | -8.4218044 | -4.7578218 | 1.4009504 |
| 77 | 2.481203 | -4.8516302 | -0.8211147 | -2.7172378 |
| 78 | 2.200017 | -3.8377048 | -3.6012114 | -11.7621848 |
| 79 | -8.986990 | 2.1634336 | -0.8544428 | 0.3446253 |
| 80 | 1.634265 | -4.1265055 | -1.3808656 | -2.8129552 |

Fig. 3 Result of SVD-PCA.

Doc # is the I.D. number of each document in the above figure. We are able to do the segmentation approach using above top 4 PCs. For segmentation of STP strategy modeling, we use an unsupervised neural networks model called SOM. We can look at the SOM result in the next figure.

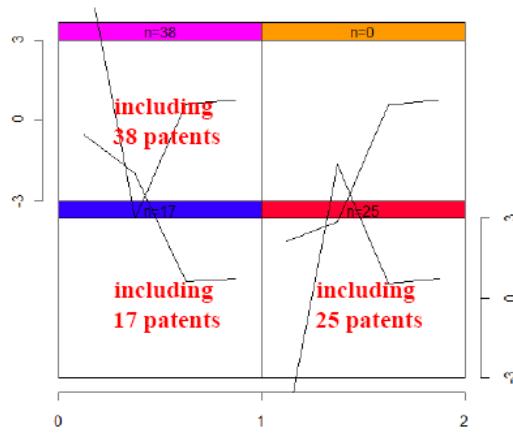


Fig. 4 SOM result.

From the SOM result, we find 3 segments for STP strategy modeling. The following table shows the segmentation result by SOM.

Table 2: Segmentation result by SOM

| Segment | Patent no. | # of patents |
|---------|---|--------------|
| 1 | 8, 18, 19, 20, 21, 22, 23, 24, 36, 40, 41, 46, 50, 55, 58, 61, 69 | 17 |
| 2 | 1, 2, 3, 4, 5, 6, 7, 12, 13, 16, 25, 26, 28, 29, 30, 32, 33, 35, 37, 38, 42, 43, 44, 45, 49, 52, 53, 54, 57, 59, 60, 62, 74, 75, 76, 77, 78, 80 | 38 |

| | | |
|---|---|----|
| 3 | 9, 10, 11, 14, 15, 17, 27, 31, 34, 39, 47, 48, 51, 56, 63, 64, 65, 66, 67, 68, 70, 71, 72, 73, 79 | 25 |
|---|---|----|

From segmentation to targeting, to define each segment is needed. So we define 3 segments by extracting top 5 keywords as the following table.

Table 3: Top 5 keywords in 3 segments

| Segments | Keywords |
|-----------|---|
| Segment 1 | Software, Configuration, Hardware, Service, Computer |
| Segment 2 | Diagnostic, bungle, Claim, Business, Simulation |
| Segment 3 | Communication, Connection, Electronic, Internet, Mail |

We did not consider the following keywords, because they showed all segments or were not meaningful. They were ‘and’, ‘the’, ‘marketing’, ‘engineering’, ‘for’, ‘which’, ‘this’, ‘that’, and so forth.

Table 4: Defining segments by top 5 keywords

| Segments | Defining segment |
|-----------|--|
| Segment 1 | Technology field for constructing software/hardware systems of marketing engineering |
| Segment 2 | Technology field for developing evaluation system of management performance and simulation tool of trial and error |
| Segment 3 | Technology field for building on-line marketing system based on Internet |

To select targeted segment, we propose recency–frequency (RF) chart as a perceptual map.

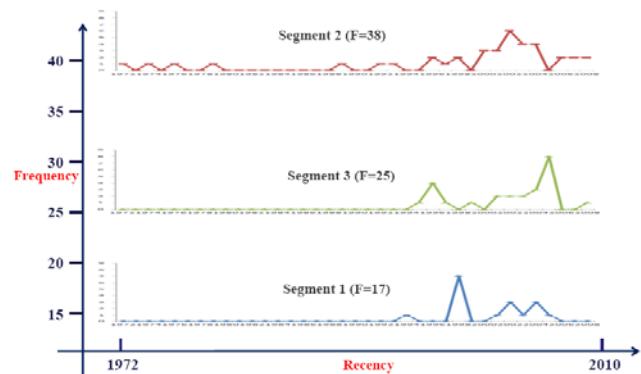


Fig. 5 RF chart.

According to vertical axis of FR chart, we knew segment 2 is higher than others. Also we can find segment 2 and segment 3 have more recency weights than segment 1.

By the frequency of RF chart, we can select segment 2 as a target technology market. Also we determined segment 3 for selecting target technology market based on recency of RF chart. But we want to determine only segment for targeting. So we need a solution of the problem. To settle this problem, we proposed an recency – frequency (RF) score function as the following.

$$RFscore(i) = \alpha \times Tf_i + (1 - \alpha) \times \left(\frac{\sum_{k=fy}^{ly} (k - (fy - 1))}{Tf_i} \right) \quad (3)$$

Where i is segment number and α is a weight. Tf_i is defined as total frequency of segment i . fy and ly represent first year and last year of the patents. The segment with the largest score becomes a target market. We selected segment 2 as a target market of the technology of marketing engineering by the following table.

Table 5: RF scores of 3 segments

| Segments | Parameters | |
|-----------|--------------|--------------|
| | $\alpha=0.5$ | $\alpha=0.4$ |
| Segment 1 | 23.21 | 24.45 |
| Segment 2 | 32.78 | 31.73 |
| Segment 3 | 27.94 | 28.53 |

Though the values of α , the RF scores of segment 2 were as large as ever. Therefore in segmentation step of STP strategy modeling, we determined segment 2 to the target market.

Next we analyzed the selected target market. The following figure shows the number of patents in segment 2 by year.

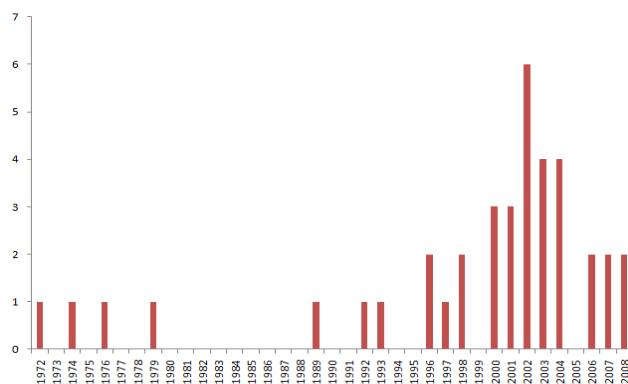


Fig. 6 # of patents in segment 2 by year.

In targeting step, we can define the technologies of segment 2 for marketing engineering as evaluation systems

of management performance and simulation tools of trial and error in diverse business environments using top 5 keywords.

In the next positioning step, we consider some positioning strategies. To begin with the needs of marketing engineering technology for segment 2 have been increased. Also we can advice companies related to marketing engineering have to do R & D for segment 2 based technologies. Therefore, the companies will be strong enterprises with good intellectual properties and business models of marketing engineering. In addition, their benefits and social contributions will be increased.

In our experiment, we used R-project statistical computing package as computing software [13].

5. Conclusions and Future works

A company's R & D strategy of marketing engineering technology is at an early state, because there is a few patent up to now in the world. So we have many chances in marketing engineering as the following. There are patents, research publications, developing new technologies, developing vacant technologies, and so forth.

We think our research has some limitation of this study. We have not any publication related to this case study, because there are few research papers about our study. So we had so difficult to work our research. In addition, we thought to need more complete gate for STP strategy modeling through technology marketing. We can expect better performance using advanced RFM score function of the segment. In this study, we used RF score except monetary (M). We have to find out a measure for M of a patent.

References

- [1] X. Chen, W. Yin, P. Tu, and H. Zhang, "Weighted k-Means Algorithm Based Text Clustering," Proceedings of International Symposium on Information Engineering and Electronic Commerce, 2009, pp. 51-55.
- [2] M. Fattori, G. Pedrazzi, and R. Turra, "Text mining applied to patent mapping: a practical business case," World Patent Information, Vol. 25, 2003, pp. 335–342.
- [3] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Advances in knowledge discovery and data mining, AAAI Press/The MIT Press, 1996.
- [4] I. Feinerer, K. Hornik, and D. Meyer, "Text Mining Infrastructure in R", Journal of Statistical Software, Vol. 25, Iss. 5, 2008, pp. 1-54.
- [5] G. H. Golub, and C. Reinsch, "Singular value decomposition and least squares solutions," Numerische Mathematik Vol. 14, No. 5, 1970, pp. 403-420.
- [6] J. F. Hair, B. Black, B. Babin, and R. E. Anderson, Multivariate Data Analysis, Prentice Hall, 1992.

- [7] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning – Data Mining, Inference, and Prediction*, Springer, 2001.
- [8] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. S., Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, 2002, pp. 881–892.
- [9] P. Kotler, and K. L. Keller, *Marketing Management*, Prentice Hall, 2009.
- [10] T. Kohonen, *Self-Organizing Maps*, Springer, 2000.
- [11] S. Lee, B. Yoon, and Y. Park, "An approach to discovering new technology opportunities: Keyword-based patent map approach", *Technovation*, Vol. 29, 2009, pp. 481-497.
- [12] G. L. Lilien, and A. Rangaswamy, *Marketing Engineering*, Prentice Hall, 2003.
- [13] R Development Core Team, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>, 2010.
- [14] Y. H. Tseng, C. J. Lin, and Y. I. Lin, "Text mining techniques for patent analysis", *Information Processing and Management*, Vol. 43, 2007, pp. 1216-1247.
- [15] United States Patent and Trademark Office (USPTO), www.uspto.gov
- [16] P. Wang, I. M. Cockburn, and M. L. Puterman, " Analysis of Patent Data-A Mixed Poisson Regression Model Approach," *Journal of Business & Economic Statistics*, Vol. 16, No. 1, 1998, pp. 27-41.
- [17] B. Yoon, and Y. Park, "Development of New Technology Forecasting Algorithm: Hybrid Approach for Morphology Analysis and Conjoint Analysis of Patent Information," *IEEE Transactions on Engineering Management*, Vol. 54, No. 3, 2007, pp. 588-599.
- [18] B. Yoon, and S. Lee, "Patent analysis for technology forecasting: Sector-specific applications," *Proceeding of IEEE International Conference on Engineering Management*, 2008, pp. 1-5.

Sunghae Jun is an associate professor in the department of Statistics, Cheongju University, Korea. He received the B.S., M.S., and Ph.D. degrees in department of Statistics from Inha University, Korea, in 1993, 1996, and 2001. Also he got his doctorate in computer science and engineering from Sogang University, Korea, 2007. He worked in NCR as a data mining consultant from 2000 to 2001. His research fields are machine learning, evolutionary computing, and management of technology(MOT).

Redesigning the user interface of handwriting recognition system for preschool children

Mohd Nizam SAAD¹, Abd. Hadi ABD. RAZAK², Azman YASIN³, Nur Sukinah AZIZ⁴

^{1,2,3}College of Arts and Sciences
Universiti Utara Malaysia, Sintok,
Kedah, Malaysia

⁴Kolej Universiti TaTi
Faculty of Computer, Media, and Technology
Information Technology Department
Kemaman, Terengganu, Malaysia

Abstract

Nowadays there are handwriting recognition systems that can be occupied to assist children learning how to write properly. However, one of the major barriers that hinders them using the system is its complex user interface where the designed is based on adult preferences. Therefore in this paper, we present the guideline to redesign the user interfaces via our experience developing a handwriting recognition system for pre-school children named Handwriting-based Learning Number (HLN). The redesign process has followed eight guidelines and rules as presented by Schniederman. The user interface satisfaction evaluation result done using Questionnaire for User Interface Satisfaction (QUIS) is very convincing where the users are almost satisfied with the redesign process that we did to the user interface. Hence we found that the guidelines are very useful and developers are all welcome to follow it if they intend to do similar system like us.

Key words - User interface redesign; Handwriting recognition system; Preschool children learning, Questionnaire for User Interface Satisfaction.

1. Introduction

A preschool children learning curve normally starts with writing, reading, and calculating simple numbers. As they growing older, they will be taught with more advance skill to master these three basics knowledge. Once they manage to get along with the learning tempo, these children are all ready to face with more challenging subject which required additional cognitive understanding. Hence, the early stage of learning how to learn especially learning literacy is crucial for these children.

Learning to write is one of the most-essential skills children will ever learn. Typing on the keyboard is obviously a very useful skill but writing by hand is more important especially for preschool children [1]. This is because writing can develop sensorimotor and it must

develop in the early aged. Research shows that in the most United Kingdom classroom, children spend between 30 % and 60 % of their school classroom time doing writing activities [2]. This is done due to the handwriting process needs attention, memory and cognition, and the motor skills. All these elements must be blended together so that the children are able write properly.

The development of writing ability is not only important in building a child's self-esteem, but is considered an essential ingredient for success in school. This is due to handwriting performance has direct effect on academic performance [3, 4].

Illegible handwriting can create a barrier to accomplishing other higher-order skills such as spelling and story composition. Moreover, academic performance studies performed by [3] showed that those who have handwriting problem such as poor handwritten and slow to write words are causally difficult to form sentences, have limited vocabularies, and cannot writing a full sentence or paragraph. They also added that these children also will face complexity in mathematics where doing the math exercise seems to be a burden job for them. As a result, they are unable to score well in their examination.

In this paper we are going to share our experience on designing the user interface for the handwriting recognition system that we developed to assist preschool children learning how to write. The handwriting recognition system that we developed is called Handwriting-based Learning Number (HLN). Having a good user interface is essential for our system since it can ease the interaction processes between the children and the system. It is a challenging task since it involves children as the user which their behavior on using the computer sometime is difficult to expect. Handwriting recognition system is a system that automated process of turning handwriting work into a

computer readable form [5]. When the handwriting is in the form of binaries, the computer able recognize them and this can ease the process of giving feedback to the children especially on how to improve their handwriting. We believed that by having the system, the children can boost up their time to learn how to write since the system can assist them whenever they learn. Additionally, we are concentrating on the offline handwriting recognition system. Off-line handwriting recognition involves the automatic conversion of text in an image into letter codes which are usable within computer and text-processing applications [6]. Most of the time, the central tasks in off-line handwriting recognition are character recognition and word recognition.

This paper is divided into six sections; we start with the introduction as the first section. Section II will describe about the challenges the preschool children face to use the system. Meanwhile Section III talks about the guideline for user interface design for children. Section IV is about the HLN development which is the prototype system. Afterward, Section V will show the analysis that we did based on data gathered using QUIS (Questionnaire for User Interface Satisfaction) questionnaire. Finally we provide the conclusion for the paper.

2. Challenges faced by pre-school children on using the handwriting recognition system.

In Malaysia, the preschool education is an informal program which is mainly established to provide the learning experience to children whose age between four to six years old. It is an early preparation before they enter the first grade in a formal school. Currently, the preschool education has already been instituted into the National Education System so that it can enhance the child's potential in all aspect as well as a good preparation before entering the school. As academicians, we want to contribute our expertise to help the preschool children education especially learning literacy. In order to do so, we create a handwriting recognition system as one of the tool that can we believe can help preschool children learning literacy faster.

When we first develop the system, we try to replicate the user interface with some of the existing handwriting system available. Although most the system is good in recognizing the handwriting, we found that the user interface is mainly design for adult. From our point of view, the children might have difficulties using the system since their mental model is different. The interaction process for them might not be as good as giving the system to adult. Our assumption is parallel with the work done by [7], where there is evidence that some children had poor understanding of what happening in the handwriting recognition system (in the study they use a software named Paragraph Pen Office 6) although they used it for months.

Secondly, most children want content that is entertaining, funny, colorful, and uses multimedia effects [8]. We found out that most the user interface for existing handwriting recognition system is too complex for the children. The system only emphasize on the complexity of the processing algorithms and the invisibility of the recognition process[5]. It is very rare to find handwriting recognition system that embedded with such entertaining features to attract children to use them. The condition might make the children to feel less attractive hence prevent them from using the system longer.

On the other hand, existing handwriting recognition system interface also does not consider on how children use the system in term of browsing, spelling, navigating, and using input devices. This is one of the three problems stated by [9] faced by children on searching and browsing the information in a system apart from limited motor skill, and inadequate knowledge criteria. The pre-school children knowledge and skills are not as good as adult since they are less experience using the system. Moreover, their cognitive way of thinking is also different than adult. Therefore they need a better user interface that can assist them using the system properly.

Based on the challenges mentioned, we strongly believe that the children will face with difficulties if we adopt the similar interface into our system, hence the need for new interface design is crucial for our handwriting recognition system.

3. User interface guidelines for developing handwriting recognition system for pre-school children

There are several researchers who have published excellent user interface guidelines for children so that other researchers can use them to conduct research. Among of them are [5],[8],[10],[11],[12] and [13]. They have provided a good path for the next researchers to do research especially on designing a good user interface for systems targeted to children. For the sake of this study, we adapted general interface design rules, presented by Schniederman. According to [5] Schniederman's general interface design rules include:

- Strive for consistency
- Enable frequent users to use shortcuts
- Offer informative feedback
- Design dialogues to yield closure
- Offer error prevention and simple error handling
- Permit easy reversal of actions
- Support internal locus of control
- Reduce short term memory load

The illustration for the guideline published can be referred in Fig. 1.

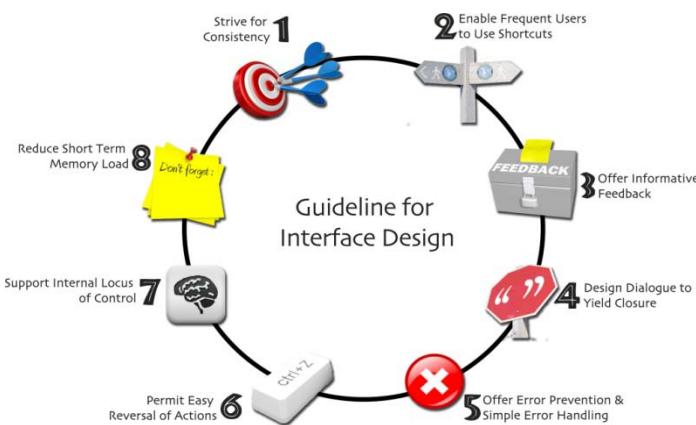


Fig. 1. Illustration for the guideline to design the user interface for children.

[5] also added that these rules are suitable to be adapted not only for ordinary type of system, but beyond than that, they are also suit when designing the user interface for children. Hence, we have followed the rules to design the user interface for our handwriting recognition system. We are also recommending readers who are interested to know further on the rules to read the book from the mention author for more detail.

4. HLN DEVELOPMENT

Based on the guidelines and rules that identified, we started our work on developing the HLN. HLN is a handwriting recognition system developed mainly for preschool children. The language that we use for the system is Malay language. We choose Rapid Application Development (RAD) as the methodology for our system development. The RAD is a system development approach encouraging and facilitating re-use of software components [14]. It incorporates prototyping and user feedback as its main mechanisms. We divide the development phase into four stages. The first phase is analysis requirement. This phase defines the functions and data subject areas that the system will support and it also determines the system's scope. We observed and interview the preschool children to get their opinion in the user interface design. Analysis for the interface of the existing handwriting recognition system such as CobWeb, Note Materials, Maths Materials and Kanji Practice Materials is also done to determine the benefits and weakness of the interface and the new system requirements.

After the requirements have been identified, the second phase is to start designing the interface. This design process is to determine the structure of the HLN and also known as the Functional Design Stage. We start the process by

making the screen design using storyboarding approach. All design rules are taken seriously to get the most best design for the system. This screen design will determine the layout and element that should be included in the screen of the system. Its include layout for text, animation, background and icon that use in HLN.

Once the design complete, we move on to the development stage. We start the development by making the prototype. When initial prototype is completed, the interview process with the children is still keep on going so that we can gather additional, more detailed requirements from them. Once we get the modified set of requirements, prototype will be updated to reflect the new set of requirements. For the HLN prototype, we used Microsoft Visual Basic as programming language to develop the recognition engine for the system. We choose Microsoft Visual Basic since it enables event-driven programming and it is very easy to use them for developing the system.

Although Microsoft Visual Basic is good for developing the recognition engine, we encountered that one of its drawback is it cannot fulfill the type of interface that we want. It has limited functions that enable us to create the system which contains entertaining features as what the children want. Hence to make the user interface more attractive, we use Macromedia Flash MX. Macromedia Flash is well known software that enables its user to create a catchy animation. The user interface of the system is in the shockwave file format. By using the software we can create dozens of the graphical element, animation, audio recording and embedded it in into the system. Our user interface design theme for the handwriting recognition system is learning in an orchard playground. Fig. 2 shows the some screen snapshot of the HLN.



Fig. 2. The HLN's snapshot

As a brief description of the HLN, it allows the children to enter one number at one time (between 0 to 9) into the box. Since the system is executed in a tablet PC, it allows

the children to use either a mouse or a tablet pen. If the children want a real writing experience, we recommend them to use the tablet pen. Hence, they can use in more meaningful way. The written number will be computed by the recognition engine in order to determine how exact the handwriting with the handwriting pattern (in template form) that we have installed into the system. The similarities will be displayed in the form of percentage value.

Apart from the learning activities, we also included several leisure activities such as simple quizzes and games. By having such features, we believe that it can enable the children to use the system in longer time. Fig. 3 shows the screenshot of the activities screen.

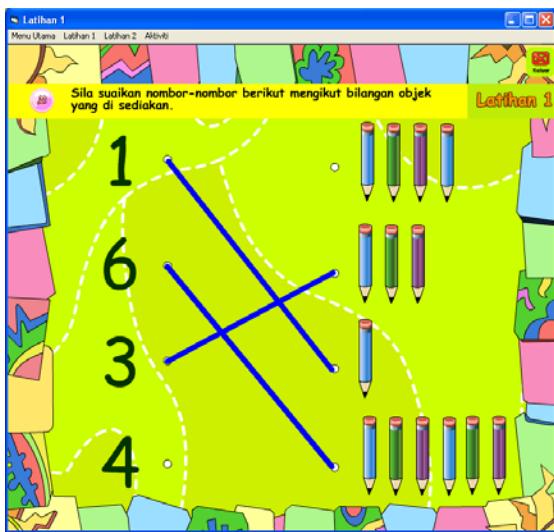


Fig.3. The screen activities in HLN

Finally, we test the prototype using to the children using the QUIS questionnaire. The next section will elaborate our finding.

5. USER SATISFACTION TEST

The evaluation is a very important since it will judge on the outcome of the development. We used the summative evaluation to test the prototype. The HLN user satisfaction evaluation was conducted on thirty children. Each of them was given brief explanation regarding the usage and the user interface of the prototype. Since we adopted the questionnaire without any modification from [15], we captured the children feedback by our own. We would be very please if we have enough time to modify the questionnaire and make it suitable for the children to express what they actually perceived about the system on their own. The QUIS can be accessed from the following website <http://oldwww.acm.org/perlman/question.html>.

The QUIS covers five dimensions which include: *Overall reaction to the software, screen, terminology and system*

information, learning and system capabilities. All questions were intended to collect data on each user opinion regarding the HLN prototype. A 10 point scales were used in the questionnaire. Table I summarizes the mean score for the users' satisfaction that we manage to gather from the users for each dimensions.

TABLE I. THE MEAN SCORE FOR USER SATISFACTION TEST

| Item | Mean | Std. Deviation |
|--|--------|----------------|
| OVERALL REACTION TO THE SOFTWARE | | |
| This software is wonderful | 4.0000 | 0.64327 |
| This software is easy to use | 3.9667 | 0.66868 |
| This software is satisfying to use | 3.5000 | 0.86103 |
| This software is adequate as needed | 3.6333 | 0.88992 |
| This software is stimulate | 3.8667 | 0.57135 |
| This software is flexible to use | 4.0667 | 0.58329 |
| SCREEN | | |
| Reading characters on the screen is easy | 3.9000 | 0.75886 |
| Highlighting simplifies task | 3.9000 | 0.66176 |
| Organization of information is clear | 3.5333 | 0.62881 |
| Sequence of screens is clear | 3.7667 | 0.56832 |
| SYSTEM INFORMATION | | |
| Use of terms throughout system is consistent | 3.5000 | 0.86103 |
| Terminology always related to task | 3.7667 | 0.50401 |
| Position of messages on screen is consistent | 4.0333 | 0.61495 |
| Prompts for input is clear | 3.7667 | 0.85836 |
| Computer always informs about its progress | 3.8000 | 0.55086 |
| Error messages is helpful | 4.0333 | 0.85029 |
| LEARNING | | |
| Easy to operate the system | 3.7667 | 0.43018 |
| Exploring new features by trial and error | 3.8667 | 0.43417 |
| Remembering names and use of commands | 4.0333 | 0.76489 |
| Performing tasks is straightforward | 3.9667 | 0.61495 |
| Help messages on the screen is helpful | 4.0667 | 0.73968 |
| Supplemental reference materials is clear | 3.7667 | 0.77385 |
| SYSTEM CAPABILITIES | | |
| The system speed is fast | 3.9000 | .71197 |
| The system is reliable | 3.9667 | .61495 |
| The system tends to be quite | 4.0667 | .73968 |
| Easy to correcting your mistakes | 3.9667 | .80872 |
| Designed for all levels of users | 4.0000 | .87099 |

The overall mean score as shown in Table 4.3 indicates that most of the users are almost satisfied using the HLN prototype. The first dimension is about the overall reaction to the software. It include questions such as is the software is wonderful, is it easy to use, is the user satisfy using it, are it adequate as needed, and is it stimulate and flexible to use. The mean score is between 3.5000 and 4.0667. This condition indicated that the users are mostly agreed that they generally satisfied with the HLN prototype. This is because the HLN prototype has music to attract the children to use and the music can turn on or off.

Meanwhile, on the second dimension; screen, the mean score is between 3.5333 and 3.9000. This is also indicating that the user is also almost agreed that the screen has managed to satisfy them using the HLN. We design the HLN's screen in a simple way yet it is so colorful and suitable for the children.

The third dimension is about system information. For this dimension, the feedback indicate that the mean score is between 3.5000 and 4.0333. This also implies that users were also almost satisfied with the terms of system information HLN.

On the other hand, for the fourth dimension; learning, the mean scores is between 3.7667 and 4.0667. This implies that users felt that the learning process to use the HLN is quite easy and highly satisfied with the system. Finally, the last dimension; system capabilities, the mean score is between 3.9000 and 4.0667. This condition is also implies that the HLN executed via Tablet PC has meet need on using the prototype. Overall, the results indicate that the users is almost agreed that the interface design for HLN prototype is good.

6. CONCLUSION

In this paper, we presented our experience on developing a handwriting recognition system named HLN that mainly developed for preschool children. At the beginning of the paper we presented our main aim to develop the system, which is to assist the children to learn writing. Later, we discuss that when developing a complex system such as the handwriting recognition system for the children, the user interface plays a big role to ensure that the children can use it easily. If it is taken for granted, the children might face some challenges to use the system. Afterward, we talk about the importance of following a well established design rules and guidelines so that the design process of the system can become successful. We also presented the process of developing the HLN and some brief information about the system description. Finally, we present the result that concluded that the users are almost satisfied with the system.

On the other hand we admit that the HLN is still far from completed. There are several limitations for the prototype such as the prototype is not fully functional due to time constraints in developing the prototype. There are also features that did not implemented to this prototype such as visual and audio as a guidance to recreate the shape of the number whereby user can perform specific learning exercises without the presence of the tutor. For future development and expansion of this research, we can implement the visual and audio guided information to attract the preschool children to use it and enhance user learning by following step by step number formation and hopefully, we can enhance the system to online handwriting recognition.

REFERENCES

- [1] M. A. Eid, M. Mansour, A. H. E. Saddik, and R. Iglesias, "A haptic multimedia handwriting learning system," in *Proceedings of the international workshop on Educational multimedia and multimedia education*. Augsburg, Bavaria, Germany: ACM, 2007, pp. 103-108.
- [2] K. McHale and S. Cermak, "Fine motor activities in elementary school:Preliminary findings and provisitional implications for children with fine motor problems," *American Journal of Occupational Therapy*, vol. 46, pp. 898-903, 1992.
- [3] A. H. Faridah, I. Naimah, Y. Hamidah, and A. Habibah, "Poverty and education: The Malay mind changes towards the excellency of academic education., " *Journal of Education Research*, vol. 7, pp. 25-56, 2005.
- [4] J. Medwell and D. Wray, "Handwriting - A Forgotten Language Skill?," *Language and Education*, vol. 22, pp. 34-47, 2008.
- [5] J. C. Read, S. MacFarlane, and P. Gregory, "Requirements for the design of a handwriting recognition based writing interface for children," in *Proceedings of the 2004 conference on Interaction design and children: building a community*. Maryland: ACM, 2004, pp. 81-87.
- [6] J. Read, S. MacFarlane, and C. Casey, " Measuring the Usability of Text Input Methods for Children," in *Proceedings of HCI International 2001*, vol. 3. New Orleans: Lawrence Erlbaum, 2001, pp. 559-572.
- [7] J. C. Read, S. MacFarlane, and C. Casey, "What's going on?: discovering what children understand about handwriting recognition interfaces," in *Proceedings of the 2003 conference on Interaction design and children*. Preston, England: ACM, 2003, pp. 135-140.
- [8] D. Grammenos, C. Paramythis, and C. Stephanidis, "Designing the User Interface of an Interactive Learning Environment for Children," presented at Proceedings of the ERCIM WG UI4ALL one-day joint workshop with i3 Spring Days 2000 on "Interactive Learning Environments for Children", Athens, Greece, 2000.
- [9] H. B. Hutchinson, B. B. Bederson, and A. Druin, "The evolution of the international children's digital library searching and browsing interface," in *Proceedings of the 2006 conference on Interaction design and children*. Tampere, Finland: ACM, 2006, pp. 105-112.
- [10] H. Gelderblom and P. Kotze, "Designing technology for young children: what we can learn from theories of cognitive development," in *Proceedings of the 2008 annual research conference of the South African Institute of Computer Scientists and Information Technologists on IT research in developing countries: riding the wave of technology*. Wilderness, South Africa: ACM, 2008, pp. 66-75.
- [11] A. Druin, B. B. Bederson, J. P. Hourcade, L. Sherman, G. Revelle, M. Platner, and S. Weng, "Designing a digital library for young children," in *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*. Roanoke, Virginia, United States: ACM, 2001, pp. 398-405.
- [12] H. Gelderblom and P. Kotze, "Ten design lessons from the literature on child development and children's use of technology," in *Proceedings of the 8th International Conference on Interaction Design and Children*. Como, Italy: ACM, 2009, pp. 52-60.
- [13] H. Niemi and S. Ovaska, "Designing spoken instructions with preschool children," in *Proceedings of the 6th international conference on Interaction design and children*. Aalborg, Denmark: ACM, 2007, pp. 133-136.
- [14] P. Beynon-Davies and S. Holmes, "Integrating rapid application development and participatory design," *Software, IEEE Proceedings*, vol. 145, pp. 105-112, 1998.
- [15] J. P. Chin, V. A. Diehl, and K. L. Norman, "Development of an instrument measuring user satisfaction of the human-computer interface," in *Proceedings of the SIGCHI conference on Human*

factors in computing systems. Washington, D.C., United States:
ACM, 1988, pp. 213-218.

AUTHORS PROFILE

Mohd Nizam Saad (Corresponding author) is a lecturer at the College of Arts and Sciences, Applied Science Division, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah. Malaysia. His research interest includes multimedia technology applied in education, preschool literacy learning, and video technology for broadcasting.

Azman Yasin is a senior lecturer at the College of Arts and Sciences, Applied Science Division, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah. Malaysia. His research interest includes software engineering education, information retrieval specifically scheduling and timetabling using artificial intelligence techniques.

Abd Hadi Abd Razak is a lecturer at the College of Arts and Sciences, Applied Science Division, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah. Malaysia. His research interest includes multimedia technology applied in education, preschool literacy learning, and video technology for broadcasting.

Nur Sukinah Aziz is a lecturer at the Faculty of Computer, Media and Technology, TATI University College (TATiUC), Teluk Kalong, 24000 Kemaman, Terengganu, Malaysia. Her research interest includes software engineering education system, web based, multimedia and HCI.

Texture Classification Using an Invariant Texture Representation and a Tree Matching Kernel

Somkid Soottitantawat¹ and Surapong Auwatanamongkol²

¹ School of Applied Statistics,
National Institute of Development Administration, Bangkok, Thailand, 10240

² School of Applied Statistics,
National Institute of Development Administration, Bangkok, Thailand, 10240

Abstract

In this paper, an alternative approach for texture classification using an invariant texture representation and a tree matching kernel is proposed. The approach identifies regions of a given texture image using a Speed-Up Robust Feature or SURF descriptor. The regions of all training texture images are then clustered into a tree of non-uniformly shaped regions based on the distribution of them using a hierarchical k-means algorithm. The tree structure forms a tree of keypoints to be used for determining similarities between two texture images. The similarity is computed based on an approximate matching kernel called a tree matching kernel. Finally, Support Vector Machines (SVMs) with the tree matching kernels are constructed to classify textures. The performances of the proposed method are evaluated through experiments performed on textures from the Brodatz and UIUCTex datasets. The experiment results demonstrate that the proposed approach is quite robust to scale, rotation, deformation and viewpoint changes and achieves higher classification rates than some other well known methods.

Keywords: *Texture Classification, Tree of Keypoints, SURF Descriptor, Support Vector Machines, Tree Matching Kernel, Hierarchical K-means Clustering.*

1. Introduction

In the visual world, textures can be regarded as the visual appearances of surfaces and may be perceived as being directional or non-directional, smooth or rough, coarse or fine, regular or irregular, etc. Several textures are observed on both artificial and natural objects and scenes. The surface characteristics of textures can be used to recognize objects in an image, to segment an image and to understand an image [27]. So, textures play an important role in many image analyses, computer vision and pattern recognition tasks. However, environment and illumination conditions can affect the appearance of textures, and so complicate the

tasks. Textures in real images can vary in scale, brightness, and rotation as imaging conditions change. Therefore, to enable texture analysis in real images, texture representation should be invariant to imaging conditions such as non-rigid deformation, viewpoint, scaling and lighting. A brief review of the invariant texture analysis methods is presented in [17].

The goal in this research is to perform texture classification that is robust to the mentioned environment and illumination conditions. A texture representation, which is invariant to the conditions, along with the new classification method is proposed. Our approach consists of the following steps: (1) Constructing an invariant texture representation step that consists of feature detection and then extraction of texture regions, which are invariant to the conditions due to both geometric and photometric transformations. We propose that Speeded Up Robust Features (SURF) is to be used as local invariant descriptors for the texture regions. (2) Building a tree of keypoints (regions) step performed hierarchical k-means clustering on all regions of all textures in the training set. A tree of keypoints is then constructed from the hierarchical k-means clustering. (3) Texture modeling step builds multi-SVM classifiers with a one-against-all tournament approach to classifying texture images. The tree matching kernel is used in the SVMs and utilizes the tree of keypoints to determine the similarity between two textures. Figure 1 shows a sketch of the process for the proposed approach.

The organization of this paper is as follows. Related works are reviewed in section 2; the proposed approach is described in section 3; experiments and results, designed to evaluate the effectiveness of the proposed approach, are presented in section 4 and finally conclusions are given in section 5.

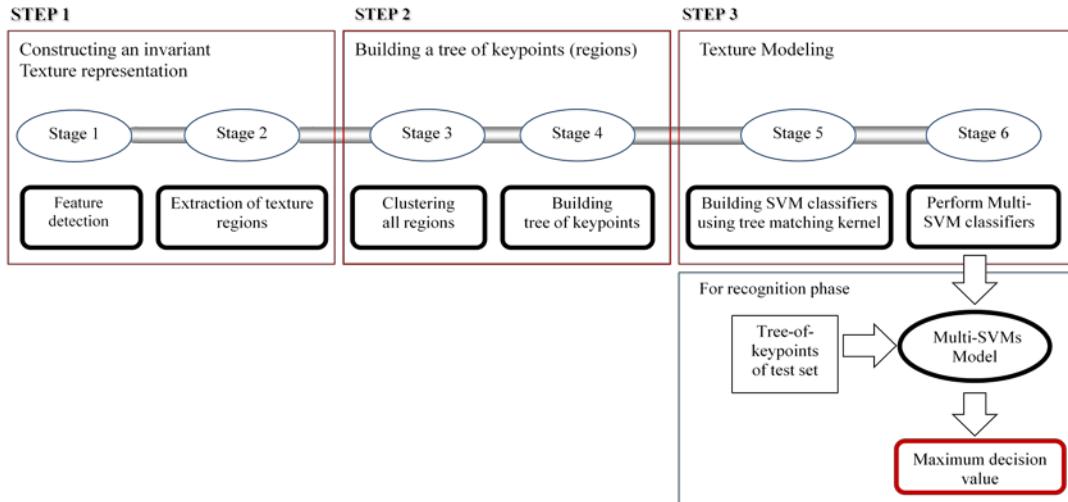


Fig 1. The framework for the proposed approach.

2. Related Works

In this section, some related works on texture classification are reviewed. Lazebnik et al. [29] proposed a sparse texture representation using local affine regions for recognizing textured surfaces under a wide range of transformations, including viewpoint changes and image nonrigid deformations. Features descriptors of each texture are clustered using standard k-means to form the texture signature $\{(m_1, u_1), (m_2, u_2), \dots (m_k, u_k)\}$, where k is the number of clusters, m_i is the center of the i^{th} cluster, and u_i is the weight of the cluster. They used Earth Mover's Distance (EMD) to measure the similarity between two signatures. During the classification stage, nearest-neighbor classification with EMD was used to classify texture images. Mellor et al. [23] described a method based on invariant combinations of linear filters. Unlike Lazebnik's methods, they proposed a novel family of filters, which provides scale invariance, resulting in a texture description invariant to local changes in orientation, contrast and scale and robust to local skew. The χ^2 similarity measure is used on histograms derived from their filter responses. Recently, Qin et al. [22] presented a novel approach to classify texture collections using an unsupervised approach. Given image database, they extracted a set of invariant descriptors from each image and the descriptors of all the images were clustered to form keypoints. A texture image can be represented by a bag-of-keypoints. Probabilistic Latent Semantic Indexing (PLSI) and Non-negative Matrix Factorization (NMF) were used for the unsupervised texture classification.

Several local features for texture representation were proposed in [9], [14], [16], [20], [30]. The local features are distinctive, invariant to many kinds of geometric and

photometric transformation. They are suitable for image classification and have been used in many applications, e.g. object recognition [1], [3], scene recognition [31], robot localization [2], and texture classification [25], [26], [29], [30]. A review of other local features can be seen in [16].

The main drawbacks of the local feature methods are that different local feature methods can produce different numbers of feature vectors and generate no obvious structural information about the vectors, for example, no ordering among the vectors is produced. Thus, to overcome these problems, similar feature vectors can be clustered to create a designated number of representative feature vectors. For example, Boughorbel et al. [28] proposed that the centroids of the computed clusters represented virtual features for images. Csurka et al. [12] used a similar clustering technique to construct bags of keypoints. Each bag is represented by a bin of a histogram. Hence, features of two images can be compared through the matching of their feature histograms.

3. Proposed Approach

3.1 Region Formation

For texture analysis, region formation is the first essential task. To form regions in a texture, we propose the recently developed Speeded-Up Robust Features (SURF) [14], [10], to use as the local descriptor. SURF has already been used in a few real world applications [2], [13], [15]. SURF is very well suited for tasks in object detection, object recognition and image retrieval [13]. SURF possesses more discriminative power than other features such as SIFT [9] and it can be computed more efficiently and yields a lower dimensional

feature descriptor resulting in faster matching [2]. SURF can be computed as follows.

3.1.1 Interesting Point Detection

To compute SURF, interesting points must be detected. The detection is performed using the Hessian-matrix approximation. Given a point $x = (x, y)$ in an image I , the Hessian matrix $H(x, \sigma)$ in x at scale σ is defined as follows:

$$H(x, \sigma) = \begin{bmatrix} L_{xx}(x, \sigma) & L_{xy}(x, \sigma) \\ L_{xy}(x, \sigma) & L_{yy}(x, \sigma) \end{bmatrix}, \quad (1)$$

Where $L_{xx}(x, \sigma)$ is the convolution of the Gaussian second order derivative $\frac{\partial^2}{\partial x^2} g(\sigma)$ with the image I in point x , and similarly for $L_{xy}(x, \sigma)$ and $L_{yy}(x, \sigma)$.

SURF approximates the second order Gaussians derivatives with box filters. Image convolutions with these box filters can be computed rapidly by using integral images. The location and the scale of the interesting points are selected by relying on the determinant of the Hessian matrix. Interesting points are localized in scale and image space by applying a non-maximum suppression in a $3 \times 3 \times 3$ neighborhood. Then, the local maxima found of the approximated Hessian matrix determinant are interpolated in the scale and image space. For more details, please refer to [14].

3.1.2 Constructing Region Descriptors

This stage consists of two steps. First, SURF constructs a circular region around the detected interesting points in order to assign a unique orientation to the former. The orientation is computed using Haar wavelets responses in both x and y directions. The Haar wavelets can be quickly computed via integral images. The dominant orientation is estimated and included as the interesting point information. Next, SURF descriptors are constructed by extracting square regions around these interesting points. These are oriented in the directions assigned in the previous step. The windows are split up in 4×4 sub-regions in order to retain some spatial information. In each sub-region, Haar wavelets responses in horizontal and vertical directions (d_x and d_y) are summed up over each sub-region. Moreover, the absolute values $|d_x|$ and $|d_y|$ are summed in order to obtain information about the polarity of the image intensity changes. Therefore, the underlying intensity pattern of each sub-region is described by a vector $V = [\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|]$. The resulting descriptor vector for all 4×4 sub-regions is of length 64, giving the standard SURF descriptor, SURF-64. An important characteristic of SURF is its fast extraction process due to the fast integral process of images and the fast non-maximum suppression algorithm. Figure 2 shows the output of the SURF regions on two sample images.

3.2 Clustering Regions

After the SURF regions of all the training texture images have been identified, the SURF regions are clustered to form a hierarchical tree. This stage is a one-time process performed before building the SVMs. The similar hierarchical tree structures have been proposed to handle descriptor matching, for instance, searching the tree for images [21] kd-tree [9] and metric trees [5], [8].

Many computer vision algorithms, including our proposed algorithm, require searching as the closest data point of a given data on a high-dimensional space [24]. The nearest neighbor search has been used to find the closest point. However, the performance of the search varies depending on properties of the datasets, such as dimensionality, correlation, clustering characteristics, and size. A better approach is to use metric tree based index methods, proposed by [8] for a fast nearest neighbor search in very large databases. This method is based on searching for the closest leaf node in a hierarchical k-means tree starting from the root down to the leaf.

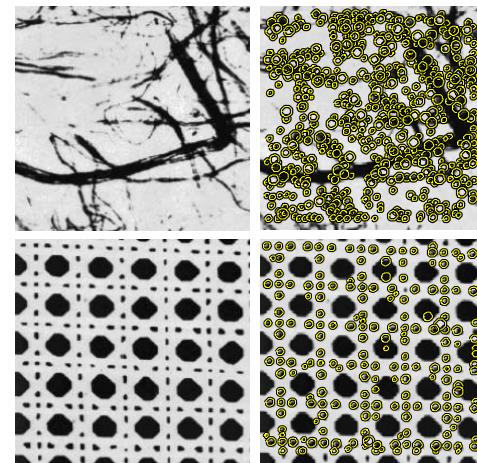


Fig. 2 Outputs of the SURF regions on two sample texture images. Left: original images, right: regions found by SURF descriptors.

The hierarchical k-mean clustering starts with a k-means process (the k-means process is described by Algorithm1) on all SURF regions of all texture images in the training dataset. Next, the same k-means process is recursively applied to each cluster derived from the previous clustering, i.e. recursively splitting each cluster into k subclusters. The recursion is stopped when the number of SURF regions in a cluster is smaller than k . The hierarchical clustering forms a corresponding tree where each leaf corresponds to each of the SURF regions, representing the keypoints of all the texture images in the training dataset. Each node in the tree is represented by the node index, the centroid and the radius (maximum distance from the centroid) of its corresponding cluster. The

centroid and the radius of a node will be used mainly to determine whether the nearest leaf node for a given data is located inside the subtree rooted at the node.

The main advantage of a k-means algorithm is its computational simplicity, which is convenient for large data sets. Its time complexity is $O(NkId)$ when clustering N data points of d dimensions with k centers and I iterations. However, the centers of clusters for k-means algorithms are often initialized randomly, which may result in different clustering solution from run to run. Several methods [5] were proposed to overcome this problem. One simple technique that we adopt to address this problem is to perform k-means for multiple runs and select the solution from the run that yields the minimum sum of squared errors (SSE). The second problem with the k-means algorithm is that empty clusters can be obtained if no SURF regions are allocated to a cluster during the assignment step. If this happens, the empty cluster will be given a new centroid selected randomly from the members of the cluster that has the highest SSE. This will split the cluster and reduce the overall SSE of the clustering.

Algorithm1: k-means process used in this paper.

- 1: Define the number of clusters k
- 2: Define the number of runs m
- 3: For i=1 to m.
- 4: Select k SURF regions randomly as initial centroids.
- 5: While (centroids do change)
- 6: Form k clusters by assigning each SURF region to its closest centroid.
- 7: Recompute the centroid of each cluster.
- 8: End
- 9: Record the k cluster centroids for the i^{th} run
- 10: End
- 11: Select the set of k clusters of the run with minimum SSE.

The computational cost for building the tree structure using hierarchical k-means clustering on a given training set is $O(LNkId)$, where L is the number of levels in the tree. The building time can be reduced significantly by limiting the number of iterations in the k-means clustering stage instead of running it until its convergence is reached [24]. Moreover, the branching factor k can affect the precision of finding the closest data in the tree as well as the building time of the tree. A higher branching factor has proven to give better precision but also a higher building time [24]. Therefore, there is a tradeoff between the precision and the building time. Figure 3 shows an example hierarchical k-means tree built from 3 texture images (with total of 19 SURF regions).

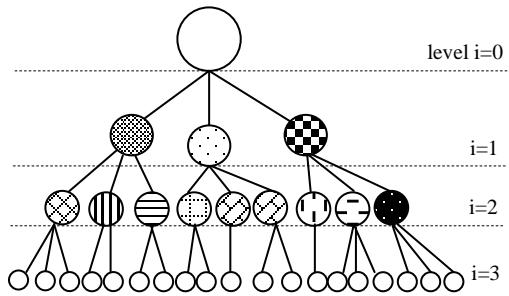


Fig. 3 Three levels of a hierarchical tree with branch factor = 3 populated to represent 3 training images with 19 SURF regions.

3.3 Tree Matching Kernel

After all the SURF regions of all the training texture images have been clustered yielding a tree of keypoints T. A tree of keypoints for any texture image can be constructed by performing hierarchical search (from root to a leaf of T) to identify which clusters of nodes at each level of T that each SURF region of the texture image belongs to i.e. determining the cluster with the closest centroid. All nodes of T, that any SURF region of the image belongs to, form the tree of keypoints of the image. Clearly, the tree of keypoints must be a subgraph of T. Let T_1 and T_2 be the trees-of-keypoints of any two texture images. To determine the similarity between the two images, the tree matching kernel score of T_1 and T_2 will be determined instead. The tree matching kernel score, inspired by Grauman et al. [18], [19], can be defined as the sum of the weighted number of SURF matches found at all common nodes of the two trees. The number of SURF matches found at a common node is defined to be the minimum between the numbers of SURF regions belonging to the two images and located within the cluster of the node minus the numbers of SURF matches already counted by all of their child nodes. The sum of these weighted counts yields the approximate tree matching kernel score.

Let $c_{ij}(T_1)$ be the number of SURF regions in T_1 located inside the cluster of the j^{th} node of level i, $c_{ij}(T_2)$ be the number of SURF regions in T_2 located inside the cluster of the j^{th} node of level i. The tree matching kernel score is computed as follows:

$$\text{Sim}(T_1, T_2) = \sum_{i=m-1}^1 \sum_{j=1}^{k^i} (w_{ij} - p_{ij}) \min(c_{ij}(T_1), c_{ij}(T_2)). \quad (2)$$

Where k is the number of sub-clusters for each cluster, m is the highest number of the levels of two trees, w_{ij} is the weight associated with the j^{th} node at level i of T, p_{ij} is the weight associated with the parent node of the j^{th} node at level i of T. The subtraction part weighted by p_{ij} corresponds to the weighted

number of SURF matches that need to be subtracted from the count of the node's parent. The weight of the j^{th} node at level i of T are set to be equal to $\exp(-2*d_{ij})$ where d_{ij} is the radius of the cluster of the node. This is according to the fact that the SURF matches found at a lower level (larger i) with a smaller value of radius should contribute more significantly to the similarity score than those of the higher levels with higher radius values. This weighting scheme meets the condition required in [18], [19] and makes the kernel, defined by the equation (2), a Mercer kernel (i.e., a symmetric positive definite kernel), which can be used with kernel based methods such as the Support Vector Machine (SVM). The matching scores are normalized by the number of SURFs of the two images in order not to favor larger images. Note that the sum in the equation (2) starts with the index $i = m-1$ because there will be no SURF region matches at the leaf level $i = m$. The time required to compute a tree matching kernel score between two trees-of-keypoints is $O(NL)$. A tree matching kernel score between a testing texture image and a training texture image representing a kernel vector can be determined in the same way as described above.

3.4 Texture Classification with SVMs

Several approaches have been proposed to generalize the binary SVM classifier to solve problems where multi-classes apply. To apply SVM for multi-class classifications, two main approaches have been suggested. The first approach is called "one against one". In this approach, a series of binary classifiers is applied to each pair of classes, and the most commonly computed class is assigned to the given object. This method requires $m(m-1)/2$ classifiers to be built. The second approach, which requires less classifiers to be built, is called "one against rest". This approach requires m binary classifiers where the i^{th} classifier is built as the samples in the i^{th} class are treated as positive examples and the rest as negative examples. In the recognition phase, a testing texture is presented to all m classifiers and is labeled according to the highest decision value among the m classifiers. Because of simplicity, the one-against-rest approach was adopted to build the multi-class SVM classifier in the experiments [7].

4. Experiments

Experiments were carried out on textures images from Brodatz [35] and UIUCTex [31] database. The experiments are similar to those in [29]. Two performance measures [22] are used to evaluate our approach, as follows.

The confusion matrix (CM):

$$CM_{ij} = \frac{|\{I_a \in Test_j : f(I_a) = i\}|}{|Test_j|} \quad (3)$$

$Test_j$ is the set of testing images which belong to class j , and $f(I_a)$ is the class label which obtains the highest

classifier score from the multi-class classifier for a given image I_a .

The mean classification rate (CR):

$$CR = \frac{\sum_{j=1}^N |Test_j| CM_{ij}}{\sum_{j=1}^N |Test_j|} \quad (4)$$

All experiments are carried on a PC Intel Core 2 Quad CPU Q6600, with a clock rate of 2.4 GHz and 4.0 GB of RAM. The classifiers are implemented using Visual C++ 2003 and Matlab (R2008a).

4.1 Datasets

The first dataset from Brodatz is a collection of texture images that features significant inter-class variability, but no geometric transformations between members of the same class. The dataset consists of 111 images. Following the same procedure as [29], we form classes by partitioning each image into nine non-overlapping fragments, for a total of 999 images. Fragment resolution is 215×215 pixels. The training set of Brodatz consists of randomly selected 333 images (3 images/class) and the other 666 images (6 images/class) are used for testing.

The second dataset from UIUCTex consists of 40 examples for each of the 25 texture types. The database is publicly available at http://www-cvr.ai.uiuc.edu/ponce_grp. The resolution of each sample is 640×480 pixels. This database includes surfaces whose texture is mainly due to albedo variations (e.g., wood and marble), 3D shape (e.g., gravel and fur), as well as a mixture of both (e.g., carpet and brick). Significant viewpoint changes and scale differences are present within each class, and illumination conditions are uncontrolled [29]. The training set of UIUCTex contains 250 images (10 images/class) and the rest of the 750 images (30 images/class) are used for testing.

Figure 4 and 5 show several example texture images from the two datasets.

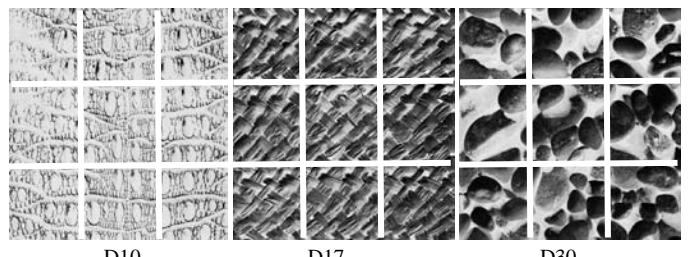


Fig. 4 Examples of three classes of textures from Brodatz dataset.

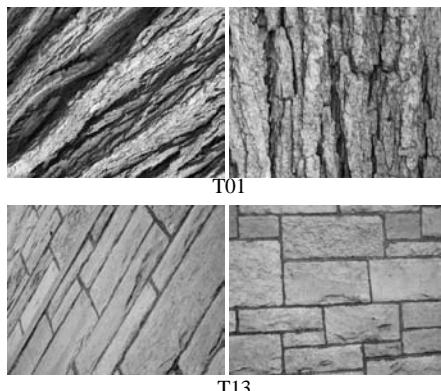


Fig. 5 Examples of three classes of textures from UIUCTex database.

4.2 Region Formation

About 220,493 SURF regions were extracted from the Brodatz training images. The average number of SURF regions extracted per image is about 663. Also, about 268,646 SURF regions were extracted from UIUCTex training images. The average number of SURF regions extracted per image is about 1,075.

4.3 Clustering SURF Regions to build a tree of keypoints

At this stage, a hierarchical tree of 64-dimensional SURF regions for each of the two datasets was built using hierarchical k-mean tree algorithm. The branching factor of 10 is used for the k-mean tree construction. Up to ten iterations of k-means were run and the tree building process can take a few hours. The tree, built from the Brodatz training set, contains 327,567 nodes (220,493 leaf nodes and 107,074 internal nodes) with $d = 64$, the number of levels = 10 and $k = 10$, while the tree, built from UIUCTex training set, contains 395,866 nodes (268,646 leaf nodes and 127,220 internal nodes) with $d = 64$, the number levels = 8 and $k = 10$.

4.4 SVM Classification

After the tree of keypoints was built, multi-class SVM classifiers were built using one-against-rest scheme. The regularization constant C was set at 1, 10, 100, 1000, and 10000. Each C was evaluated using 5-fold cross validation method on the training set only. The SVM classifiers with the parameter $C=10$ gave the best classification rates.

Once the training of SVM classifiers had been completed, the test samples were fed into the classifiers and the predicted class IDs of the test samples were compared with the true labels. The results are reported in terms of the mean classification rates.

4.5 Results

In the first experiment, all 111 textures in the Brodatz dataset were used. According to Xu et al. [11], the 111 textures can be grouped into 3 types based on the degree of regularity or type of structure of the textures. The first type consists of six textures which have degrees of regularity more than 5 (highly regular type). The second type consists of fifty five textures which have degrees of regularity between 4 and 5 (regular type). Finally, the third type consists of fifty textures which have a degree of regularity between 3 and 4 (irregular type). The results of the experiments performed on the three types of the textures are summarized in Table 1. It can be seen that the textures of the highly regular type achieve a classification rate of 100 percent, the textures of regular type achieve a mean classification rate of 90.91 percent, and the textures of irregular type achieve a mean classification rate of 87.85 percent. The mean classification rate for all textures is 90.84 percent. The results are better than those of the three previous works, see Table 3 for the comparisons. Figure 6 shows three textures that were classified incorrectly because the training and testing examples are highly non-homogeneous.

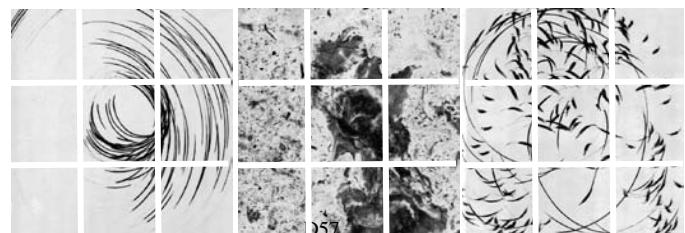


Fig.6 Examples of misclassified textures from Brodatz dataset.

In the second experiment, a set of experiments were conducted on the UIUCTex dataset with an increasing number of classes. The numbers of classes used in the set of experiments are 3, 8 and 25, the same as [22]. The results from the experiments are summarized in Table 2. For the experiments performed on the three classes of T23-T25 and the eight classes of T18-T25, Qin et al. [22] method achieved the best performance. However, as the number of classes increases to 25, the classification task becomes more challenging. The proposed methods achieved the mean classification rate of 93.60 percent for the case of 25 classes, which is better than the previous works of [22], [29].

Table 1 : The mean classification rate results of the Brodatz dataset

| <i>Categories</i> | <i>Proposed Method</i> |
|--|------------------------|
| D6, D20, D33, D51, D52, D101 (6 classes - highly regular type) | 100.00 |
| D2, D3, D4, D5, D10, D14, D16, D17, D19, D21, D23, D24, D27, D32, D34, D35, D36, D37, D38, D39, D41, D42, D43, D45, D46, D49, D50, D53, D54, D55, D56, D59, D63, D64, D65, D66, D68, D72, D75, D79, D81, D83, D84, D85, D86, D92, D94, D96, D98, D100, D102, D103, D105, D106, D110 (55 classes - regular type) | 90.91 |
| D1, D7, D8, D9, D11, D12, D13, D15, D18, D22, D25, D26, D28, D29, D30, D31, D40, D44, D47, D48, D57, D58, D60, D61, D62, D67, D69, D70, D71, D73, D74, D76, D77, D78, D80, D82, D87, D88, D89, D90, D91, D93, D95, D97, D99, D104, D107, D108, D109, D111 (50 classes - irregular type) | 87.85 |
| D1 - D111 (111 classes – all types) | 90.84 |

Table 2 : The mean classification rate results of the UIUCTex dataset

| <i>Categories</i> | <i>Lazebnik [29]</i> | <i>Qin [22]</i> | <i>Proposed</i> |
|--|----------------------|-----------------|-----------------|
| T23 – T25 (3 classes are fabric texture) | 95.89 | 100 | 94.44 |
| T18 – T25 (8 classes are fabric, wall paper, fur and two carpets) | 93.70 | 98.40 | 93.75 |
| T1-T25(25 classes) | 92.61 | 83.00 | 93.60 |

The proposed approach has shown to give high classification rates on both datasets comprising of scale, rotation, deformation and viewpoint changes hence it has proven to be robust to these environment and imaging conditions. The experiment results also show that the kernel-based learning method can yield better texture classification rates than those of the nearest-neighbor classification method with EMD [29]. Furthermore, they show that the texture representations based on the distribution of the local features like SURF are more effective than the combinations of linear filters [23] for texture classifications.

Table 3 : The mean classification rates of the four different methods for the two texture databases.

| <i>Methods</i> | <i>Brodatz</i> | <i>UIUCTex</i> |
|----------------------|-----------------------|------------------------|
| | 3 trainings per class | 10 trainings per class |
| Lazebnik et al. [29] | 88.15 | 92.61 |
| Mellor et al.[23] | 89.71 | 92.84 |
| Qin et al. [22] | 64.46 | 83.00 |
| Proposed | 90.84 | 93.60 |

5. Conclusions

In this paper, a novel texture classification method is proposed. The method uses SURF descriptors to represent the key points of a texture. All the key points of all training texture images are then clustered by a hierarchical k-means algorithm yielding a tree structure called a tree of keypoints. The tree is built to facilitate the evaluation of the tree matching kernel used by multi-class SVMs to classify a given texture. Results from experiments conducted on textures from two databases, Brodatz and UIUCTex, have shown that SURF descriptors are invariant to many kinds of geometric and photometric transformation such as scale, rotation, deformation and viewpoint changes, and can be used effectively with the tree matching kernel to achieve high classification rates on multi-class SVMs.

Acknowledgments

This research was partially funded by the National Institute of Development Administration (NIDA). We also would like to thank the School of Applied Statistics of NIDA for providing us supports on this research work.

References

- [1] A.C. Berg, T.L.Berg, and J. Malik, “Shape matching and object recognition using low distortion correspondence”, in Proceedings of the IEEE International Conference Computer Vision Pattern Recognition, Vol. 1, 2005, pp. 26-33.
- [2] A.C. Murillo, J.J. Guerrero and C. Sagües, “SURF features for efficient robot localization with omnidirectional images”, in IEEE International Conference on Robotics and Automation, Rome-Italy, 2007, pp. 3901-3907.
- [3] A.E. Johnson and M. Hebert, “Using spin images for efficient object recognition in clustered 3D scenes”, in IEEE Transaction Pattern Analysis Machine Intelligence, Vol.21, No.5, 1999, pp.433-449.
- [4] B. Caputo and L. Jie, “A Performance Evaluation of Exact and Approximate Match Kernels for Object Recognition”, in Electronic Letters on Computer Vision and Image Analysis, Vol. 8, No. 3, 2009, pp. 15-26.

- [5] B. Leibe, K. Mikolajczyk, and B. Schiele, "Effcient clustering and matching for object class recognition", in Proceedings of British Machine Vision Conference, 2006.
- [6] B. Scholkopf and A. Smola. Learning with Kernels. The MIT Press, Cambridge, MA, 2002.
- [7] C.H. Hsu and C.J.Lin, "A Practical Guide to Support Vector Classification", Department of Computer Science, National Taiwan University, Taipei, Taiwan, 2008.
- [8] D. Nister and H. Stewenius, "Scalable Recognition with a Vocabulary tree", in IEEE Conference on Computer Vision and Pattern Recognition. New York, 2006, pp. 2161-2168.
- [9] D.G. Lowe, "Distinctive image features from scale-invariant keypoints", in International Journal of Computer Vision, Vol. 60, No. 2, 2004, pp.91-110.
- [10] E. Christopher, "Notes on the open SURF Library", 2009.
- [11] F. Xu and Y.J. Zhang, "Evaluation and comparison of texture descriptors proposed in MPEG-7", in Journal of Visual Communication and Image Representation, Vol.17 (August), 2006, pp.701-716.
- [12] G. Csurka, C. Dance, L. Fan, J. Willamowsky and C. Bray, "Visual categorization with bags of keypoint", in Proceedings of the ECCV International Workshop on Statistical Learning in Computer Science, 2004.
- [13] H. Bay, B. Fasel, and L.V. Gool, "Interactive museum guide: Fast and robust recognition of museum objects", in The First International workshop on mobile vision, 2006.
- [14] H. Bay, T. Tuytelaars and L. V. Gool, "Surf: Speeded up robust features", in The ninth European Conference on Computer Vision, 2006, <http://www.vision.ee.ethz.ch/surf/>.
- [15] H. Zuo, W. Hu, O. Wu, Y. Chen, and G. Luo, "Detecting Image Spam Using Local Invariant Features and Pyramid Match Kernel", WWW 2009, April 20–24, 2009, pp. 1187-1188.
- [16] J. Li and N.M. Allinson, "A comprehensive review of current local features for computer vision", in Neurocomputing, Vol. 71, 2008, pp.1771– 1787.
- [17] J. Zhang, and T. Tan, "Brief review of invariant texture analysis methods", in Pattern Recognition. Vol 35, 2002, pp.735–747.
- [18] K. Grauman and T. Darrell, "Approximate Correspondences in High Dimensions", in Advances in Neural Information Proceeding System, 2006.
- [19] K. Grauman and T.Darrell, "The pyramid match kernel: Discriminative classification with sets of image features", in ICCV, 2005.
- [20] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors", in IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol.27, No 10, 2005, pp.1615–1630.
- [21] K. Mikolajczyk and J. Matas, "Improving Descriptors for Fast Tree Matching by Optimal Linear Projection", in iccv, 2007, pp. 1-8.
- [22] L. Qin , Q. Zheng, S. Jiang, Q. Huang and W. Gao, "Unsupervised texture classification: Automatically discover and classify texture patterns", in Image and Vision Computing, Vol.26 , 2008, pp.647-656.
- [23] M. Mellor, H. Byung-Woo and M. Brady, "Locally rotation, contrast, and scale invariant descriptors for texture analysis", in IEEE Transactions on pattern analysis and machine intelligence, 2007.
- [24] M. Muja and D. Lowe, "Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration", 2009.
- [25] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images", in International Journal of Computer Vision, Vol.62, 2005, pp.61-81.
- [26] M. Varma and A. Zisserman, "Classifying images of materials: achieving viewpoint and illumination independence", in Proceeding of the the European Conference on Computer Vision, Lecture notes in Computer Science, Vol.2352, No.3, 2002, pp.255-271.
- [27] M.Turtinen, "Learning and Recognizing Texture Characteristics using Local Binary Patterns", Ph.D. thesis, University of Oulu, 2007.
- [28] S. Bougħorbel, J.P. Tarel and N. Boujemaa, "The intermediate matching kernel for image local features", in International Joint Conference onNeural Networks, 2005, pp.889-894.
- [29] S. Lazebnik, C. Schmid and J. Ponce, "A Sparse Texture Representation using Local Affine Regions", in IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.27 (August), 2005, pp.1265-1278.
- [30] S. Lazebnik, C. Schmid and J. Ponce, "Affine-Invariant Local Descriptors and Neighborhood Statistics for Texture Recognition", in Proceedings of the Ninth IEEE International Conference on Computer Vision, 2003.
- [31] S. Lazebnik, "Local, Semi-local and Global Models for Texture, Object and Scene Recognition", Ph.D. thesis, University of Illinois at Urbana-Champaign, 2006.
- [32] S.B. Kotsiantis and P.E. Pintelas, "Recent Advances in Clustering: A Brief Survey", in WSEAS Transactions on Information Science, 2004.
- [33] S.J. Taylor and N. Cristianini, "Kernel Methods for Pattern Analysis. Cambridge University Press, 2004.
- [34] V. Vapnik and C.Cortes, "Support-Vector-Networks", in Machine Leaning, 20, 1995.
- [35] <http://www.ux.uis.no/~tranden/brodatz.html>

Somkid Sootitantawat is a PhD candidate at the School of Applied Statistics, National Institute of Development Administration, Bangkok, Thailand. She received a Master's Degree in Computer Science from the School of Applied Statistics, National Institute of Development Administration and completed her Bachelor's Degree in Education in General Science/Computer Education from Chulalongkorn University, Bangkok, Thailand. Her research interests are in Pattern Recognition and Data Mining.

Surapong Auwatanamongkol is an associate professor at the School of Applied Statistics, National Institute of Development Administration, Bangkok, Thailand. He received a Ph.D. degree in Computer Science from Southern Methodist University, USA, a Master's degree in Computer Science from Georgia Institute of Technology, USA, and a Bachelor's degree in Electrical Engineering from Chulalongkorn University, Bangkok, Thailand. His fields of research include Pattern Recognition, Data Mining, Evolutionary Computation, Parallel and Distributed Computing. He has published several papers in various international journals and conferences.

Performance Measurement of Some Mobile Ad Hoc Network Routing Protocols

Ahmed A. Radwan¹, Tarek M. Mahmoud² and Essam H. Houssein³

¹Computer Science Department, Faculty of Science, Minia University
El Minia, Egypt

²Computer Science Department, Faculty of Science, Minia University
El Minia, Egypt

³Computer Science Department, Faculty of Computers and Informatics, Benha University
Benha, Egypt

Abstract

A mobile ad hoc network (MANET) is a wireless network that uses multi-hop peer to peer routing. A user can move anytime in an ad hoc scenario and, as a result, such a network needs to have routing protocols which can adopt dynamically changing topology. To accomplish this, a number of ad hoc routing protocols have been proposed and implemented such as, Ad hoc On-Demand Distance Vector routing (AODV), Fisheye State Routing (FSR) and Location-Aided Routing (LAR). This paper compares the major characteristics of these protocols such as, routing messages overhead, throughput and end to end delay using a parallel discrete event-driven simulator, GloMoSim. The experimental results show that FSR protocol has low control overhead compared with AODV and LAR. Regarding the throughput, AODV has a high throughput compared with the other considered protocols. Considering the end to end delay, LAR protocol shows better performance over FSR and AODV protocols.

Keywords: Mobile Ad hoc networks (MANET), Ad hoc On-Demand Distance Vector routing (AODV), Fisheye State Routing (FSR) and Location-Aided Routing (LAR).

1. Introduction

Mobile Ad hoc Network (MANET) is a collection of wireless mobile nodes dynamically forming a temporary network without the use of any existing network infrastructure or centralized administration [3, 5 and 17]. In such a network, each mobile node operates not only as a host but also as a router, forwarding packets for other mobile nodes in the network that may not be within direct wireless transmission range of each other. Each node participates in an ad hoc routing protocol that allows it to discover “multi-hop” paths through the network to any other node. Some examples of the possible uses of MANET include students using laptop computers to participate in an interactive lecture, business associates sharing information during a meeting, soldiers relaying information for situational awareness on the battlefield [1], and emergency disaster relief personnel coordinating

efforts after a hurricane or earthquake. The traditional routing protocols may not be suitable for MANETs since the network topology usually changes with time. Accordingly, there are new challenges for routing protocols in MANETs. Many different protocols have been proposed to solve the routing problem in MANETs. These protocols are usually based on the graph model which implies that the mobile nodes are aware of only their connectivity with the neighbors and not the relative locations. Hence the network topology may be represented as a graph with mobile nodes occupying the vertices. The nodes between which connectivity exists are connected by an edge in the graph. Edges may be directed in case the network link is physically asymmetric. Protocols can also be defined based on the geographic model in which each node is aware of the geographical location of itself and other nodes. Protocols with such a facility, which may be provided by many mechanisms like GPS, are known as location aware protocols. Another important class of protocols is the one in which the whole networking region is divided into zones. The routing problem now is two fold consisting of inter and intra zone routing but is far easier to handle. Such a mechanism is known as zone routing [6, 9, 13 and 14].

Routing protocols may be classified into two types based on the way the route information is generated and maintained. Table Driven protocols attempt to maintain up to date route information for all the destination routes in each node of a network by maintaining routing tables [4, 12 and 16]. Each node is required to maintain these tables and also to propagate periodic updates to keep all the tables current. The need to maintain tables and the updates are overheads on the networking and hence different protocols implement different strategies to consolidate the number of routing tables required to be maintained and the method to broadcast network updates. Source Initiated or On Demand Routing makes away with the need for any tables by finding routes as per requirements [18 and 20].

Whenever a node requires a route to another node, it initiates a route discovery process in the network, which returns the routes back. The discovered routes are cached and maintained till the route is required or the destination becomes inaccessible.

In this paper, GloMoSim Simulator 2.03 version is used to simulate three ad hoc routing protocols, that is, AODV, FSR and LAR. The commonly performance metrics supported by GloMoSim for these protocols are evaluated. Since these protocols have different characteristics, the comparison of all performance differentials is not always possible. However, the following system parameters are utilized for comparative study on the protocols:

- Routing messages overhead,
- Average end to end delay,
- Throughput

The rest of the paper is organized as follows. In the following section, we briefly review about a categorization of the prominent ad hoc routing protocols and give a short introduction about the three routing protocols compared in this paper. In Section 3, we present the performance metrics of our simulation. Section 4 describes our simulation environment. Section 5 presents the result of simulation. We draw our conclusions in Section 6 followed by recommendations for future work in this regard.

2. Ad Hoc Network Routing Protocols Studied

In this section, we briefly review the AODV, FSR and LAR protocols studied in our simulations.

2.1 The Ad Hoc On-demand Distance Vector Routing (AODV)

The Ad Hoc On-demand Distance Vector Routing (AODV) protocol is a reactive routing protocol for mobile ad hoc networks [2 and 15]. As a reactive routing protocol, only routing information about the active paths is needed to maintain. In AODV, routing information is maintained in routing tables at nodes. Every mobile node keeps a next-hop routing table, which contains the destinations to which it currently has a route. A routing table entry expires if it has not been used or reactivated for a pre-specified expiration time. Moreover, AODV adopts the destination sequence number technique used by DSDV in an on-demand way [7].

2.2 Fisheye State Routing (FSR)

The Fisheye State Routing (FSR) is a proactive routing protocol based on Link State routing algorithm with effectively reduced overhead to maintain network topology information [8 and 11]. In proactive routing protocols, routing information to reach all the other nodes in a network is always maintained in the format of the routing table at every node. As indicated in its name, FSR utilizes a function similar to a fish eye. The eyes of fishes catch the pixels near the focal with high detail, and the detail decreases as the distance from the focal point

increases. Similar to fish eyes, FSR maintains the accurate distance and path quality information about the immediate neighboring nodes, and with the progressive detail as the distance increase.

2.3 Location-Aided Routing (LAR)

The Location-Aided Routing (LAR) is based on flooding algorithms. It attempts to reduce the routing overheads present in the traditional flooding algorithm by using location information. This protocol assumes that each node knows its location through a Global Positioning System (GPS). Two different LAR scheme were proposed in [10], the first scheme calculates a request zone which defines a boundary where the route request packets can travel to reach the required destination. The second method stores the coordinates of the destination in the route request packets. These packets can only travel in the direction as the relative distance to the destination becomes smaller as they travel from one hop to another. Both methods limit the control overhead transmitted through the network and hence conserve bandwidth. They will also determine the shortest path (in most cases) to the destination, since the route request packets travel away from the source and towards the destination. The disadvantage of this protocol is that each node is required to carry a GPS.

3. The performance Parameters

This section presents the performance parameters (metrics) used to evaluate the AODV, FSR and LAR protocols. The main performance parameters are routing message overhead, average end to end delay, and throughput.

3.1 Routing Message Overhead

Routing message overhead is calculated as the total number of control packets transmitted. The increase in the routing message overhead reduces the performance of the ad-hoc network as it consumes portions from the bandwidth available to transfer data between the nodes [18].

3.2 Average End to End Delay

A network's end-to-end delay is defined as the average time interval between the generation and successful delivery of data packets for all nodes in the network, during a given period of time. Packets that are discarded or lost are not included in the calculation of this metric [18].

3.3 Throughput

A network's end-to-end throughput is a measure of the network's successful transmission rate, and is usually defined as the number of data packets successfully delivered to their final destination per unit of time. However, to convert this metric to a measure of data throughput or to compare it to other networks, the network's packet size and the network's number of nodes also has to be known.

This paper therefore defines a network's end-to-end throughput as the number of data bytes successfully delivered to their final destination per unit of time, divided by the number of nodes in the network [18].

4. Simulation Environment

To compare the performance of the three routing protocols described in section (2), simulation experiments were performed. In this section, experiment modeling, design and key observations from our simulation experiments are described in that order.

Simulations were carried out with the GloMoSim library [19] which is widely used in the academic research. The GloMoSim library is a scalable simulator for wireless network and it is built using the parallel discrete-event simulation capability provided by PARSEC. The numbers of nodes used in the simulation scenarios are 100, 200, and 300, with rectangular area sizes 1500×1000 , 2000×1500 , and 3000×2000 m², respectively. The nodes placed randomly within the simulation area. The radio propagation range for each node is 376 meters and channel capacity is 2Mb/s. Each simulation is executed for 300 seconds of simulation time. IEEE 802.11 MAC protocol was used in the experiments for the MAC layer. The sources used for the simulations are CBR (constant bit rate) sources. Twenty data sessions with randomly selected sources and destinations are used in the simulations. Each source transmits data packets at 4 packets/sec rate with packet size 512 bytes until the simulation run ends.

The mobility model used is the random waypoint model [12 and 20]. In this model, a node selects a random destination within the terrain range and moves towards it at a speed between the pre-defined minimum and maximum speed. Once the node arrives at the destination, it stays for a pause time. After being stationary for the pause time, it randomly selects another destination and speed and then resumes movement. The minimum and the maximum speed for the simulations are 0 m/s and 10 m/s, respectively. Simulation runs done on variance pause time values from 0 to 300 second. The simulations have been done on a PC Pentium IV, 2GHz processor and 3GB RAM.

5. Simulation results

The following subsections represent the results of the simulation scenarios. The 100 nodes scenario results will be introduced in subsection (5.1). The 200 nodes scenario results will be introduced in subsection (5.2). The 300 nodes scenario results will be introduced in subsection (5.3).

5.1 Scenario Results with 100 Nodes

This section presents the simulation results for the 100 nodes network simulation scenario on a rectangular area 1500×1000 m².

5.1.1 Routing Message Overhead

Figure 1, shows the routing message overhead resulted from each of AODV, FSR and LAR routing protocols. As can be seen in this Figure, LAR has lower routing message overhead compared with the AODV and FSR. The pause time increases, the overhead resulted from FSR protocol tends to be zigzag. The overhead resulted from AODV protocol is increased between 0 and 60 sec and then decreases.

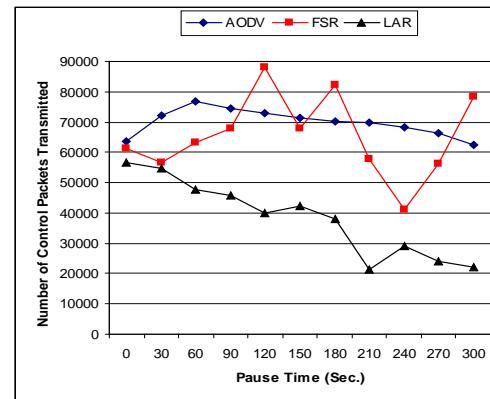


Fig. 1 Routing message overhead vs. pause time for 100 nodes

5.1.2 Average End to End Delay

Figure 2, illustrates the average end to end delay for the AODV, FSR and LAR routing protocols. The end to end delay of FSR protocol is closed to zero and increased to maximum value between pause time 200 and 300 sec. As can be seen in Figure 2 the end to end delay of the LAR protocol is decreased as the pause time is increased. For the AODV protocol, the delay is increased between pause time 0 and 100 sec and then decreased. The end to end delay of LAR protocol is better than AODV protocol.

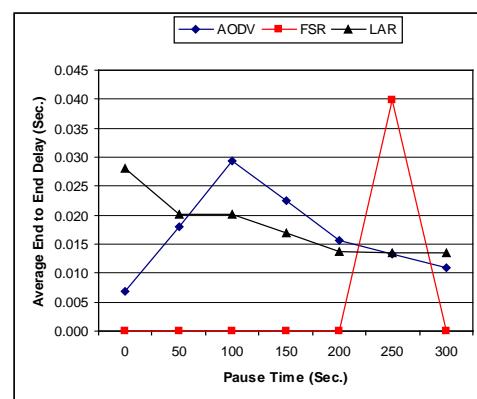


Fig. 2 Average End to End delay vs. pause time for 100 nodes

5.1.3 Throughput

Figure 3, demonstrates the throughput vs. pause time for the considered protocols. It is clear that AODV protocol has a good performance compared with both FSR and LAR.

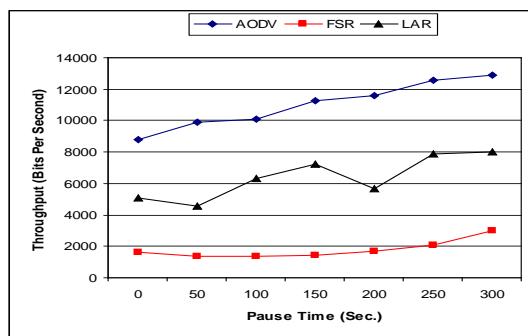


Fig. 3 Throughput vs. pause time for 100 nodes

5.2- 200 Nodes Scenario Results

This section presents the simulation results for the 200 nodes network simulation scenario on a rectangular area 2000 x 1500 m².

5.2.1 Routing Message Overhead

Figure 4, indicates the routing message overhead resulted from each of AODV, FSR and LAR routing protocols. The performance of these protocols is similar to the results obtained in the 100 nodes network simulation scenario. As shown in Figure 4, the overhead resulted from AODV protocol increases between 0 and 120 sec and then decreases. As in the 100 nodes network simulation scenario on a rectangular area 1500×1000 m 2 , the overhead resulted from FSR protocol tends to be zigzag. Here LAR protocol is more advantageous as it gives minimum overheads.

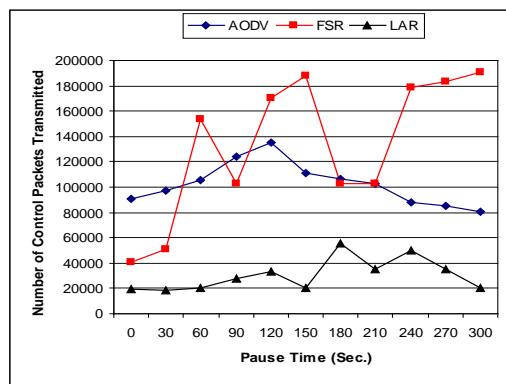


Fig. 4 Routing message overhead vs. pause time for 200 nodes

5.2.2 Average End to End Delay

Figure 5, depicts the average end to end delay for the AODV, FSR and LAR routing protocols. The end to end delay of FSR protocol is closed to zero and increased to maximum value between pause time 100 and 200 sec. The end to end delay of the LAR protocol is increased until the pause time 200 sec and then decreased. For the AODV protocol, the delay is increased between pause time 0 and 100 sec and then decreased.

5.2.3 Throughput

The throughput vs. pause time for the considered protocols is illustrated in Figure 6. It is clear that AODV protocol has a good performance compared with both FSR and LAR. The throughput of the LAR protocol decreases until pause time 100 and then increases.

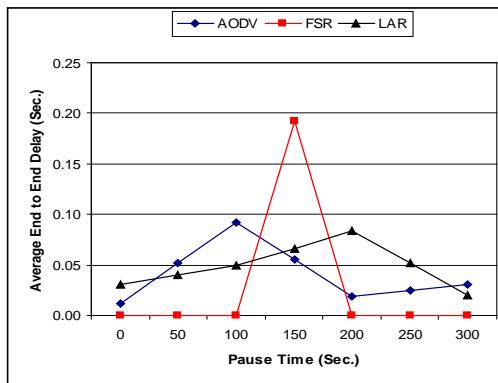


Fig. 5 Average End to End delay vs. pause time for 200 nodes

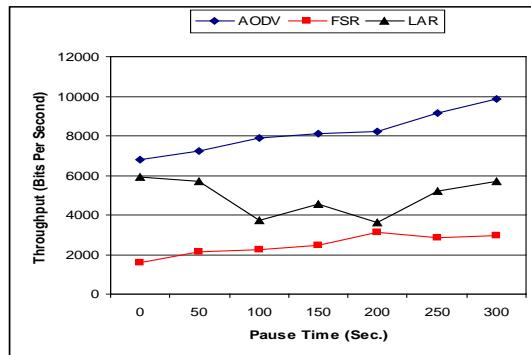


Fig. 6 Throughput vs. pause time for 200 nodes

5.3- 300 Nodes Scenario Results

This section presents the simulation results for the 300 nodes network simulation scenario on area 3000 x 2000 m².

5.3.1 Routing Message Overhead

Figure 7, illustrates the routing message overhead resulted from the three considered protocols. As the pause time increases, the number of control packets transmitted using FSR protocol increases. On the other hand the number of control packets transmitted using AODV protocol decreases. The overhead resulted from LAR protocol tends to be zigzag.

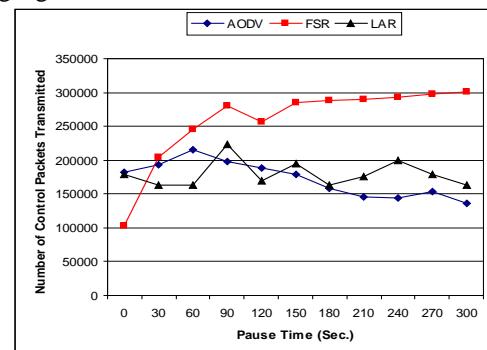


Fig. 7 Routing message overhead vs. pause time for mobicse.org

5.3.2 Average End to End Delay

Figure 8, demonstrates the average end to end delay for the AODV, FSR and LAR routing protocols. The end to end delay of FSR protocol is closed to zero all the simulation pause time. The end to end delay of the LAR protocol is increased until the pause time 250 sec and then decreased. For the AODV protocol, the delay is increased between pause time 0 and 100 sec and then decreased.

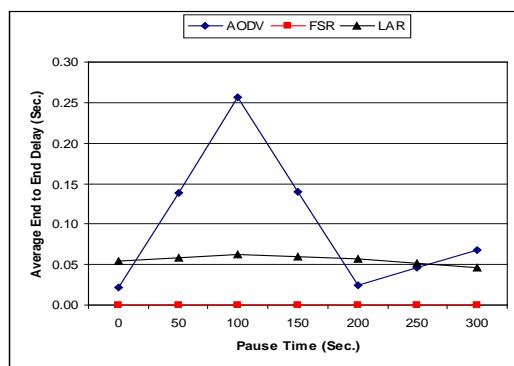


Fig. 8 Average End to End delay vs. pause time for 300 nodes

5.3.3 Throughput

Figure 9, present the throughput vs. pause time for the considered protocols. It is clear that AODV protocol has a good performance compared with both FSR and LAR.

Fig. 9 Throughput vs. pause time for 300 nodes

6. Conclusions and Future Works

6.1 Conclusions and Performance Summery

This paper attempts to determine how AODV, FSR, and LAR protocols perform under increased loads. We tested these protocols for three different scenarios (100, 200, and 300 nodes) on different rectangular areas (1500×1000 , 2000×1500 , and 3000×2000 m 2). The performance evaluation of these protocols is based on the well known GloMoSim simulator. The simulation characteristics used to evaluate the performance of these protocols are routing message overhead, end to end delay and throughput.

– Routing Message Overhead

The results of the simulation show that AODV and FSR impose a huge routing overhead compared with LAR, as shown in Figures 1, 4 and 7. This is not surprising due to, in proactive routing protocol all nodes are active and each node discovers route to other nodes in the network before the actual communication request. This leads to less time delay of route discovery during communication request, however the overhead cost is too high in this case. Moreover, as the number of nodes increases, the routing overhead clearly increases. For the FSR protocol, this problem caused by the rapid changes in network topology might overwhelm the network with control messages and flood large number route finding packets instead of buffering data packets for new route to be found and since more table updates are being sent.

– Average End-to-End Delay

From Figures 2, 5 and 8, It is clear that AODV gives average end to end delay higher than the other two protocols with high mobility due to its single path nature and inefficient manner to handle route failure. This is because when a node receives a route request for which it has the answer in its routing table, it immediately replies with the route rather than forwarding it to the destination. The source can now start to communicate with the destination. Moreover, as the number of nodes increases, the FSR protocol trends to be zero delay as shown in Figure 8. The reason for this due to the behavior of the FSR protocol, because the routes are available the moment they are needed. Also, each node consistently maintains an up-to-date route to every other node in the network, a source can simply check its routing table when it has data packets to send to some destination and begin packet transmission so, no delay occur.

– Throughput

This metric which we call the ratio of delivered packets is an important as it describes the loss rate that will be seen by the transport protocols, which in turn affects the maximum throughput that the network can support. Figures 3, 6 and 9 shows the number of bits received per second. For AODV, FSR and LAR packet delivery ratio is independent of offered traffic load. In case of AODV protocol when numbers of nodes increases, initially throughput increases as number of routes are available compared to FSR and LAR protocols. Regretfully FSR was not up to the task and it performed poorly throughout all the simulation sequences because increasing the overhead reduces the throughput.

6.2 Future Work

Recommendations for future studies that can improve the reliability of this kind of work include the following:

- This study included only one mobility model throughout the simulation. In the future work we plan to apply another mobility model may be affect on the measure performance parameters.

- protocols; routing protocols can be studied on different types of data traffic (application layer protocols) like http, ftp, telnet, and real time audio/video transmissions.
- There are several MAC protocols to be used in the simulation such as CSMA, MACA and IEEE 802.11; in this paper we applied IEEE 802.11 only. Different types of MAC may give different results for ad hoc routing protocols.
 - Finally, since we used GloMoSim, our simulation was confined to three protocols, AODV, LAR, and FSR. Additional ad hoc network protocols, such as DSDV, TORA and so on could be added in GloMoSim for comprehensive performance evaluation.

7. References

- [1] A. Rahman, S. Islam, A. Talevski, "Performance Measurement of Various Routing Protocols in Ad-hoc network", Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I, IMECS 2009, March 18 - 20, , Hong Kong, 2009.
- [2] C. Perkins, E. Belding-Royer, S. Das, "Ad hoc On-Demand Distance Vector (AODV) Routing", Network Working Group, IETF RFC, RFC 3561, July 2003.
- [3] D. B. Johnson, and D. A. Maltz, "Dynamic Source Routing in Ad-Hoc Wireless Networks", Mobile Computing, T. Imielinski and H. Korth, Eds., Kluwer, 1996, pp. 153–81.
- [4] D. Lang, "On the Evaluation and Classification of Routing Protocols for Mobile Ad Hoc Networks ", PhD thesis, 2006.
- [5] E. Ahvar, and M. Fathy, "Performance Evaluation of Routing Protocols For High Density Ad Hoc Networks based on Energy Consumption by GlomoSim Simulator", World Academy of Science, Engineering and Technology 29, 2007.
- [6] F. Yongsheng, W. Xinyu and L. Shanping, "Performance comparison and analysis of routing strategies in Mobile ad hoc networks", International Conference on Computer Science and Software Engineering, 2008.
- [7] G. He, "Destination-Sequenced Distance Vector (DSDV) Protocol", Networking Laboratory, Helsinki University of Technology, 2002.
- [8] G. Pei, M. Gerla, and T. Chen, "Fisheye State Routing in Mobile Ad Hoc Networks", ICDCS Workshop on Wireless Networks and Mobile Computing, 2000.
- [9] J. Hoebeke, I. Moerman, B. Dhoedt and P. Demeester, "An Overview of Mobile Ad Hoc Networks: Applications and Challenges", Department of Information Technology (INTEC) Ghent University – IMEC vzw, 2010.
- [10] L. Bla'zevi'c, J. Le Boudec and S. Giordano, "A Location Based Routing Method for Irregular Mobile Networks", EPFL-IC Report Number May 2003.
- [11] "Fisheye state routing protocol (FSR) for ad hoc networks", Internet Draft, draft-ietf-manet-aodv-03.txt, work in progress, 2002.
- [12] M. K. Kumar and R. S. Rajesh, "Performance Analysis of MANET Routing Protocols in Different Mobility Models", IJCSNS International Journal of Computer Science and Network 22 Security, VOL.9 No.2, Feb., 2009.
- [13] P. Van Mieghem, "Data Communications Networking", Techne Press, Amsterdam, The Netherlands, 2006.
- [14] S. Ahmed and M. S. Alam, "Performance Evaluation of Important Ad Hoc Network Protocols", EURASIP Journal on Wireless Communications and Networking, Volume 2006, Article ID 78645, Pages 1–11, 2006.
- [15] S. Das, C. Perkins, E. Royer, "Ad hoc on demand distance vector (AODV) routing", Internet Draft, draft-ietf-manetaodv- 11.txt, work in progress, 2002.
- [16] S. Rahayu Abdul Aziz, N. Adora Endut, S. Abdullah and M. Norazman Mior Daud, "Performance Evaluation of AODV, DSR, and DYMO Routing Protocol in MANET", Conference on scientific & Social Research (CSSR), March 2009.
- [17] S. S. Tyagi and R. K. Chauhan, "Performance Analysis of Proactive and Reactive Routing Protocols for Ad hoc Networks", International Journal of Computer Applications (0975 – 8887) Volume 1 – No. 14, 2010.
- [18] T. H. Abd El-Nabi, "Modeling and Simulation of a Routing Protocol for Ad Hoc Networks Combining Queuing Network Anakysis and Ant Colony Algorithms", Phd Thesis, April 2005.
- [19] The official GloMoSim Website, <http://pcl.cs.ucla.edu/projects/glomosim/> 2010.
- [20] V.C. Patil1, R. V. Biradar, R. R. Mudholkar and S. R. Sawant, "On-Demand Multipath Routing Protocols for Mobile Ad Hoc Networks Issues and Comparison", International Journal of Wireless Communication and Simulation Volume 2 Number 1, pp. 21–38, 2010.

Ahmed A. A. Radwan the Dean of the Faculty of Computers and Information, Minia University, Egypt. He received his Ph. D. degree from Liverpool University, England in Computer Science. His main areas of interests are parallel processing, multicast communication, Algorithm design and analysis, Optimal Routing Algorithms for Computer Networks, Parallel Systems, Graph Drawing and Mobile Networks.

Tarek M. Mahmoud received his Ph.D. degree in engineering (computer science) from Bremen University (Germany) in 1997. Since 1997, he has been with the Department of computer science of Minia University. Currently, he is an associate professor there. His research interests include wired and wireless networks, pattern recognition, combinatorial optimization, metaheuristics algorithms, web and semantic mining.

Essam H. Houssein received his M. Sc. degree in CS from the CS Department at Minia University. His research interests included routing in wireless networks, Ad-Hoc networks, and metaheuristics algorithms. He currently Lecturer Assistant, CS Department, Faculty of Computers & Informatics, Benha University to teaching and assisting in the delivery of academic education courses for undergraduates.

Evolution of Computer Virus Concealment and Anti-Virus Techniques: A Short Survey

Babak Bashari Rad¹, Maslin Masrom² and Suhaimi Ibrahim³

¹ Faculty of Computer Science and Information System, University Technology Malaysia
Skudai, 81310 Johor, Malaysia

² Razak School of Engineering and Advanced Technology, University Technology Malaysia
Kuala Lumpur, 54100 Selangor, Malaysia

³ Advanced Informatics School, University Technology Malaysia
Kuala Lumpur, 54100 Selangor, Malaysia

Abstract

This paper presents a general overview on evolution of concealment methods in computer viruses and defensive techniques employed by anti-virus products. In order to stay far from the anti-virus scanners, computer viruses gradually improve their codes to make them invisible. On the other hand, anti-virus technologies continually follow the virus tricks and methodologies to overcome their threats. In this process, anti-virus experts design and develop new methodologies to make them stronger, more and more, every day. The purpose of this paper is to review these methodologies and outline their strengths and weaknesses to encourage those are interested in more investigation on these areas.

Keywords: Computer Virus, Computer Antivirus, Evolution of Computer Viruses, Antivirus Techniques, Virus Concealment.

1. Introduction

Since the first days of appearance of early malwares, there is a big contest between virus creators and anti-virus experts and it is becoming more complicated every day, and will continue afterward. At the same time as anti-virus softwares are advancing their methods, on the other hand, the virus writers are seeking for new tactics to overcome them. They utilize various techniques to put their products out of sight of the scanners, because if antivirus programs can easily find their viruses, they cannot sufficiently spread far in the wild. Hence, virus authors always struggle to create new code evolution tactics to beat against the detectors. Accordingly, virus techniques grew increasingly throughout all the years, from plainest methods to some more advanced strategies.

Although, the anti-virus specialists generally follow the new techniques used in advanced malwares and attempt to overcome them, however, all the new defense techniques are not sufficient and there is an extremely necessity for more researches. The most important thing is to analyze and understand all the previous methods, well.

In this paper, firstly, we give a short description on evolution of computer viruses and their classifications in aspect of concealment tactics. Then we survey the most common scanning and detection methods used in anti-virus software. Anti-virus software employed different methodologies in analyzing, scanning, and detecting viruses to provide sufficient safety for computer systems. In the next section, we present a comparison table that shows these detection methods and their features. It helps us to understand the advantages and disadvantages of each method and compare them. In the last section, we terminate with a conclusion and some recommendations.

2. Evolutionary Concealment

Computer malwares can be classified according to their different characteristics in several various manners, such as classification by target or classification by infection mechanism. One of these classification types is according to concealment techniques employed.

2.1 Encrypted Viruses

Encryption is practically the most primitive approach to take cover the operation of the virus code [1]. The ultimate aim of encrypted viruses is change of the virus body binary

codes with some encryption algorithms to hide it from simple view and make it more difficult to analyze and detect [2]. The first encrypting virus, CASCADE, appeared in 1988 [3].

Normally, encrypted viruses are made of two key parts: the encrypted body of the virus, and a small decryption code piece [4]. When the infected program code gets to run, firstly, the decryption loop executes and decrypts the main body of the virus. Then, it moves the control to the virus body. In some viruses, decryption loop performs something more, in addition to its main task. For instance, it may calculate the checksum to make sure that the virus code is not tampered, but as a general principle, the decryptor should be created as small as possible to avoid the anti-virus software, which is trying to exploit the decryptor loop's string pattern for scanning purpose.

Encryption hides the virus body from those who like to view the virus code or tamper the infected files using code viewers or hexadecimal editors [5]. However, virus programmers use the encryption for some reasons. Four of the major motivations as described by Skulason [4] are:

- 1. To avoid static code analysis:** Some programs try to analyze code automatically and generate warning if suspect code is found. Encryption is used to disguise suspicious codes and prevent static analysis.
- 2. To delay the process of inspection:** It can make the analysis process a bit more difficult and time-consuming, however it usually can increase the time of process only a few minutes.
- 3. To prevent tampering:** Many new variants of a virus can be produced with a minor change in the original virus code. Encryption makes it difficult to change the virus by non-experts.
- 4. To escape from detection:** an encrypted virus cannot be detected through simple string matching before decryption, because only decryptor loop has identical string in all variants. Hence, signature for an encrypted virus is limited and must be selected precisely.

2.2 Oligomorphic Virus

Although virus creators attempted to conceal the first generation of viruses with encryption methods, the decryptor loops were remained constantly in new infected files, so anti-virus software normally had no trouble with such virus that was inspected and for which a signature string was obtained. To overcome this vulnerability, virus writers employed several techniques to create a mutated body for decryptors. These efforts caused the invention of new type of concealment viruses, named as oligomorphic viruses.

Oligomorphic viruses are willing to substitute the decryptor code in new offspring. The easiest method to apply this idea is to provide a set of different decryptor loops rather than one. Signature based detection depending on byte pattern of decryptor, though it is a achievable solution, but it is not a practical way [1]. Therefore, oligomorphic viruses make the detection process more difficult for signature based scanning engines.

2.3 Polymorphic Virus

The most usual approach developed in anti-virus softwares and tools to identify the viruses and malwares is signature-based scanning [6]. It makes use of small strings, named as signatures, results of manual analysis of viral codes. A signature must only be a sign of a specific virus and not the other viruses and normal programs. Accordingly, a virus would be discovered, if the virus related signatures were found. To avoid this detection, virus can change some instructions in new generation and cheat the signature scanning. Polymorphic viruses exploit this concept. When the virus decides to infect a new victim, it modifies some pieces of its body to look dissimilar. As encryption and oligomorphism, scheme of polymorphism is to divide the code into two sections, the first part is a code decryptor, which its function is decryption of the second part and passes the execution control to decrypted code. Then, during the execution of this second part, a new different decryptor will be created, which encrypts itself and links both divisions to construct a new copy of the virus [7, 8].

In fact, polymorphism is a newer and progressive variety of oligomorphism. Concerning of encryption, polymorphic virus, oligomorphic and encrypted viruses are similar, but the exception is the polymorphic virus has capability to create infinite new decryptors [2, 9]. Polymorphic virus exploits mutation techniques to change the decryptor code. Furthermore, each new decryptor may use several encryption techniques to encrypt the constant virus body, as well.

2.4 Metamorphic Virus

Virus writers like to make the lifetime of their produced viruses longer, so they constantly challenge to make the detection as more difficult as possible for antivirus specialists. They have to spend a plenty of time to produce a new polymorphic virus that it may not be able to spread out broadly, but an anti-virus expert may handle the detection of such a virus in a short time [10].

Even for the most complicated polymorphic viruses, after code be emulated sufficiently, the original code will

become visible and can be detected by a simple string signature scanning [11].

Peter Szor quoted in [9], the shortest definition of the metamorphic virus, which defined by Igor Muttik, is “Metamorphics are body-polymorphics.” Because metamorphic viruses are not encrypted, they do not require decryptor. Metamorphic virus is similar to polymorphic virus in aspect of making use of an obfuscation engine. Metamorphic virus mutate all of its body, rather it changes the code of decryption loop. All possible techniques applicable by polymorphic virus to produce new decryptor can be used by a metamorphic virus on whole virus code to create a new instance.

3. Anti-Virus Techniques

3.1 First-Generation Scanners

Scanners of first-generation employed not complicated techniques in order to find known computer viruses. Earliest scanners typically looked for certain patterns or sequences of bytes called string signatures.

Once a virus is detected, it can be analyzed precisely and a unique sequence of bytes extracted from the virus code. This string often called signature of the virus and is stored in the anti-virus scanner database. It must be selected such that not likely is appeared in benign programs or other viruses, optimistically. This technique uses this signature to detect the previously analyzed virus. It searches the files to find signatures of the viruses. It is one of the most basic and simplest methods employed by antivirus scanners. Some examples of virus signature strings, which are published in Virus Bulletin [12], are given in Table 1.

Table 1: Examples of viruses string signature

| Virus Name | String Pattern (Signature) |
|-------------------|---|
| Accom.128 0 | 89C3 B440 8A2E 2004 8A0E 2104 BA00 05CD 21E8 D500 BF50 04CD |
| Die.448 | B440 B9E8 0133 D2CD 2172 1126 8955 15B4 40B9 0500 BA5A 01CD |
| Xany.979 | 8B96 0906 B000 E85C FF8B D5B9 D303 E864 FFC6 8602 0401 F8C3 |

The anti-virus engine scans the binary code of files to find these strings; if it encounters with a known pattern, it alerts detection of the matching virus. Normally, a string contain of sixteen unique bytes is properly adequate in size to distinguish a 16-bit viral code with no false positive. However, for 32-bit malicious codes to be recognized accurately, more enlarged strings are required, specifically

if the malware is created through high level programming languages [1].

3.1.1 Special Cases in String Scanning

Sometimes signature string scanning need some special conditions in bytes comparison process. Some of most useful exceptional cases in string scanning are [1]:

(i) Wildcards: Use of wildcards allows excluding some constant byte values or ranges of values from comparison. For example, in the following string, bytes identified by the “?” symbol are not considered in matching. The wildcard %2 indicates the scanner attempt to find the following byte value in the next two places.

8B96 09?? B000 E85C %2 FF8B D5B9 D303 E864

Some of earlier encrypted and polymorphic viruses are detectable by wildcards. Furthermore, W32/Regswap metamorphic virus could be detected using this method [9].

(ii) Mismatches: It allows negligible values for non-specified quantity of bytes inside a string, regardless of their position. For example, the string 0A 32 B3 17 80 F6 24 with mismatch 2, is compatible with these strings:

13 0A 32 01 17 80 4D 24 72
 0A D9 1E 17 80 F6 24 AA 92

(iii) Generic Degree: when a virus has more than one variant, the variants are analyzed to extract one unique string that indicates all of them. This kind of string scanning uses this common pattern to find any previously identified variants of a virus family. It often exploits wildcards and mismatches, as well, to cover several different patterns of virus family variants.

3.1.2 Bookmarks

Use of Bookmarks is a simple way to ensure a more reliable detection and decrease the risk of false positive. For example, the number of bytes located between the beginning of the virus code and the first byte of the signature can be use as a suitable bookmark. Choosing a reliable bookmark is virus host specific. For example, for boot viruses, appropriate bookmarks may show addresses of boot sectors placed on the disk. In the file-infecting viruses, a proper bookmark can point to the offset of original program header. In addition, the length of the virus could be a very helpful bookmark.

3.1.3 Speed up Techniques

Almost all anti-virus scanners overspend most of the search time matching input data with previously discovered virus signatures. Normally, scanners employ various types of multi signature string comparing algorithm. In 2005, more than 100,000 virus signatures was known and is increasing continuously [13]. Therefore, the algorithms need to be performed as faster as possible. There are several techniques to make the string scanning faster. Some of most common methods to speed up the searching algorithm are:

Hashing: Generally, hashing techniques are commonly employed in searching algorithms as data structure to make the access to elements based on nonnumeric or large value keys faster and improve the speed of the process, generally. In anti-virus scanners, hashing is exploited in order to decrease the number of searching strings within the file. The methods may use 1 byte or 16-bit or 32-bit words of the scan strings to produce a hash value, which is the index in the hash table [14].

Top-and-Tail Scanning: Because virus codes are sited usually at the beginning or end of the victim files, scanning only the first and the last parts, instead of whole file is a useful idea to raise the speed of signature detection procedure more. This procedure is known as top-and-tail scanning. It reduces the number of disk accesses and optimizes scanning speed. However, since the scanners will search only in some specific areas, it may produce false negative results and diminish the accuracy and reliability [14].

Entry-Point and Fixed-Point scanning: These techniques also help the scanning engines to execute more rapidly. They use the concept of the program execution entry-point, which is achievable by the headers of executable files. Because viruses usually seek entry-point of the file as a target, search can be started from this point. In order to keep the execution of a file in normal manner, the virus has to get the executing control from the original start point and pass it again to the infected file original entry point after it terminates its code subroutine.

Fixed-point scanning is useful when there is no sufficient helpful string in the entry point. The scanner firstly specifies an initial location M. Then it tries to find every string that is compatible with signature, at positions M + X. These scanners reduce considerably the number of Input/Output access on disk and speed up the algorithm running process [1].

3.2 Second-Generation Scanners

The second-generation scanners started to develop, when the simple pattern scanning techniques lost their efficiency for detecting newer and more complicated viruses, appropriately. In addition, this generation of scanners introduced exact and almost exact recognition that caused the antivirus scanners became more trustable.

3.2.1 Smart Scanning

Smart scanning refers to a defense optimizing method for the newer generation of viruses, which try to conceal their code within a sequence of worthless instructions such as no operation NOP instructions.

When virus-mutating kits started to develop, simple signature-based scanning was not so effectual way because these pre-supplied kits could produce viruses much different visually from their original appearance. The mutation prepared tools can insert junk instructions, which have no effect on the execution process, among the program source instructions.

Smart scanning skip junk instructions, like NOPs, and do not consider them as the virus signature bytes. In addition, to improve the detection possibility of variants of a virus, a region of the virus body is chosen which does not include any addresses of data or other subroutines. Furthermore, smart scanning is utilized for detection of macro viruses that are written in text formats. It can ignore some characters employed to transform the appearance of the virus code, such as Space and TAB characters, and consequently improve the detection procedure quality, as a result.

3.2.2 Skeleton Detection

Skeleton detection is especially effective in order to detection of macro viruses. It does not utilize strings or checksums for detection purpose [14].

Eugene Kaspersky, Russian virus researcher and founder of the Kaspersky Anti-Virus, invented this technique and presented it for the first time. It reduces the searching zone inside a target file by removing each instruction that does not probably belong to the virus code before the scanning procedure starts. Firstly, the procedure parses statements of the macro virus one-by-one and removes any unimportant statements and all blanks gaps. Hence, the skeleton of the codes will be remained containing of only fundamental macro code, which the scanner exploit it to detect the virus [1].

3.2.3 Nearly Exact Identification

The purpose of nearly exact identification is more accurately detection of the viruses. One common method is to employ two strings as the signature of the virus, rather than only one. The virus is nearly exact identified, if both strings are existed in the file. Therefore, it makes disinfection process more reliable and risk-free, and ensures that the detected virus is not probable to be an unverified alternative of the primary version of the virus that maybe requires non-similar disinfection manner. Combination with bookmarks makes this technique more dependable.

Exploitation of a checksum range chosen from a virus code is also an alteration of nearly exact identification method that computes a checksum of the byte values in a specific area of the virus body. It brings about better accuracy, because a larger section of the virus body can be selected, with no need to overload the antivirus database.

In addition, it is not required to employ search strings, in order to implement nearly exact identification. In Kaspersky anti-virus algorithm, its creator, Eugene Kaspersky, does not make use of signature strings, instead employs two cryptographic checksums. These checksums are calculated at two specific locations, with given sizes inside the object.

3.2.4 Exact Identification

The exact identification technique utilizes non-variable bytes in the virus code as many as required to find a checksum of all bytes in the virus program, which contains constant value. The variable bytes of the virus body are ignored and a map of every constant byte is produced. This is the only method, which can promise an accurate detection of virus variants. It is often used as combination with the techniques of the first generation scanners. Exact identification method can discriminate exactly among various types of a virus, as well.

Even though it has many profits, but implementation of this technique make the scanners slow, slightly. In addition, really it is not easy to implement it for the outsized computer viruses.

3.2.5 Heuristics Analysis

The heuristics analysis is a useful method for detection of new unknown malwares [15]. It is especially helpful for detection of macro viruses too. It can be so worthwhile for binary viruses, as well, but it may extremely produce false positive output that is a major drawback of scanners [16].

Users cannot trust and will not purchase such anti-virus software that frequently produces tremendously false positives.

However, there are many situations, where a heuristic analyzer can be very valuable, and detect variants of a known virus family, as well. Heuristic analysis can be classified as two categories: static or dynamic [17]. Static heuristic is founded on the analysis of file structure and the code organization of the virus. While the static heuristic scanner is based on plain signs and code analysis to recognize the behavior of programs, the dynamic heuristic scanner performs CPU emulation of virus code, and tries to gather its information.

Some examples of heuristic flags are as following items, which express specific structural problems, may not be included in benign Portable Executables that are compiled using a 32-bit compiler [1]:

- Possible Gap between Sections
- Code Execution Starts in the Last Section
- Suspicious Section Characteristics
- Suspicious Code Section Name
- Virtual Size is Incorrect in Header of PE
- Multiple PE Headers
- Suspicious Imports from KERNEL32.DLL by Ordinal
- Suspicious Code Redirection

3.3 Virus-specific Detection

Sometimes the general virus detection algorithm may not be able to deal with a particular virus. In such conditions, a virus specific detection algorithm must be developed to carry out detection procedure. Actually, this kind of detection is not a regular method, but it denotes any special method that is specifically designed for a given particular virus. This approach is also called algorithmic scanning, but because it can be misleading [1], we use virus-specific detection term instead of algorithmic scanning.

This technique may bring about many problems such as portability of the scanner on different platforms and stability of the code. To overcome these problems, virus-scanning languages have been developed that in their plainest form, seeking and reading operations in scanned objects are allowed.

3.3.1 Filtering

This technique is used to optimize the performance of anti-virus engine regarding of scanning speed. It is especially useful in virus-specific detections because those are very time-consuming and high complexity in performance.

As a virus normally infects a particular or a set of known objects, signatures can be classified according to the infection type, such as .COM and .EXE files, boot sector, scripts, or macros, and so on. For example, executable viruses infect only programs such as .EXE and .COM, which are executable, macro viruses only attack to files or documents that can perform macro statements, and boot viruses place on the boot areas of disks. Through this exclusion action, when a specific file is searched for scan purpose, only the signatures relevant to its category are checked to keep scanning time down.

3.3.2 Static Decryptor Detection

As mentioned above, several types of viruses encrypt their body to prevent string scanning detection. In encrypted virus, the number of bytes, which can be used for string matching by scanners, is less. It makes trouble for string signature scanning engines. Therefore, anti-virus products have to use decryptor detection specific to a particular virus, which is not a very high quality method since it may produce many false negatives and false positives. In addition, because the virus body will not be decrypted during scanning, this technique cannot promise for a full disinfection.

However, the method can be a bit faster when it is employed together with an efficient filtering. It can also be employed to find other kinds of encrypted virus, such as oligomorphic or polymorphic viruses.

3.3.3 X-RAY Scanning

The X-RAY scanning method is also a virus-specific approach that is used to detect viruses of encrypted category, as well. X-ray scanning attacks the encryption of the virus rather than searching for the decryptor. It works based on a previously identified plaintext of the virus, and applies all encryption methods singly on special parts of files, such as top or tail of the file or supposed entry-point, to find the given plain text in decrypted virus body. X-raying exploits weaknesses of the virus encryption algorithm [18]. This scanning method is able to find advanced polymorphic viruses, as well [1].

The following example from [18], explain a simple X-raying. One of the most common encryption functions usually used in viruses is XOR operation. Each byte of the encoded text is resulted by applying XOR function on a byte of the plain text with a fixed 8-bit value between 0 and $0xFF$, which is called the encryption key. For instance, consider a plain text $T = E8\ 00\ 00\ 5D$ is encrypted with XOR operator and encryption key $k = 0x99$. After performing encryption, T will be converted as:

$$S = 71\ 99\ 99\ C4$$

We can decide whether S is an encrypted form of T or not, in two stages. Firstly, we can find the value of k , encryption key, a simple calculation using value of the first byte of S . In this example, if we assume that $71 = E8 \text{ XOR } k$, then we can infer that $k = 0x99$. Secondly, we can check and verify whether the remaining bytes of the cipher text S can be decoded correctly using the assumed key k or not.

The weakness of this method is that it is very time-consuming when the start of the virus is not placed at a fixed location, so the encryption methods have to be applied on a large section of the file. The considerable advantage of this scanning is that it decrypts completely the virus body, and consequently makes disinfection possible, even the necessary information for removing is in encrypted form.

3.4 Code Emulation

This is one of the strongest detection techniques. It simulates the computer central processor, main memory, storage resources and some necessary functions of operating system by a virtual machine to run the malware virtually and investigate its behavior and performance. The malicious code does not execute on actual machine and it is controlled by the virtual machine precisely, therefore there is no risk for unintentionally propagation of malware.

The emulator imitates instructions of the machine by simulating CPU registers and flags, virtually. It resembles the execution of programs and detection procedure analyzes all instructions, individually.

For polymorphic viruses or other types of encrypted codes, after a given quantity of iterations or after a pre-defined stop situation, the scanner checks the contents of memory of the virtual machine. After sufficient iterations, polymorphic virus will decrypt its encrypted body and the real code will be revealed in the virtual memory. Scanner may use the following methods to choose when it breaks off the emulation loop: *Stopping with break points*, *Tracking of decryptor using profiles*, and *Tracking of active instructions*. When the emulation terminates, the virus will be checked by using string pattern matching or other scanning techniques [1].

Veldman in [19] called more generally this method as Generic Detection, in the case of any kind of encrypted malwares. He describes it as a way to decrypt an encrypted virus. Essentially, a generic detection consists of four parts: *processor emulator*, *memory emulator*, *system emulator*, and *decision mechanism* [2, 19].

Some more intelligent malwares alter their behavior or does not allow to be executed at all, if they perceive that there is an emulator. More about emulators and methods used to detect and attack emulators can be found in [20, 21, 22].

3.4.1 Dynamic Decryptor Detection

This is an attempt to detect the decryptor via emulation of the code. Actually, it is a method made of joining static decryptor detection and code emulation. It is helpful when the decryption loop is very long and time-consuming and code emulation merely is not suitable [1].

For example, it may identify the probable entry-point of the virus. Then, during the process, a specific algorithmic detection can examine the memory of virtual machines to find which areas have been modified. If it discovered any not reliable changes, extra scanning can verify the executed instructions and profile them, so the fundamental instructions set of decryptor loop can be recognized. Later this set can be exploited in order to detection of the virus. However, for the purpose of a perfect disinfection, because the complete decryption of the virus is required, the emulator has to simulate and execute the virus for a protracted time; accordingly, this procedure is not a practical approach [1].

For detecting more complicated polymorphic viruses, dynamic technique can be used, which employs code optimization procedure to make the decryptor routine smaller and transform it to a limited essential set of instructions by eliminating the dead code and non essential or garbage instructions, like NOPs and ineffective jumps that have no result. It helps to make emulation process faster and gives a signature for polymorphic decryptor.

4. Comparison

Table 2 summarized more common virus detection methods, which are explained above. Some more useful properties of detection methods are given in the table for a brief comparison. Symbol ✓ means the method can support the property or may affect on the property positively. Actually, symbol ✓ dedicates an advantage for the method, while symbol ✗ shows a weakness of the method. In scanning speed column, ✓ denotes that the method can improve the scanning speed and reduce the time complexity.

For example, from the table it can be seen that hashing techniques in first-generation scanners can improve the

scanning speed and supports complete disinfection of the infected host, but it cannot used for detection of variants of a virus family, or unknown viruses or macro viruses. It has no effects on the false negative or false positive alarm, as well, in comparison to simple string signature scanning.

5 Conclusion and Future Recommendations

In this paper, though we try to review all most conventional antivirus techniques, but not all of them can be covered in a short survey.

Although the anti-virus software attempt to become updated and overcome the malwares threats, however we have to accept that virus authors are one step more ahead, because they decide how to attack first and anti-virus technologies have to only defense against their attacks. Therefore, computer virology area needs more researches and investigation to be able to guess the future coming threats.

There are many weaknesses in both viruses and anti-virus technologies, which must be studied and known well. Viruses usually look for the Achilles' heels in the defense system and attempt to attack them. Some major problems in detection methods are:

- 1 Most of detection methods are not powerful against evolutionary advanced or new viruses
- 2 Scanning process usually takes a considerable amount of time to search a system or networks for the patterns.
- 3 An anti-virus and its virus database need to be updated continually and extremely, otherwise it cannot be reliable.

So, interested researchers on the area of computer virology and anti-virus technologies are strictly recommended to work on these most important vulnerabilities.

Table 2: Comparison table of virus detection methods according to their features

| | | virus family detection | | | | | | | macro viruses | metamorphic viruses | encrypted/polymorphic viruses | false positive | false negative | | |
|--|-----------------------------|----------------------------|---|---|---|---|---|---|----------------------------------|------------------------------|-------------------------------|------------------------|----------------------------------|------------------------------|-----------|
| | | scanning speed improvement | | | | | | | new or unknown viruses detection | promise perfect disinfection | scanning speed improvement | virus family detection | new or unknown viruses detection | promise perfect disinfection | |
| first-generation scanners (string signature scanning) | simple scanning | | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | Low | Low |
| | optimizing techniques | wildcards | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | Low | Low |
| | | mismatch | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | Low | Low |
| | | generic degree | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | Low | Low |
| | bookmarks | | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | Very Low | Low |
| | speed-up techniques | hashing | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | Low | Low |
| | | top-and-tail scanning | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | Low | High |
| | | entry-point/fixed-point | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | Low | Low |
| second- generation scanners | smart scanning | | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Low | Low |
| | skeleton detection | | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | Low | Low |
| | nearly-exact identification | | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | Very Low | Very Low |
| | exact-identification | | ✓ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | Zero | Zero |
| | heuristic analysis | | ✗ | ✗ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | Very High | Low |
| virus-specific detection | general | | ✓ | ✗ | ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Low | Low |
| | optimizing techniques | filtering | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | Low | Low |
| | | static decryptor detect | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | Very High | Very High |
| | X-RAY scanning | | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | Low | Low |
| code emulation | Generic Detection | | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | Low | Low |
| | dynamic decryptor detection | | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | Low | Low |

References

- [1] Szor, P., The Art of Computer Virus Research and Defense, Addison-Wesley Professional, 2005.
- [2] Aycock, J., Computer Viruses and Malware, AB, Canada: Springer, 2006.
- [3] Beaucamps, P., "Advanced Polymorphic Techniques", Proceedings of World Academy of Science, Engineering and Technology, Vol 25, Vol. 25, No. 2007, pp. 400-411.
- [4] Skulason, F., "Virus Encryption Techniques", Virus Bulletin, Vol. No. 1990, pp. 13-16.
- [5] Johansson, K., COMPUTER VIRUSES: The Technology and Evolution of an Artificial Life Form, 1994.
- [6] Zhang, Q., "Polymorphic and metamorphic malware detection", Ph.D. thesis, North Carolina State University, United States -- North Carolina, 2008.
- [7] Bonfante, G., M. Kaczmarek, and J.Y. Marion, "Toward an Abstract Computer Virology", 2005.
- [8] Ludwig, M., The Giant Black Book of Computer Viruses, Arizona: American Eagle Publications, 1995.
- [9] Szor, P. and P. Ferrie, "Hunting for Metamorphic", in 11th International Virus Bulletin Conference, Year, Vol. pp. 123-144.
- [10] Szor, P., "The new 32-bit medusa", Virus Bulletin, Vol. No. 2000, pp. 8-10.
- [11] Jordan, M., "Dealing with Metamorphism", Virus Bulletin, Vol. No. 2002, pp. 4-6.
- [12] Vb, "IBM Viruses (Update)", Virus Bulletin, Vol. No. 1999, pp. 5-6.
- [13] Erdogan, O. and P. Cao, "Hash-AV: Fast virus signature scanning by cache-resident filters", in IEEE Global Telecommunications Conference (GLOBECOM'05), Year, Vol. 3, pp. 1767-1772.
- [14] Catalin, B. and A. Vi Oiu, "Optimization of Antivirus Software", Informatica, Vol. 11, No. 2007, pp. 99-102.
- [15] Bidgoli, H., Handbook of information security, Wiley, 2006.
- [16] Arnold, W. and G. Tesauro, "Automatically generated Win32 heuristic virus detection", in VIRUS BULLETIN CONFERENCE, Year, Vol. 51, pp.
- [17] Nachenberg, C., "Understanding heuristics: Symantec's bloodhound technology", 1998.
- [18] Perriot, F. and P. Ferrie, "Principles and practise of x-ray imaging", in Virus Bulletin Conference (VB2004), Year, Vol. pp. 51-56.
- [19] Veldman, F., "Generic Decryptors Emulators of the future", in IVPC conference, 1998, Vol. pp.
- [20] Ferrie, P., "Attacks on Virtual Machine Emulators", in AVAR Conference, 2006, Vol. pp. 128-143.

- [21] Raffetseder, T., C. Kruegel, and E. Kirda, "Detecting system emulators", *Information Security*, Vol. No. 2007, pp. 1-18.
- [22] Josse, S., "Secure and advanced unpacking using computer emulation", *Journal in Computer Virology*, Vol. 3, No. 3, 2007, pp. 221-236.



Babak Bashari Rad is currently a PhD candidate in Computer Science, at University Technology of Malaysia International Campus, Kuala Lumpur. He has completed his Master degree (2002) in Computer Engineering-Artificial Intelligence and Robotic as the outstanding student of Computer Science and Engineering (CSE) department, faculty of engineering at Shiraz University, Iran. He has been working as a faculty lecturer for nine past years in Azad University branches at Iran. His interests and research areas are computer virology, malwares, information security, code analysis, and machine learning methodologies. He is studying on metamorphic virus analysis and detection methodologies for his PhD thesis.



Maslin Masrom received the Bachelor of Science in Computer Science (1989), Master of Science in Operations Research (1992), and PhD in Information Technology/Information System Management (2003). She is an Associate Professor, Razak School of Engineering and Advanced Technology Malaysia, Universiti Technology Malaysia International Campus, Kuala Lumpur. Her current research interests include information security, ethics in computing, e-learning, human capital and knowledge management, and structural equation modeling. She has published articles in both local and international journals such as *Information and Management Journal*, *Oxford Journal*, *Journal of US-China Public Administration*, *MASAUM Journal of Computing*, *ACM SIGCAS Computers & Society* and *International Journal of Cyber Society and Education*.



Suhaimi Ibrahim received the Bachelor in Computer Science (1986), Master in Computer Science (1990), and PhD in Computer Science (2006). He is an Associate Professor attached to Advanced Informatics School (AIS), Universiti Teknologi Malaysia International Campus, Kuala Lumpur. He is an ISTQB certified tester and currently being appointed a board member of the Malaysian Software Testing Board (MSTB). He has published articles in both local and international journals such as the *International Journal of Web Services Practices*, *Journal of Computer Science*, *International Journal of Computational Science*, *Journal of Systems and Software*, and *Journal of Information and Software Technology*. His research interests include software testing, requirements engineering, Web services, software process improvement and software quality.

Performance Evaluation of Two Node Tandem Communication Network with Dynamic Bandwidth Allocation having Two Stage Direct Bulk Arrivals

K. Srinivasa Rao¹, Kuda Nageswara Rao², P.Srinivasa Rao³

¹ Department of Statistics, Andhra University, Visakhapatnam-530003, A.P., India

^{2, 3} Dept. of Computer Science & Systems Engineering, College of Engineering
Andhra University, Visakhapatnam-530003, A.P., India

Abstract

A two node tandem communication network with dynamic bandwidth allocation (DBA) having two stage direct bulk arrivals is developed and analyzed. The messages arriving to the source are packetized and stored in the buffers for forward transmission. Dynamic bandwidth allocation strategy is proposed by adjusting the transmission rate at every node just before transmission of each packet. The arrival and transmission processes at each node are characterized through compound Poisson and Poisson processes such that several of the statistical characteristics of communication networks identically matches. Using the difference – differential equations, the performance measures like the joint probability generating function of the content of two buffers, average buffer content, mean delays and throughput of nodes are derived and analyzed. It is observed that the bulk arrivals at two nodes and DBA have significant influence on performance measures. This network is much useful in Tele and Satellite communications.

Keywords: *Communication networks, Dynamic bandwidth allocation, Two- stage Bulk arrivals and Performance measures*

1. Introduction

It is generally known that packet switching gives better utilization over circuit or message switching and yields relatively short network delay. In packet switching, the message is divided into a random number of small packets each having an independent header for routing. This phenomenon is visible in Tele and Satellite communications where packet switching is effectively deployed. It can be characterized through statistical multiplexing by approximating the arrival process with a compound Poisson process (Kin K. Leung, 2002; K.Srinivasa Rao et al. 2006).

To have an efficient transmission, some algorithms have been developed with various protocols and allocation strategies for optimum utilization of the bandwidth (Emre and Ezhan, 2008; Gundale and Yardi, 2008; Hongwang and Yufan, 2009; Fen Zhou et al. 2009; Stanislav, 2009). These strategies are developed based on flow control or bit dropping techniques. But utilization of the idle bandwidth by adjusting the transmission rate instantaneously just before transmission of a packet is more important to maintain quality of service (QoS).

Dynamic bandwidth allocation strategy of transmission considers the adjustment of transmission rate of the packet depending upon the content of the buffer connected to transmitter at that instant. Recently, P.S.Varma et al. (2007) have utilized this strategy for a

two node communication network. However, they assumed that the arrivals to the source node are single packets. But, in communication systems, the packet arrivals to the source node are

in bulk since the message is converted into a random number of packets depending upon the message size. Hence, the Poisson assumption made for arrival process of packets may lead to inaccurate prediction of performance measures in the communication networks. Therefore, it is needed to develop and analyze the tandem communication network models with bulk arrivals having dynamic bandwidth allocation. Very little work has been reported in the literature regarding tandem communication networks with bulk arrivals which are quite common in places like Tele and Satellite communications. Kuda Nageswara Rao et al. (2010) have developed a communication network with dynamic bandwidth allocation having bulk arrivals. They approximated the arrival process with a compound Poisson process which characterizes the bulk arrival nature of the communication networks. However, they assumed that the arrivals are only to the initial node. But, in Tele communication systems, the messages may arrive directly to the second node also in addition to the packets forwarded through the first node. This phenomenon of direct bulk arrivals for both nodes has significant influence on buffer management and optimal utilization of resources in general and particularly with dynamic bandwidth allocation. With this motivation, in this paper, a two node communication network with dynamic bandwidth allocation having direct bulk arrivals to two nodes is developed and analyzed using mathematical modeling. Conducting laboratory experiments with variable load conditions for communication networks are complicated and time consuming, a mathematical model provide a basic frame work for performance evaluation of communication networks (Yukuo,1993; Ushio Sunita et al. 1997; Gaujal et al. 2002; Anney et al. 2010).

Using the difference – differential equations, the performance measures of the communication network like the joint probability generating function of the content of buffers, average contents of buffers, mean delays in transmissions, throughput etc., are derived. The sensitivity

of the performance measures with respect to the model parameters is also studied through numerical illustration.

2. Communication Network Model

Consider a communication network model with two nodes in tandem having bulk arrivals and dynamic bandwidth allocation. The arrivals to node 1 and node 2 are assumed to follow a compound Poisson process with parameters λ_1 and λ_2 respectively. The compound Poisson process is capable of portraying the bulk arrival nature of the communication network. Here, it is considered that the messages that arrive to both nodes are converted into random number of packets and form a batch. The batch size distribution of packets are assumed to follow rectangle (uniform) distribution probability distribution functions C_{k1} and C_{k2} with parameters (a_1, b_1) and (a_2, b_2) respectively for buffer 1 and buffer 2.

It is also assumed that the number of transmissions at each transmitter follow Poisson processes with parameters μ_1 and μ_2 respectively. The transmission rates of packets in each node are instantaneously adjusted depending on the content of the buffers just before its transmission. The queue discipline is First-In-First-Out (FIFO). There is no termination of packets after the transmission of first node. The schematic diagram representing the communication network is shown in Figure 1.

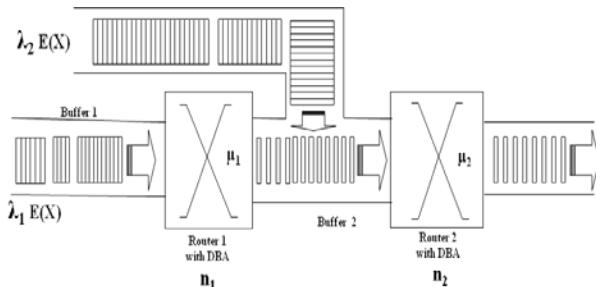


Fig.1 Communication network with two stage Bulk arrivals and dynamic bandwidth allocation

Let $P_{n_1, n_2}(t)$ be the probability that there are n_1 packets in the first buffer and n_2 packets in the second buffer at time t . Then, the difference-differential equations governing the network are

$$\frac{\partial P_{n_1, n_2}(t)}{\partial t} = -(\lambda_1 + n_1 \mu_1 + n_2 \mu_2 + \lambda_2) P_{n_1, n_2}(t) + (n_1 + 1) \mu_1 P_{n_1+1, n_2-1}(t) + (n_2 + 1) \mu_2 P_{n_1, n_2+1}(t) + \lambda_1 \left[\sum_{i=1}^{n_1} P_{n_1-i, n_2}(t) C_{i1} \right] + \lambda_2 \left[\sum_{j=1}^{n_2} P_{n_1, n_2-j}(t) C_{j2} \right] \quad (1)$$

$$\frac{\partial P_{n_1, 0}(t)}{\partial t} = -(\lambda_1 + n_1 \mu_1 + \lambda_2) P_{n_1, 0}(t) + \mu_2 P_{n_1, 1}(t) + \lambda_1 \left[\sum_{i=1}^{n_1} P_{n_1-i, 0}(t) C_{i1} \right] \quad (2)$$

$$\frac{\partial P_{0, n_2}(t)}{\partial t} = -(\lambda_1 + n_2 \mu_2 + \lambda_2) P_{0, n_2}(t) + \mu_1 P_{0, n_2-1}(t) + (n_2 + 1) \mu_2 P_{0, n_2+1}(t) + \lambda_2 \left[\sum_{j=1}^{n_2} P_{0, n_2-j}(t) C_{j2} \right] \quad (3)$$

$$\frac{\partial P_{1, 0}(t)}{\partial t} = -[\lambda_1 + \mu_1 + \lambda_2] P_{1, 0}(t) + \mu_2 P_{1, 1}(t) + \lambda_1 [P_{0, 0}(t) C_{11}] \quad (4)$$

$$\frac{\partial P_{0, 1}(t)}{\partial t} = -(\lambda_1 + \mu_2 + \lambda_2) P_{0, 1}(t) + \mu_1 P_{0, 0}(t) + 2\mu_2 P_{0, 2}(t) + \lambda_2 [P_{0, 0}(t) C_{12}] \quad (5)$$

$$\frac{\partial P_{0, 0}(t)}{\partial t} = -(\lambda_1 + \lambda_2) P_{0, 0}(t) + \mu_2 P_{0, 1}(t) \quad (6)$$

Assuming that the system is under steady state

$$\lim_{t \rightarrow \infty} P_{n_1, n_2}(t) = P_{n_1, n_2} \quad \text{and} \quad \lim_{t \rightarrow \infty} \frac{dP_{n_1, n_2}(t)}{dt} = 0$$

The steady state equations of the model are

$$-(\lambda_1 + n_1 \mu_1 + n_2 \mu_2 + \lambda_2) P_{n_1, n_2} + (n_1 + 1) \mu_1 P_{n_1+1, n_2-1} + (n_2 + 1) \mu_2 P_{n_1, n_2+1} + \lambda_1 \left[\sum_{i=1}^{n_1} P_{n_1-i, n_2} C_{i1} \right] + \lambda_2 \left[\sum_{j=1}^{n_2} P_{n_1, n_2-j} C_{j2} \right] = 0$$

(7)

$$-(\lambda_1 + n_1 \mu_1 + \lambda_2) P_{n_1, 0} + \mu_2 P_{n_1, 1} + \lambda_1 \left[\sum_{i=1}^{n_1} P_{n_1-i, 0} C_{i1} \right] = 0$$

(8)

$$-(\lambda_1 + n_2 \mu_2 + \lambda_2) P_{0, n_2} + \mu_1 P_{0, n_2-1} + (n_2 + 1) \mu_2 P_{0, n_2+1} + \lambda_2 \left[\sum_{j=1}^{n_2} P_{0, n_2-j} C_{j2} \right] = 0$$

(9)

$$-[\lambda_1 + \mu_1 + \lambda_2] P_{1, 0} + \mu_2 P_{1, 1} + \lambda_1 [P_{0, 0} C_{11}] = 0$$

$$(10) \quad -(\lambda_1 + \mu_2 + \lambda_2) P_{0, 1} + \mu_1 P_{0, 0} + 2\mu_2 P_{0, 2} + \lambda_2 [P_{0, 0} C_{12}] = 0$$

(11)

$$-(\lambda_1 + \lambda_2) P_{0, 0} + \mu_2 P_{0, 1} = 0 \quad (12)$$

Let $P(Z_1, Z_2) = \sum_{n_1=0}^{\infty} \sum_{n_2}^{\infty} P_{n_1, n_2} Z_1^{n_1} Z_2^{n_2}$ and

$C(Z) = \sum_{k=1}^{\infty} C_k Z^k$ be the probability generating functions of P_{n_1, n_2} and C_k respectively.

Multiplying the equations 7 to 12 with corresponding $Z_1^{n_1}$ and $Z_2^{n_2}$ and summing over all n_1, n_2 , we get the joint probability generating function of n_1 packets in the first buffer and n_2 packets in the second buffer at any time when the system is under equilibrium as

$$P(Z_1, Z_2) = \exp \left[\lambda_1 \sum_{k_1=1}^{\infty} \sum_{r=1}^{k_1} \sum_{j=0}^{r-1} (-1)^{2r-j} C_{k_1}^{(k_1)} C_r^{(r)} \left(\frac{\mu_1}{\mu_2 - \mu_1} \right)^j (Z_2 - 1)^j \left((Z_1 - 1) + \frac{\mu_1 (Z_2 - 1)}{\mu_2 - \mu_1} \right)^{r-j} \left(\frac{1}{j \mu_2 + (r-j) \mu_1} \right) \right] + \lambda_2 \left[\sum_{k_2=1}^{\infty} \sum_{s=1}^{k_2} C_{k_2}^{(k_2)} C_s^{(s)} (Z_2 - 1)^s \left(\frac{1}{\mu_2 s} \right) \right] \quad (13)$$

3. Performance Measures of the Communication Network

The probability generating function of the first buffer size distribution is

$$P(Z_1) = \exp \left[\lambda_1 \sum_{k_1=1}^{\infty} \sum_{r=1}^{k_1} (-1)^{2r} C_{k_1}^{(k_1)} C_r^{(r)} (Z_1 - 1)^r \left(\frac{1}{r \mu_1} \right) \right] \quad (14)$$

The probability that the first buffer is empty as

$$P_{0, 0} = \exp \left[\lambda_1 \sum_{k_1=1}^{\infty} \sum_{r=1}^{k_1} C_{k_1}^{(k_1)} C_r^{(r)} (-1)^{3r} \left(\frac{1}{r \mu_1} \right) \right] \quad (15)$$

The mean number packets in the first buffer is

$$L_1 = \frac{\lambda_1}{\mu_1} \left[\sum_{k_1=1}^{\infty} C_{k_1}^{(k_1)} k_1 \right] \quad (16)$$

The utilization of the first node is

$$U_1 = 1 - p_0.$$

$$= 1 - \exp \left[\lambda_1 \sum_{k_1=1}^{\infty} \sum_{r=1}^{k_1} C_{k_1}^{k_1} C_r (-1)^{3r} \left(\frac{1}{r\mu_1} \right) \right] \quad (17)$$

The throughput of the first node is

$$\text{Thp}_1 = \mu_1 \cdot U_1 \\ = \mu_1 \left[1 - \exp \left[\lambda_1 \sum_{k_1=1}^{\infty} \sum_{r=1}^{k_1} C_{k_1}^{k_1} C_r (-1)^{3r} \left(\frac{1}{r\mu_1} \right) \right] \right] \quad (18)$$

The average delay in the first buffer is

$$W(N_1) = \frac{L_1}{\text{Thp}_1} = \frac{\frac{\lambda_1}{\mu_1} \left[\sum_{k_1=1}^{\infty} C_{k_1} k_1 \right]}{\mu_1 \left[1 - \exp \left[\lambda_1 \sum_{k_1=1}^{\infty} \sum_{r=1}^{k_1} C_{k_1}^{k_1} C_r (-1)^{3r} \left(\frac{1}{r\mu_1} \right) \right] \right]} \quad (19)$$

The variances of the number of packets in the first buffer is

$$\text{Var}(N_1) = E[N_1^2] - E[N_1]^2 \\ = \frac{\lambda_1}{2\mu_1} \left[\sum_{k_1=1}^{\infty} C_{k_1} k_1 (k_1 - 1) \right] + \frac{\lambda_1}{\mu_1} \left[\sum_{k_1=1}^{\infty} C_{k_1} k_1 \right] \quad (20)$$

The coefficient of variation of the number of packets in the first buffer is

$$cv(N_1) = \frac{\sqrt{\text{Var}(N_1)}}{L_1} \quad (21)$$

The probability generating function of the second buffer size distribution is

$$P(Z_2) = \exp \left[\lambda_2 \sum_{k_2=1}^{\infty} \sum_{r=1}^{k_2} \sum_{j=0}^r (-1)^{2r-j} C_{k_2}^{k_2} C_r (-1)^j C_j \left(\frac{\mu_1}{\mu_2 - \mu_1} \right)^r (Z_2 - 1)^j \left(\frac{1}{J\mu_2 + (r-j)\mu_1} \right) \right] \\ + \lambda_2 \left[\sum_{k_2=1}^{\infty} \sum_{s=1}^{k_2} C_{k_2}^{k_2} C_s (-1)^s \left(\frac{1}{\mu_2 s} \right) \right] \quad (22)$$

The probability that the second buffer is empty as

$$p_{00} = \exp \left[\lambda_2 \sum_{k_2=1}^{\infty} \sum_{r=1}^{k_2} \sum_{j=0}^r (-1)^{3r-j} C_{k_2}^{k_2} C_r (-1)^j C_j \left(\frac{\mu_1}{\mu_2 - \mu_1} \right)^r \left(\frac{1}{J\mu_2 + (r-j)\mu_1} \right) \right] \\ + \lambda_2 \left[\sum_{k_2=1}^{\infty} \sum_{s=1}^{k_2} C_{k_2}^{k_2} C_s (-1)^s \left(\frac{1}{\mu_2 s} \right) \right] \quad (23)$$

The mean number packets in the second buffer is

$$L_2 = \frac{\lambda_2}{\mu_2} \left(\sum_{k_2=1}^{\infty} C_{k_2} k_2 \right) + \frac{\lambda_2}{\mu_2} \left(\sum_{k_2=1}^{\infty} C_{k_2} k_2 \right)$$

The utilization of the second node is

$$U_2 = 1 - p_{00} \\ = 1 - \exp \left[\lambda_2 \sum_{k_2=1}^{\infty} \sum_{r=1}^{k_2} \sum_{j=0}^r (-1)^{3r-j} C_{k_2}^{k_2} C_r (-1)^j C_j \left(\frac{\mu_1}{\mu_2 - \mu_1} \right)^r \left(\frac{1}{J\mu_2 + (r-j)\mu_1} \right) \right] \\ + \lambda_2 \left[\sum_{k_2=1}^{\infty} \sum_{s=1}^{k_2} C_{k_2}^{k_2} C_s (-1)^s \left(\frac{1}{\mu_2 s} \right) \right] \quad (25)$$

The throughput of second node is

$$\text{Thp}_2 = \mu_2 \cdot U_2$$

$$= \mu_2 \cdot \left\{ 1 - \exp \left[\lambda_2 \sum_{k_2=1}^{\infty} \sum_{r=1}^{k_2} \sum_{j=0}^r (-1)^{3r-j} C_{k_2}^{k_2} C_r (-1)^j C_j \left(\frac{\mu_1}{\mu_2 - \mu_1} \right)^r \left(\frac{1}{J\mu_2 + (r-j)\mu_1} \right) \right] \right. \\ \left. + \lambda_2 \left[\sum_{k_2=1}^{\infty} \sum_{s=1}^{k_2} C_{k_2}^{k_2} C_s (-1)^s \left(\frac{1}{\mu_2 s} \right) \right] \right\} \quad (26)$$

The average delay in the second buffer is

$$W(N_2) = \frac{L_2}{\text{Thp}_2} = \frac{\frac{\lambda_2}{\mu_2} \left(\sum_{k_2=1}^{\infty} C_{k_2} k_2 \right) + \lambda_2 \left[\sum_{k_2=1}^{\infty} \sum_{s=1}^{k_2} C_{k_2}^{k_2} C_s (-1)^s \left(\frac{1}{\mu_2 s} \right) \right]}{\mu_2 \cdot \left\{ 1 - \exp \left[\lambda_2 \sum_{k_2=1}^{\infty} \sum_{r=1}^{k_2} \sum_{j=0}^r (-1)^{3r-j} C_{k_2}^{k_2} C_r (-1)^j C_j \left(\frac{\mu_1}{\mu_2 - \mu_1} \right)^r \left(\frac{1}{J\mu_2 + (r-j)\mu_1} \right) \right] \right. \\ \left. + \lambda_2 \left[\sum_{k_2=1}^{\infty} \sum_{s=1}^{k_2} C_{k_2}^{k_2} C_s (-1)^s \left(\frac{1}{\mu_2 s} \right) \right] \right\}} \quad (27)$$

The variances of the number of packets in the second buffer is

$$\text{Var}(N_2) = E[N_2^2] - E[N_2]^2 \\ = \variance{N_2} = \lambda_2 \left\{ \left(\sum_{k_2=1}^{\infty} C_{k_2} k_2 (k_2 - 1) \right) \left(\frac{\mu_1}{\mu_1 - \mu_2} \right)^2 \left[\left(\frac{1}{2\mu_1} \right) - 2 \left(\frac{1}{\mu_1 + \mu_2} \right) + \left(\frac{1}{2\mu_2} \right) \right] \right. \\ \left. + \left\{ \lambda_2 \sum_{k_2=1}^{\infty} C_{k_2} k_2 (k_2 - 1) \left(\frac{1}{2\mu_2} \right) \right\} + \left\{ \lambda_2 \left(\sum_{k_2=1}^{\infty} C_{k_2} k_2 \right) \left(\frac{1}{\mu_2} \right) \right. \right. \\ \left. \left. + \left\{ \frac{\lambda_2}{\mu_2} \left(\sum_{k_2=1}^{\infty} C_{k_2} k_2 \right) \right\} \right\} \quad (28)$$

The coefficient of variation of the number of packets in the second buffer is

$$cv(N_2) = \frac{\sqrt{\text{Var}(N_2)}}{L_2} \quad (29)$$

The probability that the network is empty is

$$p_{00} = \exp \left[\lambda_1 \sum_{k_1=1}^{\infty} \sum_{r=1}^{k_1} \sum_{j=0}^r (-1)^{2r-j} C_{k_1}^{k_1} C_r (-1)^j C_j \left(\frac{\mu_1}{\mu_2 - \mu_1} \right)^r \left(\frac{1}{J\mu_2 + (r-j)\mu_1} \right) \right] \\ + \lambda_2 \left[\sum_{k_2=1}^{\infty} \sum_{s=1}^{k_2} C_{k_2}^{k_2} C_s (-1)^s \left(\frac{1}{\mu_2 s} \right) \right] \quad (30)$$

The mean number packets in the network is

$$L_N = L_1 + L_2 \quad (31)$$

where,

L_1 = the mean number of packets in the first node

L_2 = the mean number of packets in the second node

4. Performance Measures with Uniform Batch Size Distribution

In this section, the performance of the communication network under steady state conditions is discussed with the assumption that the number of packets that each message can be converted follows a uniform (rectangular) distribution with parameters (a_1, b_1) and (a_2, b_2) for buffer 1 and buffer 2 respectively. Then the joint probability generating function of the buffers size when the system is under equilibrium is

$$P(Z_1, Z_2) = \exp \left[\lambda_1 \sum_{k_1=a_1}^{b_1} \sum_{r=1}^{k_1} \sum_{j=0}^r (-1)^{2r-j} \left(\frac{1}{b_1 - a_1 + 1} \right) {}^{k_1} C_r ({}^r C_j) \left(\frac{\mu_1}{\mu_2 - \mu_1} \right)^j (Z_1 - 1)^r \left((Z_1 - 1) + \frac{\mu_1 (Z_2 - 1)}{\mu_2 - \mu_1} \right)^{r-j} \right] \\ + \lambda_2 \left[\sum_{k_2=a_2}^{b_2} \sum_{s=1}^{k_2} \left(\frac{1}{b_2 - a_2 + 1} \right) {}^{k_2} C_s (Z_2 - 1)^s \left(\frac{1}{\mu_2 s} \right) \right] \quad (32)$$

The probability generating function of the first buffer size distribution is

$$P(Z_1) = \exp \left[\lambda_1 \sum_{k_1=a_1}^{b_1} \sum_{r=1}^{k_1} (-1)^{2r} \left(\frac{1}{b_1 - a_1 + 1} \right) {}^{k_1} C_r (Z_1 - 1)^r \left(\frac{1}{r \mu_1} \right) \right] \quad (33)$$

The probability that the first buffer is empty is

$$p_0 = \exp \left[\lambda_1 \sum_{k_1=a_1}^{b_1} \sum_{r=1}^{k_1} \left(\frac{1}{b_1 - a_1 + 1} \right) {}^{k_1} C_r (-1)^{3r} \left(\frac{1}{r \mu_1} \right) \right] \quad (34)$$

The mean number packets in the first buffer is

$$L_1 = \frac{\lambda_1 (a_1 + b_1)}{2 \mu_1} \quad (35)$$

The utilization of the first node is

$$U_1 = 1 - p_0.$$

$$= 1 - \exp \left[\lambda_1 \sum_{k_1=a_1}^{b_1} \sum_{r=1}^{k_1} \left(\frac{1}{b_1 - a_1 + 1} \right) {}^{k_1} C_r (-1)^{3r} \left(\frac{1}{r \mu_1} \right) \right] \quad (36)$$

The throughput of the first node is

$$Thp_1 = \mu_1 \cdot U_1 \\ = \mu_1 \left\{ 1 - \exp \left[\lambda_1 \sum_{k_1=a_1}^{b_1} \sum_{r=1}^{k_1} \left(\frac{1}{b_1 - a_1 + 1} \right) {}^{k_1} C_r (-1)^{3r} \left(\frac{1}{r \mu_1} \right) \right] \right\} \quad (37)$$

The average delay in the first buffer is

$$W(N_1) = \frac{L_1}{Thp_1} = \frac{\frac{\lambda_1 (a_1 + b_1)}{2 \mu_1}}{\mu_1 \left\{ 1 - \exp \left[\lambda_1 \sum_{k_1=a_1}^{b_1} \sum_{r=1}^{k_1} \left(\frac{1}{b_1 - a_1 + 1} \right) {}^{k_1} C_r (-1)^{3r} \left(\frac{1}{r \mu_1} \right) \right] \right\}} \quad (38)$$

The variances of the number of packets in the first buffer is

$$\text{Var}(N_1) = E[N_1^2] - E[N_1]^2 \\ = \frac{\lambda_1}{2 \mu_1} \left[\sum_{k_1=a_1}^{b_1} \left(\frac{1}{b_1 - a_1 + 1} \right) k_1 (k_1 - 1) \right] + \frac{\lambda}{\mu_1} \left[\sum_{k_1=a_1}^{b_1} \left(\frac{1}{b_1 - a_1 + 1} \right) k_1 \right] \quad (39)$$

The coefficient of variation of the number of packets in the first buffer is

$$cv(N_1) = \frac{\sqrt{\text{Var}(N_1)}}{L_1} \quad (40)$$

The probability generating function of the second buffer size distribution is

$$P(Z_2) = \exp \left[\lambda_1 \sum_{k_1=a_1}^{b_1} \sum_{r=1}^{k_1} \sum_{j=0}^r (-1)^{2r-j} \left(\frac{1}{b_1 - a_1 + 1} \right) {}^{k_1} C_r ({}^r C_j) \left(\frac{\mu_1}{\mu_2 - \mu_1} \right)^r (Z_2 - 1)^r \right] \\ + \lambda_2 \left[\sum_{k_2=a_2}^{b_2} \sum_{s=1}^{k_2} \left(\frac{1}{b_2 - a_2 + 1} \right) {}^{k_2} C_s (Z_2 - 1)^s \left(\frac{1}{\mu_2 s} \right) \right]$$

The probability that the second buffer is empty is

$$p_0 = \exp \left[\lambda_1 \sum_{k_1=a_1}^{b_1} \sum_{r=1}^{k_1} \sum_{j=0}^r (-1)^{2r-j} \left(\frac{1}{b_1 - a_1 + 1} \right) {}^{k_1} C_r ({}^r C_j) \left(\frac{\mu_1}{\mu_2 - \mu_1} \right)^r \left(\frac{1}{J \mu_2 + (r - J) \mu_1} \right) \right] \\ + \lambda_2 \left[\sum_{k_2=a_2}^{b_2} \sum_{s=1}^{k_2} \left(\frac{1}{b_2 - a_2 + 1} \right) {}^{k_2} C_s (-1)^s \left(\frac{1}{\mu_2 s} \right) \right] \quad (42)$$

The mean number packets in the second buffer is

$$L_2 = \frac{\lambda_1 (a_1 + b_1)}{2 \mu_2} + \frac{\lambda_2 (a_2 + b_2)}{2 \mu_2}$$

(43)

The utilization of the second node is

$$U_2 = 1 - p_0$$

$$= 1 - \exp \left[\lambda_1 \sum_{k_1=a_1}^{b_1} \sum_{r=1}^{k_1} \sum_{j=0}^r (-1)^{2r-j} \left(\frac{1}{b_1 - a_1 + 1} \right) {}^{k_1} C_r ({}^r C_j) \left(\frac{\mu_1}{\mu_2 - \mu_1} \right)^r \left(\frac{1}{J \mu_2 + (r - J) \mu_1} \right) \right] \\ + \lambda_2 \left[\sum_{k_2=a_2}^{b_2} \sum_{s=1}^{k_2} \left(\frac{1}{b_2 - a_2 + 1} \right) {}^{k_2} C_s (-1)^s \left(\frac{1}{\mu_2 s} \right) \right] \quad (44)$$

The throughput of second node is

$$Thp_2 = \mu_2 \cdot U_2$$

$$= U_2 \cdot \left\{ 1 - \exp \left[\lambda_1 \sum_{k_1=a_1}^{b_1} \sum_{r=1}^{k_1} \sum_{j=0}^r (-1)^{2r-j} \left(\frac{1}{b_1 - a_1 + 1} \right) {}^{k_1} C_r ({}^r C_j) \left(\frac{\mu_1}{\mu_2 - \mu_1} \right)^r \left(\frac{1}{J \mu_2 + (r - J) \mu_1} \right) \right] \right. \\ \left. + \lambda_2 \left[\sum_{k_2=a_2}^{b_2} \sum_{s=1}^{k_2} \left(\frac{1}{b_2 - a_2 + 1} \right) {}^{k_2} C_s (-1)^s \left(\frac{1}{\mu_2 s} \right) \right] \right\} \quad (45)$$

The average delay in the second buffer is

$$W(N_2) = \frac{L_2}{Thp_2} = \frac{\frac{\lambda_1 (a_1 + b_1)}{2 \mu_2} + \frac{\lambda_2 (a_2 + b_2)}{2 \mu_2}}{U_2 \cdot \left\{ 1 - \exp \left[\lambda_1 \sum_{k_1=a_1}^{b_1} \sum_{r=1}^{k_1} \sum_{j=0}^r (-1)^{2r-j} \left(\frac{1}{b_1 - a_1 + 1} \right) {}^{k_1} C_r ({}^r C_j) \left(\frac{\mu_1}{\mu_2 - \mu_1} \right)^r \left(\frac{1}{J \mu_2 + (r - J) \mu_1} \right) \right] \right. \\ \left. + \lambda_2 \left[\sum_{k_2=a_2}^{b_2} \sum_{s=1}^{k_2} \left(\frac{1}{b_2 - a_2 + 1} \right) {}^{k_2} C_s (-1)^s \left(\frac{1}{\mu_2 s} \right) \right] \right\}} \quad (46)$$

The variances of the number of packets in the second buffer is

$$\text{Var}(N_2) = E[N_2^2] - E[N_2]^2$$

$$= \lambda_1 \left\{ \left[\sum_{k_1=a_1}^{b_1} \left(\frac{1}{b_1 - a_1 + 1} \right) k_1 (k_1 - 1) \left(\frac{\mu_1}{\mu_2 - \mu_1} \right)^2 \left[\left(\frac{1}{2 \mu_1} \right) - 2 \left(\frac{1}{\mu_1 + \mu_2} \right) + \left(\frac{1}{2 \mu_2} \right) \right] \right] \right. \\ \left. + \left[\lambda_2 \sum_{k_2=a_2}^{b_2} \left(\frac{1}{b_2 - a_2 + 1} \right) k_2 (k_2 - 1) \left(\frac{1}{2 \mu_2} \right) \right] + \left[\lambda_1 \left(\frac{a_1 + b_1}{2} \right) \left(\frac{1}{\mu_2} \right) \right] \right. \\ \left. + \left[\lambda_2 \left(\frac{a_2 + b_2}{2} \right) \right] \right\} \quad (47)$$

The coefficient of variation of the number of packets in the second buffer is

$$cv(N_2) = \frac{\sqrt{\text{Var}(N_2)}}{L_2} \quad (48)$$

The probability that the network is empty is

$$p_{00} = \exp \left[\lambda_1 \sum_{k_1=a_1}^{b_1} \sum_{r=1}^{k_1} \sum_{j=0}^r (-1)^{2r-j} \left(\frac{1}{b_1 - a_1 + 1} \right) {}^{k_1} C_r ({}^r C_j) (\theta \mu_1)^r \left(\frac{(\mu_2)^{r-j}}{(\mu_2 - \mu_1)^r} \right) \left(\frac{1}{J \mu_2 + (r - J) \mu_1} \right) \right] \\ + \lambda_2 \left[\sum_{k_2=a_2}^{b_2} \sum_{s=1}^{k_2} \left(\frac{1}{b_2 - a_2 + 1} \right) {}^{k_2} C_s (-1)^s \left(\frac{1}{\mu_2 s} \right) \right] \quad (49)$$

The mean number packets in the network is

$$L_N = L_1 + L_2 \quad (50)$$

Where,

L_1 = the mean number of packets in the first node
 L_2 = the mean number of packets in the second node

5. Performance Evaluation of the Communication Network

The performance of the proposed network is discussed through numerical illustration. Different values of the parameters are considered for bandwidth allocation and arrival of packets. λ_1 and λ_2 are the message arrival rates at node 1 and node 2 respectively. The number of packets that can be converted into a message varies from 1 to 10 depending on the length of the message. The number of arrivals of packets to the buffers is in batches of random size. The batch size is assumed to follow uniform (rectangle) distribution with parameters (a_1, b_1) and (a_2, b_2) for first and second buffers respectively. μ_1 is the transmission rate of node 1 which varies from 10×10^4 packets/sec to 14×10^4 packets/sec. The packets leave the second node with a transmission rate of μ_2 which varies from 16×10^4 packets/sec to 20×10^4 packets/sec. In both the nodes, dynamic bandwidth allocation is considered i.e. the transmission rate of each packet depends on the number of packets in the buffer connected to it at that instant.

The following set of values of the model parameters are considered in computing the performance measures like, Probabilities of emptiness of the network, first and second buffers, Mean number of packets in first and second buffers, Utilization of the nodes, Throughput of the nodes, Mean delays in first and second buffers and are given tables.

$a_1 = 1, 2, 3, 4, 5 ; b_1 = 6, 7, 8, 9, 10 ; a_2 = 1, 2, 3, 4, 5;$
 $b_2 = 6, 7, 8, 9, 10 ; \lambda_1 = 0.5, 1.5, 2.0, 2.5$ (with multiplication of 10^4 messages/sec), $\lambda_2 = 0.5, 1.5, 2.0, 2.5$ (with multiplication of

126
 10^4 messages/sec), $\mu_1 = 10, 11, 12, 13, 14$ (with multiplication of 10^4 packets/sec) and $\mu_2 = 16, 17, 18, 19, 20$ (with multiplication of 10^4 packets/sec)

From equations (34), (42) and (49), the probability of network emptiness and buffers emptiness are computed for different values of $a_1, b_1, a_2, b_2, \lambda_1, \lambda_2, \mu_1, \mu_2$ observed that when the values of the network parameters $a_1, b_1, a_2, b_2, \lambda_1$ and λ_2 increase, there is a decrease in the emptiness of the network, first and the second buffers. The emptiness of the first buffer remains constant for an increase in the parameter values a_2, b_2 and λ_2 . When the transmission rates of first node (μ_1) and second node (μ_2) increase, the network and second buffer emptiness decrease.

From equations (35), (36), (43), 442 and (50), mean number of packets of the network are computed for different values of a_1, b_1 ,

$a_2, b_2, \lambda_1, \lambda_2, \mu_1$ and μ_2 and are given in Table 1. The relationship between mean number of packets in the network, buffers and the parameters $a_1, b_1, a_2, b_2, \lambda_1, \lambda_2, \mu_1$ and μ_2 is shown in Figures2.

It is observed that when the values of network parameters $a_1, b_1, a_2, b_2, \lambda_1$, and λ_2 increase, the mean number of packets in the network, mean number of packets in the second buffer and the utilization of the second node increase. The mean number of packets in the first buffer and utilization of the first node increase when a_1, b_1 and λ_1 increase and remain constant when a_2, b_2 and λ_2 vary. When the transmission rate of node 1 (μ_1) varies from 10 to 14, the mean number of packets in the first buffer, utilization of the first and second

Table 1: Values of Mean Number of Packets, Average Delay and Throughput of Nodes

| a_1 | b_1 | a_2 | b_2 | λ_1 # | λ_2 # | μ_1 \$ | μ_2 \$ | L_1 | L_2 | L_N | Thp_1 | $W(N_1)$ | Thp_2 | $W(N_2)$ | Thp_1 |
|----------|-----------|----------|-----------|---------------|---------------|------------|------------|---------|---------|--------|---------|----------|---------|----------|---------|
| 1 | 6 | 5 | 10 | 1 | 2 | 5 | 20 | 0.7 | 0.925 | 1.625 | 1.55208 | 0.45101 | 6.76047 | 0.13682 | 1.55208 |
| 2 | 6 | 5 | 10 | 1 | 2 | 5 | 20 | 0.8 | 0.950 | 1.750 | 1.66845 | 0.47949 | 7.02045 | 0.13532 | 1.66845 |
| 3 | 6 | 5 | 10 | 1 | 2 | 5 | 20 | 0.9 | 0.975 | 1.875 | 1.75558 | 0.51265 | 7.25833 | 0.13433 | 1.75558 |
| 4 | 6 | 5 | 10 | 1 | 2 | 5 | 20 | 1.0 | 1.000 | 2.000 | 1.82600 | 0.54765 | 7.47674 | 0.13375 | 1.82600 |
| 5 | 6 | 5 | 10 | 1 | 2 | 5 | 20 | 1.1 | 1.025 | 2.125 | 1.88539 | 0.58343 | 7.67794 | 0.13350 | 1.88539 |
| 1 | 6 | 5 | 10 | 1 | 2 | 10 | 20 | 0.35000 | 0.925 | 1.275 | 1.69588 | 0.20638 | 6.57102 | 0.14077 | 1.69588 |
| 1 | 7 | 5 | 10 | 1 | 2 | 10 | 20 | 0.40000 | 0.950 | 1.35 | 1.78256 | 0.22440 | 6.73770 | 0.14100 | 1.78256 |
| 1 | 8 | 5 | 10 | 1 | 2 | 10 | 20 | 0.45000 | 0.975 | 1.425 | 1.85971 | 0.24197 | 6.88857 | 0.14154 | 1.85971 |
| 1 | 9 | 5 | 10 | 1 | 2 | 10 | 20 | 0.50000 | 1.000 | 1.500 | 1.92918 | 0.25918 | 7.02588 | 0.14233 | 1.92918 |
| 1 | 10 | 5 | 10 | 1 | 2 | 10 | 20 | 0.55000 | 1.025 | 1.575 | 1.99234 | 0.27606 | 7.15145 | 0.14333 | 1.99234 |
| 1 | 6 | 1 | 10 | 1 | 2 | 10 | 20 | 0.35000 | 0.725 | 1.075 | 1.69588 | 0.20638 | 6.00647 | 0.12070 | 1.69588 |
| 1 | 6 | 2 | 10 | 1 | 2 | 10 | 20 | 0.35000 | 0.775 | 1.125 | 1.69588 | 0.20638 | 6.19517 | 0.12510 | 1.69588 |
| 1 | 6 | 3 | 10 | 1 | 2 | 10 | 20 | 0.35000 | 0.825 | 1.175 | 1.69588 | 0.20638 | 6.34237 | 0.13008 | 1.69588 |
| 1 | 6 | 4 | 10 | 1 | 2 | 10 | 20 | 0.35000 | 0.875 | 1.225 | 1.69588 | 0.20638 | 6.46503 | 0.13534 | 1.69588 |
| 1 | 6 | 5 | 10 | 1 | 2 | 10 | 20 | 0.35000 | 0.925 | 1.275 | 1.69588 | 0.20638 | 6.57102 | 0.14077 | 1.69588 |
| 1 | 6 | 5 | 6 | 1 | 2 | 10 | 20 | 0.35000 | 0.725 | 1.075 | 1.69588 | 0.20638 | 6.20764 | 0.11679 | 1.69588 |
| 1 | 6 | 5 | 7 | 1 | 2 | 10 | 20 | 0.35000 | 0.775 | 1.125 | 1.69588 | 0.20638 | 6.31124 | 0.12280 | 1.69588 |
| 1 | 6 | 5 | 8 | 1 | 2 | 10 | 20 | 0.35000 | 0.825 | 1.175 | 1.69588 | 0.20638 | 6.40529 | 0.12880 | 1.69588 |
| 1 | 6 | 5 | 9 | 1 | 2 | 10 | 20 | 0.35000 | 0.875 | 1.225 | 1.69588 | 0.20638 | 6.49147 | 0.13479 | 1.69588 |
| 1 | 6 | 5 | 10 | 1 | 2 | 10 | 20 | 0.35000 | 0.925 | 1.275 | 1.69588 | 0.20638 | 6.57102 | 0.14077 | 1.69588 |
| 1 | 6 | 5 | 10 | 0.5 | 2 | 10 | 20 | 0.17500 | 0.83750 | 1.0125 | 1.69588 | 0.19723 | 5.63362 | 0.14866 | 0.88731 |
| 1 | 6 | 5 | 10 | 1.0 | 2 | 10 | 20 | 0.35000 | 0.92500 | 1.2750 | 1.69588 | 0.20638 | 6.57102 | 0.14077 | 1.69588 |
| 1 | 6 | 5 | 10 | 1.5 | 2 | 10 | 20 | 0.52500 | 1.01250 | 1.5375 | 2.43271 | 0.21581 | 7.44725 | 0.13596 | 2.43271 |
| 1 | 6 | 5 | 10 | 2.0 | 2 | 10 | 20 | 0.70000 | 1.10000 | 1.8000 | 3.10416 | 0.2255 | 8.26631 | 0.13307 | 3.10416 |
| 1 | 6 | 5 | 10 | 2.5 | 2 | 10 | 20 | 0.87500 | 1.18750 | 2.0625 | 3.71603 | 0.23547 | 9.03192 | 0.13148 | 3.71603 |
| 1 | 6 | 5 | 10 | 1 | 0.5 | 10 | 20 | 0.35000 | 0.36250 | 0.7125 | 1.69588 | 0.20638 | 3.63835 | 0.09963 | 1.69588 |
| 1 | 6 | 5 | 10 | 1 | 1.0 | 10 | 20 | 0.35000 | 0.55000 | 0.9000 | 1.69588 | 0.20638 | 4.68093 | 0.11750 | 1.69588 |
| 1 | 6 | 5 | 10 | 1 | 1.5 | 10 | 20 | 0.35000 | 0.73750 | 1.0875 | 1.69588 | 0.20638 | 5.65708 | 0.13037 | 1.69588 |
| 1 | 6 | 5 | 10 | 1 | 2.0 | 10 | 20 | 0.35000 | 0.92500 | 1.2750 | 1.69588 | 0.20638 | 6.57102 | 0.14077 | 1.69588 |

| 1 | 6 | 5 | 10 | 1 | 2.5 | 10 | 20 | 0.35000 | 1.11250 | 1.4625 | 1.69588 | 0.20638 | 7.42672 | 0.14980 | 1.69588 |
|---|---|---|----|---|------------|-----------|------|---------|---------|---------|---------|---------|---------|---------|---------|
| 1 | 6 | 5 | 10 | 1 | 2 | 10 | 20 | 0.35 | 0.925 | 1.275 | 1.69588 | 0.20638 | 6.57102 | 0.14077 | 1.69588 |
| 1 | 6 | 5 | 10 | 1 | 2 | 11 | 20 | 0.31818 | 0.925 | 1.24318 | 1.70984 | 0.18609 | 6.54281 | 0.14138 | 1.70984 |
| 1 | 6 | 5 | 10 | 1 | 2 | 12 | 20 | 0.29167 | 0.925 | 1.21667 | 1.72159 | 0.16942 | 6.51695 | 0.14194 | 1.72159 |
| 1 | 6 | 5 | 10 | 1 | 2 | 13 | 20 | 0.26923 | 0.925 | 1.19423 | 1.73162 | 0.15548 | 6.49316 | 0.14246 | 1.73162 |
| 1 | 6 | 5 | 10 | 1 | 2 | 14 | 20 | 0.25 | 0.925 | 1.175 | 1.74028 | 0.14366 | 6.47121 | 0.14294 | 1.74028 |
| 1 | 6 | 5 | 10 | 1 | 2 | 16 | 0.35 | 1.15625 | 1.50625 | 1.69588 | 0.20638 | 6.2151 | 0.18604 | 1.69588 | |
| 1 | 6 | 5 | 10 | 1 | 2 | 17 | 0.35 | 1.08824 | 1.43824 | 1.69588 | 0.20638 | 6.31524 | 0.17232 | 1.69588 | |
| 1 | 6 | 5 | 10 | 1 | 2 | 18 | 0.35 | 1.02778 | 1.37778 | 1.69588 | 0.20638 | 6.40731 | 0.16041 | 1.69588 | |
| 1 | 6 | 5 | 10 | 1 | 2 | 19 | 0.35 | 0.97368 | 1.32368 | 1.69588 | 0.20638 | 6.49229 | 0.14998 | 1.69588 | |
| 1 | 6 | 5 | 10 | 1 | 2 | 20 | 0.35 | 0.925 | 1.275 | 1.69588 | 0.20638 | 6.57102 | 0.14077 | 1.69588 | |

= Multiples of 10,000 messages/Second, \$ = Multiples of 10,000 Packets

nodes and the mean number of packets in the network decrease while the mean number of packets in the second buffer remain constant. As the transmission rate of node 2 (μ_2) varies from 16 to 20, the mean number of packets in the second node, utilization of the second node and the mean number of packets in the network decrease and the mean number of packets in the first buffer and utilization of the first node remain constant.

From equations (37), (38), (45) and (46), mean delays in the buffers and throughput of the nodes are computed for different values of a_1 , b_1 , a_2 , b_2 , λ_1 , λ_2 , μ_1 and μ_2 and are given in Table 1. The relationship between throughput of the nodes, mean delays in the buffers and the parameters a_1 , b_1 , a_2 , b_2 , λ_1 , λ_2 , μ_1 and μ_2 is shown in Figure 3 and Figure 4 respectively. From Table1, it is observed that when the batch size distribution parameters a_1 and b_1 varies from 1 to 5 and 6 to 10 respectively, the throughput of the first and

b_2 increase from 1 to 5 and 6 to 10 respectively. Similarly, when the message arrival rate λ_1 varies from 0.5×10^4 messages/sec to 2.5×10^4 messages/sec, the throughput of first and second nodes and also the mean delay in first buffer increase.

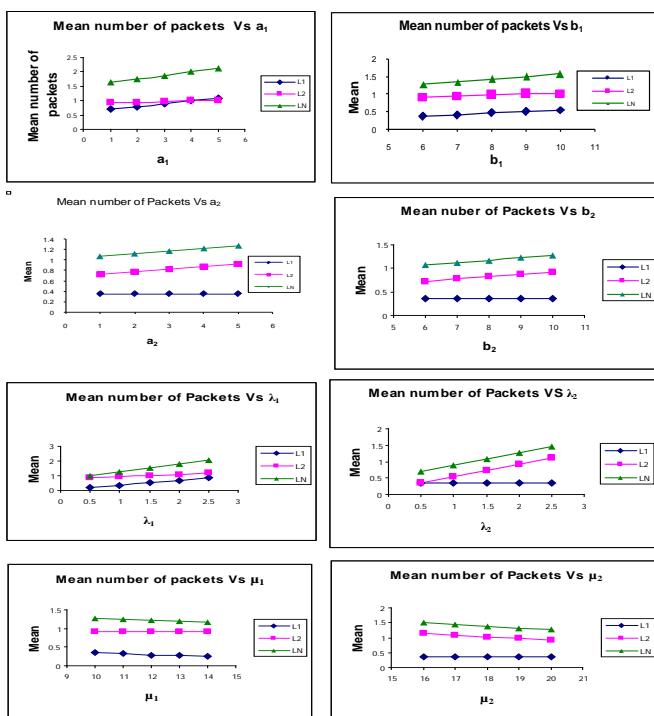


Fig. 2 Relation between Mean number of Packets and various input Parameters

second nodes and mean delay in the first buffer increase and the mean delay in second buffer decrease. The throughput of the first node and mean delay in the first buffer remain constant and the throughput of the second node and mean delay in the second buffer increase when the batch size distribution parameters a_2 and

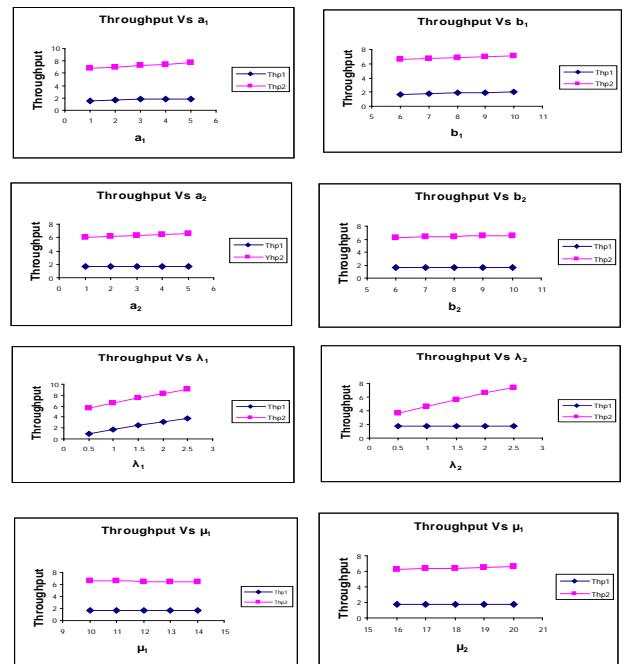


Fig. 3 Relation between Throughput of the nodes and various input Parameters

When the message arrival rate λ_2 varies from 0.5×10^4 messages/sec to 2.5×10^4 messages/sec, the throughput of the first node and mean delay in first buffer remain constant while the throughput of the second node and mean delay in the second buffer increase. Similarly, the impact of variation in other parameters on throughput and mean delay can be observed from the Table 1.

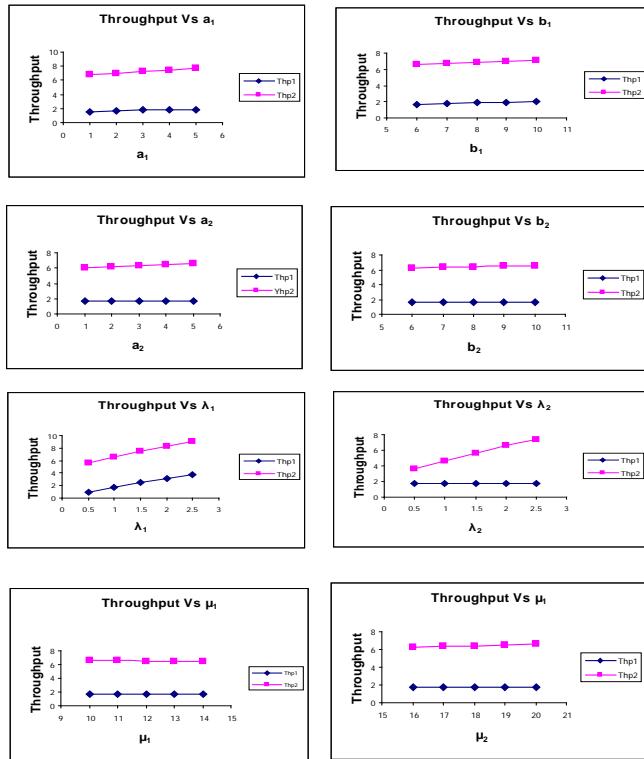


Fig. 4 Relation between Mean delays in the buffers and various input Parameters

6. Sensitivity Analysis

Table 2: Sensitivity Analysis

| Parameter | Performance Measure | % of Change in Parameters | | | | | | |
|------------------|----------------------|---------------------------|---------|---------|---------|---------|---------|---------|
| | | -15 | -10 | -5 | 0 | +5 | +10 | +15 |
| $\lambda_1 (=1)$ | L_1 | 0.595 | 0.63 | 0.665 | 0.7 | 0.735 | 0.77 | 0.805 |
| | L_2 | 0.89875 | 0.9075 | 0.91625 | 0.925 | 0.93375 | 0.9425 | 0.95125 |
| | Thp_1 | 1.35441 | 1.42152 | 1.48741 | 1.55208 | 1.61556 | 1.67788 | 1.73904 |
| | Thp_2 | 6.46091 | 6.56151 | 6.66136 | 6.76047 | 6.85884 | 6.95648 | 7.0534 |
| | $W(N_1)$ | 0.43931 | 0.44319 | 0.44709 | 0.45101 | 0.45495 | 0.45891 | 0.4629 |
| | $W(N_2)$ | 0.13911 | 0.13831 | 0.13755 | 0.13682 | 0.13614 | 0.13549 | 0.13486 |
| $\lambda_2 (=2)$ | L_1 | 0.35 | 0.35 | 0.35 | 0.35 | 0.35 | 0.35 | 0.35 |
| | L_2 | 0.8125 | 0.85 | 0.8875 | 0.925 | 0.9625 | 1 | 1.0375 |
| | Thp_1 | 1.69588 | 1.69588 | 1.69588 | 1.69588 | 1.69588 | 1.69588 | 1.69588 |
| | Thp_2 | 6.02989 | 6.21265 | 6.39301 | 6.57102 | 6.7467 | 6.92008 | 7.09119 |
| | $W(N_1)$ | 0.20638 | 0.20638 | 0.20638 | 0.20638 | 0.20638 | 0.20638 | 0.20638 |
| | $W(N_2)$ | 0.13475 | 0.13682 | 0.13882 | 0.14077 | 0.14266 | 0.14451 | 0.14631 |
| $\mu_1 (=10)$ | L_1 | 0.41176 | 0.38889 | 0.36842 | 0.35 | 0.33333 | 0.31818 | 0.30435 |
| | L_2 | 0.925 | 0.925 | 0.925 | 0.925 | 0.925 | 0.925 | 0.925 |
| | Thp_1 | 1.66922 | 1.67903 | 1.68787 | 1.69588 | 1.70317 | 1.70984 | 1.71596 |
| | Thp_2 | 6.61841 | 6.60187 | 6.58609 | 6.57102 | 6.55661 | 6.54281 | 6.52961 |
| | $W(N_1)$ | 0.24668 | 0.23162 | 0.21828 | 0.20638 | 0.19571 | 0.18609 | 0.17736 |
| | $W(N_2)$ | 0.13976 | 0.14011 | 0.14045 | 0.14077 | 0.14108 | 0.14138 | 0.14166 |
| $\mu_2 (=20)$ | L_1 | 0.35 | 0.35 | 0.35 | 0.35 | 0.35 | 0.35 | 0.35 |
| | L_2 | 1.08824 | 1.02778 | 0.97368 | 0.925 | 0.88095 | 0.84091 | 0.80435 |
| | Thp_1 | 1.69588 | 1.69588 | 1.69588 | 1.69588 | 1.69588 | 1.69588 | 1.69588 |
| | Thp_2 | 6.31524 | 6.40731 | 6.49229 | 6.57102 | 6.64418 | 6.71238 | 6.77613 |
| | $W(N_1)$ | 0.20638 | 0.20638 | 0.20638 | 0.20638 | 0.20638 | 0.20638 | 0.20638 |
| | $W(N_2)$ | 0.17232 | 0.16041 | 0.14998 | 0.14077 | 0.13259 | 0.12528 | 0.1187 |
| Parameter | Performance Measures | % Change in Parameters | | | | | | |
| | | -60 | -40 | -20 | 0 | +20 | +40 | +60 |
| | L_1 | 1.2 | 1.3 | 1.4 | 1.5 | 1.6 | 1.7 | 1.8 |
| | L_2 | 1.05 | 1.075 | 1.1 | 1.125 | 1.15 | 1.175 | 1.2 |

Sensitivity analysis of the network model is performed with respect to the parameters a_1 , b_1 , a_2 , b_2 , λ_1 , λ_2 , μ_1 and μ_2 on the mean number packets in the first and second buffers, the mean delays in the first and second buffers and also throughput of the first and second nodes. The computed values of the performance measures are given in Table 2. The following data has been considered for the sensitivity analysis.

$a_1=5$, $a_2=5$, $b_1=10$, $b_2=10$, $\lambda_1=1 \times 10^4$ messages/sec, $\lambda_2 = 2 \times 10^4$ messages/sec $\mu_1 = 10 \times 10^4$ packets/sec, $\mu_2 = 20 \times 10^4$ packets/sec.

The performance measures of the model are computed with variation of -15%, -10%, 0%, +5%, +10% and +15% on the input parameters λ_1 , λ_2 , μ_1 and μ_2 . A variation of -60%, -40%, -20%, 0%, +20%, +40% and +60% on the batch size distribution parameters a_1 and a_2 and -30%, -20%, -10%, 0%, +10%, +20% and +30% b_2 to retain them as integers. As λ_1 increases to 15%, the average number of packets in the two buffers increasing, the the average delay in the first buffer is increasing and the the average delay in the second buffer is decreasing. Similarly, as the batch size distribution parameter a_1 increase by 60%, the average number of

| | | | | | | | | |
|----------------------|--------------------|---------|---------|---------|---------|---------|---------|---------|
| a ₁ (=5) | Thp ₁ | 1.87975 | 1.94594 | 2.00055 | 2.04734 | 2.08843 | 2.12514 | 2.15835 |
| | Thp ₂ | 7.75525 | 7.95604 | 8.14075 | 8.31121 | 8.46902 | 8.61554 | 8.75196 |
| | W(N ₁) | 0.63838 | 0.66806 | 0.69981 | 0.73266 | 0.76613 | 0.79995 | 0.83397 |
| | W(N ₂) | 0.13539 | 0.13512 | 0.13512 | 0.13536 | 0.13579 | 0.13638 | 0.13711 |
| a ₂ (=5) | L ₁ | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 |
| | L ₂ | 0.975 | 1.025 | 1.075 | 1.125 | 1.175 | 1.225 | 1.275 |
| | Thp ₁ | 2.04734 | 2.04734 | 2.04734 | 2.04734 | 2.04734 | 2.04734 | 2.04734 |
| | Thp ₂ | 7.98407 | 8.11219 | 8.21896 | 8.31121 | 8.39283 | 8.46622 | 8.53304 |
| | W(N ₁) | 0.73266 | 0.73266 | 0.73266 | 0.73266 | 0.73266 | 0.73266 | 0.73266 |
| | W(N ₂) | 0.12212 | 0.12635 | 0.1308 | 0.13536 | 0.14 | 0.14469 | 0.14942 |
| | | -30 | -20 | -10 | 0 | +10 | +20 | +30 |
| b ₁ (=10) | L ₁ | 1.2 | 1.3 | 1.4 | 1.5 | 1.6 | 1.7 | 1.8 |
| | L ₂ | 1.05 | 1.075 | 1.1 | 1.125 | 1.15 | 1.175 | 1.2 |
| | Thp ₁ | 1.932 | 1.97402 | 2.01226 | 2.04734 | 2.07975 | 2.10984 | 2.13793 |
| | Thp ₂ | 7.85319 | 8.01636 | 8.16868 | 8.31121 | 8.44491 | 8.57061 | 8.68905 |
| | W(N ₁) | 0.62112 | 0.65856 | 0.69574 | 0.73266 | 0.76932 | 0.80575 | 0.84194 |
| | W(N ₂) | 0.1337 | 0.1341 | 0.13466 | 0.13536 | 0.13618 | 0.1371 | 0.1381 |
| b ₂ (=10) | L ₁ | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 | 1.5 |
| | L ₂ | 0.975 | 1.025 | 1.075 | 1.125 | 1.175 | 1.225 | 1.275 |
| | Thp ₁ | 2.04734 | 2.04734 | 2.04734 | 2.04734 | 2.04734 | 2.04734 | 2.04734 |
| | Thp ₂ | 8.0851 | 8.16696 | 8.24197 | 8.31121 | 8.37553 | 8.43558 | 8.49192 |
| | W(N ₁) | 0.73266 | 0.73266 | 0.73266 | 0.73266 | 0.73266 | 0.73266 | 0.73266 |
| | W(N ₂) | 0.12059 | 0.12551 | 0.13043 | 0.13536 | 0.14029 | 0.14522 | 0.15014 |

packets in the two buffers increasing and the the average delay in the first buffer is increasing. Overall analysis of the parameters reflects that dynamic bandwidth allocation strategy for congestion control tremendously reduces the delays in communication and improves quality of service by reducing burstiness in buffers.

7. Conclusions

In this paper, we developed and analyzed a two node Communication network model with Dynamic Bandwidth

Allocation (DBA) for bulk arrivals at both buffers. The statistical multiplexing of the Communication network is developed by characterizing the arrivals at both buffers connected to the two nodes in tandem as compound Poisson processes and transmission times with Poisson processes. This representation accurately matches the arrival and service process at Internet and Telecommunication processes. Using the Chapman Kulkmogrov transition equations, the joint probability generating function of the number of packets in each buffer is derived. The behavior of the network is analyzed by obtaining the system performance measures for any general bulk size arrival distribution and in particular uniformly distributed bulk arrivals of packets. It is observed that the bulk size arrival distribution parameters are significantly influencing the congestion, mean delays in buffers and throughput of the transmitters. The sensitivity analysis through the numerical studies reveals that the DBA strategy can reduce the burstness in buffers and improves the quality of service (QoS). This numerical model also includes several of the earlier Communication network models as particular cases for specific values of the input parameters. This Communication network model is much useful in analyzing the performance of several communication networks at Tele and Satellite communications, Computer communications, ATM scheduling, Bandwidth allocation etc. It is also possible to extend this network with non-Marchovian transmission times and priority structures which requires further investigation.

References

- [1]. Kin K. Leung (2002), Load-dependent service queues with application to congestion control in broadband networks. Performance Evaluation, Vol.50, Issue 1-4, pp 27-40.
- [2]. K. Srinivasa Rao, Prasad Reddy and P.SureshVarma (2006). Inter dependent communication network with bulk arrivals, International Journal of Management and Systems, Vol.22, No.3, pp. 221-234
- [3]. Emre Yetginer and Ezhan Karasan (2008), dynamic wavelength allocation in IP/WDM metro access networks, IEEE Journal on selected areas in Communications, Vol.26, No.3, pp 13-27.
- [4]. Gunadle, A.S and Yadri, A.R (2008), Performance evaluation of issues related to video over broadband networks, Proceedings of World Academy of Sciences, Engineering and technology, Vol.36, pp 122-125
- [5]. Hongwang Yu and yufan Zheng (2009), Global behavior of dynamical agents in direct network, Journal of control theory and applications, Vol.7, No.3, pp 307-314.
- [6]. Fen Zhou, Miklos Molnar and Bernard Cousin (2009), Avoidance of multicast incapable branching nodes for multicast routing in WDM, Photonic network communications, Vol.18, No.3, pp378-392.
- [7]. Stanislav Angelov, Sanjeev Khanna and Keshav Kunal (2009), The network as a storage device:Dynamic routing with bounded buffers, Algorithmica, Vol.55, No.1, pp 71-94.
- [8]. P.Suresh Varma and K.Srinivasa Rao (2007), A Communication network with load dependent transmission, International Journal of Matnemtical Sciences, Vol.6, No.2, pp 199-210.
- [9]. Kuda.Nageswara Rao et al. (2010), A Tandem Communication Network with Dynamic Bandwidth Allocation and Modified Phase Type Transmission having Bulk Arrivals, International journal of computer science issues, Vol.7, No.3, pp.18-26.

- [10]. Yukuo Hayashida (1993), Throughput analysis of tandem type go-back NARQ scheme for satellite communications, IEEE Transactions on Communications, Vol.41,pp 1517-1524.
- [11]. Ushio Sunita and Yasushi Masuda (1997) Tandem queues with Bulk arrivals, Infinitely many servers and correlated service times, Journal of applied probability, vol.34, No.1, pp.248-257.
- [12]. Gaujal, B. and Hyon, E.(2002), Optimal routing policies in deterministic queues in tandem, Proceedings of Sixth International Workshop on Discrete Event Systems (WODES'02),pp 251-257.
- [13]. Anyne Chen-Phil pollett-jumping Li-Hanjun Zhang (2010), Markovian Bulk arrivals and bulk-service queues with state-dependent control, queuing systems vol.64, pp.267-304.

Authors Profile



Dr. K.Srinivasa Rao is presently working as Professor and head, Department of Statistics, Andhra University, Visakhapatnam. He is elected chief editor of Journal of ISPS and elected Vice-President of Operation Research of India. He guided 22 students for Ph.D in Statistics, Computer Science, Electronics and Communications and Operations Research. He published 75 research papers in national and International journals with high reputation. His research interests are Communication Systems, Data Mining and stochastic models.



Mr. Kuda Nageswara Rao is presently working as Associate professor in the Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam. He presented several research papers in national and International conferences and seminars. He published a good number of papers in national and International journals. He guided several students for getting their M.Tech degrees in Computer Science and Engineering. His current research interests are Communication networks, Internet Technologies and Network security.



Dr. Peri. Srinivasa Rao is presently working as Professor in the Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam. He got his Ph.D degree from Indian Institute of Technology, Kharagpur in Computer Science in 1987. He published several research papers and delivered invited lectures at various conferences, seminars and workshops. He guided a large number of students for their M.Tech degrees in Computer Science and Engineering and Information Technology. His current research interests are Communication networks, Data Mining and Computer Morphology.

Fast Scalar Multiplication in ECC Using The Multi Base Number System

G. N. Purohit¹, Asmita Singh Rawat²

¹ Aim & Act, Department of Mathematics, Banasthali University
Jaipur, Rajasthan,304022, India

² Aim & ACT, Department of Computer Science, Banasthali University
Jaipur, Rajasthan,304022, India

Abstract

As a generalization of double base chains, multibase number system is very suitable for efficient computation of scalar multiplication of a point of elliptic curve because of shorter representation length and hamming weight. In this paper combined with the given formulas for computing the 7- Fold of an elliptic curve point P an efficient scalar multiplication algorithm of elliptic curve is proposed using 2,3 and 7 as basis of the multi based number system . The algorithms cost less compared with Shamirs trick and interleaving with NAFs method.

Keywords: *Scalar multiplication, Elliptic curve, Double base number system, Multibase number system, Double chain, Septupling.*

1. Introduction

Public key cryptography has been widely studied and used since Rivest, Shamir and Adleman invented the cryptography or cryptosystem RSA [1] in 1975. The system heavily depends on integer factorization problem [IFB] using large key bits of the order 1024 bits or 2048 bits . Later on Diffie- Hellman [2] developed the public key exchange algorithm using the discrete logarithmic problem [DLP]. Elgammel also used DLP in encryption and digital signature authentication [DSA] scheme. However, these conventional public key cryptographic systems, such as RSA and DSA are impractical in WSNs due to low processing power of sensor nodes. Koblitz [3] and Miller [4] independently used elliptic curves for cryptography using Elliptic curve Discrete Logarithmic Problem [ECDLP] and provided elliptic curve cryptographic [ECC].

In recent years ECC has received increased acceptance and has been included in standards room bodies such as ANSI,

IEEE, ISO and NIST. Compared to traditional cryptographic systems like RSA, ECC offers smaller key sizes and more efficient arithmetic, which results in faster computation, lower power consumption as well as memory and band width savings. Thus ECC is especially useful for mobile constrained devices like WSN, which enables wireless mobile devices to perform secure communication efficiently and establishes secure end to end connections.

In ECC, points on elliptic curves over finite fields are used to generate finite abelian groups to implement public key cryptographic primitives. Cryptosystems in ECC are based on the group of points on an elliptic curve over a finite field. They rely on the difficulty of finding the value of a scalar, given a point and the scalar multiple of that point. This corresponds to solving the discrete logarithm problem. However, it is more difficult to solve the Elliptic curve DLP than its original counterparts. Thus elliptic curve cryptosystems provide equivalent security as the existing public key cryptosystems, but with much smaller key lengths. In addition another benefit is that each user may select a different curve E even though the underlying field K remains the same for all users. Thus the hardware which depends on the field remains the same and the curve E can be changed periodically for extra security. Traditionally ECCs has been developed over finite fields which have either prime order or binary fields of order 2^m . The fundamental operation for generating a finite abelian group over an elliptic curve is the addition of two points on it. If point P on EC is added to itself $(k-1)$ times then we obtain a new point kP on elliptic curve and kP is termed as the scalar multiplication of point P by scalar k . Among the many arithmetic operations like addition, inversion, scalar multiplication involved in ECC, the scalar multiplication is the most important, energy and time consuming operation. A key factor for its fast implementation in ECC is to

compute the scalar multiplication efficiently, when k is a large integer. Various fast algorithms have been proposed for this purpose. Traditionally the integer k is represented in binary form and the double and add method is applied to calculate kP .

In this paper we first compute the 7-fold of an elliptic curve point P , i.e. $7P$. The formulas of doublings ($2P$), tripling($3P$), triple and add($3P+P$), quadrupling($4P$) , quadruple and add($4P+P$ and quintupling($5P$) are available in literature. Double base number representation of an integer in bases $\{2,3\}$, $\{2,5\}$ and $\{3,5\}$ and their generalizations to triple base representation base $\{2,3,5\}$ was recently reported in [5].

In this paper, an efficient scalar multiplication algorithms of a point P on an elliptic curve is proposed using triple base representation of the scalar using 2,3 and 7 as basis of the multibase number system. We obtain a sparser representation of the scalar, and the present algorithm costs less compared to the existing algorithms. We restrict our work on non super-singular elliptic curves defined over the field F_2m , however this can be suitably modified for any other type of elliptic curve.

The rest of the paper is organized as following: In the next section we report the related work. In Section-3 we evaluate sep-tupling $7P = (x_7, y_7)$ of a point

$P = (x, y)$ and calculate its cost in terms of multiplications, squaring and inversions. The costs of addition and subtraction are ignored which are negligible in comparison to other costs. The triple base representation of an integer is in section 4. Multi base number representation (MBNR) and multi base chain representation and their implementation in scalar multiplication are discussed in section 5 and 6 respectively. Concluding remarks are given in the end.

3. Related Work

The classical approach of representing the integer k in binary form and then performing the scalar multiplication by a standard double and add method has efficient triple ($3P$) and double($2P$) of point P , a ternary (binary approach for fast scalar multiplication is presented in [6]. For general curves a DBNS representation of the scalars using 2 and 3 as bases has been proved quite efficiently [7]. For last couple of years double base number system [DBNS] has been proposed to be used in this context by several authors [8, 9, 10, 11]. In search of sub linear scalar multiplication algorithms, authors of [8] have been used complex bases, 3 and τ for Koblitz curves. As a new approach for fast scalar multiplication, point halving was proposed independently by Knudsen [12] and Schroepel

[13]. They suggested that point doubling in the double and add method can be replaced by a faster point halving operation. A detailed analysis of the speed advantage of employing point halving instead of point doubling is available in [9]. Further point halving can be combined with frobenius endomorphism so as to speed up the corresponding operation in Koblitz curve by 25 percent [14, 15]. In yet another development the double base number representation of integer was generalized to multibase number representation with 2, 3 and 5 as basis elements and is included in [16,17]. The efficient scalar multiplication using multibase number representation included in [16] which also includes quintuple formula. Multibase multiplication using MBNR is included in [17], Scalar multiplication combining MBNR with point halving is discussed in [18].

Our contribution in this paper is computing 7 fold ($7P$) of an elliptic curve point P for a curve over binary field and using the same in scalar multiplication. The scalar multiplication uses the representation of the scalar as sum/ difference of product of powers 2, 3 and 7.

4. Septupling

In this section we consider Sep-tupling ($7P$) of a point P on an elliptic curve. We begin with a discussion of an elliptic curve.

4.1: Elliptic Curve

An elliptic curve over a finite field GF (Galois field) K is defined by an equation

$$E : y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6 \quad , \quad (1)$$

where $a_1, a_2, a_3, a_4, a_5, a_6 \in K$ are the parameters of the curve and $\Delta \neq 0$, Δ being the discriminant of the curve E .

In the case of binary field $K = F_2m$, the non- super singular curves are used for cryptography, whose Weierstrass equation can be simplified to the form.

$$y^2 + xy = x^3 + ax^2 + b \quad \text{Where } a, b \in F_2m \quad \text{and } \Delta = b \neq 0 . \quad (2)$$

If $P = (x_1, y_1)$ and $Q = (x_2, y_2)$ are two points on E (F_2m) then their sum $P + Q$ is also a point (x_3, y_3) on E , where x_3 and y_3 are given by.

$$\begin{aligned} x_3 &= \lambda^2 + \lambda + x_1 + x_2 + a \\ y_3 &= \lambda(x_1 + x_3) + x_3 + y_1 , \end{aligned} \quad (2)$$

where $\lambda = \frac{y_1 + y_2}{x_1 + x_2}$

Further double of P i.e. $2P$ is also a point (x_4, y_4) on curve E, where

$$x_4 = \mu^2 + \mu + a = x_1^2 + \frac{b}{x_1^2}$$

$$y_4 = x_1^2 + \mu x_4 + x_4,$$

$$\mu = x_1 + \frac{y_1}{x_1}$$

The usual scalar multiplication kP of P by scalar k is obtained using the above described two operations add and double. For example $25P$ is calculated as

$$25P = 2(2(2(P + 2P)) + P)$$

or

$$2(2(2(2P) + P)) + (2(2P) + P)$$

These group operations in affine coordinates required field inversion besides multiplication and squaring. We denote by i 's and m the cost of one inversion , one squaring and one multiplication respectively. The cost of additions of two points $P + Q$ and of double of a point P , $2P$ are equal and equals to $i+2m$. However we shall neglect the cost of field additions in case of elliptic curves over binary fields. It may be noted that cost of squaring in case of binary fields is almost free. The cost of a repeated doubling $w - DBLP = 2^w P$ is $+ (4w - 2)m$ as reported in [8]. The costs of : (i)double and add, $DA(P, Q) \rightarrow 2P \pm Q$,

repeated tripling, $wTPR(P) \rightarrow 3^w P$ and (iii) triple and add $TA(P, Q) \rightarrow 3P \pm Q$ are given in [8] as i) $i+am$, ii) $i+7m$ and (iii) $2i+am$, respectively.

4.2 Point Septupling

Let P be (x, y) be a point on an elliptic curve given by equation () over a binary field. We shall calculate the 7-fold of P given by, $7P = (x_7, y_7)$, that is we shall obtain expression for x_7 and y_7 in terms of x and y .

For non-super singular curves over a binary field, the division polynomials are given by

$$\Psi_1 = 1$$

$$\Psi_2 = x$$

$$\Psi_3 = x^4 + x^3 + a = A$$

$$\Psi_4 = x^6 + ax^2 = x^2(A - x^3) = B$$

The higher degree division polynomials are obtained using the following recurrence relations:

$$\Psi_{2n+1} = \Psi_{n+2}\Psi_n^3 - \Psi_{n-1}\Psi_{n+1}^3$$

$$\Psi_2\Psi_{2n} = \Psi_{n+2}\Psi_n\Psi_{n-1}^2 - \Psi_{n-2}\Psi_n\Psi_{n+1}^2$$

Using these relations we obtain

$$\Psi_5 = \Psi_4x^3 - \Psi_3^3 = Bx^3 - A^3 = C$$

$$\Psi_6 = \frac{\Psi_5\Psi_3x^2 - \Psi_3\Psi_4^2}{x} = \frac{CAx^2 - AB^2}{x} = D$$

$$\begin{aligned} \Psi_7 &= \Psi_3 + 2\Psi_3^3 - \Psi_2\Psi_4 \\ &= A + 2A^3 - xB^3 = E \end{aligned}$$

$$\begin{aligned} \Psi_8 &= \frac{\Psi_6\Psi_4\Psi_3^2 - \Psi_2\Psi_4\Psi_5^2}{\Psi_2} \\ &= \frac{DBA^2 - xBC^2}{x} = F \end{aligned}$$

For any point $P(x, y)$ on E, its n-fold $n(3P)$ is given by

$$\begin{aligned} [n]P &= (x + \frac{\Psi_{n+1}\Psi_{n-1}}{\Psi_n^2}, y + x + \frac{\Psi_{n+1}\Psi_{n-1}}{\Psi_n^2} + \frac{\Psi_{n+1}^2\Psi_{n-2}}{\Psi_2\Psi_n^3} + \\ &\quad (x^2 + y)\frac{\Psi_{n+1}\Psi_{n-1}}{\Psi_2\Psi_n^2}) \end{aligned}$$

So the value of (x_7, y_7) for the 7-fold over binary field.

Thus one can be computed from the above equation as follows:

$$x_7 = x + \frac{\Psi_8\Psi_6}{\Psi_7^2}$$

$$x_7 = x + \frac{FD}{E^2}$$

$$y_7 = x + y + \frac{\Psi_8\Psi_6}{\Psi_7^2} + \frac{\Psi_8^2\Psi_5}{\Psi_2\Psi_7^3} + (x^2 + y)\frac{\Psi_8\Psi_6}{\Psi_2\Psi_7^2}$$

$$y_7 = x + y + \frac{FD}{E^2} + \frac{F^2C}{xE^3} + (x^2 + y)\frac{FD}{xE^2}$$

The cost of evaluating various polynomials defined above:

| Polynomials | Operations |
|---|----------------------|
| $A = x^4 + x^3 + a$ | $2[s] + 1[m]$ |
| $B = x^6 + ax^2 = x^2(A - x^3)$ | $1[m]$ |
| $C = \Psi_5 = \Psi_4x^3 - \Psi_3^3$ $= Bx^3 - A^3$ | $1[s] + 2[m]$ |
| $D = \frac{CAx^2 - AB^2}{x}$ $+ 1[s]$ | $4[m] + 1[i]$ |
| $E = A + 2A^3 - xB^3$ | $1[m] + 1[s]$ |
| $F = \frac{DBA^2 - xBC^2}{x}$ | $3[m] + 1[s]$ |
| $\frac{FD}{E^2}$ $1[i]$ | $1[m] + 1[s] +$ |
| $\frac{F^2C}{xE^3}$ | $1[s] + 4[m] + 1[i]$ |

Thus the total cost of the hepta tupling is $3[i] + 7[s] + 18[m]$. Neglecting the cost of squaring (in case of EC over binary fields) the total cost turns out to be $3[i] + 18[m]$. We can also compute $7P$ as $2(2P)+3P$ or $2(3P)+P$. Using the generic method the costs of $TPL(P)$ and $DBL(P)$ are respectively $i+7m$ and $i+2m$. Further the costs of $DA(P, Q)$ are $2i+9m$. Hence the total cost $7P=2(3P)+P=4i+18m$. If we consider $7P$ as $2(2P)+3P$ then the total cost is $5i+20m$. Hence cost calculated by us is the least.

The following table represents the costs of different operations used for the efficient scalar multiplication using the binary field method.

Table 2 : table of costs for different operations

| S.No. | Operations | Binary field costs |
|-------|------------|--------------------|
| 1 | $P+Q$ | $1I + 1S + 2M$ |
| 2 | $2P$ | $1I + 1S + 2M$ |
| 3 | $2P+Q$ | $1I + 2S + 9M$ |
| 4 | $3P$ | $1I + 4S + 7M$ |
| 5 | $3P+Q$ | $2I + 3S + 9M$ |
| 6 | $4P$ | $1I + 5S + 8M$ |
| 7 | $5P$ | $1I + 5S + 13M$ |
| 8 | $7P$ | $3I + 7S + 18M$ |

5.Multibase Number Representation (MBNR)

First we review double base number system (DBNS)

5.1 DBNS

Improving the classical methods of double and add for scalar multiplication a new method (DBNS), using bases besides 2, were introduced [2, 3, 5]. In this system one can represent k as the sum of terms of the form $s_i 2^{bi} 3^{ci}$, where $s_i \in \{-1, 1\}$, such representation always exists and in fact this number system is quite redundant. One of the most interesting properties of the representation is that among all the possible representation for a given integer, some of them are really sparse, that is to say that the number of non-zero terms is quite low.

To compute DBNS representation of an integer, one usually uses a greedy algorithm. It consists of the following: find the closest integer of the form $2^{bi} 3^{ci}$ to k , subtract it from k and repeat the process with $k' = k - 2^{bi} 3^{ci}$ till it is equal to zero. Performing a point scalar multiplication using this number system is relatively easy. Letting k be equal to $\sum_{i=1}^n s_i 2^{bi} 3^{ci}$ one just needs to compute $[s_i 2^{bi} 3^{ci}] P$ for $i=1$ to n and then add all the points.

Example:

$$895712 = 2^{10} 3^6 + 2^9 3^5 + 2^8 3^4 + 2^7 3^3 + 2^6 3^2 + 2^5 3^1 + 2^4 3^0 + 2^1 3^0.$$

Even if the number of additions is quite low, in practice such a method requires too many doublings and triplings. For this reason the general DBNS representation has been considered to be not suitable for scalar multiplication.

To overcome this problem the concept of double base chains was introduced in [3]. In this system, an integer k is still represented as $\sum_{i=1}^n s_i 2^{bi} 3^{ci}$, but with the restriction that allowing a Horner like evaluation of kP using only doublings and triplings, however, with significantly increase in the number of point additions.

5.2 Multibase Number Representations (MBNR)

Let $B = \{b_1, b_2, \dots, b_l\}$ be set of small integers. A representations of integer k as a sum of powers of elements of B of the form $k = \sum_{j=1}^m s_j b_1^{cj_1} \dots b_l^{cj_l}, s_j \in \{-1, 1\}$ is called a multibase representations of k using the base B . The integer m is the length of the representation of k using the base B . The integer m is the length of the representation or Double base number system (DBNS) or double base number representation discussed in previous section.

(DBNR) is a special case with with $|B| = 2$. In this paper we are particularly interested in multibase representation with $B = \{2, 3, 7\}$. The multi base representations with $B = \{2, 3, 5\}$ have been discussed by many authors [4, 6]. Authors in [17] combined with MBNR with point halving.

The double base number system is highly redundant. Further these representation are very short in length, a 160bit integer can be represented using around 23 terms using the base $B = \{2,3\}$. The results on length of DBNS representation are included in [2]. The multi base representation is even shorter and more redundant than the DBNS. The same 160 bit integer can be represented using around 15 terms using a triple base $B = \{2, 3, 5\}$.

Example:

$$895712 = 2^4 3^7 5^2 + 2^4 3^5 5^1 + 2^4 3^4 5^0 + 2^1 3^4 5^0 + \\ 2^0 3^2 5^0 + 2^0 3^1 5^0$$

The multi base representation of a number using a triple base $B = \{2, 3, 7\}$ is even shorter and sparse as compared to its representation using the triple base $\{2,3,5\}$.

Example:

$$895712 = 2^9 3^5 7^1 + 2^7 3^3 7^1 + 2^5 3^1 7^1 + 2^5 3^1 7^0$$

In this article, unless otherwise stated, by a multi base representation of k , we mean a representation of the form.

$$k = \sum_i s_i 2^{bi} 3^{ci} 7^{di}$$

Where $s_i \in \{-1,1\}$ and the terms of the form $2^{bi} 3^{ci} 7^{di}$ will be termed as 3-integers. A general multibase representation although very short is not suitable for a scalar multiplication algorithm. So we include a special representation with restricted exponents.

Definition: A multi base representation $k = \sum_i s_i 2^{bi} 3^{ci} 7^{di}$ using the base $B = \{2, 3, 7\}$ is called a step multibase representation (SMBR) if the exponents $\{b_i\}, \{c_i\}$ and $\{d_i\}$ form three separate monotonic decreasing sequence.

We consider an example illustrating this definition for the same number.

Example:

$$895712 = 2^5 3^4 7^3 + 2^4 3^2 7^2 - 2^3 3^0 7^2 - 2^3 3^0 7^0$$

An integer k has several SMBR, the simplest one being the binary representation. If k is represented in SMBR, then we can write it using Horner's rule and an addition chain, like double base chain in [1], for scalar multiplication can easily be developed. In case of our base system $\{2,3,7\}$, we require b_1 doublings, c_1 tripling and d_1 sep tuplings. An integer can be converted to a multi base representationwith base $\{2,3,7\}$ using the Greedy Algorithm as:

GREEDY ALGORITHM:

while $k > 0$

let z be the largest integer $2^b 3^c 7^d$

Output(b, c, d)

replace k by $k-z$

$k - z \leftarrow 0$

else

end.

In this process the pre-computed points are extensively used to accelerate the scalar multiplication in applications where extra memory is available. So we have used a new method multi base chain representation in which one does not require any pre-computations but in this method the expansion of the scalar reduces the cost of the scalar multiplication making it faster.

The important contribution in [7] was the new ternary-binary method to perform the efficient scalar multiplication. Ciet et al.[7] have proposed a ternary-binary method for fast ECC scalar multiplication. It makes use of efficient doubling (2P), tripling (3P), quadrupling (4P). In this paper a new septenary /ternary /binary approach for fast ECC scalar multiplication is proposed, which makes the use of septupling (7P) for the efficient scalar multiplication.

In this base system only b_1 doublings, c_1 tripling and d_1 sep- tuplings are needed for the scalar multiplication; in the next section we give implementation of this method and develop Septupling, 7P, for point P .

6. Scalar Multiplication Implementation and Algorithm.

We have already suggested that an integer k can be represented in multi base number system as the sum or difference of the mixed powers of 2, 3 and 7, as given in the following equation

$$k = \sum_i s_i 2^{bi} 3^{ci} 7^{di} \quad \text{with } s_i \in \{-1,1\} \quad \text{and} \\ b_i, c_i, d_i \geq 0$$

The sequence of the binary and ternary exponents decreases monotonically, i.e.

$$b_1 \geq b_2 \geq b_3 \dots \geq b_m \geq 0, \quad c_1 \geq c_2 \geq c_3 \dots \geq c_m \geq 0$$

and $d_1 \geq d_2 \geq d_3 \dots \geq d_m \geq 0$, and thus a multi base chain is formed.

For implementing the scalar multiplication we use a recursive formula for the fast computation of scalar multiplication using following equation for recursive calculations.

$$K_1 = 1, \quad K_i = 2^u 3^v 7^w K_{i-1} + s_i \quad \text{with } i \geq 2,$$

$$s_i \in \{-1, 1\},$$

where u is the difference of two consecutive binary exponents, v is the difference of two consecutive ternary exponents and w is the difference of two consecutive septenary exponents.

To implement it we have used the following algorithm.

An integer k , can be converted to a multi base representation

$$k = \sum_i s_i 2^{b_i} 3^{c_i} 7^{d_i} \quad \text{with } s_i \in \{-1, 1\} \quad \text{and}$$

$$b_i, c_i, d_i \geq 0 \text{ using greedy algorithm as already explained in Section 5. Now we describe the algorithm:}$$

ALGORITHM:

Input: An integer $k = \sum_{i=1}^m s_i 2^{b_i} 3^{c_i} 7^{d_i}$,

$$s_i \in \{-1, 1\}$$

And such that $b_1 \geq b_2 \geq b_3 \dots \geq b_m \geq 0$,
 $c_1 \geq c_2 \geq c_3 \dots \geq c_m \geq 0$ and
 $d_1 \geq d_2 \geq d_3 \dots \geq d_m \geq 0$, and a point $P \in E(F_2 m)$.

Output: the point $kP \in E(F_2 m)$

$$Z \leftarrow s_1 P$$

```

for i = 1, ..., m - 1 do
    u <- b_i - b_{i+1}
    v <- c_i - c_{i+1}
    w <- d_i - d_{i+1}
    if u = 0 then
        Z <- 7^w Z
    if v ≠ 0 then
        Z <- 3(3^{v-1} Z) + s_{i+1} P      // TA used here
    else
        Z <- Z + s_{i+1} P
    else

```

```

Z <- 7^w Z
Z <- 3^v Z
Z <- 2^{u-1} Z           // DA is used here
Z <- 2Z + s_{i+1} P

```

Return Z

As an example for illustration of this algorithm we consider computing $895712P$. We first develop the multi base chain as given below.

$$895712 = 2^5 3^4 7^3 + 2^4 3^2 7^2 - 2^3 3^0 7^2 - 2^3 3^0 7^0$$

and compute $127P$, $2285P$, $111964P$ and finally $895712P$ successively.

Table2: Method of calculating $895712P$ in different iterations for sep-tupling

| i | K | s | u | v | w |
|---|-------------------------------|----|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 0 |
| 2 | $126 K_1 + 1 = 127$ | 1 | 1 | 2 | 1 |
| 3 | $18 K_2 - 1 = 2285$ | -1 | 1 | 2 | 0 |
| 4 | $\frac{49}{K_3 - 1} = 111964$ | -1 | 0 | 0 | 2 |
| 5 | $8 K_4 = 895712$ | 0 | 3 | 0 | 0 |

This algorithm has used a multibase representation of the scalar with 2, 3 and 7 as the base numbers and it uses group operation like ADD, DBL, w-DBL, DA, TA for efficient computation. The new multi base chain method and proposed septenary/ternary/binary method is much faster than any other methods mentioned above for scalar multiplication for the binary fields without requiring any pre computations.

7. Conclusion

In this paper we have presented fast and secure scalar multiplication algorithms. In our work we have proposed a new algorithm for MBNR representation of an integer and combining with the scalar multiplication. We have shown that the length of the MBNR is shorter than the

DBNR and is also more redundant, since the number of representation grows faster as the number of base element is higher. For the MBNR representation we have used 2, 3 and 7 as the bases which makes the representation sparser

8. References

- [1]. Rivest, R. Shamir,A, & Adleman A. (1978). "A method for obtaining digital signature and public key cryptosystems". Communication of the ACM, vol.21, pp120-126
- [2]. Diffie, W.Hellman, M.E. (1976). "New directions in cryptography", IEEE Transactions Information theory, IT-22(6).
- [3]. Koblitz N.(1987). "Elliptic curve cryptosystems.Maths computing 48", (177) pp. 203-209
- [4]. Miller, V. (1986). "Use of elliptic curves in cryptography .Advances in cryptology"- Crypto'85 pp 417-426.
- [5] P. K. Mishra, V. S.Dimitrov. "Efficient Quintuple Formulas for Elliptic Curves and Efficient Scalar Multiplication Using Multibase Number Representation". Springer-Verlag, 2007, volume 4779, pages 390-406.
- [6].F.Morian, J.olivos, (1990). "Speedong up computation on an elliptic curve using addition – subtraction chains", Information theory applications, vol.24, pp 531-543.
- [7].M. Ciet, M. Joye, K. Lauter, P.L. Montgomery,(2003) "Trading Inversions for Multiplications in Elliptic Curve Cryptography",Cryptology ePrint Archive, Report 2003/257 . Also to appear in Design, Codes and Cryptography
- [8] V. Dimitrov, L. Imbert and P.K. Mishra,(2005) "Efficient and Secure Elliptic Curve Point Multiplication using Double-Base Chains," Advances in Cryptology - ASIACRYPT'05, LNCS Vol. 3788, pp. 59-78, Springer-Verlag, 2005.
- [9]. K.W. Wong, Edward, C.W.Lee, L.M.Cheng, Xiaofeng Liao,(2006), "Fast Scalar Multiplication using new Double Base Chain and Point Halving", Applied Mathematics and Computation.
- [10].Avanzi, R.M and Sica, F(2006). "Scalar multiplication on Koblitz curves using Double bases". Technical Report Available at <http://eprint.iacr.org/2006/067>.
- [11].V. Dimitrov ;K.V.Jarvanian ,M.J.Jacobian , W.F.Chan and Z.Huang,(2006). "FGPA implementation of point multiplication on Koblitz curves Using Kleinian integers". LNCS 4249 pp. 445-459, Springer Verlag.
- [12].C.Doche and L.Imbert. "Extended Double Base Number Systems with applications to elliptic Curve Cryptography" , Available at <http://eprint.iacr.org/2006/330>.
- [13].M.Ciet and F.Sica (2005). "An Analysis of Double base Number system and a sub linear scalar multiplication Algorithm". LNCS Vol.3715 pp. 171-182. Springer Verlag.
- [14]. E.W.Knudsen (1999). "Elliptic scalar multiplication using point halving", LNCS Vol.1716 pp 135-149.
- [15]. R.Schroeppel.(2000) "Elliptic curve Point Ambiguity Resolution Apparatus and method ." International patent Application Number PCT/US00 31014, field November 9.
- [16].R.M.Avanzi,C.Henbergerand,,H.Prodinger(2005)."Minimality of the hamming weight of the τ -NAF for Koblitz Curve and Improves combination with point halving". Cryptology eprint archive, Report 2005/225.
- [17].R.M.Avanzi, M. Ciet, F.Sica (2004). "Faster Scalar multiplication on Koblitz Curves Combining Point halving with the Frobenius Endomorphism" , LNCS Vol.2947, pp.28-40.
- [18].A.M.Ismail, M.R.MD Said, K.A.Mohd Atan , I.S.Rakhimov(2010). "An Algorithm to enhance Elliptic Curves scalar Multiplication Combinig MBNR with point halving", Applied Mathematical sciences, Vol.4,pp.1259-1272.
- [18].D.Bernstein, P.Birkner, P.Longa, C.Peters. "Optimizing Double Base Elliptic Curve Single Scalar Multiplication". LNCS Vol. 4859, pp.167-182, Springer Verlag.
- [19].P.Longa (2007): "Accelerating the scalar multiplication on Elliptic curve Cryptosystems over prime fields" . Master thesis University Of Ottawa, <http://patriconiga.bravehost.com/publications.html>.
- [20] R. Dahab and J. Lopez (1998), "An Improvement of Guajardo-Paar Method for Multiplication on non Super Singular Elliptic Curves".In Proceedings of the XVIII International Conference of the Chilean Computer Science Society (SCCC'98), IEEE CS Press, November 12-14, Anto Fagasta, Chile, pp.91-95.
- [21].V.S.Dimitrov,L.Imbert, and P.K.Mishra,(2005) " Fast elliptic Curve Point Multiplication using Double-Base Chain", Cryptology ePrint Archive , Report 2005/069.

Prof. G. N. Purohit. He is a Professor in Department of Mathematics & Statistics at Banasthali University (Rajasthan).Before joining Banasthali University, he was Professor and Head of the Department of Mathematics, University of Rajasthan, Jaipur. He had been Chief-editor of a research journal and regular reviewer of many journals. His present interest is in O.R., Discrete Mathematics and Communication networks. He has published more 40 research papers in various journals.

Asmita Singh Rawat received the B.Sc degree from University Of Lucknow and M.C.A degree from U.P Technical University in 2006 and 2009, respectively. She is currently working towards a PhD degree in computer Science at the Banasthali University of Rajasthan. Her research interests include wireless sensor network security with a focus on the elliptic curve cryptography.

Accurate methods of calculating the Coronary Sinus Pressure plateau

Loay Alzubaidi

Department of Computer Science, Prince Mohammad bin Fahd University
 AL-Khobar, Saudi Arabia

Abstract

Salvage of ischemic myocardium during the intervention with Pressure Controlled Intermittent Coronary Sinus Occlusion (PICSO) has been described to be effective during experimental models of coronary artery occlusions but the standard calculation method of the CSP quantities in current use gives an approximate calculation. A novel computation method (dp/dt method) was introduced and evaluated to describe the rise/release of CSP. The new method is a more accurate way of calculating the systolic and diastolic plateau and the rise time by determining the slope of CSP. This method results in a time derivative that describes the changes in CSP measurements. The results are shown to bear a close resemblance to the clinical effect of coronary sinus occlusion. The Haemodynamic Quantities which calculated using the CSP method (T90 method) 'evaluated by Schreiner' will be compared with the Haemodynamic Quantities which calculated using the new method (dp/dt method).

Keywords: coronary sinus pressure, PICSO, mathematical model, intermittent occlusion

1. Introduction

Pressure controlled intermittent coronary Sinus occlusion (PICSO) is considered to be a physiological procedure for salvaging myocardium at risk after infarction or intra operative arrest of the heart. Intermittent occlusion of the coronary sinus by means of an inflatable balloon catheter temporarily obstructs the outflow from the cardiac veins in right atrium and thus leads to an increase of CSP (systolic as well as diastolic) in the course of a few heart beats. This back-pressure retrograde forces blood into regions deprived of regular perfusion after coronary artery infarction. After a few seconds CSP comes to a plateau (systolic higher than diastolic) the height of which is determined by the coronary artery input pressure as well as the myocardial power to squeeze blood into the coronary venous system. Because the beneficial effect of this intervention appears to be closely linked to exact application, mathematical models have been developed over many years in order to put the estimation of occlusion and release times on a quantitative basis.

$$P_{CSP}(t) = \begin{cases} A * \exp(B * [1 - \exp(-C * t)]) - 1 & \text{when } 0 < t < T1 \\ D * \exp(E * [1 - \exp(-\frac{F}{t})]) - 1 & \text{when } T1 \leq t < T2 \end{cases} \quad (1)$$

Where

P_{CSP}(t) = Coronary sinus pressure (mmHg)

t = Time (s), measured from the start of occlusion

A, D = fitting parameter in (mmHg)

B, E = fitting parameter (dimensionless)

C, F = Fitting parameter (1/s)

T1 = Time that mark the end of the CSP occlusion phase (s)

T2 = Time that mark the end of the CSP release phase (s)

The first part (equation 1a) describes the rise of the CSP during the Inflation (occlusion) time.

$$P_{CSP}(t) = A * \exp(B * [1 - \exp(-C * t)]) - 1 \quad (1a)$$

The second part (equation 1b) describes the release of the CSP during the deflation (release) time.

$$P_{CSP}(t) = D * \exp(E * [1 - \exp(-\frac{F}{t})]) - 1 \quad (1b)$$

The systolic and diastolic peaks were fitted with the nonlinear least-square algorithms as shown in the Fig. (1)

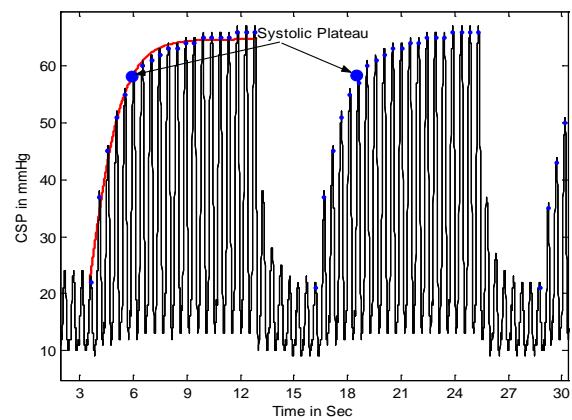


Fig. 1 Systolic Plateau of CSP using T90 method, the solid curve represents the fitted three parameter model functions, and the circle represents the systolic plateau

2. Method

The time derivative of CSP (dp/dt) describes the changing in the CSP quantities. The derivative of equation (1a) results in

$$\frac{dp}{dt} = A * B * C * \exp(-C * t) * \exp(B * (1 - \exp(-C * t))) - 1 \quad (2)$$

The time derivative of CSP has two types of peaks, positive

peaks (dp/dt Max) and negative peaks (dp/dt Min) as shown in the Fig. (2). These peaks were fitted separately with the least square algorithms using new fitting parameters α , β and γ

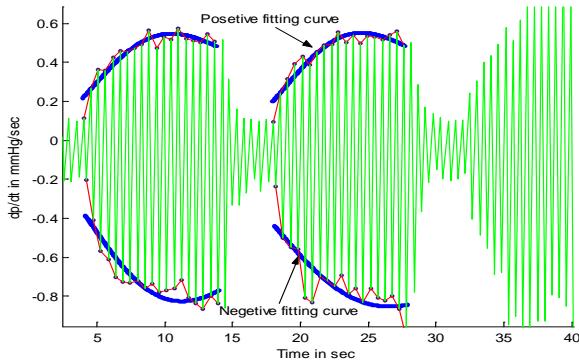


Fig. 2 Fitting of the dp/dt of CSP using the the least square algorithms

The first derivative equation (2) will be fitted with new fitting parameters α , β and γ using the non linear least square algorithms.

$$dp/dt = \alpha * \beta * \gamma * \exp(-\gamma * t) * \exp(\beta * (1 - \exp(-\gamma * t)) - 1) \quad (3)$$

Where

dP/dt = first derivative of CSP to time (mmHg/s)

α =fitting parameter in (mmHg)

β = fitting parameter (dimensionless)

γ = Fitting parameter (1/s)

Due to the intrinsic shape of the model for dp/dt rise there is an inflection point marking the steepest slope (represent the rise time). This inflection point can be calculated during algebraic manipulation by setting the 2nd derivative to zero

$$\frac{d^2 P_{csp}(tr)}{dt^2} = 0 \quad (4)$$

The rise time (tr) will be calculated by solving equation (4), this leads to

$$tr = (1/\gamma) * \ln(\beta) \quad (5)$$

Fig. (3) Shows the inflection point and its relationship to the fitted curve of the CSP and the fitted curve of the dp/dt , this point represent the rise time, that the CSP needs to reach its systolic plateau.

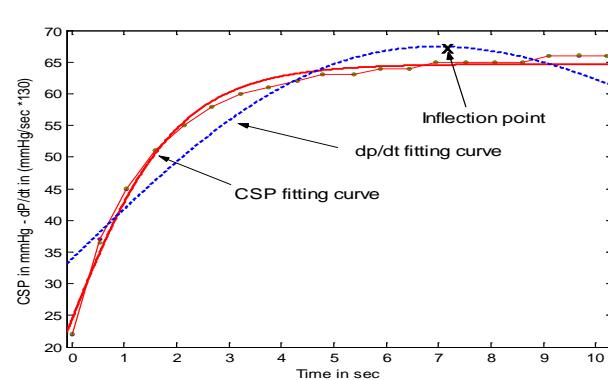


Fig. 3 Relationship between CSP and the dp/dt the solid curve represent the fitting of the CSP peaks during the inflation time, and the dash curve represents the fitting of the positive peaks of the dp/dt

The systolic plateau will be calculated by inserting tr into equation (1a), and it will be expressed in terms of two fitted parameters groups, this will gives

$$P_{csp}(tr) = A * \text{Exp}\{B * [1 - \text{Exp}(-C * (1/\gamma) * \ln(\beta))] - 1\} \quad (6)$$

Fig. (4) Shows systolic plateau and the rise time calculated from the mathematical model of the systolic rise using T90 method and dp/dt method. The derived quantities serve as diagnostic parameters for a quantitative assessment of physiological condition and as predictors for an optimal adjustment of coronary sinus cycles.

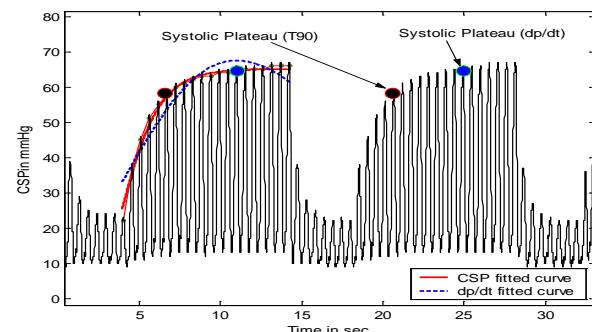


Fig. 4 The CSP rise fitted curve and the dp/dt positive peaks fitted curve with the Systolic plateau of T90 method

3. Results

The model parameters and the derived quantities will change with time. For any diagnostic value it is essential to establish ranges which can be used as reference intervals for the normal state. The following results comprise a preliminary investigation of the spread of the derived quantities observed during PICSO. Over 1000 PICSO was calculated using an automatic computation module, the Haemodynamic quantities was calculated for each PICSO. The rise time describes how

long it takes to reach a pressure plateau after a prolonged occlusion cycle. The systolic plateau and its rise time were used to compare between T90 and dp/dt method. Fig. (5) shows the result of both calculations, it's very clear that the systolic plateau of dp/dt method higher than the T90 method.

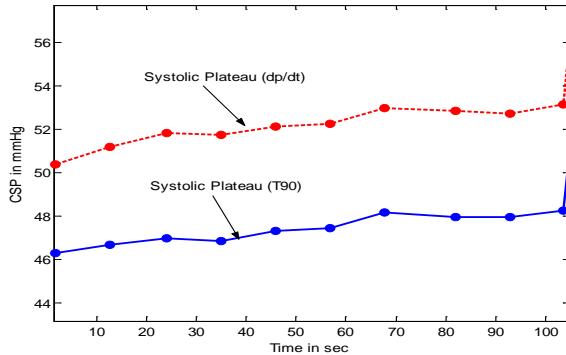


Fig. 5 Compare between the calculated systolic plateau of the coronary sinus pressure using T90 method and dp/dt method.

Fig. (6) Shows a part of the calculated data, it compare between the rise time of both methods, the rise time calculated using the dp/dt method is longer than the rise time of T90 method.

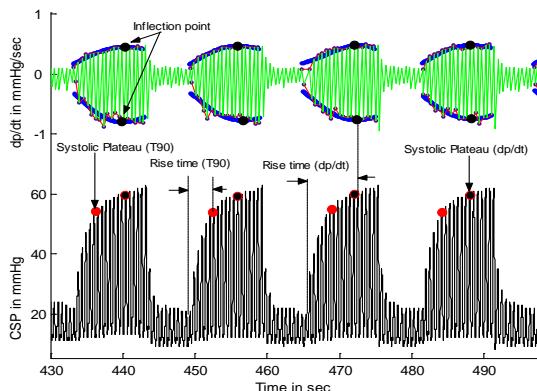


Fig. 6 Relation between systolic plateau and rise time of dp/dt and T90

4. Conclusion

The mathematical model equation (1) served as useful tool for describing the rise and the release of the coronary sinus pressure during the inflation and deflation periods. The CSP was expressed in the term of fitted parameters, three for inflation and three for deflation. Several hemodynamic quantities such as systolic and diastolic plateau, rise time, heart rate per PICS0 and the mean integral of CSP were derived from this model using two methods, T90 method and dp/dt method. During the comparison of the results of both methods,

we found that the results of dp/dt method are more exact and stable than the results of T90 method.

The physiological correlate of high systolic values is a good contractility of the myocardium and the systolic plateau can be seen as a monitoring parameter for myocardial strength. Myocardial strength is also reflected in the systolic rise time, the shorter is better. However, the rise time is also influenced by capillary and venous capacity as well as by the average coronary artery flow. The systolic rise time can be used as a calculated parameter for the closed loop regulation of PICS0.

References

- [1] Alzubaidi L, Mohl W, Rattay F. Automatic Computation for Pressure Controlled Intermittent Coronary Sinus. J IJCSI 2010; Vol. 7, Issue 6, pp.285-289;
- [2] Mohl W, Gueggi M, Haberzeth K, Losert U, Pachinger O, Schabart A. Effects of intermittent coronary sinus occlusion (ICSO) on tissue parameters after ligation of LAD. Bibliotheca Anatomica 1980; 20: 517-521.
- [3] Glogar D, Mohl W, Mayr H, Losert U, Sochor H, Wolner E. Pressure-controlled intermittent coronary sinus occlusion reduces myocardial necrosis (Abstract). Am J Cardiol 1982; 49: 1017.
- [4] Schreiner W, Neumann F, Schuster J, Froehlich KC, Mohl W. Computation of derived diagnostic quantities during intermittent coronary sinus occlusion in dogs. Cardiovasc Res 1988; 22(4): 265-276.
- [5] Schreiner W, Mohl W, Neumann F, Schuster J. Model of the haemodynamic reactions to intermittent coronary sinus occlusion. J Biomed Eng 1987; 9(2): 141-147.
- [6] Kenner T, Moser M, Mohl W, Tied N. Inflow, outflow and pressures in the coronary microcirculation. In: CSI - A New Approach to Interventional Cardiology. Mohl W, Faxon D, Wolner E (editors). Darmstadt: Steinkopff; 1986; 15.
- [7] Neumann F, Mohl W, Schreiner W. Coronary sinus pressure and arterial flow during intermittent coronary sinus occlusion. Am J Physiol 1989; 256(3 Pt 2): H906-915.
- [8] Moser M, Mohl W, Gallasch E, Kenner T. Optimization of pressure controlled intermittent coronary sinus occlusion intervals by density measurement. In: The Coronary Sinus, Vol. 1. Mohl W, Glogar D, Wolner E (editors). Darmstadt: Steinkopf; 1984; pp.529-536.
- [9] Mohl W, Glogar D, Kenner T, Klepetko W, Moritz A, Moser M. Enhancement of washout induced by pressure controlled intermittent coronary sinus occlusion (PICS0) in the canine and human heart. In: The Coronary Sinus, Vol. 1 Mohl W, Glogar D, Wolner E (editors). Darmstadt: Steinkopf; 1984; pp.537-548.

QUESEM: Towards building a Meta Search Service utilizing Query Semantics

Neelam Duhan¹ and A. K. Sharma²

¹ Department of Computer Engineering, YMCA University of Science & Technology
Faridabad, Haryana, India

² Department of Computer Engineering, YMCA University of Science & Technology
Faridabad, Haryana, India

Abstract

Current Web Search Engines are built to serve needs of all users, independent of the special needs of any individual. The documents are returned by matching their queries with available documents, with no emphasis on the semantics of query. As a result, the generated information is often very large and inaccurate that results in increased user perceived latency. In this paper, a Semantic Search Service is being developed to help users gather relevant documents more efficiently unlike traditional Web search engines. The approach relies on the online web resource such as dictionary based sites to retrieve possible semantics of the query keywords, which are stored in a definition repository. The service works as a meta-layer above the keyword-based search engine to generate sub-queries based on different meanings of user query, which in turn are sent to the keyword-based search engine to perform Web search. This approach relieves the user in finding the desired information content and improves the search quality for certain types of complex queries. Experiments depict its efficiency as it results in reduced search space.

Keywords: World Wide Web, Information Retrieval, Search Engine, Query Processing, Semantic Search Techniques.

1. Introduction

Search Engines enable users to navigate through the web information content incrementally and interactively. The outcomes of search typically depend on the submitted queries. But, the effectiveness of queries cannot be guaranteed as they vary from user to user, although exposing common information needs. In contrast, the same user query may convey different interpretations or different information needs. So, most of the users get irrelevant results or the effort for reaching needed information becomes very high due to vague or ambiguous formation of their queries. In other words, users often do not know how to combine the right words and in what order. Due to the lack of domain knowledge, users tend to post very

short queries, which generally do not express their information needs clearly and thus, the precision and recall of the search results get decreased. Query refinement comes in handy in these situations.

Keyword based indexing in search engines is another factor towards irrelevancy of search results. They are unable to associate the query words with the related fields of our daily world. We have a magnitude of words, out of which many possess more than one meaning. The meanings are sometimes totally unrelated; for instance, how can “lead” be a verb meaning *to go first* and also the name of a *heavy metal*? Consider a user, who intends to search any one term out of “mouse”, “cloud”, “cluster” or “tree” on a search engine. Most commonly, the terms return the results concerning computer field only or the one which is most popular. In case, user did not find the required information, the modified query may be resubmitted and even then, it is not guaranteed that he will find the exact required pages. This process is very time consuming and irritating.

There are following issues, which need to be addressed in modern day search engines:

1. A term can have several synonyms, which are not considered while returning the search results to the user as a lack of their availability.
2. A term may have several different meanings in different contexts. One may be interested in a particular field and other contexts, which are not needed, may increase the volume of search results for nothing good and just become a hurdle in finding the appropriate URLs. Thus, the problem of “Information Overkill” arises.
3. Search engines are unable to provide different descriptions of query terms to users so as to assist them in searching in the right direction.

Most of the information about the synonyms, contexts and descriptions (or definitions) exists in *definition based* or *dictionary based* sites, which can be utilized by the search engines to resolve the above said issues. This paper proposes the concept of *QUESEM*, a Meta search service over the keyword based search, which utilizes the online web resources to provide semantic and context-oriented web search to the users. The rest of the paper has been organized as follows. Section 2 describes the related research done towards the semantic and context based searching. Section 3 explains the proposed approach in detail, while in Section 4, the system architecture, various components and the definition repository have been discussed. Section 5 gives the practical evaluation of the proposed approach and finally, Section 6 concludes the paper with a discussion of the future research.

2. Related Work

The development of semantic web search systems has been an emerging area of research since the last few years and many researchers have shown their interest in this particular field. Query Refinement and expansion has become an essential information retrieval approach that interactively recommends new terms related to a particular query. As keyword based queries are likely to miss important results due to unstructured and semantically heterogeneous nature of the Web, therefore query expansion is considered an effective method to bridge the gap between users' internal information needs [11,15] and external query expressions.

Thesaurus-based query expansion [12,13] generally relies on statistics to find certain correlations between query terms. Cui et al. [14] mine click-through records of search results from query logs to establish mappings from query terms to strongly correlated document terms which are then used for query expansion.

Giorgos Akrivas et al. [9] used the semantic entities, rather than normal terms and for this, they used the knowledge stored in a semantic encyclopedia, specially the ordering relations, in order to perform a semantic expansion of the query. Their work of query expansion also considered the query context, which is defined as a fuzzy set of semantic entities created by an inclusion function. Furthermore, they integrated the approach with the user's profile also.

In another approach [10], sentence context is used to perform web searches rather than keywords. In this method, contexts replace specific words in the search request with other predetermined words. The authors reduced false positives with an intelligent search based on

grammar and English sentence structure. In their work, intelligent sentence searching converts each document into a set of simple sentences using only words in the predefined dictionary. These simple sentences capture the essence of the document. The conversion methodology uses synonyms, idiomatic expressions, grammar, patterns of speech and word location to create a searchable index. Because of the limited dictionary and elimination of most ambiguities, they claimed that such searches can be free of false positives.

Yaling Lie et al. [6] proposed a process-based search engine to handle certain type of queries that have an implicit process in it. The authors proposed to extract the process step by step of any given query as they assumed that most of the queries are in the form of a process and it would be better to find sub steps and then find the related URLs. To do this task they have taken help of process handbook to get the possible steps of any given query.

Y. Li et al. [8] proposed a solution to solve Web search queries having transactional intent, called *transactional queries*; for example, *Download software or fill in an online study-plan*. They claimed that many Web searches belong to this category. Existing search engines do not recognize these specific queries. In their work, a hand-crafted rule-based classifier is developed to recognize a collection of transactional Web pages for a set of transactions. General users are not able to contribute to or access the classifier. Therefore, it is difficult to build rules covering most of the transactional needs.

In [4], a novel method, Q-Rank, has been proposed to leverage the implicit feedbacks from the logs about the users' search intents. The authors claimed that to improve the relevance ranking for underspecified queries requires better understanding of users' search goals. By analyzing the semantic query context extracted from the query logs, they invented Q-Rank to effectively improve the ranking of search results for a given query. Experiments showed that Q-Rank outperforms the current ranking system of large-scale commercial Web search engines, improving the relevance for 82% of the queries with an average increase of 8.99% in terms of discounted cumulative gains. Because Q-Rank is independent of the underlying ranking algorithm [17, 18], it can be integrated with existing search engines.

The mechanisms proposed in the literature have their own advantages in particular areas of applications, but a critical look at the available literature indicates that most of existing semantic and context-based search techniques suffer from a couple of the following limitations:

- They are not able to capture the full set of synonyms for each type of query submitted by the users.

- Most of the techniques require complex analysis involving natural language processing and linguistic preprocessing to discover the context and semantics of query terms.
- The techniques utilizing external web resources for query expansion require complex retrieval efforts and the resource integration.
- Many of the techniques are targeted towards relevant page retrieval, but the produced results may not be presented in an easy user navigational manner. Moreover, the techniques do not aim to provide the user with a list of choices related to the query, from which the user can browse according to his interest.

This paper contributes towards developing a search service for semantic and context-based information retrieval, while at the same time, keeping in view the above limitations. The proposed technique has been made to take the advantage of dictionary based information available on the Web to gather possible meanings of a term and generate the query responses accordingly. The results will be represented in the form of clusters according to the newly found sub-terms. The next section explains in detail the proposed approach.

3. Proposed Approach of Semantic Search

In order to address the problems associated with keyword based search engines [16], a system called Query Semantic Search System (abbreviated as QUESEM, pronounced /'Qu-sem/) to improve searching quality and reduce searching time is proposed. *QUESEM* maintains a database of definitions (referred to as Definition Repository), as the core of the system to accomplish its desired task.

In an abstract form, the approach to be followed by the system can be shown diagrammatically as in Fig. 1. Given a query, *QUESEM* first analyses it, after which, it matches terms in the query with the data (term_title and/or descriptive fields) stored in Definition Repository (explained in subsequent sections) to locate the relevant definitions also called sub-terms related to query. The *definition* here is meant to describe a semantic description of the term in question. For every distinct definition of the query, it uses an existing keyword-based search engine (e.g., Google or Yahoo) to search the Web pages on the WWW. *QUESEM* then displays the results in the form of clusters corresponding to each extracted definition.

A Topical Crawler is developed here to download Web sites, which specialize in definition-related or dictionary-related content. A Definition-Generator/Annotator with

machine learning techniques is designed to automatically extract the relevant definitions from the crawled Web pages.

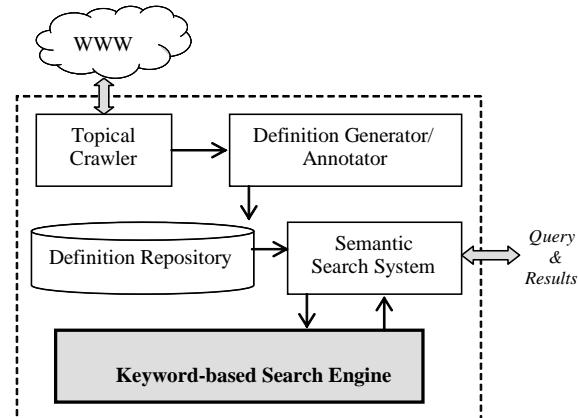


Fig. 1 An Abstract Architecture of QUESEM

An example scenario of a common search is given below to better illustrate the proposed approach:

A User 'X' wants to gain knowledge about "cluster". He is totally new to this term and has no idea about this.

As shown in Fig. 2, instead of displaying the URLs matching the query keywords or their combinations as in traditional search engines, *QUESEM* displays the matched terms having distinct meanings (definitions) related to the term 'cluster'. It may also include the related term descriptions.

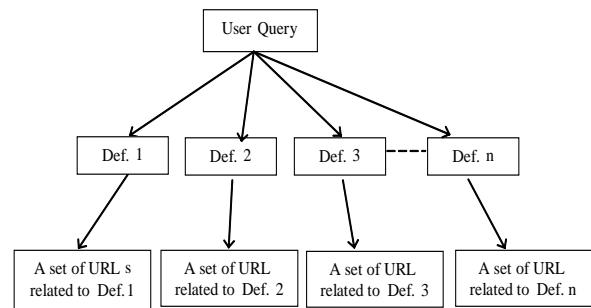


Fig. 2 The Expected output of QUESEM

'X' chooses the matched term which is of his interest, reads the description and then if found interesting, further explores the related URLs.

The aim here is to solve the information overkill problem with the help of dictionary based sites. The terms that are

found closer to the given search query on *yourdictionary.com* or other such sites are extracted by the definition generator module, annotated by annotator and stored in the Definition Repository for further use by the system. The next section describes the detailed system architecture and the functioning of various components involved over there.

4. System Architecture

The detailed system architecture of *QUESEM* is shown in Fig. 3, where the dashed line represents the proposed meta-search service. In order to achieve the required task, architecture is divided into two major sub-systems as given below:

1. Definition Repository Generation
2. Definition based Search

The basic definitions and the working of these two subsystems are explained under.

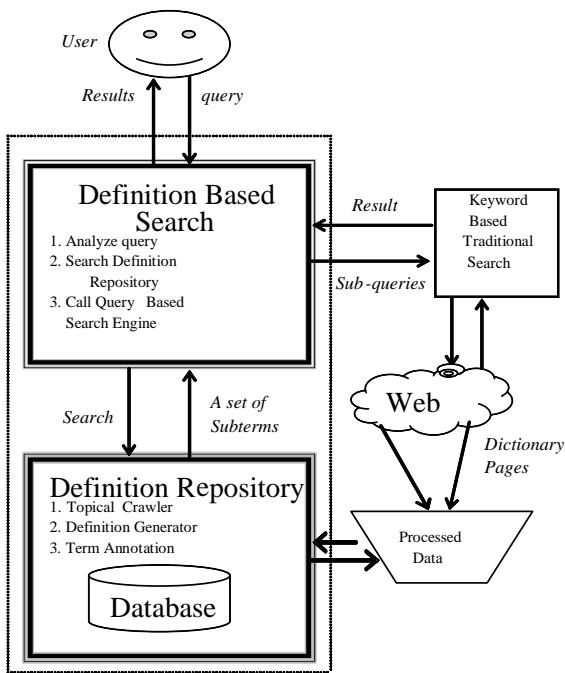


Fig. 3 High-Level System Architecture of QUESEM

4.1 Basic Definitions

Some definitions are formulated here, which are related to the proposed search system.

Definition of “Definition”:

A *definition* is a phrase or set of symbols that define the meaning of a term or similar kind of things. A term may

have many different senses or meanings in different contexts. For each such specific sense, a definition is a set of words that defines it. Its existence in a particular field defines a terminology of that field.

Definition of “Definition Repository”:

A database for storing terms, their related definitions and a group of programs, which provide means to collect and access the data. The schema must contain at least the two relations as shown in Fig. 4, and it may further contain optional fields or relations to facilitate term hierarchy. It collects and manages definitions, which are used by Definition based search sub-system to expand the initial query into multiple sub-queries or definitions.

| TERM | | | |
|------------------------------|-------------------|-------------------------|----------------|
| <u>Term_ID</u> | <u>Term_Title</u> | <u>Term_Description</u> | |
| DEFINITION ANNOTATION | | | |
| <u>Def_ID</u> | <u>Def_Title</u> | <u>Def_Description</u> | <u>Term_ID</u> |

Fig. 4 Schema of Definition Repository

The schema is very simple and is made to handle only single level of hierarchy, however whenever there is a need to extend the schema complexity to handle the large size of definition repository. The fields with a solid underline represent the primary key and a dashed underline represents a foreign key. Table 1 gives description of various fields in the Definition Repository.

Table 1. Various Fields in Definition Repository

| Field | Description |
|-------------------------|--|
| <u>Term_Id</u> | An attribute that gives an identity number to the term under consideration. |
| <u>Term_Title</u> | It contains the actual query terms that are entered by the user. |
| <u>Term_Description</u> | It contains a limited length snippet to have a small description of the term. |
| <u>Def_Id</u> | It represents the Identity numbers of the semantic Definitions (subqueries) extracted for each query term. |
| <u>Def_Title</u> | This column specifies the definitions for query terms that are found relevant in the dictionary-based sites after parsing. |
| <u>Def_Description</u> | Small description of each definition. |

Fig. 5 gives the state of the repository for the term “Cluster” and its various definitions viz. “Cluster Headache”, “Cluster Bomb” etc.

| TERM | | |
|-----------------------|--------------------|--|
| 1 | Cluster | A group of similar objects... |
| DEFINITION ANNOTATION | | |
| 1.1 | Cluster headache | It is less common than migraine headache..... |
| 1.2 | Cluster bomb | Cool crust band..... |
| 1.3 | Cluster analysis | Cluster analysis classifies a set of observations..... |
| 1.4 | Cluster bean | Drought tolerant herb..... |
| 1.5 | Cluster (computer) | It's a technique to categorize ... |
| | | 1 |

Fig. 5 Example Illustration of Definition Repository

Definition of “*Definition based Search*”:

It is an extension to the traditional Web search, which discovers the implied definitions d_1, d_2, \dots, d_n , of the initial query q (if they exist) using the definition repository. It performs traditional Web search on each definition d_i , $i=1, 2, \dots, n$. Let r_i denote the obtained URL list by searching on d_i , then $\langle r_1, r_2, \dots, r_n \rangle$ represents the result of the Query Semantic search on the initial query q .

An example of this type of search was briefly described (illustrated in Fig. 2) earlier. The current assumption for this search is that the titles of definition in the *definition annotation* relation represent potential definitions for the initial user query term. For instance, for a given query “cluster”, definitions extracted are ‘cluster headache’, ‘cluster beans’ etc. and the results for the query “cluster” now contain the result URLs of “cluster headache”, ‘cluster beans’ etc.

4.2 Definition Repository Generation

Definition Repository is maintained by a series of steps as illustrated in Fig. 6. Various inputs, outputs and the components involved in the process with their mode of working are explained below in detail.

4.2.1 Information Resource: Dictionary Based Sites

For automatically populating the repository with high-quality definitions, the input resource taken is the dictionary based sites, which are a rich store of semantic definitions of terms explicitly published on the Web. In general, the web pages, which possibly contain this type of information, are very sparsely distributed on the Web. In order to efficiently extract more sub terms related to the query terms, while downloading relatively fewer pages, a topical crawler is needed to crawl only those Web sites, which are specialized in dictionary-related content.

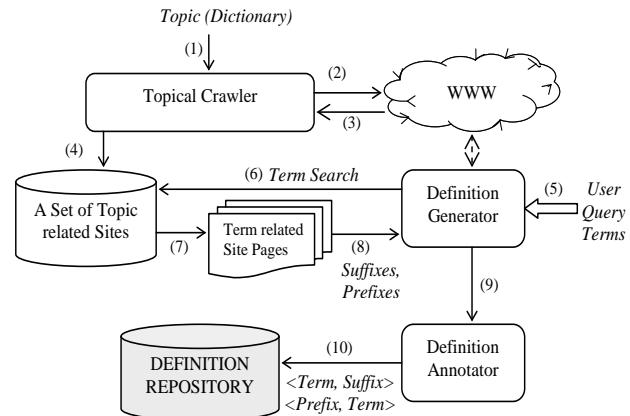


Fig. 6 Definition Repository Generation Process

4.2.2 Topical Crawler

A typical crawler for a generic web search engine recursively traverses through hyperlinks to explore the undiscovered portion of the Web [1]. The basic idea behind topical crawling is to estimate the relevance of undiscovered documents by the relevance of fetched documents, which directly or indirectly link to the undiscovered ones [2, 3]. To start the crawl process, some topic of interest such as “dictionary” or related content is needed. Topical crawlers maintain priority queues, where most possibly related documents have the highest priority to be downloaded under the constraints of limited computing resources. Thus, topical crawlers traverse the topic-related portion of the Web.

A publicly available Web search service (Google SOAP API) can be used to filter dictionary-related Web sites by searching dictionary-related keywords (e.g. *dictionary*, *thesaurus*, *synonyms*, *answers*, *definition* etc.) in the title of the home page of a Web site. Here, we have taken only one site *yourdictionary.com*, but a number of sites can be used statically to do this task.

The topical crawler will output a set S of sites related to dictionary based content, which is stored in a local repository to be referred further by *definition generator*.

4.2.3 Local-Site Search by Definition Generator

When the user first time submits a query to *QUESEM*, its keywords would be passed to *definition generator* module. This component consults the set S of dictionary-based sites and passes these query terms over their interfaces to perform a local-site search. After that, the resultant set of pages related to query terms are retrieved by this component and placed in a set D . The resultant pages are

further parsed to extract the linked pages related to query terms. The extracted pages are also kept in D .

An approach [5], which uses local site-searches to estimate the relevance of the documents before fetching them, is used here to help getting the definitions. The algorithm for local-site searching is shown in Fig. 7.

Algorithm: Local_Site_Searching (Initial query, S)

I/P= Initial query q and Set S of Sites stored in a repository
O/P= Unstructured document set D containing documents that are response pages against the initial query q .
// Start of Algorithm
 Begin
 For (every site $si \in S$) // perform the local-site search
 Begin
 Step1: fetch the homepage of si
 Step2: find the local-site search form
 Step3: Submit the query terms (q)
 Step4: Fetch the response pages in local Repository D
 Step5: For (every response Page) //Find linked pages
 Begin
 Parse the Page;
 Fetch all result links;
 If synonyms link exist
 Begin
 Fetch synonyms page and place in D
 End
 End
 End
 Return the fetched page set D

Fig. 7 Algorithm for Local Site Search for Query Terms

The set D of documents d_i for $i = 1 \dots n$ is the returned set of pages fetched from dictionary-based sites, which contain the relevant information regarding the semantics or context of query terms. This set D is used for *definition generation* and *annotation*.

4.2.4 Definition Generation & Annotation

As the response pages are the result of the dictionary based sites, it is assumed that the pages will contain the direct thesaurus and synonyms of the query terms. The approach to be followed to find semantic definitions is given below:

“Extract the prefix and suffix tokens of query terms from pages d_i , annotate them with query terms and store the result in Definition Repository”

The *Definition Generator* simply extracts the *prefix* and *postfix* present consecutively in combination with the query terms from the pages belonging D , while *Definition Annotator* combines them suitably with the query terms to result in proper annotations of definitions. We can expect

that the result will be the required modified input for our search goal. The algorithm for definition generation and annotation is shown in Fig. 8.

Algorithm: Definition_Generator_Annotator (D)

I/P= Unstructured document set D of pages containing semantics of query terms.
O/P= Term and their associated Annotated Definitions
//Start of Algorithm
 Begin
 Term_title= q
 For (every d_i in D)
 Begin
 Set_before= Consecutive tokens occurring before q in d_i ;
 Set_after= Consecutive tokens occurring after q in d_i ;
 End
 For (every token t_i in Set_before)
 $t_i_Title \leftarrow t_i + q >$ // t_i_Title is the title of definition
 For (every t_i in Set_after)
 $t_i_Title \leftarrow q + t_i >$
 Place term_titles, definition_titles t_i in the Definition Repository
 End

Fig. 8 Algorithm for Definition Generation/Annotation

The prefixes and postfixes of the query terms play an important role in semantic definition generation. It is assumed that most of the definition based sites manage the data in the thesaurus and synonyms form. This assumption may restrict the number of definitions that could be found, but for a precise search, this assumption may be much more relevant as in dictionary based sites, the most relevant sub terms can be found nearby to the basic term.

Therefore, prefix and suffix represent the synonyms/context, which have some how a meaning equivalent to query term. To extract query semantics from the pages d_i , which are in the form of prefix and suffix, parsing is required. A parser is used to do tokenization of the response pages and the query terms are kept in track to find their relevant prefixes or suffixes, which are in turn annotated with the query terms by the *Definition Annotator* to generate the definition titles. For example, as was shown in Fig. 5, “bean”, “bomb”, “headache”, “analysis”, “computer”, “controller” etc. all represent the prefixes or suffixes related to the term “cluster”. The figure also shows the definition titles after annotations e.g. “cluster headache” is the annotated definition.

4.2.5 Populating the Definition Repository

Initially definition repository would be empty and it will be maintained as the user queries would be submitted to *QUESEM*. The definitions generated by the *Definition Generator & Annotator* with respect to each new query will be populated into the Definition Repository. It may be

possible that the definition repository may contain some inadequate definitions, but as user explores a cluster of URLs relative to his interest, some abrupt clusters may not affect the search.

4.3 The Definition based Search

When user submits a keyword based query to *QUESEM*, the keywords are passed to “Definition Repository Generation” subsystem to build the definition Repository as well as to the “Definition based Search” subsystem to respond to the user in the form of cluster of result URLs. Contrary to the traditional keyword based search, semantic or definition based search requires slightly complex query processing. Various modules which are used in definition based query processing are given below and are outlined in Fig. 9.

1. Query Analyzer
2. Definition Searcher
3. Query Transformer and Processor

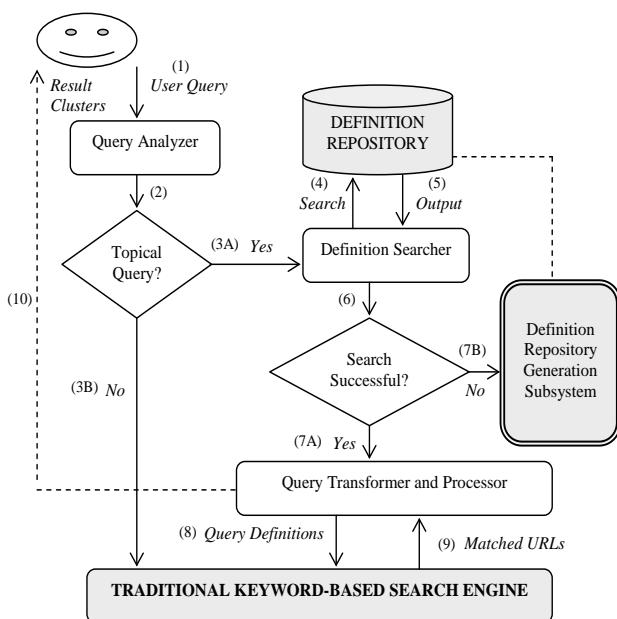


Fig. 9 Definition Based Search Subsystem

Query processing involves analysis of user query to perform extended search for it. Given a query, a *Query Analyzer* first decides whether to perform definition based search or a traditional search for the query? The *QUESEM* works only for topical queries. The topical user queries, which may be solved by this system, can observe two characteristics:

- They can have multiple meanings.
- They can have a number of synonyms.

After the analysis, the next step is to find the related definitions by searching the definition repository. This is performed by the *Definition Searcher* module. Finally, the initial query is transformed into a sequence of sub-queries or definitions by the *Query Transformer & Processor*. Obtained sub-queries are then sent individually to the traditional keyword-based search engine (like Google) to find the matched pages. Now the *Query Processor* represents the results obtained by different sub-queries in the form of clusters to the user.

The functioning of different components is described briefly in following subsections:

4.3.1 Query Analyzer

Query analyzer is responsible to check whether definition based search is applicable to the query or not? Analyzer first examines the query to decide whether a query is topical or not? A topical is the query, which is generally framed by simple keywords. For example “mouse”, “waiter”, “data mining”, “HCL laptop” etc. are topical queries, while “how to drive a car”, “when monsoon will come” etc. are not topical queries.

For analyzing the queries, openNLP functions [7] are used by the system. A query which contains a combination of <predicate, object> is considered a goal based query [6, 7], otherwise queries containing either <predicate> or <object> is considered a topical query, for which *QUESEM* gives better results. It is assumed that topical search query either have a *verb phrase* or a *noun phrase*. The queries, which are not topical, are made to be searched for using the traditional search system.

4.3.2 Definition Searcher

The job of Definition Searcher is to check whether the topical query already exists in the system’s Definition Repository. If it is there, that means some user has already queried it and as result of which its definitions are already stored in the database. Now, it is not required again to follow the same procedure, but simply extracting the definitions from the definition repository. If, on the other side, if definition repository doesn’t return any set or say return NULL then it is the functional responsibility of the topical crawler of *Definition Generation Subsystem* to become active and perform all the tasks necessary to get the new definitions matching the new query and expand the definition repository.

In searching the definitions, literal term matching is done. It is simple keyword-to-keyword matching which return a set of definitions corresponding to the term. As the user queries tend to contain common words, punctuation marks,

stop words, case sensitivity etc, handling all these aspects comes in preprocessing the query which is sent to the definition searcher. The related definitions are passed to Query Transformer and Processor to process the queries.

4.3.3 Query Transformer & Processor

Query transformer is responsible for making the retrieved definitions as a sequence of well-defined queries so that they could be searched individually as independent queries. The original query and the matched set of definitions will be combined to form a new set of sub-queries. Query Processor searches these sub-queries individually with the help of a traditional search engine and then represents the resultant URLs in the form of a cluster with cluster label being the *definition_Title*.

5. Performance Evaluation

QUESEM was implemented in asp.net 2.0, C# and HTML with MS SQL Server 2005 at the back end to support definition repository. For the experimental purposes, only *yourdictionary.com* is examined for the current scenario. A group of 25 users from different domains were asked to search on *QUESEM* and other keyword based search engines like Google, Yahoo etc. The net performance of *QUESEM* in terms of quality of search results and reduced navigation time is observed to be higher than normal traditional web search. Fig. 10 shows interface of the search service.

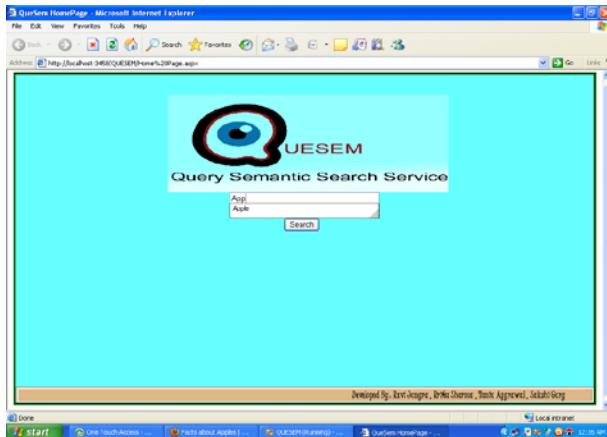


Fig. 10 Interface of QUESEM

Fig. 11 shows the result screen after submitting a query “apple”. It can be observed that QUESEM displays related terminologies in terms of definitions like “apple mobile”, “apple computer”, “apple fruit” etc. Similarly, Fig. 12 displays the terminologies related to query “stand”. When user clicks on a definition, corresponding results are

displayed using a traditional keyword based search engine. Here Fig. 13 and 14 show the results of queries “apple” and “history of apple” after redirecting them to Google.

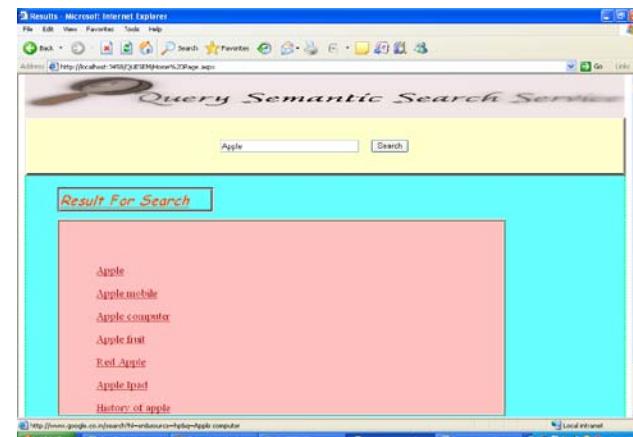


Fig. 11 The Definitions after submitting query “apple”



Fig. 12 The Definitions after submitting query “stand”

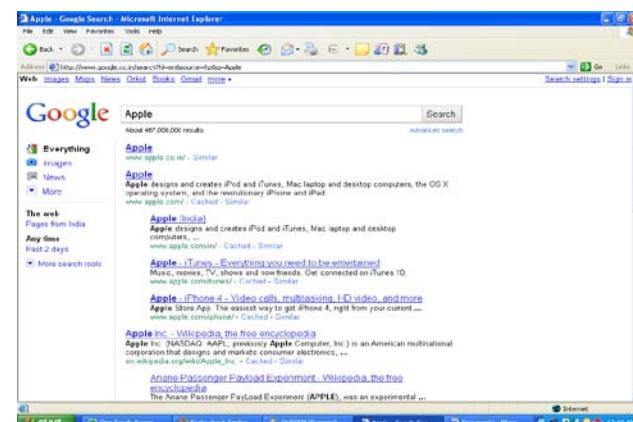


Fig. 13 Results for query “apple”

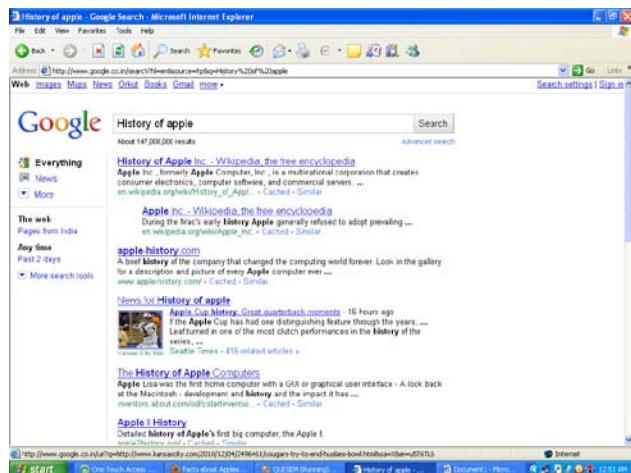


Fig. 14 Results for query “History of Apple”

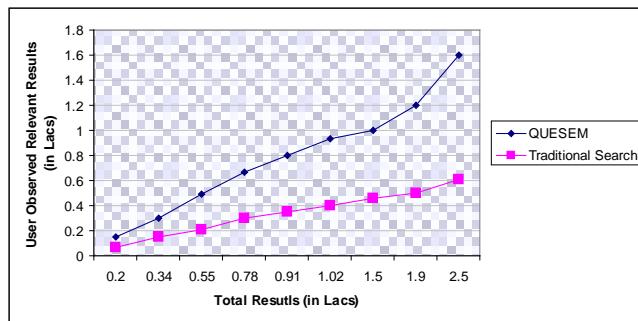


Fig. 15 Performance of QUESEM Vs Normal Search

The performance of QUESEM in terms of quality of search results and reduced navigation time is observed to be higher than the normal traditional web search. It is also assumed to be less complex as compared to other meta-search engines. Fig. 15 shows the performance comparison of QUESEM with the keyword-based search engines. It may be observed from the figure that in a normal web search, as the number of results increases, the user observed relevancy of documents decreases, while it remains approximate constant in semantic web search.

5. Conclusion

In traditional query/keyword based web search, some times, there are situations where a document is very much relevant to the user query but doesn't contain any similar term as described in his query and thus, does not appear in the search results. In this paper, a Query Semantic Search system to address these types of situations and “information overkill problem” has been developed. It is made to utilize the existing Web resources to automatically extract the synonyms (or semantics, thesaurus and

contexts) related to user queries and enhance the user search efficiency. The proposed system QUESEM is designed to serve as a Meta layer above the traditional Web search engines. The different components are integrated to form a complete system towards effective search. NLP techniques are utilized for examining the user queries to be solved by the proposed system. Assisted by the information of definitions related to semantics of query terms, QUESEM is able to understand users' queries in a better way to perform more meaningful searches.

The future research includes enhancing the system towards serving different types of complex user queries, which may involve multiple keywords and even the disordered keywords. This will require building efficient query analyzers so as to direct the user search towards the right direction.

References

- [1] Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., and Raghavan, S., “Searching the Web”. ACM Transactions on Internet Technology. Aug. 2001, pp: 2-43.
- [2] Aggarwal, C. C., Al-Garawi, F., and Yu, P. S., “On the design of a learning crawler for topical resource discovery”. ACM Transactions on Information Systems. Vol. 19, No. 3, Jul. 2001, pp: 286-309.
- [3] Chakrabarti, S., Vandenberg, M., and Dom, B. “Focused crawling: a new approach to topic-specific Web resource discovery”. In Proceedings of the Eighth International Conference on World Wide Web, Toronto, Canada, 1999. pp: 1623-1640.
- [4] Ziming Zhuang, Silviu Cucerzan, “Exploiting Semantic Query Context to Improve Search Ranking”. IEEE International Conference on semantic Computing, 2008 DOI: <http://doi.ieeecomputersociety.org/10.1109/ICSC.2008.8>
- [5] Liu, Y. and Agah, A, “Crawling and Extracting Process Data from the Web”. In Proceedings of the 5th International Conference on Advanced Data Mining and Applications, Beijing, China, August 17-19, 2009, Springer-Verlag, Berlin, Heidelberg, LNCS 5678, pp: 545-552.
- [6] Liu, Y. and Agah, A, “A Prototype Process-Based Search Engine”. In Proceedings of the third IEEE International Conference on Semantic Computing, Berkeley, CA, September 14-16, 2009.
- [7] OpenNLP, <http://opennlp.sourceforge.net/>
- [8] Y. Li, R. Krishnamurthy, S. Vaithyanathan, and H. V. Jagadish, “Getting work done on the web: supporting transactional queries,” In Proceedings of the 29th ACM SIGIR, Seattle, Washington, 2006, pp. 557- 564.
- [9] Giorgos Akrivas, Manolis Wallace, Giorgos Andreou, Giorgos Stamou and Stefanos Kollias, “Context Sensitive Semantic Query Expansion”. Proceedings of the 2002 IEEE International Conference on Artificial Intelligence Systems (ICAIIS'02).
- [10] Alan Chickinsky, Chief Scientist, “Intelligent Searching using Sentence Context”. IEEE International Conference on

- Technologies for Homeland Security, 2008, May 2008.
Greater Boston.
- [11] Broder, A, "A Taxonomy of Web Search". SIGIR Forum, pp. 3-10, 2002.
- [12] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T.K., and Harshman, R, "Indexing by Latent Semantic Analysis". Journal of American Society of Information Sciences, 41 (6), pp. 391-407. 1990.
- [13] Qiu, Y. and Frei, H, "Concept-based Query Expansion". In Proceedings of the 16th ACM SIGIR Conference, pp. 160-169. 1993.
- [14] Cui, H., Wen, J. Nie, J., and Ma, W, "Probabilistic Query Expansion Using Query Logs". In Proceedings of the 11th WWW Conference, 2002.
- [15] Jens Graupmann, Jun Cai and R. Schenkel, "Automatic Query Refinement Using Mined Semantic Relations", International Workshop on Challenges in Web Information Retrieval and Integration, April 2005, pp. 205-213.
- [16] A. K. Sharma, Neelam Duhan, Bharti Sharma, "A Semantic Search System using Query Definitions". Proceedings of First ACM International Conference on Intelligent Interactive Technologies and Multimedia, Allahabad, India, Dec. 28-30, 2010, pp: 269-272.
- [17] Neelam Duhan, A. K. Sharma, Komal Kumar Bhatia, "Page Ranking Algorithms: A Survey". In proceedings of the IEEE International Advanced Computing Conference (IACC'09), Patiala, India, 6-7 March 2009, pp: 1530-1537.
- [18] Neelam Duhan, A. K. Sharma. "A Novel Approach for Organizing Web Search Results using Ranking and Clustering". International Journal of Computer Applications 5(10):1-9, August 2010. Published By Foundation of Computer Science.

Neelam Duhan received her B.Tech. degree in Computer Science & Engineering with Hons. from Kurukhetra University, Kurukshetra in 2002 and M.Tech. degree with Hons. in Computer Engineering from Maharshi Dayanand University, Rohtak in 2005. Presently, she is working as Assistant Professor in Computer Engineering Department in YMCA University of Science & Technology, Faridabad and has a teaching experience of 7 years. She is pursuing Ph.D. in Computer Engineering from Maharshi Dayanand University, Rohtak and her areas of interest are Databases, Data Mining, Search Engines and Web Mining.

Prof. A. K. Sharma received his M.Tech. in Computer Science & Technology with Hons. from University of Roorkee (Presently I.I.T. Roorkee) in 1989 and Ph.D (Fuzzy Expert Systems) from JMI, New Delhi in the year 2000. He obtained his second Ph.D. in IT from IIITM, Gwalior in 2004. His research interests include Fuzzy Systems, Object Oriented Programming, Knowledge representation and Internet Technologies. Presently he is working as the Dean, Faculty of Engineering and Technology & Chairman, Dept of Computer Engineering at YMCA University of Science and Technology, Faridabad. His research interest includes Fuzzy Systems, OOPS, Knowledge Representation and Internet Technologies. He has guided 9 Ph.D. thesis and 8 more are in progress with about 175 research publications in International and National journals and conferences. He is the author of 7 books. Besides being member of many BOS and Academic councils, he has been Visiting Professor at JMI, IIITM, and I.I.T. Roorkee.

Multihop Routing In Self-Organizing Wireless Sensor Networks

Rajashree.V.Biradar¹, Dr. S. R. Sawant², Dr. R. R. Mudholkar³, Dr. V.C .Patil⁴

¹Department of Information Science and Engineering, Ballari Institute of Technology and Management, Bellary-583104, Karnataka, India.

²Department of Electronics, Shivaji University, Kolhapur-416004, Maharashtra, India.

³Department of Electronics, Shivaji University, Kolhapur-416004, Maharashtra, India.

⁴Department of Electronics and Communication Engineering, Ballari Institute of Technology and Management, Bellary-583104, Karnataka, India.

Abstract

Wireless sensor networks have emerged in the past decade as a result of recent advances in microelectronic system fabrication, wireless communications, integrated circuit technologies, microprocessor hardware and nano-technology, progress in ad-hoc networking routing protocols, distributed signal processing, pervasive computing and embedded systems. As routing protocols are application specific, recent advances in wireless sensor networks have led to many new protocols specifically designed for routing. Efficient routing in a sensor network requires that the routing protocol must minimize energy dissipation and maximize network life time. In this paper we have implemented several multihop flat based routing protocols like Flooding, Gossiping and a cluster based protocol Multihop-LEACH which does inter-cluster and intra-cluster multihopping using TinyOs and TOSSIM simulator. Evaluation and comparison reveals that Multihop-LEACH protocol utilizes less power and least delay compared to other protocols. We further evaluated the Multihop-LEACH protocol with varying probability of clustering to extend the network life time.

Keywords: Multihop, Flooding, Gossiping, Multihop-LEACH, TinyOS, nesC, TOSSIM and Probability.

1. Introduction

Sensor networks have emerged as a promising tool for monitoring (and possibly actuating) the physical worlds, utilizing self-organizing networks of battery-powered wireless sensors that can sense, process and communicate. Wireless sensor networks [1,2] consist of small low power nodes with sensing, computational and wireless communications capabilities that can be deployed randomly or deterministically in an area from which the users wish to collect data. Typically, wireless sensor networks contain hundreds or thousands of these sensor nodes that are generally identical. These sensor nodes have the ability to communicate either among each other or directly to a base station (BS). The sensor network is

highly distributed and the nodes are lightweight. Intuitively, a greater number of sensors will enable sensing over a larger area. As the manufacturing of small, low-cost sensors become increasingly technically and economically feasible, a large number of these sensors can be networked to operate cooperatively unattended for a variety of applications. The features of sensor networks [3] are as depicted below.

Varying network size: The size of a sensor network can vary from one to thousands of nodes.

Low cost: For the deployment of sensor nodes in large numbers, a sensor node should be inexpensive.

Long lifetime network: An important characteristic of a sensor network is to design and implement efficient protocols so that the network can last as long as possible.

Self-organization: Sensor nodes should be able to form a network automatically without any external configuration.

Query and re-tasking: The user should be able to query for special events in a specific area, or remove obsolete tasks from specific sensors and assign them with new tasks. This saves a lot of energy when the tasks change frequently.

Cooperation/Data aggregation: Sensor nodes should be able to work together and aggregate their data in a meaningful way. This could improve the network efficiency.

Application awareness: A sensor network is not a general purpose network. It only serves specific applications.

Data centric: Data collected by sensor nodes in an area may overlap, which may consume significant energy. To

prevent this, a route should be found in a way that allows in-network consolidation of redundant data.

Recent advances in wireless sensor networks have led to many new protocols specifically designed for sensor networks. Most of the attention, however, has been given to the routing protocols since they might differ depending on the application and network architecture [4]. To prolong the lifetime of the sensor nodes, designing efficient routing protocols is critical. Even though sensor networks are primarily designed for monitoring and reporting events, since they are application dependent, a single routing protocol cannot be efficient for sensor networks across all applications. Multihop routing, clustering and data aggregation are important techniques in minimizing the energy consumption in sensor networks [12, 13, 14].

In this paper we describe and implement several multihop routing protocols for sensor networks and present a critical analysis and evaluation of these protocols. The performance comparison considering all the characteristics that should be possessed by routing protocols reveals the important features that need to be taken into consideration while designing new routing protocols for sensor networks. The remainder of this paper is organized as follows. Section 2 contains classification of routing protocols, section 3 contains description of routing protocols implemented, Section 4 contains implementation and simulation, section 5 contains simulation matrices and results and, finally section 6 contains conclusion and future work.

2. Classification of routing protocols

Broadly speaking, almost all of the routing protocols can be classified according to the network structure; as flat, hierarchical or location-based. Further, these protocols can also be classified according to operation mode; multipath-based, query-based, negotiation-based, QoS-based, and coherent-based [5]. Figure 1 illustrates classification of WSN routing protocols.

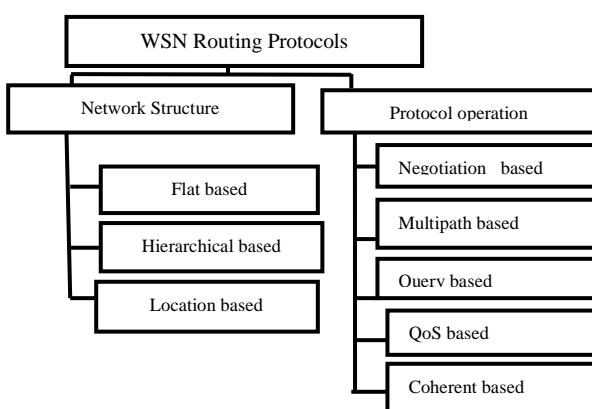


Fig. 1: Classification of WSN Routing Protocols

2.1 Network Structure

Based on the structural orientation of a network, which includes structural orientation of base stations and the structural orientation of sensor nodes we classify routing protocols as flat based, hierarchical based and location based.

Flat based: In these networks, all nodes play the same role and there is absolutely no hierarchy. Flat routing protocols distribute information as needed to any reachable sensor node within the sensor cloud [6]. No effort is made to organize the network or its traffic, only to discover the best route hop by hop to a destination by any path.

Hierarchical based: This class of routing protocols sets out to attempt to conserve energy by arranging the nodes into clusters as shown in Figure 2. Nodes in a cluster transmit to a head node within close proximity which aggregates the collected information and forward this it to the base station [6, 7].

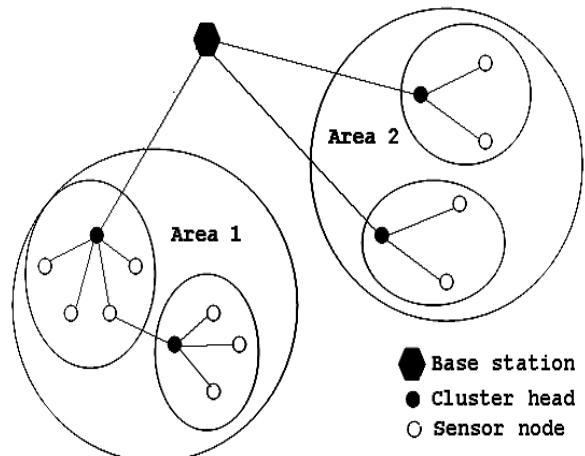


Fig. 2: Clustering Nodes

Good clustering protocols play an important role in network scalability as well as energy efficient communication. On the negative side of it, clusters may lead to a bottleneck. This is because only one head communicate on behalf of the entire cluster. Energy depletion will be strongest in that head.

Location based: Most of the routing protocols for sensor networks require location information for sensor nodes. In most cases location information is needed to calculate the distance between two particular nodes so that energy consumption can be estimated. Since there is no addressing scheme for sensor networks like IP-addresses, location information can be utilized in routing data in an energy efficient way [6].

2.2 Protocol Operation

It describes the main operational characteristics of a routing protocol; in terms of communication pattern, hierarchy, delivery method, computation, next-hop.

Multipath based: In this case, the network derives benefit from the fact that there may be multiple paths between a source node and the destination. Using different paths ensures that energy is depleted uniformly and no single node bears the brunt [12, 13].

Query based: Here the focus lies on propagation of queries throughout the network by the nodes which require some data. Any node which receives a query and also has the requested data, replies with the data to the requesting node. This approach conserves energy by minimizing redundant or non-requested data transmissions [8].

Negotiation based: The nodes here exchange a number of messages between themselves before transmission of data [9, 10]. The benefit of this is that redundant data transmissions are suppressed. It should however be ensured that the negotiation transmissions are not allowed to exceed an extent that the energy saving benefit is offset by the negotiation overhead.

QoS-based: QoS based protocols have to find a trade-off between energy consumption and the quality of service [11]. A high energy consumption path or approach may be adopted if it improves the QoS. So when interested in energy conservation, these types of protocols are usually not very useful.

Coherent-based : Coherence based protocols focus on how much data processing takes place at each node[11]. In coherent protocols, data is sent to an aggregator node after minimum possible processing, and processing is then done at the aggregator. Coherent processing is usually adopted for energy efficient routing because they reduce the computation steps per node. However, the aggregator nodes must have more energy than the other ordinary nodes, or else they will be depleted rapidly.

3. Description of routing protocols implemented

3.1 Flooding

Flooding [1] starts with a source node sending its data to all of its neighbors. Upon receiving a piece of data, each node then stores and sends a copy of the data to all of its neighbors. Only packets which are destined for the node itself or packets whose hop count has exceeded a preset limit are not forwarded. This is therefore a straight forward protocol requiring no protocol state at any node, and it disseminates data quickly in a network where bandwidth is not scarce and links are not loss-prone. The

main benefit of Flooding is that it requires no costly topology maintenance or route discovery. Once sent a packet will follow all possible routes to its destination. If the network topology changes sent packets will simply follow the new routes added. Flooding does however have several problems. One such problem is implosion. Implosion is where a sensor node receives duplicate packets from its neighbors. Figure 3 illustrates the implosion problem. Node A broadcasts a data packet ([A]) which is received by all nodes in range (nodes B and C in this case). These nodes then forward the packet by broadcasting it to all nodes within range (nodes A and D). This results in node D receiving two copies of the packet originally sent by node A. This can result in problems determining if a packet is new or old due to the large volume of duplicate packets generated when flooding. Overlap is another problem which occurs when using Flooding. If two nodes share the same observation region both nodes will witness an event at the same time and transmit details of this event. This results in nodes receiving several messages containing the same data from different nodes. Figure 4 illustrates the overlap problem. Nodes A and B both monitor geographic region Y. When nodes A and B flood the network with their sensor data node C receives two copies of the data for geographic region Y as it is included in both packets. Another problem with Flooding is that the protocol is blind to available resources. Messages are sent and received by a node regardless of how much power it has available. In addition to this the number of packets generated by the Flooding protocol causes a lot of network traffic and causes a large network wide energy drain across the network. This can shorten the life of the network.

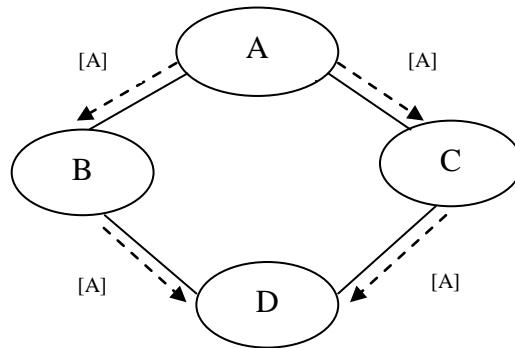


Fig. 3: Implosion problem

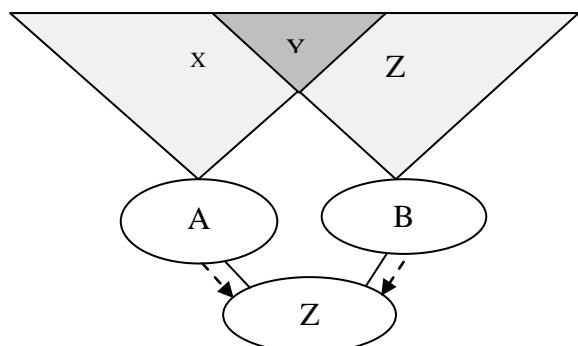


Fig. 4: Overlap problem

3.2 Gossiping

The Gossiping protocol is based on the Flooding protocol. Gossiping is proposed to address some critical problems of the Flooding scheme [1, 2]. Instead of broadcasting each packet to all neighbors the packet is sent to a single neighbor chosen at random from a neighbor table. Having received the packet the neighbor chooses another random node to send to. This can include the node which sent the packet. This continues until the packet reaches its destination or the maximum hop count of the packet is exceeded.

Gossiping avoids the implosion problem experienced by Flooding as only one copy of a packet is in transit at any one time. However, it may cause another problem, the long packet delay. Because the sender randomly selects the subset of the result in a router neighbors to transmit data, the selected sensors may result farther than the shortest path between the sender and the sink. Hence, this may extend the packet delay time. While gossiping distributes information slowly, it dissipates energy at a slow rate as well. Consider the case where a single data source disseminates data using gossiping. Since the source sends to only one of its neighbors, and that neighbor sends to only one of its neighbors, the fastest rate at which gossiping distributes data is 1 node/round. Finally, we note that, although Gossiping largely avoids implosion, it does not solve the overlap problem.

3.3 Multihop Low Energy Adaptive Clustering (Multihop-LEACH)

Multihop-LEACH is a cluster based routing algorithm in which self-elected cluster heads collect data from all the sensor nodes in their cluster, aggregate the collected data by data fusion methods and transmit the data through an optimal path between the cluster head (CH) and the base station(BS) through other intermediate CHs and use these CHs as a relay station to transmit data through them as shown in figure 5.

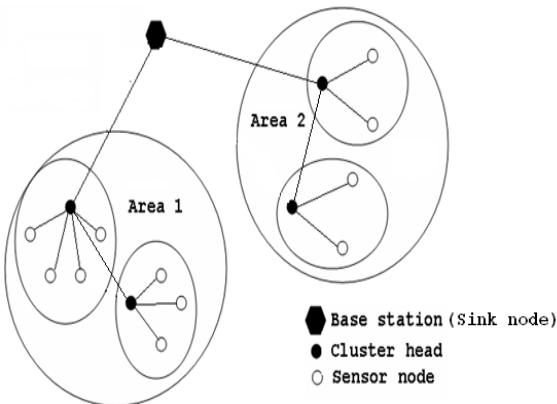


Fig. 5: Nodes communicate to Base Station through an optimal path of Cluster Heads

These self elected cluster heads continue to be cluster heads for a period referred to as a round. At the beginning of each round, every node determines if it can be a cluster head during the current round by the energy left at the node. In this manner, a uniform energy dissipation of the sensor network is obtained. If a node decides to be a cluster head for the current round, it announces its decision to its neighbors. Other nodes which choose not to be cluster heads determine to which cluster they want to belong by choosing the cluster head that requires the minimum communication energy. Multihop-LEACH was mainly proposed for routing data in wireless sensor networks which have a fixed base station to which recorded data needs to be routed. All the sensor nodes are considered static, homogenous and energy constrained. The sensor nodes are expected to sense the environment continuously and thus have data sent at a fixed rate. These assumptions make it unsuitable for sensor networks where a moving source needs to be monitored.

The operation of Multihop-LEACH is separated into two phases: the setup phase and the steady state data transfer phase. In the set up phase, the clusters are organized and cluster heads selected. During the setup phase, the cluster heads are selected based on the suggested percentage of probability of clustering for the network and the number of times the node has been a cluster-head so far. This decision is made by each node n choosing a random number between 0 and 1. If the number is less than a threshold $T(n)$, the node becomes a cluster-head for the current round. The threshold is set as follows:

$$T[n] = \begin{cases} \frac{P}{1-P(r \bmod 1/p)} & \text{if } n \in G \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Where P is the desired cluster-head probability, r is the number of the current round and G is the set of nodes that have not been cluster-heads in the last $1/P$ rounds. Once the nodes have elected themselves to be cluster heads they broadcast an advertisement message (ADV). Each non cluster-head node decides its cluster for this round by choosing the cluster head that requires minimum communication energy, based on the received signal strength of the advertisement from each cluster head.

After each node decides to which cluster it belongs, it informs the cluster head by transmitting a join request message (Join-REQ) back to the cluster head. After receiving all the messages from the nodes that would like to be included into the cluster and based on the number of nodes in the cluster, the cluster head creates and announces a TDMA schedule, assigning each node a time slot when it can transmit. Each cluster communicates using different CDMA codes to reduce interference from

nodes belonging to other clusters. The CDMA code to be used in the current round is transmitted along with the TDMA schedule.

In the steady state phase, the actual data transfer to the base station takes place. Upon receiving all the data, the cluster head node aggregates it before sending it to the other cluster head nodes. After a certain time, determined a priori, the network goes back to the set up phase and enters another round of selecting new cluster heads.

Inter-cluster and intra-cluster multi-hop communication are the two major concepts considered in Multihop-LEACH protocol.

Multihop inter-cluster operation: In this model network is grouped into different clusters. Each cluster is composed of one cluster head (CH) and cluster member nodes. The respective CH gets the sensed data from its cluster member nodes, aggregates the sensed information and then sends it to the Base Station through an optimal multihop tree formed [21] between cluster heads (CHs) with base station as root node as shown in figure 5.

Multihop intra-cluster operation: However, we note that in general using single hop communication within a cluster for communication between the sensor nodes and the cluster heads may not be the optimum choice. When the sensor nodes are deployed in regions of dense vegetation or uneven terrain, it may be beneficial to use multi-hop communication among the nodes in the cluster to reach the cluster head. As it is possible for nodes to remain disconnected from the network due to a cluster head not being in range, each node is able to request another connected node to become a cluster head. This occurs after a timeout period and is done through a normal advertisement message.

4. Implementation and simulation

All routing protocols are implemented with TinyOS [17] using the nesC [16] programming language. A complete application utilizing the library components of TinyOS is developed to test the protocol. The TOSSIM [15] simulator, which builds directly from the TinyOS components is used to simulate implemented protocols. TOSSIM is provided free with TinyOS. It is designed to emulate a sensor network running TinyOS on a PC. TOSSIM also provides a graphical front end to a TinyOS simulation through the TinyViz program written in Java.

4.1 Introduction to TinyOS

TinyOS is an event driven operating system designed for sensor networks, where demands on concurrency and low power consumption are high but the hardware resources are limited [17]. The main strength of TinyOS is that it has a very small footprint – the kernel which occupies

approximately 100kb of memory. This means that most of the precious available memory can be allocated to application needs. TinyOS can also be executed on microprocessors that support clock speeds of 5MHz or less as is the case wireless sensor network hardware. Aside from the kernel, TinyOS comes equipped with many support tools, library routines and sample applications and full source is provided. This archive contains the following components:

TinyOS core: Operating System Kernel and Run-time routines.

nesC compiler: An extension to the GNU compiler system.

Sample Applications inc. nesC source code: Applications written in nesC which demonstrate the capabilities of the system and provide a base for extension and adaptation to specific requirements.

Library Routines and System Components (inc. nesC source code): Most importantly, the TinyOS package contains a well-defined hierarchy of system library components. These components provide an abstraction layer for communication and components such as sensors etc.

TOSSIM (TinyOS Simulator): This program is a WSN simulator allowing the simulation of 1000's of motes (discussed in more detail later)

Debugging Tools: There are a number of debugging tools available, including TOSSIM which allow the programs to be interrogated during execution and program states and system calls to be echoed to a PC terminal screen.

Documentation: Documentation is provided for all components although fairly limited. The nesC compiler can also be invoked to produce documentation from source code.

Tutorial: A tutorial in HTML format is also available within the TinyOS downloadable archive and on the web.

4.2 Introduction to nesC

The Network embedded system C (nesC) is an open source programming language is specialized for sensor networks [16]. It is an extension of the C programming language which was designed to facilitate the implementation of the structuring concepts and execution model of TinyOS. nesC was primarily designed for use with embedded systems such as sensor networks. nesC defines a component based model in order to make it possible to split applications into separate parts which communicates with each other using bidirectional interfaces. nesC does not permit separate compilation as C does. This is because nesC uses whole program analysis to

improve the performance and make the source code safer. Because the size of the application often is relatively small the need for separate compilation is not very critical.

In nesC there is a separation of construction and composition. Programs are built out of components which are 'wired' together to form whole programs. In nesC components provide and use bidirectional interfaces which are the only way to access a component. An interface declares a set of commands which must be implemented by the interface provider and a set of events which must be implemented by the user of that interface. If a component wishes to call a command in an interface it must implement the events associated with that interface. The only communication between components is by commands and events. Commands and events are similar to functions and methods in other languages and are used in the same way.

4.3 Introduction to TOSSIM

TOSSIM is a discrete event simulator for TinyOS sensor networks [15]. Instead of compiling a TinyOS application for a mote, users can compile it into the TOSSIM framework, which runs on a PC. This allows users to debug, test, and analyze algorithms in a controlled and repeatable environment. As TOSSIM runs on a PC, users can examine their TinyOS code using debuggers and other development tools.

The main aim of TOSSIM is to provide a high fidelity simulation of TinyOS applications. In order to achieve this, the focus of TOSSIM is to simulate the execution of TinyOS as opposed to simulating the real world. TOSSIM is very flexible and allows the simulation of thousands of motes with differing behavior in a variety of environments.

The advantage of TOSSIM over alternative simulators is that it is native to TinyOS and nesC source code. TinyViz is a Java visualization and actuation environment for TOSSIM. TinyViz provides an extensible graphical user interface for debugging, visualizing, and interacting with TOSSIM simulations of TinyOS applications. Using TinyViz, we can easily trace the execution of TinyOS apps, set breakpoints when interesting events occur, visualize radio messages, and manipulate the virtual position and radio connectivity of motes. TinyViz supports a simple "plugin" API that allows us to write our own TinyViz modules to visualize data in an application-specific way, or interact with the running simulation.

4.4 Implementation

Implementation is carried out in two stages. In the first stage two flat based multihop routing protocols namely Flooding & Gossiping and one cluster based protocol Multihop-LEACH are implemented, analyzed and compared. The results clearly show that cluster based protocol Multihop-LEACH is more energy efficient than

Flooding and Gossiping. In the second stage Multihop-LEACH is further modified and evaluated with varying probability of clustering to improve success rate and to extend network life time. It is proved that increasing the probability of clustering will improve the energy consumption of Multihop-LEACH routing protocol.

As all protocols use multihop routing technique as shown in figure 6, they use MHEngine (multiop engine) module of TinyOS to broadcast and route packets. Selected routing protocol will enable route select module and Path Selection Module (PSM) to select route for data forwarding between sensor nodes. The selected path is sent to MHEngine. The multihop component architecture is shown below.

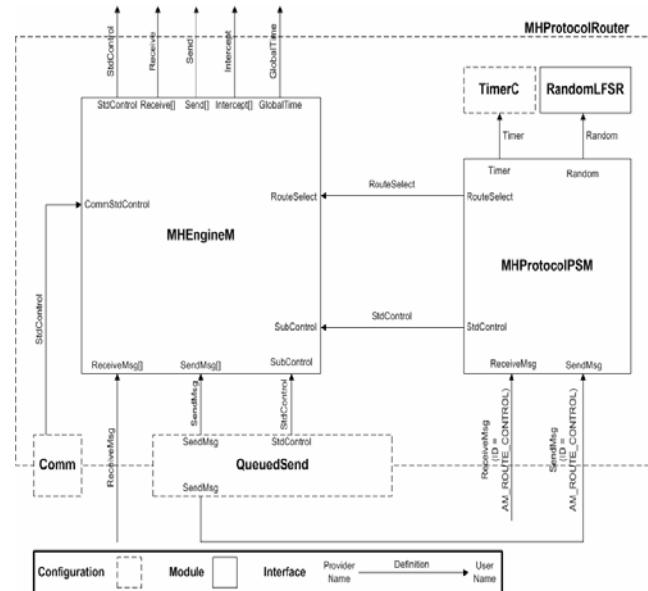


Fig. 6: Multihop component architecture

5. Simulation metrics and results

5.1 Evaluation metrics [18, 19, 20]

Latency: This performance metric is used to measure the average End-to-End delay of data packet transmission. The End-to-End delay implies the average time taken between a packet initially sent by the source, and the time for successfully receiving the message at the

destination. Measuring this delay takes into account the queuing and the propagation delay of the packets. The time taken to deliver a packet to the base station from the origin node will be looked at when evaluating the protocols. In addition the per hop time delay will also be looked at. Lower latency is preferable to higher latency.

Battery usage: The power consumption is the sum of used power of all the nodes in the network, where the used

power of a node is the sum of the power used for communication, including transmitting (P_t), receiving (P_r), and idling (P_i). The amount of power used during the simulation will be monitored and used for evaluating the protocols. Batteries have a finite amount of power and nodes die once power runs out. For this reason lower power usage is preferable to higher power usage. In addition the distribution of power usage across the network will be looked at. Uniform drain is preferable.

Success rate: The number of packets received from a node at the base station will be compared with the number of packets sent by a node in order to calculate the Success rate.

Connectivity: The number of nodes that have a route to the base station will be used to assess the node connectivity provided by a particular routing protocol. More connected nodes in a network are preferable to fewer connected nodes.

5.2 Simulation and implementation parameters

The parameters used in simulating the protocols are given below in table 1.

Table 1: Summary of the parameters used in the simulation

| parameters | Value |
|-------------------------------|--|
| Simulation time | 1800 sec |
| Number of node | 10,20,30,40,50 |
| Routing protocols | Flooding, Gossiping, Multihop-LEACH |
| Nodes distribution | randomly distributed |
| Network topology | Loss topology generated using LossyBuilder of TOSSIM . |
| Max. Packet(message) size | 29 bytes |
| Transmitting power per packet | 3 power units |
| Receiving power per unit | 2 power units |
| Message generation rate | 1 message per 5 seconds. |
| CH probability | 10%,25%,40%,50% |
| Nodes distribution | Nodes are randomly distributed |
| Operating system | TinyOs |
| Simulator | TOSSIM |
| Programming language | nesC |

5.3 Simulation test cases

The various simulation test cases used in evaluating the three routing protocols that are implemented in two stages are given below in table 2. Multihop-LEACH with 50 nodes is evaluated by varying the probability of clustering.

Table 2: Simulator test cases

| The test cases used in the I stage of implementation | | | |
|---|----------------|-----------------|---------------------------|
| Test | Protocol | Number of Nodes | Probability of clustering |
| i. | Flooding | 10 | |
| ii. | Gossiping | 10 | |
| iii. | Multihop-LEACH | 10 | 25% |
| iv. | Flooding | 20 | |
| v. | Gossiping | 20 | |
| vi. | Multihop-LEACH | 20 | 25% |
| vii. | Flooding | 30 | |
| viii. | Gossiping | 30 | |
| ix. | Multihop-LEACH | 30 | 25% |
| x. | Flooding | 40 | |
| xi. | Gossiping | 40 | |
| xii. | MULTIHOP-LEACH | 40 | 25% |
| xiii. | Flooding | 50 | |
| xiv. | Gossiping | 50 | |
| xv. | Multihop-LEACH | 50 | 25% |
| The test cases used in the II stage of implementation | | | |
| xvi. | Multihop-LEACH | 50 | 10% |
| xvii. | Multihop-LEACH | 50 | 25% |
| xviii. | Multihop-LEACH | 50 | 40% |
| xix. | Multihop-LEACH | 50 | 50% |

5.3 Results

Simulated results obtained using TOSSIM can be viewed and tested in two ways. One way of visualizing the output by using a graphical tool TinyViz and the other way is by storing the results in a output text file.

Output from graphical tool TinyViz: A sample graphical display depicted in figure 7 shows that all the cluster head nodes send a packet to the base station using Multihop-LEACH protocol with a network of 50 nodes.

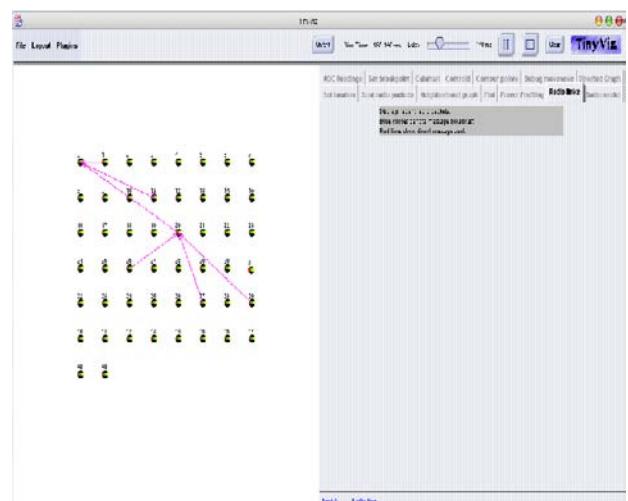


Fig. 7: Cluster head nodes send a packet to the base station.

Results obtained from output file in I stage of implementation: The output from the simulator stored in an output file has been processed in order to evaluate the protocols based on the metrics specified. The processed results are depicted below.

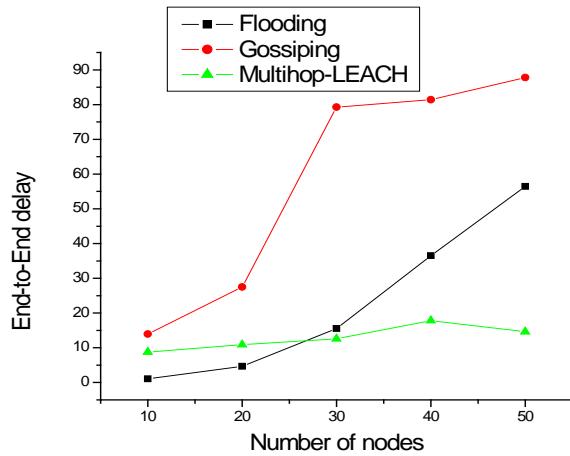


Fig. 8: Number of Nodes Vs End-to-End Delay

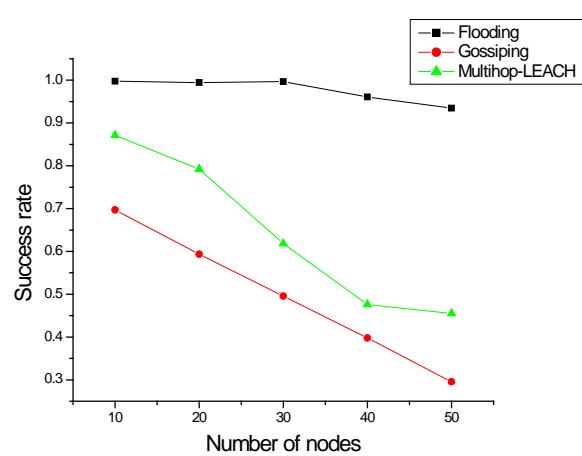


Fig. 11: Number of Nodes Vs Success rate

Results obtained from output file in the II stage of implementation: The probability of becoming Cluster Head in every node is further increased to improve the latency, success rate and connectivity of Multihop-LEACH protocol.

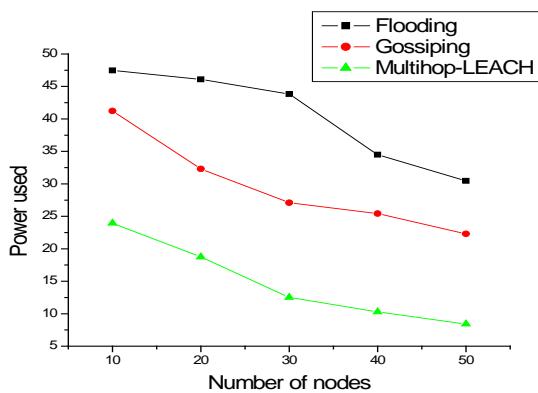


Fig. 9: Number of Nodes Vs Power usage per message

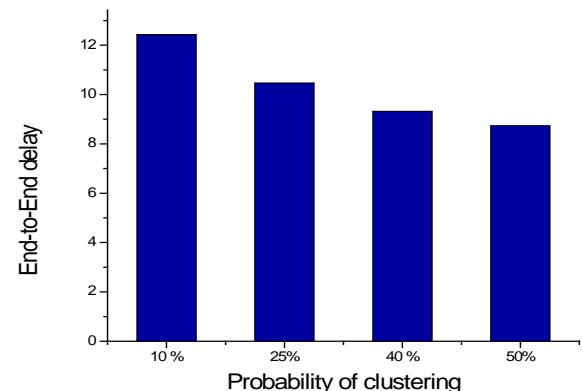


Fig.12: Probability of clustering Vs Latency (End-to-End dealy)

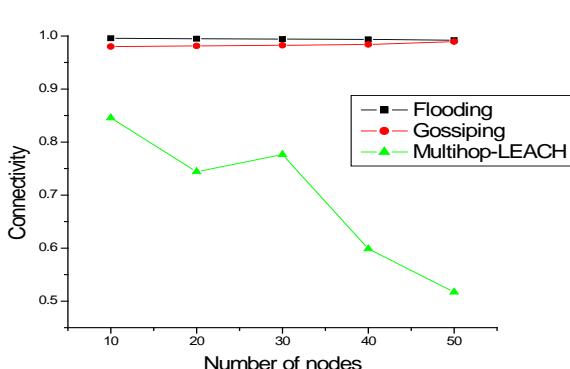


Fig.10: Number of Nodes Vs Connectivity

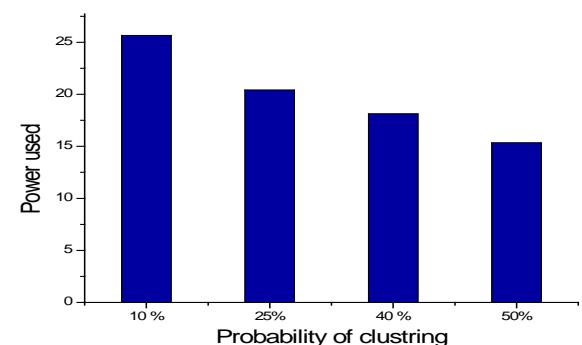


Fig. 13: Probability of clustering Vs Power usage per message

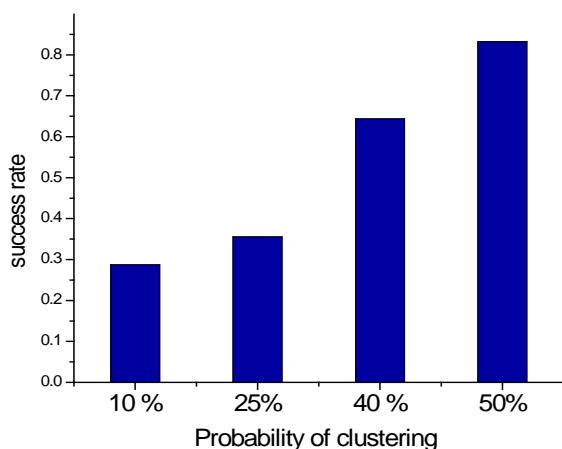


Fig. 14: Probability of clustering Vs Success rate

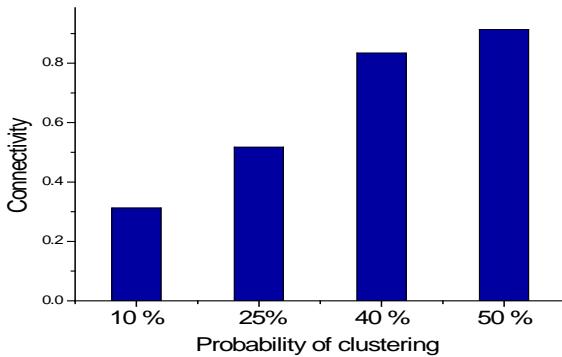


Fig. 15: Probability of clustering Vs Connectivity

Evaluation of results: Result of evaluation clearly indicates that Flooding is the worst in case of power efficiency. Gossiping provides some improvement over Flooding in terms of power usage per message. Power usage per message is less for Multihop-LEACH when compared to other two protocols. Connectivity and End-to-End delay are more in Flooding and gossiping compared Multihop-LEACH as Multihop-LEACH depends on distribution of cluster head nodes around the network. Success rate in gossiping is less because many packets will get dropped when the packets hop count reaches a maximum limit. In Multihop-LEACH success rate and connectivity are improved by increasing the probably of clustering for every nodes.

6. Conclusion and future work

The overall conclusion is that Multihop-LEACH routing protocol is best choice to move towards a network with less energy consumption as it involves energy minimizing techniques like multihop communication, clustering and data aggregation. For applications like military where

energy consumption is not much to be bothered and more performance is required, Flooding is the best choice as at is simple to construct. For applications where network subjected to more scalability like environmental monitoring, Gossiping is the best choice as it uses a medium amount of power and no matter how large the network is, each node uses roughly the same amount of power. For applications where energy utilization is more critical like health monitoring, Multihop-LEACH is the best choice. Multihop-LEACH uses both inter cluster as well as intra cluster communication. The power usage, latency and success rate in Multihop-LEACH can further improved by increasing probability of clustering. We can still minimize the energy consumption and extend the network life time by improving the clustering technique. The main limitation of using TinyOs simulator is that multihop engine requires at least 2 sec between two message generations to avoid congestion and hence it requires more simulating time for evaluating the protocols with increase in network size.

References

- [1] Jamal N. Al-Karaki Ahmed E. Kamal, “Routing Techniques in Wireless Sensor Networks: A Survey”, IEEE Wireless Communications, 2004, vol.11 pp: 28.
- [2] Sarjoun S. Doumit, Dharma P. Agrawal, “Self-Organizing and Energy-Efficient Network of Sensors”, IEEE, 2002, pp. 1-6.
- [3] I.F. Akyildiz, W Su, Y. Sankarasubramaniam and E Cayirci, “Wireless Sensor Networks, A Survey,” Communication Magazine, IEEE, August 2002, Vol. 40, Issue 8, pp. 102-114.
- [4] W.Heinzelman, “Application specific protocol architectures for wireless networks”, *PhD Thesis*, MIT, 2000.
- [5] K. Akkaya, M. Younis,“A Survey on Routing Protocols for Wireless Sensor Networks”, Ad-hoc Networks, May 2005, Vol. 3, No. 3, pp. 325-349.
- [6] F. Ye, A. Chen, S. Liu, L. Zhang, “A scalable solution to minimum cost forwarding in large sensor networks”, Proceedings of the tenth International Conference on Computer Communications and Networks (ICCCN) , 2001, pp. 304-309.
- [7] M. Chu, H. Hausscker, and F. Zhao, “Scalable Information-Driven Sensor Querying and Routing for ad hoc Heterogeneous Sensor Networks”, in the International Journal of High Performance Computing Applications, Vol. 16, No. 3,, August 2002.
- [8] D. Braginsky and D. Estrin, “Rumor Routing Algorithm For Sensor Networks”, International Conference on Distributed Computing Systems (ICDCS’01), November 2001.
- [9] W.Heinzelman, J. Kulik, and H. Balakrishnan, “Adaptive Protocols for Information Dissemination in Wireless Sensor Networks”, Proc. 5th ACM/IEEE Mobicom Conference (MobiCom ’99), Seattle, WA, August 1999. pp. 174-85.

- [10] J. Kulik, W. R. Heinzelman, and H. Balakrishnan, “Negotiation-based protocols for disseminating information in wireless sensor networks”, *Wireless Networks*, 2002, Vol. 8, pp. 169-185.
- [11] K. Sohrabi, J. Pottie, “Protocols for self-organization of a wireless sensor network”, *IEEE Personal Communications*, 2000, Vol. 7, Issue 5, pp 16-27.
- [12] J.-H. Chang and L. Tassiulas, “Maximum Lifetime Routing in Wireless Sensor Networks”, Proc. Advanced Telecommunications and Information Distribution Research Program (ATIRP2000), College Park, MD, Mar. 2000, pp 334-335.
- [13] C. Rahul, J. Rabaey, “Energy Aware Routing for Low Energy Ad Hoc Sensor Networks”, *IEEE Wireless Communications and Networking Conference (WCNC)*, vol.1, March 17-21, 2002, Orlando, FL, pp. 350-355.
- [14] W.R.Heinzelman ,A.Chandrakasan and H.Balakrishnan, “Energy-Efficient Communication Protocol for Wireless Microsensor Networks”, *33rd Hawaii International Conference on System Sciences*, Volume 8, January 2000.
- [15] P. Levis and N. Lee, “TOSSIM A Simulator for TinyOS Networks”, Included with the TinyOS 1.1.0 software, September 2003
- [16] D. Gay, P. Levis, D. Culler and E. Brewer, nesC 1.1 Reference Manual, Included with the TinyOS 1.1.0 software, May 2003
- [17] TinyOS Home Page,
<http://webs.cs.berkeley.edu/tos/index.html>, April 2004
- [18] Rajashree.V.Biradar, V.C Patil, Dr. R. R. Mudholkar , Dr. S. R. Sawant , “Classification And Comparison Of Routing Protocols In Wireless Sensor Networks”, *Ubiquitous Computing and Communication Journal*, 2009, volume 4, pp.704-711.
- [19] C. F. Chaisserini, M. Garetto, “Modeling the Performance of Wireless Sensor Networks”, *IEEE INFOCOM 2004, 23 Annual Joint Conference of the IEEE Computer and Communications Societies*, Hong Kong, China, 7-11 March 2004, Vol. 1, pp. 231.
- [20] S. Dai, X. Jing, and L. Li, “Research and analysis on routing protocols for wireless sensor networks Communications, Circuits and Systems Proceedings”, *International Conference on IEEE*, May 2005, vol. 1, pp. 407–411.
- [21] Dissertation, Hang Zhou, Zhe Jiang and Mo Xiaoyan, “Study and Design on Cluster Routing Protocols of Wireless Sensor Networks”, 2006.

Service-Oriented Architecture and model for GIS Health Management: Case of cancer in Morocco

Zouiten Mohammed¹, Harti Mostafa² and Nejari Chakib³

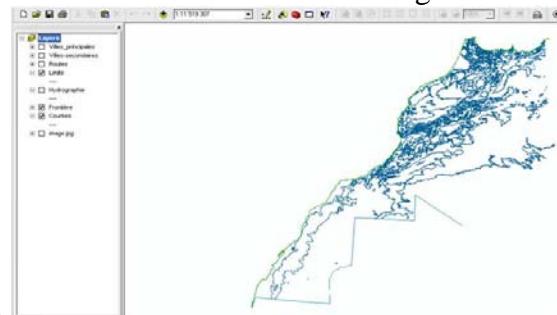
¹ Department of computer Sciences , U.S.M.B.A , Faculty of Sciences Dhar El Mehraz
Fez, 30 000, Morocco

² Department of computer Sciences , U.S.M.B.A , Faculty of Sciences Dhar El Mehraz
Fez, 30 000, Morocco

³ Epidemiology Department, U.S.M.B.A, Faculty of Medecine
Fez, 30 000, Morocco

Abstract

In morocco, the prevalence on cancer cases increased and became ranked as the second cause of death. Therefore, forming a cancer control program and putting strategic action plans into practice became an important matter for health industry. The correlation of variations in different societies and environmental factors should be examined spatially with reliable data. To do this, cancer occurrence density maps have to be created. In this study, a database was built with the use of GIS to examine the distribution of cancer cases, and maps relating to cancer events in allocation units were created. Cancer cases data registered in 2010 by Ministry of health, and lalla Salma Association to fight against cancer and National Cancer Prevention and Control Plan were used. Using ArcGIS 10



[1],
Figure 0: Moroccan map region prevalence

The distribution of cancer cases was presented on cancer maps including allocation units and incidence values, which were calculated for each town-based region. According to the world standards, cancer rates were determined by the special analysis power of GIS.

Keywords: Oriented Service Architecture; GIS Health, profiles, Cancer, Morocco

1. Introduction

Nowadays according to the National Cancer Prevention and Control Plan 2010-2019; 30 000 new cancer cases per year are registered including 1200 cases of childhood cancer which means 4% of the overall rate. Among Women: breast cancer is the most common cancer representing 36% of cases. Rates of cervical cancer are substantially lower than breast cancer with a ratio of 13%. Among Men: lung cancer represents 24% of cases and prostate cancer represents 8% of the overall rate [3].

Geographical Information Systems (GIS) has strong capabilities in mapping, analyzing not only spatial data, but also non-spatial data, and integrating many kinds of data to greatly enhance disease surveillance. It can render disease data along with other kinds of data like environmental data, representing distribution contagious disease with various cartographical styles.

Meanwhile, the rapid development of the internet influences the popularity of web-based GIS, which itself shows a great potential for sharing disease information through distributed networks.

1-1 GIS and decision making

Distributing and sharing disease maps via the web could help decision makers across health jurisdictions and authorities collaborate in preventing[2], controlling and responding to a specific disease outbreak and it's time factor analyzing. By comparing the thematic maps at different time intervals, the spatial-temporal change of disease could be projected, including temporal cluster shift and vector transmission rates and mobility of susceptible populations.

2. Related Work

There are a number of research projects related to telemedicine. Here we will focus on works proposing context-aware systems to support assistance of patients outside hospital. Most attempts have focused on such as health status monitoring and alert (e.g. medicine taking, training activities, etc.), patient behavior and daily activities modeling. Our reference scenarios include management and immediate accounting of cases for predicting risk [1].

Main objectives of pilot project are: defining a model for early detection of breast cancer fully adapted to Moroccan specifications, and defining each component of this model.

3. Method

This method has 8 steps. Firstly the study of a breast cancer registry: ages 45 to 65 years. Then identify recruitment strategy via primary health providers. After, a method is choosing for screening adapted to our resources: Clinical Breast Examination. Then develop a curriculum

for CBE by organizing training for health workers: doctors and nurses. Then a diagnostic with mammography must be done. Then organize a taking charge at the third health level from an oncology center for women with a positive diagnostic. After, developing an Information system, including GIS functionalities to automate the process of early detection of breast cancer and provide relevant data to each step respecting the frame work presented in figure 1.

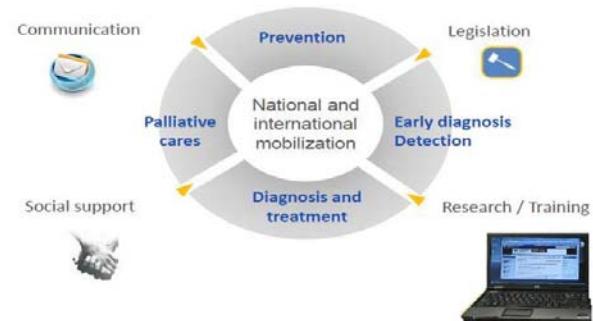


Figure 1: Conceptual frame work .
 The last step consists to develop a strategic health communication tools based on an architecture oriented services [11] like exposed on figure 2.

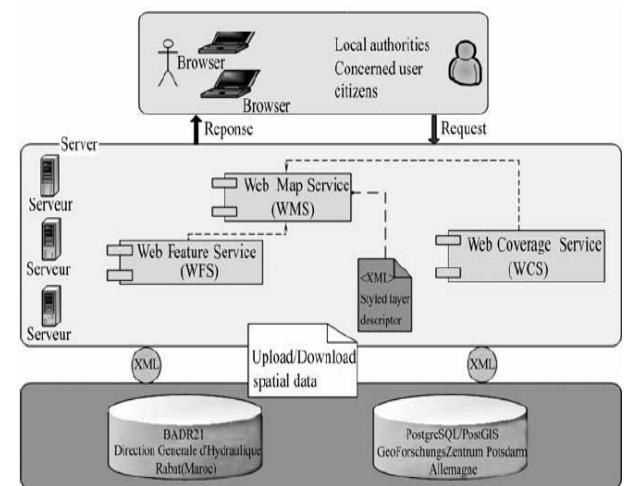


Figure 2: project program for early detection [7].

To identify vulnerable populations and areas at risk of tuberculosis in morocco and in order to

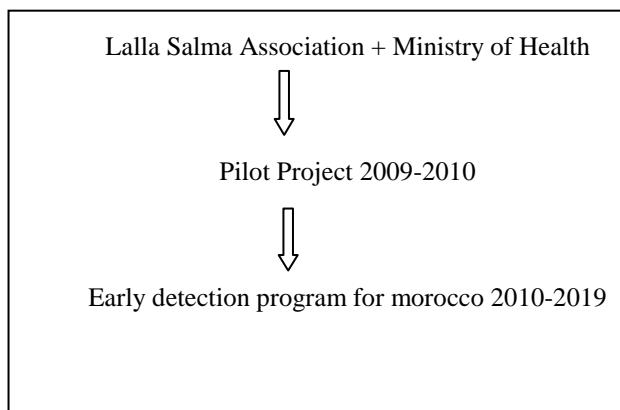
optimize the surveillance and prevention in relation to populations and medical human resources and cancer cases. The aim is to understand the geographic distribution and dynamic of this pandemic of cases, to characterize risk factors and human environmental impact that may explain this distribution and to identify these sensitive areas.

3-1 Ontology based cancer modeling

In this section, we will describe the main concepts of the ontology-based disease modeling approach for patient assistance and awareness scenario. In morocco and especially in fez, no geographical parameter is taken into account according to a standard sheet UHC Hassan II of fez [4].

Over 10 years (2010 to 2019) the National cancer prevention Control Plan envisages 74 operational measures.

In Morocco the breast cancer is the most frequent cancer. Women arrive at the late stage of this disease. Unfortunately , there is no early detection program in Morocco, that is why it gives high priority for the NCPCP 2010-2019 according to the below schema :



As already mentioned, we extend an ontology-based cancer representing main general concepts and relations for context representation. Our work moves from the widely accepted definition of context, provided in [3]: "Context is any information that can be used to characterize the situation of an entity." Therefore, an entity is a patient, place, computational entity, or object

which is considered relevant for determining the behavior of an application [5].

Hereafter we describe the following ontologies: Patient personal domain ontology and awareness ontology. This tatter ontology represent care networks resources coming from different organizations (health teams, social community members, etc.).

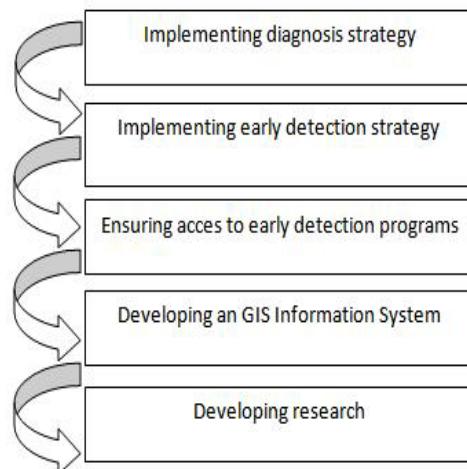


Figure 3: general context ontology

In this application scenario, context includes data items describing patient geographical localization and disease context. Context reasoning is used mainly for awareness, and management and fight.

As already mentioned, we extend an ontology-based context model representing main general concepts and relations for context representation. Our work consists of defining a model for early detection of breast cancer and defining each component for this model fully adapted to Morocco:

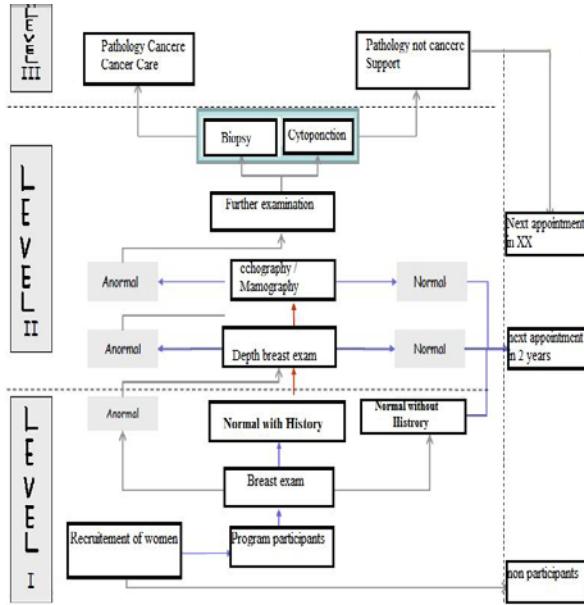


Figure 4: Model for early detection of cancers

The context model has been written in OWL (web Ontology Language) [4]. OWL fragments are hereafter represented by means of UML class diagrams. UML classes represent OWL classes, attributes represent OWL data-type properties and associations among classes are used for OWL Object Properties representation [8].

Figure 4 illustrates a fragment in the context of ontology specialized for the patient personal domain. A specialization of medical treatments which are monitored (tests and positions) is sufficient for early detection of breast cancer ,but model can be easily extended to include further levels and used for other type of cancers and diseases. In order to overcome user context requirement, we focused on preference-driven formalization: In fact, users are described in terms of profile and preferences. User Profile is composed of both dynamic and static metadata [9]. Dynamic properties include, for example user locality and town, while static properties are grouped into three categories: identification such as an ID code, a string or an URL, is used to name and identify the user. Capabilities: represent the user's abilities and capacities. User requirements: describe user's parameters that

must be always satisfied during service provisioning like presented in figure 5:

```

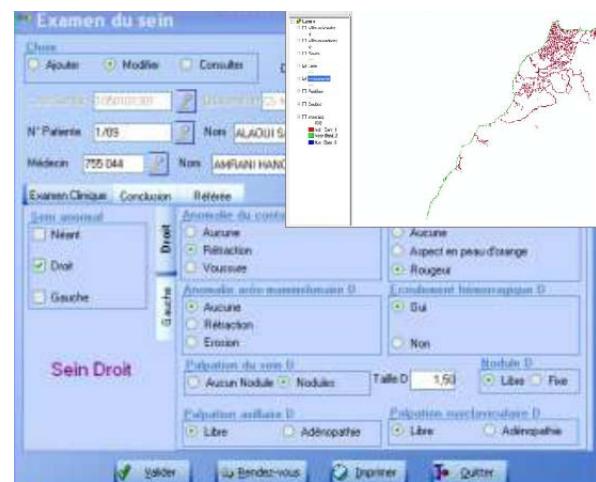
<rdf:RDF>
<profile:User rdf:ID="med">
    <profile:hasProfile>
        <profile:Profile rdf:ID= "med_Profile">
            <id:Name
                rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
                    Alice Brown</id:Name>
            </profile:profile_id>
            <profile:profile_cap>
                <user_cap:LanguageCapability rdf:ID="LanguageCap_1">
                    <user_cap:speaks rdf:resource="&language-ont;arabic "/>
                    <user_cap:speaks rdf:resource="&language-ont;french "/>
                </user_cap:LanguageCapability>
            </profile:profile_cap>
            <profile:profile_req>
                <user_cap:Requirement rdf:ID="Requirement_1">
                    <profile:requires>...</profile:requires>
                </user_cap:Requirement>
            </profile:profile_req>
        </profile:Profile>
    </profile:hasProfile>
</profile:User>
</rdf:RDF>

```

Figure 5: Static user profile specifications

4 . Results

This pilot project was launched in the téma-skhirat in September 2009. This region includes rural and urban population. All steps of the project were precisely implemented. After 3 months, 1200 women were screened, and 4 women were diagnosed and treated for breast cancer. It was too early to have accurate statistics, but the process was triggered and the health workers involved and used software tools as shown in Figure 6.



Cervical cancer is the leading cause of cancer-related mortality among women in developing

countries and accounts for more than 290,000 deaths worldwide each year. In Morocco, mortality is presented in the below figure [13]

| Indicator | Morocco | Northern Africa | World |
|--|---------|-----------------|--------|
| Crude mortality rate ¹ | 7.2 | 3.0 | 8.2 |
| Age-standardized mortality rate ¹ | 8.4 | 4.0 | 7.8 |
| Cumulative risk (%) ages 0-74 years ¹ | 0.9 | 0.5 | 0.9 |
| Annual number of deaths | 1152 | 3101 | 275128 |

Standardized rates have been estimated using the direct method and the World population as the reference.

¹ Rates per 100,000 women per year.

Data sources:

IARC. Globocan 2008. (Specific methodology for Morocco: The number of cancer deaths in 2008 was estimated from incidence estimates and site specific survival, estimated by the GPD method. For further details refer to http://globocan.iarc.fr/DataSource_and_Methods.asp and <http://globocan.iarc.fr/method/method.asp?country=504>.)

Figure 6: Mortality of cervical cancer in morocco, northern Africa and the world

5. Conclusion

Health authorities in Morocco have begun a campaign to fight cancer by opening new treatment centres and expanding health-care coverage. Morocco launched the 8-billion dirham campaign, which aims to make treatment, detection and preventive care more accessible, on March the 23rd. Four regional health-care centres will be opened in Safi, Laayoune, Meknes and Tangier, in addition to two special cancer centres for women in Rabat and Casablanca, and two paediatric cancer centres in Fes and Marrakech. Palliative care units will be added to several provincial hospitals, while existing oncology centres in Morocco will be expanded. Morocco currently has five state-run cancer centres and four private-sector facilities to treat the disease, which accounts for 7.2% of all deaths annually, with 30,000 new cases diagnosed each year. "The plan has come at just the right time to address the growing need to combat cancer at the national and regional level, and reflects Morocco's commitment to adopting a regional strategy on the issue," said the WHO regional director for the Eastern Mediterranean, Hussein Gezairy.

The anti-cancer campaign will also expand cancer patients' right to receive health-care benefits to offset the high costs of treatment, which is especially critical in a country where two-thirds of citizens have no health-care coverage. According to the Health Ministry, up to 90% of the treatment costs for certain types of cancer are borne by the patient, which in turn impoverishes them and their families. Fatiha, a 52-year-old housekeeper, knows first-hand how steep the costs of treatment can be after undergoing a mastectomy and chemotherapy. "Each session costs me 2,600 dirham," she told Magharebia. "Benefactors are helping me to get treatment. Without them, I'd have been dead long ago." A significant portion of the anti-cancer campaign will focus on prevention and early detection. To further this aim, the Health Ministry will build more than 30 screening centres throughout the country over the next 10 years to screen women for early signs of breast and cervical cancers. The campaign will also highlight preventive measures individuals can take to prevent the onset of the disease by living a healthier lifestyle, stopping smoking and avoiding other carcinogenic products. Around 40% of all cancers are preventable, cancer specialists claim. Health Minister Yasmina Baddou praised the plan for its "ambitious yet realistic response to cancer" and its efforts to provide affordable, high-quality care for those who suffer from long-term illnesses.

References

[1] M. ZOUTEN, M. HARTI, and C.NEJJAR, “Adaptation of covertes’ care and alert systems favoring localization in cases declaration via Internet: Tuberculosis in Morocco”, presented at the GIS 2010 francophone conference ESRI Versailles 30 September 2010

[2] M. ZOUTEN, M. HARTI, and C. NEJJARI, “Relationship between GIS and Data warehouse for decision making”, presented at the GIS 2010 francophone conference ESRI Versailles 30 September 2010

[3] Early detection for breast cancer in morocco, program and information system. World cancer congress, August 2010 China

[4] Ministry of Health- Morocco

<http://www.sante.gov.ma/>

[5] F.Paganeli, and D. Giuli “An ontology-based context model for home health monitoring and alerting in chronic patient care networks”

[6] S.Gao, D. Mioc, F. Anton, X. Yi, and D. Coleman, “Online GIS services for mapping and sharing disease information” IJHG 2008

[7] R. Stones, N. Mathew (2001) Beginning databases with postgresQL [M]. Chicago: Wrox press

[8] J. Oulidi, and H. Benaabidate, L. Löwner, and al. (2008) Management strategies of water resources in the arid zone of South-Eastern Morocco. New York: Springer

[9] Copyright © 2009 Formation SAS - Stage SASBI - Data warehouse - Datamart.

<http://www.formations-sas.fr/data-warehouse>

[10] Donnay, Jean-Paul*Pantazis, Dimos N.

« La conception de SIG, méthode et formalisme » – Paris : Hermès, 1996.

[11] Li Shiming, Saborowski J, Nieschulze J, et al. (2007) Web service based spatial forest information system using an open source software approach[J]. Journal of Forestry Research, 18(2): 85-90

[12] Poth A, Müller M, Schmitz A, et al.(2007) Deegree web map service .2.1[OL].
<http://www.deegree.org>

[13] Human Papillomavirus and related cancers.
Summary report update. September 15, 2010

Acknowledgments

I would like to express my gratitude to my mother for her sacrifices, to Mr Farih for his help and to my supervisors, Dr. Harti Mostafa and Dr. Nejjar chakib; for their expertise, understanding, and patience, added considerably to my graduate experience. I appreciate them vast knowledge and skill in many domains, and their assistance.

Zouiten Mohammed PhD student in GIS for Health, he has a Masters degree in Computer Engineering, professor of computer engineering in IGA fez. He published an article (IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.11, November 2010 218-222). He is a student in the Department of Computer Science of Faculty of Sciences Dhar El Mehraz of Fez; University Sidi Mohammed Ben Abdellah. Supervised by Professor Mostafa HARTI: Professor at the Faculty of Sciences Dhar El Mehraz of Fez ; University Sidi Mohammed Ben Abdellah and Professor Chakib NEJJARI Professor Epidemiologist at the Faculty of Medicine of fez in the same University. His research interests are: GIS for health, modelization object oriented and Ontology Engineering.

Token Ring Algorithm To Achieve Mutual Exclusion In Distributed System – A Centralized Approach

Sandipan Basu

Post Graduate Department of Computer Science, St. Xavier's College, University of Calcutta
Kolkata-700016, INDIA

Abstract

This paper presents an algorithm for achieving mutual exclusion in Distributed System. The proposed algorithm is a betterment of the already existing Token Ring Algorithm, used to handle mutual exclusion in Distributed system. In the already existing algorithm, there are few problems, which, if occur during process execution, then the distributed system will not be able to ensure mutual exclusion among the processes and consequently unwanted situations may arise. The proposed algorithm will overcome all the problems in the existing algorithm, and ensures mutual exclusion among the processes, when they want to execute in their critical section of code.

Keywords: *TIMESTAMP_OF_REQUEST_GENERATION, PID, NEW, REQUEST_QUEUE, EXISTS, COORDINATOR, UPDATE.*

1. Introduction

In most of the distributed systems it is very common that, resources are being shared among various processes, with the condition that a single resource can be allocated to a single process at a time. Therefore, mutual exclusion is a fundamental problem in any distributed computing system. So, the goal is to find a solution that will synchronize the access among shared resources in order to maintain their consistency and integrity.

In this paper, the proposed algorithm is able to handle the problems of mutual exclusion in a distributed system. It is also able to handle all other problems that may arise, while a process is executing in its critical section.

2. Existing Work

Till now, several token-based algorithms have been proposed. Some of them are –

- | | |
|-------------------------------|------------|
| 1. Ricart - Agrawala | algorithm. |
| 2. Suzuki - Kazami | algorithm. |
| 3. Mizuno - Neilsen - Rao | algorithm. |
| 4. Neilson - Mizuno | algorithm. |
| 5. Helary – Plouzeau – Raynal | algorithm. |
| 6. Raymon'd | algorithm. |
| 7. Singhal's | algorithm. |
| 8. Maimi – Trehel | algorithm. |
| 9. Misra- Srimani | algorithm. |
| 10. Nishio – Li – Manning | algorithm. |

3. Preconditions

The effectiveness of an algorithm depends on the validity of the assumptions that are made. In distributed mutual exclusion environment certain assumptions must be considered to make the work successful. The following assumptions are made for this algorithm:-

1. All nodes in the system are assigned unique identification numbers from 1 to N.
2. There is only one requesting process executing at each node. Mutual exclusion is implemented at the node level.
3. Processes are competing for a single resource.
4. At any time, each process initiates at most one outstanding request for mutual exclusion.
5. All the nodes in the system are fully connected.

6. A process can enter only one critical section when it receives the token. If it wants to enter another critical section, the process must send another request to the coordinator.
7. Instead of using globally synchronized physical clocks, Lamport's concept of logical clock is used in distributed system that we are considering. The concept of logical clock is the way to associate a timestamp (simply a number independent of any clock time) with each system event so that events that are related to each other by the ***happened-before*** relation (directly or indirectly) can be properly ordered in that sequence. To ensure that all events that occur in a distributed system can be totally ordered, the use of **any arbitrary total ordering of events**, proposed by Lamport, is applied in the distributed system considered in this algorithm. For instance, if events a and b happen respectively in processes P1 and P2, and both events will have a timestamp of (say) 40, then according to Lamport's proposal the timestamp associated with events a and b will be 40.1 and 40.2 respectively, where the process identity number (PID) of processes P1 and P2 are 1 and 2 respectively. Using this technique, we will be able to assign unique timestamp to each event in a distributed system to provide a total ordering of all events in the system.

As we are considering distributed systems, some assumptions also need to make about the communications network. This is very important because nodes communicate only by exchanging messages between them. The following aspects about the reliability of the distributed communications network should be considered.

1. Messages are not lost or altered and are correctly delivered to their destination in a finite amount of time.
2. Messages reach their destination in a finite amount of time, but the time of arrival is variable.
3. Nodes know the physical layout of all nodes in the system and know the path to reach each other.

4. Algorithm

In this algorithm, we consider that a set of processes are logically organized in a ring structure. Between several

processes, one process acts as a coordinator. It is the coordinator's task to generate a token and circulate the token around the ring, as needed. In this algorithm, we consider that the token can move in any direction as per the necessity. In the ring structure, every process maintains the current ring configuration of the system. If a process is removed or added into the system, then the updating must be reflected to all the processes' ring configuration table. A ring configuration table is something that contains information about process id (PID), state of the processes and their status in the current system.

The algorithm works as follows:-

Consider, the processes are numbered as P1, P2, P3...Pn. For simplicity of our discussion, consider process P1 wants to enter in its critical section. P1 will send a request [REQUEST (PID, TIMESTAMP_OF_REQUEST_GENERATION)] to the coordinator to acquire the token. A REQUEST message contains two parameters- a) PID (i.e. process id of the requesting process), b) Time Stamp of request generation. Here, two cases may appear:

Case 1:-

At this particular time, if no other process is executing in its critical section, and the coordinator retains the token, then the coordinator will send the token to the requesting process (P1). After acquiring the token, the process keeps the token and enters in its critical section. After exiting from the critical section the process returns the token to the coordinator.

Case 2:-

But, if some other process P_i ($i \neq 1$) is executing in its critical section, then the coordinator sends a WAIT signal, to the requesting process (P1) and the request from process P1 is stored in the REQUEST QUEUE, maintained by the coordinator only. In between, if any other processes want to enter into critical section and send request to the coordinator, then those requests will also be stored in the REQUEST QUEUE. When the coordinator gets back the token, then it sends the token to one of the waiting process in the REQUEST_QUEUE, which has smallest TIMESTAMP_OF_REQUEST_GENERATION.

The algorithm will overcome the drawbacks of the general token ring algorithm, in the following manner.

Loss of Token:-

When a process (say) P1 wants to enter into its critical section, it sends request to the coordinator. If the coordinator retains the token, it then sends the token to the requesting process (P1). After getting the token, the process will send an acknowledgment to the coordinator, and enters the critical section. During the execution of the critical section, P1 will continually send an EXISTS signal to the coordinator at certain time interval, so that, the coordinator becomes acquainted that the token is alive and it has not lost. As a reply to every EXISTS signal, the coordinator sends back an OK signal to that particular process (P1), so that the process that is executing in its critical section (P1), gets to know that the coordinator is alive also.

Now, suppose, the coordinator is not receiving the EXISTS signal from that process P1. Here two cases may appear:-

Case 1:-

The coordinator assumes that the token has lost. Then the coordinator will regenerate a new token and sends it to that process (P1) and again it starts executing its critical section.

Case 2:-

The process P1 may crash or fail while executing in its critical section and consequently, the coordinator does not receive any EXISTS signal from P1. Hence, the coordinator will identify it as a crashed process and update the ring configuration table and send the UPDATE signal with update information to other processes to update their own ring configuration tables.

Again it may be the case, that the process P1 (which is currently executing in its critical section), is not receiving the OK signal from the coordinator. So, P1 would assume that the coordinator is somehow crashed. At this moment, the process P1 will become the new coordinator and complete the critical section execution. The new coordinator will send a message [COORDINATOR (PID)] to every other process, that it becomes the new coordinator and send the UPDATE signal with update information, to update the ring configuration tables maintained by all other processes.

The algorithm also overcomes the overhead of token circulation in the ring. If no processes in the ring want to enter in its critical section, then there is no meaning of circulating the token throughout the ring. Rather, in this approach, the coordinator will keep the token, until any other process requests it.

- The algorithm guarantees mutual exclusion, because at any instance of time only one process can hold the token and can enter into the critical section.
- Since a process is permitted to enter one critical section at a time, starvation cannot occur.

5. Performance Analysis

The performance of the algorithm will be evaluated in terms of the total number of messages required for a node to enter into a critical section. The number of messages exchanged for an entry into a critical section to take effect will be used as a complexity measure. In this algorithm the number of messages per critical section entry varies from 3 (when the coordinator possesses the token and no other process is executing in its critical section. REQUEST->GRANT->RELEASE) to 4 (when some other process is already executing in its critical section. REQUEST->WAIT->GRANT->RELEASE).

For a total no. of n processes in the ring, the waiting time from the moment a process wants to enter a critical section until its actual entry, may vary from 0 to n-1; 0, in the case, when a process wants to enter a critical section and acquires the token immediately from the coordinator; (n-1), in the case, when the process sends the request after all other process has already requested for the token.

6. Illustration

The above algorithm can best be explained by an example.

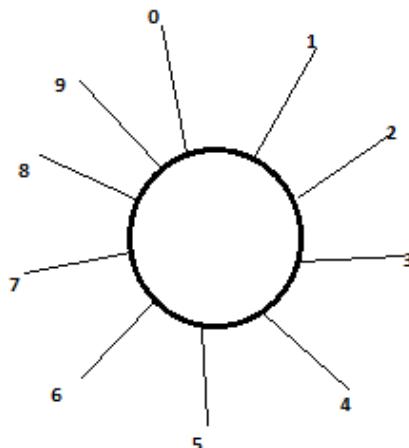


Figure 1: Ring Configuration

Consider a distributed system consists of 10 processes (P0-P9). Between these processes, one process is selected as a coordinator. Suppose when the ring is initialized, process P0 is selected as the coordinator. As soon as P0 is selected as the coordinator, it generates a token and retains the token. Now, process P5 wants to enter in its critical section. So, P5 sends a request [REQUEST (PID, TIMESTAMP_OF_REQUEST_GENERATION)] to the coordinator (P0).

Table 1: Ring Configuration Table

| Process ID | State | Status |
|-------------------|--------------|--------------------|
| P0 | Alive | Coordinator |
| P1 | Alive | Normal |
| P2 | Alive | Normal |
| P3 | Alive | Normal |
| P4 | Alive | Normal |
| P5 | Alive | Normal |
| P6 | Alive | Normal |
| P7 | Alive | Normal |
| P8 | Alive | Normal |
| P9 | Alive | Normal |

Now, if no other process is executing in its critical section, and the token has been kept by P0, then immediately the token is send to process P5. After completing the execution

of its critical section, process P5 releases the token and gives it back to the coordinator.

In this case, the situation may appear that, while process P5 is executing in its critical section, process P7 wants to enter its critical section and sends a request to the coordinator[REQUEST(PID,TIMESTAMP_OF_REQUEST_GENERATION.)]. As process P5 possesses the token and executing in its critical section, the coordinator sends a WAIT signal to process P7, and stores the request in the REQUEST_QUEUE. Now suppose, immediately after, process P2 also wants to enter in its critical section, and sends a request [REQUEST (PID, TIMESTAMP_OF_REQUEST_GENERATION)] to the coordinator. As process P5 is still executing in its critical section, so the coordinator sends a WAIT signal to process P2, and stores the request in the REQUEST_QUEUE. After process P5 has exited from its critical section, it releases the token and sends it back to the coordinator. Then the coordinator selects the process with smallest TIMESTAMP_OF_REQUEST_GENERATION and sends the token to the corresponding process.

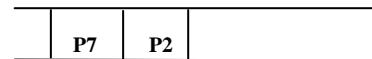


Figure 2: REQUEST QUEUE

One thing that we have to keep in mind, while a process is executing in its critical section, during that, with a certain time interval the process continually sends an EXISTS signal to the coordinator, to indicate that the token is alive. In reply of each EXISTS signal, the coordinator sends an OK signal to that process, indicating, the coordinator is alive also.

In the previous case, when process P5 is executing in its critical section and the processes P7 and P2 (or any other processes), send request to the coordinator and wait to acquire the token, then it could happen that the coordinator might crash. After this incident, when process P5 sends the EXISTS signal but does not receive the OK signal within a fixed timeout period, then process P5 assumes that the coordinator has crashed, hence it becomes the new coordinator, and process P5 announces this by sending a message [COORDINATOR (PID)] to the rest of the alive processes in the ring. Then all the other processes update their ring configuration table to the current state of the ring.

Table 2: Ring Configuration Table

| <i>Process ID</i> | <i>State</i> | <i>Status</i> |
|-------------------|--------------|--------------------|
| P0 | Dead | <i>Unknown</i> |
| P1 | Alive | Normal |
| P2 | Alive | Normal |
| P3 | Alive | Normal |
| P4 | Alive | Normal |
| P5 | Alive | Coordinator |
| P6 | Alive | Normal |
| P7 | Alive | Normal |
| P8 | Alive | Normal |
| P9 | Alive | Normal |

Consequently, each process including P7 and P2 get to know that the process P5 became the new coordinator. As processes P7 and P2 are still not being able to enter their critical section, they both send requests to the new coordinator P5, to acquire the token. Now, there is a very significant point to notice that, previously P7 was the first one to sends its request and then P2 sends its request. But, when, again they send their requests to the coordinator, it may happen, that the request from process P2 reaches to the coordinator prior process P7's request. In this situation one may think that, as the request form process P2 reaches first to the coordinator (P5), so, the coordinator may send the token to process P2 and not process P7. But, it is not the case. As previously mentioned, when a coordinator selects a process from its REQUEST QUEUE, (when multiple requests are in the queue) it always makes the selection based on the smallest 'time stamp of request generation'. So accordingly, process P7 will get the chance to acquire the token.

After a while, it may be the case, process P0 has restarted. Then it will send a message NEW to every other process in the ring. Hence, every other process will update their corresponding ring configuration table. In this situation the present coordinator gets to know that a new entry has been done. As a result the coordinator will send a message [COORDINATOR (PID)] and also send the current ring configuration table to the revived process P0; so that revived process (P0) gets to know who the current coordinator is and also maintains the ring configuration table.

Another situation may appear, that when no other processes are executing in its critical section, then somehow the coordinator has crashed. In this case, the process that will notice it first, it will be the new coordinator in the ring and announces it by sending the [COORDINATOR (PID)] message to every other process

and consequently every other process updates their ring configuration table.

7. Conclusions

In this paper, the proposed algorithm does not allow the circulation of the token along the ring, when there is no need (i.e. when no process wants to enter in its critical section). Loss of a token in the ring can easily be detected, and regeneration of token can be done easily in this algorithm. And process crash and recovery of crashed process can easily be managed using this algorithm. And there is no chance of creation of duplicate tokens in the ring.

Hence, the proposed algorithm overcomes all the drawbacks that may appear in the existing Token Ring Algorithm for handling Mutual Exclusion in Distributed System.

References

- [1] Andrew S. Tanenbaum, *Distributed Operating System*, Pearson Education, 2007.
- [2] Pradeep K. Sinha, *Distributed Operating Systems Concepts and Design*, Prentice-Hall of India private Limited, 2008.
- [3] H.Attiya and J.Welch, *Distributed Computing Fundamentals, Simulation and Advanced Topics*, Second Edition, A John Wiley & Sons, Inc., Publication, 2004.
- [4] Martin G. Velazquez, "A survey of Distributed Mutual Exclusion Algorithms", Department of Computer Science, Colorado State University.
- [5].Ajay D. Kshemkalyani and Mukesh Singhal, *Distributed Computing principles, Algorithms, and Systems*, Cambridge University Press,2008.

Sandipan Basu is final year student of M.Sc. Computer Science, St. Xavier's College, University of Calcutta. He completed B.Sc (Honours) degree in Computer Science from Asutosh College, University of Calcutta. His research interests include Distributed Systems, Networking, Operating System and Cryptography.

University Grades System Application using Dynamic Data Structure

Nael Hirzallah¹, Dead Al-Halabi² and Baydaa Al-Hamadani³

¹ Computer Engineering, Fahad Bin Sultan University
Tabuk, Saudi Arabia

² What is Next
Amman, Jordan

³ University of Huddersfield
Oldham, UK

Abstract

Most e-Learning platforms implemented in educational institutes provides a tool for instructors to enter the grades and for students to view them. This tool with the appropriate workflow is considered one of the most sensitive and important applications in any University Information Management System. Consequently, implementing this tool should consider the fact that it must be flexible and adaptable from time to time. This paper focuses on evaluating few different approaches to implement such a tool that belong to a so-called Static Approach. It also discusses the limitations of this approach. The paper then introduces a different, yet dynamic approach to implementing a Grades System Tool. An analytical study of the efficiency of the suggested system is also presented.

Keywords: Dynamic Grades Tool, Static Grades Tool, Data Structure, Grades System.

1. Introduction

Many educational institutes are moving towards implementing a full e-Learning platform that offers state-of-the-art tools for students, academic and administrative staff, as well as the university community at large. Among the tools that are considered important is the virtual driving force for students, which is the one used to manage the Grades. This tool is used to record the worthiness of students' efforts during a semester. Therefore, there is a demand to increase the trust in a university Grades Tool (GT) by most parties involved in the higher education process, such as, Teachers, Students, Managers, and Administrators, [1, 2]. A GT has to be able to adopt new development strategies and accompany the modernization in its background and planning, in order to achieve its objectives, [2, 3]. It has to be built using a strong, efficient, and customized system to enable efficiency in time and

efforts to all high education partners. Furthermore, saving time and money by an institution is one of its top priority requirements. Besides, recognizing new techniques and continuous development in the university sub systems indicates the growth of the institution reputation in local and global societies. This is considered by executives as an important marketing feature, [4, 5].

In this paper, two GT running in two different universities will be discussed. These tools are labeled as static due to their inability to adopt new Grading System policies to a certain extent. The paper then introduces a dynamic grades tool approach that depends on linked list structures in its implementation. The advantages of such approach over static GT will be discussed, as well.

The paper is organized as follows: section two summarizes the Grading System policies used in Jordanian universities. Section 3 presents a study of two existing GT's that are using the static data structure approach. In section 4, the paper introduces the Dynamic Grades Tool Approach (DGTA) followed by a discussion on the requirements, preparation, implementation, and performance of a DGTA. Finally, in section 6, the paper draws its conclusion.

2. Grading systems

2.1 University's Grading Policy

All Jordanian universities use either percentage or point-for-weight grading systems (i.e. letter grade system) [2, 5]. Table 1, [3], shows the basic transfer scheme from a

percentage grading system to a point grading system that exists in Jordan.

Table 1: Basic grade scheme used in Jordan.

| Scale | U.S. Grade Equiv. |
|--------|-------------------|
| 80-100 | A |
| 70-79 | B |
| 50-69 | C |
| 0-49 | F |

The Jordanian Grading Systems (GS) policy states the following issues, which should be considered by any GT [4]:

1. GS applies a numerical grade system in addition to the letter grade.
2. Every instructor is responsible for the following: entering student grades, evaluating students work, judging the course progress for her / his courses, and changing or modifying the final reported grades.
3. Every instructor is responsible for evaluating student's written documents or oral discussions.
4. Instructor (s) can make the grade more specialized, pursuant to his/her rights and authority given by the educational institute s/he works in.
5. All the oral evaluations have to be graded.
6. All sub grades have to be entered into a GT.
7. Course instructor (s) has the responsibility to enter valid data. Data have to be in the range of [minimum, maximum]. Also, s/he must make sure that the data are verified, accurate, and consistent.
8. Grades have to be accepted and verified at the department level and then at the faculty level.

2.2 The Value-Driven Meaning of Grades

GS is changeable and variant from time to time according to a given course specific policy. The transfer between a numeric GS to a letter GS or vice versa is possible and often needed.

There are two major paths for evaluating student's work in terms of percentages: either summative or formative. Both metrics should give indication of the imagination, creativity, and skills necessary for the rapidly changing requirements of modern social life.

Thus, any assessment criteria should guarantee the following:

1. Fairness
2. **Validity**
3. **Reliability**

The summative assessment is based on the overall summation of sub-activities that had occurred during a

semester for a particular course. It involves written paper, such as 1st, 2nd, and Mid exams, assignments, essays, tutorials, quizzes, self reading materials, and class projects. On the other hand, the formative assessment is a self-reflective process for a student. It is based on class discussions, questions, and seminars. The Final grade can be a mix between summative and formative; its assessment shall be at the end of the semester [8].

3. Existing Grads Tools

The observations discussed in this section are based on the experience gained by working on various systems in different institutions as a user with the role of an instructor. Static GT (SGT) is a client/ server application. Client sends and receives the required class information that belongs to an instructor. The user screen will be filled with the required information by opening a channel with the server.

There are three types of universities in Jordan: public, private, and distance higher education, [5]. The discussion will focus on two of these universities, labeled in this paper as "A" and "B". University "A" is a public university and "B" is a private one. University "A" uses letter grades, while "B" uses percentage grades. University "A" has two policies for obtaining the final grade. The first one states that the final grade is divided into: 1st Exam, 2nd Exam, course-work, and a Final exam. While in the second policy, the final grade is divided into: a Mid exam, course-work, and a Final exam. On the other hand, University "B" has only one policy to obtain the final grade: 1st exam, 2nd exam, course-work, and a Final exam.

The detailed weight distribution for each sub grade was left flexible and usually set by either the department or the instructor. Different course weight division may exist. One example is (15, 25, 10 and 50), while another one has (20, 20, 10 and 50).

In University "A", which uses letter grades system, the course weight distribution may be changed from one semester to another.

Figure 1 shows a snapshot of a weight distribute of one course offered by University "A" in one semester. The screen is divided into two blocks; the above one acts as a list of templates. One may select a course template then fill the sub weights values in the second block below it.

The screenshot shows a software application window titled "النحوين" (Arabic) and "المواد" (Subjects). At the top, there are dropdown menus for "رقم المادة" (Subject Number), "النحوين" (Section), and "رقم المدرب" (Instructor Number). Below these are buttons for "حذف" (Delete), "تعديل" (Edit), and "خروج" (Exit). The main area contains a table with columns: "صفة المادة" (Subject Type), "نصف الفصل" (Mid-term), "وتحصيل العلامة" (Grade Achieved), "النحوين" (Section), "وصف المادة" (Subject Description), and "رقم المادة" (Subject Number). The table lists five rows of data. At the bottom, there is a summary table with columns: "النحوين" (Section), "وصف المادة" (Subject Description), and "النحوين" (Section).

Figure 1: distribute grades weight in “A” under SGT.

Figure 2 shows another snapshot of an SGT screen from university “A” that allows instructors to enter grades for a specific course section. The screen illustrates the relationship between the course section and the student list that belongs to that section. The screen includes the following information: the teacher name, academic year, course number, section number, lecture room, semester number, and lecture time. It also includes student number, student name, 1st, 2nd, and 3rd grades, course work, and the Final exam. The last two columns are for the total grade, one in percent and the other in letters.

The screenshot shows a software application window titled "رسيد علامات المواد" (Grade Entry for Materials). At the top, there are dropdown menus for "اسم صاحب هيئة التدريس" (Instructor Name), "الفصل الأول" (First Semester), "النحوين" (Section), and "رقم المادة" (Subject Number). Below these are buttons for "طباعة" (Print), "مسح الكلمة" (Clear), and "خروج" (Exit). The main area contains a table with columns: "رقم الطالب" (Student Number), "النحوين" (Section), "نحوين" (Section), "نحوين" (Section), "نحوين" (Section), "نحوين" (Section), and "نحوين" (Section). The table lists 10 rows of data.

Figure 2: university “A” SGT entering grade screen.

University “B” uses percentages for evaluating the work of students in a semester. Sub grade weights are to be figured out from the data rather than from the system. There was no checking for exceeding the sub grade limit weight, if any, except at the end, when computing for the total grade out of 100. For example, if all entered grades of one exam are between [0, 20] then one may conclude that the maximum grade for that exam is 20. Figure 3 shows a snapshot of a screen of the SGT that allows instructors to enter grades for a course section in a semester.

The screenshot shows a software application window titled "Course No. 1202240" and "Course Arabic Name إلكترونيات". At the top, there are dropdown menus for "Course En. Name Electronics" and "Section No. 1". Below these are buttons for "Calculate All" and "Exit". The main area contains a table with columns: "Student No.", "Student Name", "1st Exam", "2nd Exam", "Assigns.&Projects", "Final Exam", and "Sum". The table lists 12 rows of data, each representing a student's performance across different evaluation metrics.

Figure 3: snapshot of screen of SGT of university “B”

The screen in Figure 3 is like that of Figure 2. It includes the following information: semester number, course number, section number, and the credit hours for the given course. It also contains: students' numbers, students' names, first and second grades, course work, final exam and total grade. The second last column describes the students situation in the course in terms of Withdraw, Absent from Final Exam, or Denied.

Generally speaking, static data structure implementation is easy to deal with and fast to implement. Its data access is a straight forward process; only a direct location is needed to obtain the data, such as the index value. There is no time wasted and an indexing schema can be used to organize its access time.

The main disadvantage is the waste of unused memory. Take for example the following scenario: if a course has its evaluation metric (Exam1, Exam2, Course-Work, and a Final exam) for 80 students, this means that there is a need for four columns multiplied by 80 records, which equals to 320 memory fields. Assume another course that is distributed as: mid-exam, Course-work, and a Final exam, for 80 students. This would need 3 columns multiplied by 80 records which equals to 240 fields. Therefore, if the system is set to have statically 320 fields, this would result in 80 wasted fields. In other words, the system creates k-columns even if the number of needed ones is less than k, in order to accommodate the worst case scenario. This is of course for one course. Now, assume that you have N-courses, then there will be 80 X N wasted fields.

Moreover, a waste in memory would result in delays in the access times when retrieving data under heavy load conditions. This would affect application ranking as it considers strongly page-load speeds.

4. Suggested proposal for DGT

4.1 Dynamic features

The theory behind dynamic allocation is based on the following statement: only what is necessary to build will be built. Thus, there is no need for unusable storage to be created, and no memory to waste.

Access and retrieving times are the most important features to be considered. The difficulty in implementing a system based on dynamic approach comes from analyzing and considering the risks that may occur due to scenarios that are rare to occur.

Dynamic approach takes in its consideration time and storage factors. Storage retrieving mechanism should exist, in addition to focusing on time scheduling.

4.2 DGT Procedures

The main two features concerned in dynamic approach that depend on each other are space and time. Figure 4 illustrate this relationship. Assume the x-axis represents time and the y-axis represents memory size allocated. The figure shows ascending relationship between them. That is, by time the memory space needed increases. Yet there is no fixed rhythm for dynamic approach as it is in the case of the static approach. Note that the memory size needed by the end of a semester for the same class in both approaches are the same. However, in the static approach the memory gets allocated at early times, while in the dynamic approach, it gets allocated by time. This will have a good effect on the complexity and access time.

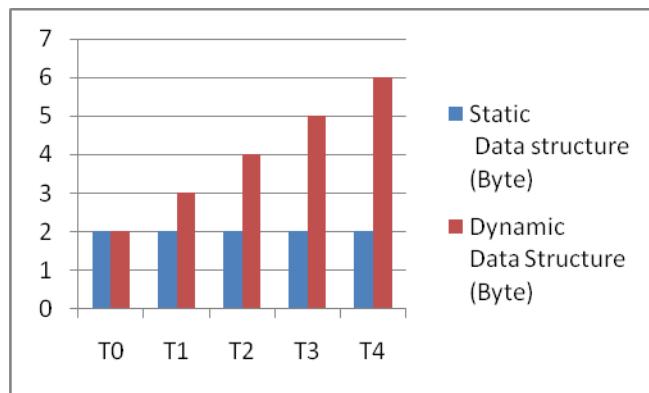


Figure 4: data size for static and dynamic structures

The main steps to populate a DGT with data are as follows:

- Preparation: Identify the essential data and build the corresponding data structure.

- Grading: For sub-grade 1, build the dynamic data structure associated with it and fill it with the proper data.
- Repeat step two for subsequent sub-grades, say second exam, first quiz, first assignment, and so on, till the Final exam.

For step one, prior to a semester, students register in a section for a course. The course coordinator usually sets the weight of each sub-grade. For example the first exam gets a weight of W_1 , the second exam gets W_2 , and so on up to W_k , where k is the number of sub-grades. One scenario could be as follows: ($W_1=10$, $W_2= 20$, $W_3=5$, $W_4=15$ and $W_5= 40$) where $k =5$.

Step two can be accomplished automatically at its previously assigned time, as stated in the course syllabus, or manually by the course coordinator.

Figure 5 shows how the data structure of such a system would look like. It has an array of pointers that has the size of N , where N represents the number of students registered in the class. Each pointer links the array with a structure that contains the student's number and a pointer to a link list for the student's grades. In what follows, the C++ notations will be used.

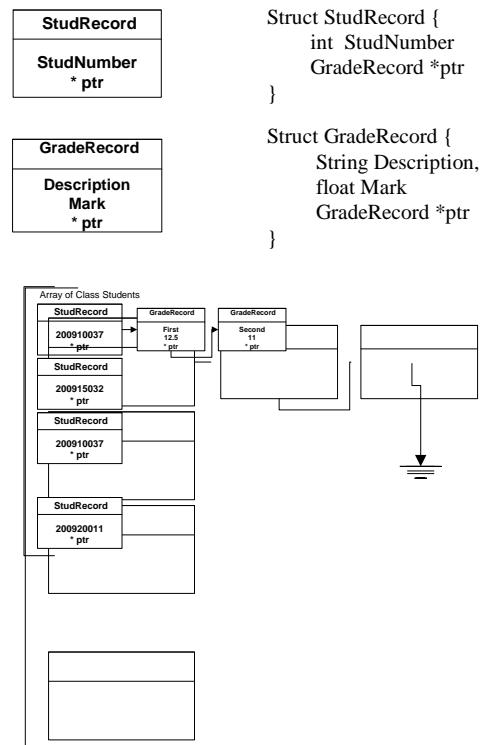


Figure 5: DGT data structure

Step two procedure is illustrated in Figure 6. The procedure gets the number of sub-grade to be entered and assigns it to i. For all the students in the class (1 to N) the sub-grade value is entered separately. Each value is checked against its maximum value, which is Wi. If it is validated, a new node of type GradeRecord will be created and initialized.

4.3 Performance Analysis

The rapid changes in the GS policy have to be reflected in the tool or application. A classical evaluation of the student work, that is, using three written exams namely: first, second and final, would be probably better implemented if using a GT that uses static data structures, SGT. This suitability comes from the ascending relationship between the mostly fixed measurements requirements and static data structures, [5]. The problems arise when there are many student works evaluation schemes rather than just one or two. For example, TMA (Tutor-Marked Assignment) is a vital example in student work evaluation environment that usually varies in number and weight from one section to another, from one course to another, and from one semester to another [7].

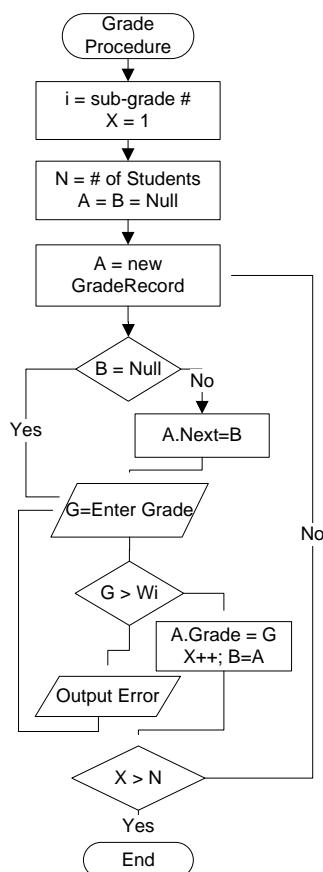


Figure 6: Flow chart of Grading procedure

Therefore, the GT is a changeable tool that has to provide a capability to understand the new non-functional and functional requirements in order to be able to support efficiency, reliability, portability, usability, performance and space, in addition to validation, accuracy, and consistency of data (grades - functional requirements that depend on the system domain). For that DGT is a more suitable solution under the requirements changing condition.

In this section we present an analytical module to study the worthiness of implementing the suggested solution methodology of adopting dynamic data structure in GT in terms of complexity (big-O and memory size).

Big-O analysis depends on the run time of the application. DGT is a client-server tool or application, and it is assumed that the computation for a DGT is done on the server side. It is also assumed that the network infrastructure is well built to eliminate communication negative factors, as well as has negligible page-load timings.

For the analysis, the following specific assumptions were considered:

- A course has four weights, $k=4$, that is W_1, W_2, W_3 and W_4 .
- All algorithms had been run and the computation of big-O is based on the fact that the procedures have reached the final exam. (i.e. complete course evaluation).

The $O(F(N)) = \text{Big-O}$ for preparation algorithm added to it the Big-O for the mid-term (W_1) grade, added to it the Big-O for course work (W_2) grade, and so on. The preparation procedure will pass on every cell in the array and fill it with the student number and a null pointer for the grade list. This process will take $O(N)$. While the Grading procedure will pass on every student in the list and add the corresponding grade as required, (such as adding item in a linked list). This process will take $O(i)$ where i is a constant to indicate the number of sub-grades for one student to evaluate. For N students, this will yield $O(N)$. Thus, $O(F(N))$ will result in the following:

$$O(F(N)) = O(N)$$

A brief comparison was made between SGT (existing system) and DGT (suggested solution) based on the processing time and memory size. In SGT with respect to memory, it remains the same throughout the application life time. Thus, the memory size is of the order of $O(c \times N)$, where c is a constant that represents the number of fields to be entered (5 in our earlier assumption).

While in the suggested solution, the amount of memory used depends on time. As an example, the number of fields to be created for the first exam at time t1 will be N. Later, another N fields will be created for the second exam at time t2.

Memory complexity in DGT is observed to be reduced. Also there is an obvious reduction in the processing time as well, due to less memory being used.

5. Conclusion

The paper has presented the importance of the right implementation to a Grades Tool. Couple of real implementation examples was discussed that belong to traditional static programming habit (array of records). Such approach was labeled Static Grades Tool (SGT), and its limitations were presented. The paper then presented the use of dynamic data structures (array of link-lists) in such applications, labeling them as Dynamic Grades Tool, (DGT). A reduction in storage and processing times were the driving factors. Moreover, an analysis on the run time using big-O method gave good indicators on the superiority of DGT over SGT.

References

- [1] <http://www.ju.edu.jo/units/registration/Home.aspx>, Last visit 20-4-2010
- [2] <http://www.uop.edu.jo/admission/Default.aspx?lang=en&location=admission>, last visit 20-4-2010
- [3] <http://www.wes.org/gradeconversionguide/>
- [4] <http://www.uop.edu.jo/Admission/Grading.aspx?lang=en&location=FS>, last visit 20-4-2010
- [5] Software & Systems Requirements Engineering: In Practice, Brian Berenbac, daniel j. paulish, juergen kazmeier, arnold rudorfer, Mnc raw hell, 2009.
- [6] Data Structures and Algorithms in Java, Michael T. Goodrich, Roberto Tamassia, John Wiley & Sons, 4TH edition, 2006
- [7] <http://www.open.ac.uk/assessment/pages/tma-submission-methods.php>, last visited 30-5-2010
- [8] <http://www.nmsa.org/Publications/WebExclusive/Assessment/tabid/1120/Default.aspx>, last visit 30-5-2010
- [9] Distributed systems: Principles and Paradigms, Andrew S. Tanenbaum and Maarten van steen, prentice Hall, 2nd edition, 2002
- [10] Principles of Distributed database systems, M. Tamer Ozsu and Patrick Valduriez, 2nd edition, 1999

A Paper Presentation on Software Development Automation by Computer Aided Software Engineering (CASE)

Nishant Dubey

School of Computer and Electronics, IPS Academy
Indore, MP, PIN 452012, India

Abstract

Now a day, system developers are faced to produce complex, high quality software to support the demand for new and revised computer applications. This challenge is complicated by strict resource constraints, forcing management to deploy new technologies, methods and procedures to manage this increasingly complex environment. Often the methods, procedures and technologies are not integrated. Therefore, they achieve less than desired improvements in productivity, or force management to make tradeoff decisions between software quality and developer efficiency. Thus the production lines have to be developed faster, too. A very important role in this development is *Software Engineering* because many production processes are 'computer aided', so software has to be designed for this production system. It seems very important to do the software engineering right and fast.

Keywords: CASE, Software Engineering, Tools, Process,

1. Introduction

Software Engineering is still a relatively new area of engineering. Indeed the phrase itself gained widespread use after a 1968 NATO-sponsored conference. As the name suggests, Software Engineering deals with the exposing of the process of designing, creating and maintaining software. It is perhaps useful to bear in mind that a structured approach to a solution is in no way a barrier to creativity. Indeed, much modern architecture, although bound by rigid guidelines and specifications can be truly breathtaking.

Today everything has to go faster. Because of the increasing speed of changing market-demands new products replace old ones much earlier than before, so the development of new products has to go faster. Thus the production lines have to be developed faster, too. In the past, software systems were build using traditional development techniques, which relied on hand-coding applications.

Computer-Aided Software Engineering (CASE) helps system developers meet their challenge by providing a new generation of

integrated system development tools which provides an automated environment in which to design and implement system projects. CASE technology enables system developers to improve both quality and efficiency, resulting in a net improvement in maintenance and development productivity.

The objectives of this paper are to provide the reader with sufficient information about CASE technology to develop an evaluation and implementation strategy for utilizing CASE to improve systems development productivity.

2. Computer Aided Software Engineering

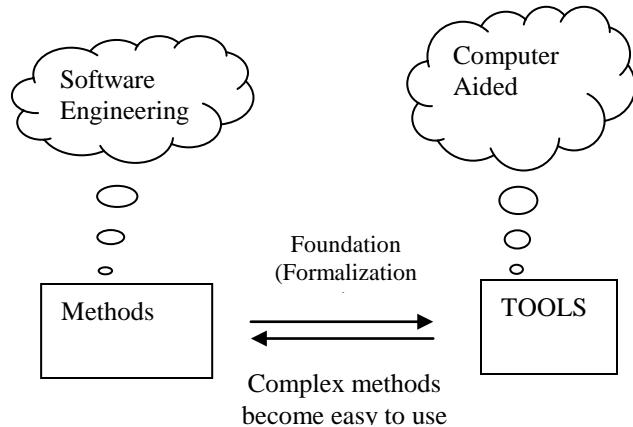


Fig 1 CASE

This term is used for a new generation of tools that applies engineering principles to the development and analysis of software specifications. Simply, computers develop software for other computers in a fast way by using specific tools. When implanted in a concurrent engineering environment, this process is taking place while some parts of the developed software are running already. It's a sort of on-line software engineering. There are a lot of problems with these kinds of systems because they are very complex, not easily maintainable and fragile.

The tools developed right now are evolutional products out of earlier tools. The first tools, developed in the mid '70s, were

mainly introduced to automate the production and to maintenance structured diagrams. When this automation covers the *complete life-cycle process*, it is called Integrated Computer Aided Software Engineering (I-CASE). When only one specific part of the life-cycle process is covered we call just- Computer Aided Software Engineering (CASE).

Recently, CASE tools have entered a third phase: the introduction of new methodologies based on capabilities of I-CASE tools. These new methodologies utilize Rapid Prototyping techniques to develop applications faster, at lower cost and higher quality. By using Rapid Prototyping a prototype can be made fast, so the developed system can be tested more often in between the development-phases because it doesn't cost much time to create a prototype. Mistakes can be detected and corrected earlier this way. The earlier this can be done, the better because correcting these mistakes gets harder and more expensive when the system is developed further. So a lot of time and money can be saved using Rapid Prototyping.

As said above, a new set of tools is necessary. These tools should automate each phase of the life-cycle process and tie the application development more closely to the strategic operations of the business. A lot of different tools have been developed over the years and are being developed right now. There are so many tools, that we could easily get confused. To survey all these CASE tools we divide them in the following categories:

- i. *Business Process Engineering Tools* – The primary objectives of these tools is to represent business data objects, their relationships and how these data objects flow between different business areas within a company.
- ii. *Requirement Tracing Tools* – The objective of these tools is to provide a systematic approach to the isolation of requirements that begins with the proposal or specification of customer's request. The typical requirement of these tools includes human interactive text evaluation with the database management system which stores and categorize every system need that is parse from the original specification.
- iii. *Software Configuration Management Tools* – it is lies in the kernel of every CASE environment. These tools help in identification, version control, change control, auditing and status accounting.
- iv. *Process Modeling and Management Tools* – Process Modeling Tools represent the key elements of a process such that it can be better understood. Process management tools provide the links to other tools for providing support defining process activities.
- v. *PRO/SIM Tools* – Prototyping and Simulation tools provide the software engineer with the ability for predicting the behavior of a real time system prior to the time that it is made, and these tools enable for developing mock ups of real time systems.
- vi. *Project Planning Tools* – These tools focus on Software Project Effort and Cost Estimation and Project Scheduling.
- vii. *Matrix and Management Tools* – Software metrics improve the ability of manager to control and coordinate the process of software engineering. Management tools capture project specific metrics which provide the overall indication of productivity or quality.
- viii. *Analysis and Design Tools* – These tools enable to create the models of system to be made. This model has a representation of data, function and behavior and characterization of data architectural, component-level and design interface.
- ix. *Project Management Tool* – On a continuing basis the project plan and schedule must be tracked and monitored. These tools are extension for project planning tools.
- x. *Risk Analysis Tools* – These tools enable a manager to make a risk table by providing detailed guidance in the risk identification and risk analysis.
- xi. *Documentation Tools* – These tools provide an important opportunity to improve productivity. Document production tools support every aspect of software engineering and represent a substantial leverage opportunity for all developers.
- xii. *Prototyping Tools* – Prototyping tools enables the creating of data design, coupled with both screen and report layouts.
- xiii. *Quality Assurance Tools* – CASE tools majority claims on majority claim on quality assurance to focus are the metrics tools for auditing source code to determine compliance with language standard. For building the quality of software other tools extract technical metrics in an effort to the project.
- xiv. *Programming Tools* – These tools encompass the compilers, editors and debuggers to support most conventional programming languages. In this category OOP environment, application generators and database query language resides.
- xv. *Interface Design & Development Tools* – These tools are software component tool kit which contained menus, buttons, window, icon, device drivers etc. These tool kits replaced by prototyping tools.
- xvi. *System Software Tools* – CASE is workstation technology and high quality network system software, distributed component support, bulletin board, email, object management services and other capabilities of communication.
- xvii. *Database Management Tools* – These tools are include on the emphasis of configuration object from RDBMS to Object Oriented DBMS for CASE.
- xviii. *Integration and Testing Tools* – in this the following testing tools are categorized
 - a. Data Acquisition – These acquire the data that is to be used during testing.
 - b. Statement Management – These analyze the source code without executing the test cases.
 - c. Dynamic Measurement – These analyze the source code during execution.
 - d. Simulation – These simulate the function of hardware or other externals.
 - e. Text Management – These help in the planning development and the control of testing.
 - f. Cross Functional Tools – These cross the bounds of the preceding categories.

- xix. *Web Development Tool* – These tools help in the generation of text, graphics forms, scripts, applets and other elements of a webpage.
- xx. *Client Server Testing Tools* – The client/server environment demands the testing tools that exercise the graphical user interface and the network communication requirements for client and server.
- xi. *Test management Tools* – These tools are used to control and coordinate software testing. These manage and coordinate regression testing, platform comparison and conduct the batch testing of programs with human/computer interfaces interactively.
- xxii. *Reengineering Tools* – In these tools for legacy software address a set of maintenance activities that absorb significance percentage of all software related effort.
- xxiii. *Static Analysis Tool* – Static testing tools help in deriving test cases. In this code based testing tools takes source code and perform number of analyses that results in test case generation. Specialized testing language enable to write detailed test specifications that describe each test case and logistics for its execution. Requirement based testing tools isolate specific user requirements and suggests test cases for exercising the requirement.
- xxiv. *Dynamic Analysis Tools* – Dynamic tools can be either intrusive or non intrusive. Intrusive tools change the software that is to be tested by inserting extra instruction that perform the activities. Non intrusive testing tools use a separate hardware processor to contain the program being tested.

3. CASE Building Blocks

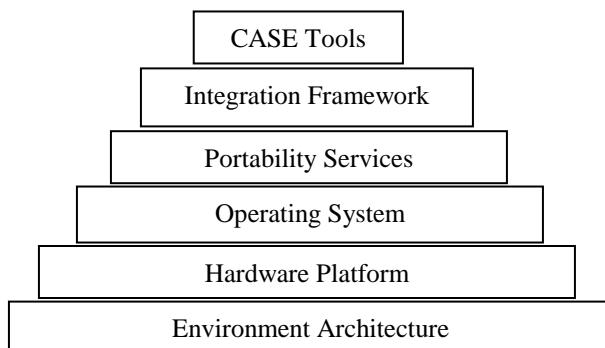


FIG 2 CASE BUILDING BLOCKS

Each building block forms a foundation for the next, with tools sitting at the top of the heap. The environment architecture composed of hardware platform and operating system lays the ground work for the CASE. The operating system includes the networking software, database management and object management services. Portability services provide a bridge between CASE tools and their integration framework and the environment architecture. Integration framework is a collection of specialized programs that enables individual CASE tools to

communicate with one another, to create a project database and to exhibit the same look and feel to the software engineer (end user).

4. Conclusion

CASE technology is now available for most routine activities in the software process. This has led to some improvements in software quality and productivity although these have been less than predicted by early advocates of CASE.

Various companies offer CASE software capable of supporting some or all of these activities. While many CASE systems provide special support for object-oriented programming, the term CASE can apply to any type of software development environment.

5. References

1. Sommerville, Ian (2004). *Software Engineering*, Second. Pearson Education. ISBN 0321210263.
2. Pressman, Roger S. (2005). *Software Engineering: A Practitioner's Approach*, Sixth. McGraw-Hill. ISBN 0072853182.
3. Jones C.B., Software Development – A Rigorous Approach, London, Prentice Hall
4. Fairly R,(2000) Software Engineering Concept, Tata Mc Graw Hill Pub.
5. Pfeleger S.L.,(1988) Software Engineering, Prentice Hall International Inc.
6. Caine S & E Gordon, (1975), A tool for Software Design, Vol 47,AFIPS Press, Montvale, NJ
7. Halstead M. H.,(1977) Elements of Software Science, New York, Elsevier North Holland
8. Bennett K ,(1991), Software Maintenance, Butterworth – Heinemann Ltd, Ox Ford
9. Sage A and Palmer J.D.,(1990), Software Engineering, John Wiley & Sons
10. Capers Jones(2009), Software Engineering Best Practices, Edition-1, ISBN: 9780070683594
11. David Gustafson,(2003), Schaum's Outline of Software Engineering, Edition: 1, ISBN: 9780071377942
12. Carma L. McClure,(1988), Prentice Hall Publication, ISBN : 9780131193307

Simulation and Performance Analysis of Adaptive Filtering Algorithms in Noise Cancellation

Lilatul Ferdouse¹, Nasrin Akhter², Tamanna Haque Nipa³ and Fariha Tasmin Jaigirdar⁴

**1 Department of Computer Science, Stamford University Bangladesh
Dhaka, Bangladesh**

**2 Department of Computer Science, Stamford University Bangladesh
Dhaka, Bangladesh**

**3 Department of Computer Science, Stamford University Bangladesh
Dhaka, Bangladesh**

**4 Department of Computer Science, Stamford University Bangladesh
Dhaka, Bangladesh**

Abstract

Noise problems in signals have gained huge attention due to the need of noise-free output signal in numerous communication systems. The principle of adaptive noise cancellation is to acquire an estimation of the unwanted interfering signal and subtract it from the corrupted signal. Noise cancellation operation is controlled adaptively with the target of achieving improved signal to noise ratio. This paper concentrates upon the analysis of adaptive noise canceller using Recursive Least Square (RLS), Fast Transversal Recursive Least Square (FTRLS) and Gradient Adaptive Lattice (GAL) algorithms. The performance analysis of the algorithms is done based on convergence behavior, convergence time, correlation coefficients and signal to noise ratio. After comparing all the simulated results we observed that GAL performs the best in noise cancellation in terms of Correlation Coefficient, SNR and Convergence Time. RLS, FTRLS and GAL were never evaluated and compared before on their performance in noise cancellation in terms of the criteria we considered here.

Keywords: Adaptive Filter, Noise, Mean Square Error, RLS, FTRLS, GAL, Convergence

1. Introduction

A Digital communication system consists of a transmitter, channel and receiver connected together. Typically the channel suffers from two major kinds of impairments: Intersymbol interference and Noise. The principle of noise cancellation is to obtain an estimate of the interfering signal and subtract it from the corrupted signal. Adaptive noise cancellation [1]-[2], a specific type of interference cancellation, relies on the use of noise cancellation by subtracting noise from a received signal, an operation controlled in an adaptive manner for the purpose of improved signal to noise ratio. It is basically a dual-input, closed loop adaptive control system as illustrated in fig 1.

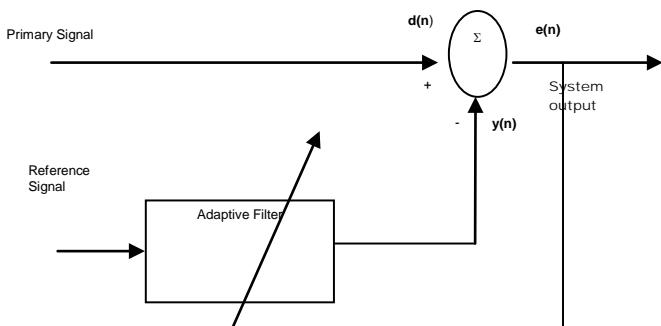


Fig. 1 Noise Cancellations.

Here the adaptive filter [2] is used to cancel unknown interference contained in a primary signal, with the cancellation being optimized in some sense. The primary signal serves as the desired response for the adaptive filter. The reference signal is employed as the input to

the filter. This paper studies and analyzes the performances of three adaptive algorithms in noise cancellation. We present simulations based on different types of signals mixed with various types of noise. Despite the theoretical nature of the study, efforts have been made to emphasize signals of practical use. Therefore, audio and electrical signals that are subject to noise in real world have been considered. Audio files were read and microphones connected real audio signals. The analysis of the results offered useful insight into the characteristics of the algorithms

2. Literature Review

Different applications of the adaptive digital filters were studied as early as the 1950s. Now adaptive filters are ubiquitous tools for numerous real-world scientific and industrial applications. The initial works on adaptive echo cancellers started around 1965. It appears that Kelly of Bell Telephone Laboratories was the first to propose the use of an adaptive filter for echo cancellation, with the speech signal itself utilized in performing the adaptation [2]-[3]. Yuu-Seng Lau, Zahir M.Hossain and Richard Harrisa explained Performance of Adaptive Filtering Algorithms: A comparative study. It showed a cooperative performance study between the time-varying LMS (TV-LMS) and other two main adaptive approaches: The Least Mean Square (LMS) algorithm and the Recursive Least Square (RLS) algorithm. Their study disclosed the algorithm execution time, the minimum Mean Square Error (MSE) and required filter order [4]. Hyun-Chool Shin, Ali H.Sayed and Woo-Jin Song described the Mean Square Performance of Adaptive filters using averaging theory. This paper uses averaging analysis to study the mean-square performance of adaptive filters, not only in terms of stability conditions but also in terms of expressions for the mean-square error and the mean-square deviation of the filter, as well as in terms of the transient performance of the corresponding partially averaged systems [5]. R.C. North J.R. Zeidler , T.R. Albert , W.H. Ku presented the Comparison of adaptive lattice filters to LMS transversal filters for sinusoidal cancellation. This paper compare the performance of the recursive least squares lattice (RLSL) and the normalized step-size stochastic gradient lattice (SGL) algorithms to that of the least mean square (LMS) transversal algorithm for the cancellation of sinusoidal interference. It is found that adaptive lattice filters possess a number of advantages over the LMS transversal filter, making them the preferred adaptive noise cancellation (ANC) filter structure if their increased computational costs can be tolerated [6]. Syed .A.Haider and M.Lotfizad developed a new approach for canceling attenuating noise in speech signals. This paper

presented a nice tradeoff between convergence properties and computational complexity and showed that the convergence property of fast affine projection (FAP) adaptive filtering algorithm is superior to that of usual LMS, NLMS, and RLS algorithm [7].

For the learning of FIR filters using linear adaptive filtering algorithms ,it is well known that recursive-least-squares(RLS) algorithms produce a faster convergence speed than stochastic gradient descent techniques, such as the basic least-mean-squares(LMS) algorithms, or even gradient-adaptive-lattice LMS(GAL)[2]-[3].In this paper we present an implementation of adaptive noise canceller using Recursive Least Square (RLS), Fast Transversal Recursive Least Square (FTRLS) and Gradient Adaptive Lattice (GAL) algorithms with the intention to compare their performance in noise cancellation in terms of convergence behavior, convergence time, and correlation coefficients and signal to noise ratio.

3. Adaptive Algorithms

Recursive Least Squares (RLS) algorithm is capable of realizing a rate of convergence that is much faster than the LMS algorithm, because the RLS algorithm utilizes all the information contained in the input data from the start of the adaptation up to the present.

3.1 The Standard RLS Algorithms

In the method of least squares, at any time instant $n > 0$ the adaptive filter parameter (tap weights) are calculate so that the quantity of the cost function

$$\zeta(n) = \sum_{k=1}^n \rho_n(k) e_n^2(k) \quad (1)$$

is minimized and hence the name least squares. In $k=1$ is the time at which the algorithm starts, $e_n(k), k = 1, 2, \dots, n$, are the samples of error estimates that would be obtained if the filter were run from time $k = 1$ to n , using the set of filter parameters that is computed at time n , and $\rho_n(k)$ is a weighting function. Actually the RLS algorithm performs the following operations:

- Filters the input signal $x(n)$ through the adaptive filter $w(n-1)$ to produce the filter output $y(n)$
- Calculates the error sample $e(n) = d(n) - y(n)$

- Recursively updates the gain vector $k(n)$
- Updates the adaptive filter coefficients

3.1.1 The algorithm can be summarized through following steps:

Input Parameters

Tap-weight vector estimate, $\hat{w}(n-1)$

Input vector, $x(n)$

Desired output, $d(n)$

And the matrix, $\psi_\lambda^{-1}(n-1)$

Output:

Filter output, $y_{n-1}(n-1)$

Tap weight vector update, $\hat{w}(n)$

And the updated matrix, $\psi_\lambda^{-1}(n)$

Procedure:

1. Computation of the gain vector:

$$u(n) = \psi_\lambda^{-1}(n-1)x(n)$$

$$k(n) = \frac{1}{\lambda + x^T(n)u(n)}u(n)$$

2. Filtering:

$$\hat{y}_{n-1}(n) = \hat{w}^T(n-1)X(n)$$

3. Error-estimation:

$$\hat{e}_{n-1}(n) = d(n) - \hat{y}_{n-1}(n)$$

4. Tap-weight vector adaptation:

$$\hat{w}(n) = \hat{w}(n-1) + k(n)\hat{e}_{n-1}(n)$$

5. $\psi_\lambda^{-1}(n)$ Update:

$$\psi_\lambda^{-1}(n) = \lambda^{-1}(\psi_\lambda^{-1}(n-1) - k(n)[x^T(n)\psi_\lambda^{-1}(n-1)])$$

3.2 Fast Transversal RLS Algorithm

Fast transversal filter (FTF) algorithm involves the combined use of four transversal filters for forward and backward predictions, gain vector computation and joint process estimation. The main advantage of FTF algorithm is reduced computational complexity as compared to other available solutions. The derivation of the algorithm follows hereafter.

3.2.1 Summary of the FTRLs Algorithm

The FTRLs algorithm is summarized below by collecting together the relevant equations.

Input parameters:

Tap-input vector $x_{N-1}(n-1)$, desired output $d(n)$

Tap-weight vectors $\bar{a}_N(n-1), \bar{g}_N(n-1)$ and $\hat{w}_N(n-1)$

Normalized gain vector, $\bar{k}_N(n-1)$

Least squares sums or auto correlations $\zeta_N^{ff}(n-1), \zeta_N^{bb}(n-1)$

Output:

The updated values of

$\bar{a}_N(n), \bar{g}_N(n), \hat{w}_N(n), \bar{k}_N(n), \zeta_N^{ff}(n), \zeta_N^{bb}(n)$

Prediction:

$$f_{N,n-1}(n) = \hat{a}_N^T(n)x_{N+1}(n)$$

$$f_{N,n}(n) = \gamma_N(n-1)f_{N,n-1}(n)$$

$$f_{N,n}(n) = \gamma_N(n-1)f_{N,n-1}(n)$$

$$\zeta_N^{ff}(n) = \lambda\zeta_N^{ff}(n-1)f_{N,n-1}(n)$$

$$\begin{aligned} \gamma_{N+1}(n) &= \lambda \frac{\zeta_N^{ff}(n-1)}{\zeta_N^{ff}(n)} \gamma_N(n-1) \\ \bar{K}_{N+1}(n) &= \left[\begin{array}{c} 0 \\ \bar{K}_N(n-1) \end{array} \right] + \lambda^{-1} \frac{f_{N,n-1}(n)}{\zeta_N^{ff}(n)} \tilde{a}_N(n-1) \end{aligned}$$

$$\tilde{a}_N(n) = \tilde{a}_N(n-1) - \left[\begin{array}{c} 0 \\ \bar{K}_N(n-1) \end{array} \right] f_{N,n}(n)$$

$$b_{N,n-1}(n) = \lambda\zeta_N^{bb}(n-1)\bar{k}_{N+1,n-1}(n)$$

$$\beta(n) = 1 - b_{N,n-1}(n)\gamma_{N+1}(n)\bar{K}_{N+1,n-1}(n)$$

$$\gamma_N(n) = \beta^{-1}(n)\gamma_{N-1}(n)$$

$$b_{N,n}(n) = \gamma_N(n)b_{N,n-1}(n)$$

$$\zeta_N^{bb}(n) = \lambda\zeta_N^{bb}(n-1) + b_{N,n}(n)b_{N,n-1}(n)$$

$$\begin{bmatrix} K_N(n) \\ 0 \end{bmatrix} = \bar{K}_{N+1}(n) - \bar{K}_{N+1,n-1}(n)\bar{g}_N(n-1)$$

$$\bar{g}_N(n) = \bar{g}_N(n-1) - \begin{bmatrix} \bar{k}_N(n) \\ 0 \end{bmatrix} b_{N,n}(n)$$

Filtering:

$$e_{N,n-1}(n) = d(n) - \hat{w}_N^T(n-1)x_N(n)$$

$$e_{N,n}(n) = \gamma_N(n)e_{N,n-1}(n)$$

$$\hat{w}_N(n) = \hat{w}_N(n-1) + \bar{k}_N(n)e_{N,n}(n)$$

3.3 Gradient Adaptive Lattice

The *gradient-adaptive lattice(GAL)* filter is due to Griffiths(1977,1978) and may be viewed as a natural extension of the normalized least-mean-square(LMS) filter in that both types of filter rely on a stochastic gradient approach for their algorithmic implementations.

3.3.1 Summary of the GAL Algorithm

Parameters: M =final prediction order

β =constant, lying in the range (0.1)

$\hat{\mu} < 0.1$

δ : small positive constant

a : another small positive constant

Multistage lattice predictor:

$$f_0(m) = b_0(n) = u(n)$$

$$e_{m-1}(n) = \beta e_{m-1}(n-1) + (1-\beta)(|f_{m-1}(n)|^2 + |b_{m-1}(n-1)|^2)$$

$$f_m(n) = f_{m-1}(n) + k_m b_{m-1}(n-1)$$

$$b_m(n) = b_{m-1}(n-1) + k_m f_{m-1}(n)$$

$$\hat{k}_m(n) = \hat{k}_m(n-1) - \frac{\hat{\mu}}{\varepsilon_{m-1}(n)} (f_{m-1}^*(n)b_m(n) + b_{m-1}(n-1)f_m^*(n))$$

Filtering:

$$y_m(n) = y_{m-1}(n) + \hat{h}_m^*(n)b_m(n)$$

$$e_m(n) = d(n) - y_m(n)$$

$$\|b_m(n)\|^2 = \|b_{m-1}(n)\|^2 + |b_m(n)|^2$$

$$\hat{h}_m(n+1) = \hat{h}_m(n) + \frac{\tilde{\mu}}{\|b_m(n)\|^2} b_m(n)e_m^*(n)$$

4. Simulation Results

Simulation based on four different types of signals mixed with various types of noise. Signals are periodic signal, audio signal, chirp signal and saw-tooth signal. Each signal has been subjected to some noise. Then the convergence behaviors of the RLS, FTF and GAL algorithms for these signals have been analyzed. Audio files were read and microphones connected real audio signals. The signals were then polluted by white, pink, grey and burst noise. We also apply AWGN channel model and take low, moderate and high signal-to-noise ratio. The signals were then passed through the simulation of the adaptive filter, and their error recovery rate, correlation coefficient and time were calculated. The analysis of the results offered useful insight into the characteristics of the algorithms. For the RLS algorithm, two parameters were varied to find their effect on the performance. One of them is the filter length, and the other is the forgetting factor. In Fast RLS algorithms, FTF and GAL, the performances were analyzed by varying different filter length, forgetting factor and step size parameter.

4.1 Comparison based on Noise Cancellation Performance

In order to compare noise cancellation capability, three methods of presentation have been shown. One of them is the plotting of Mean Square Error with number of samples. Error convergence characteristics of the three algorithms have been shown on the same graph to attain visual comparison.

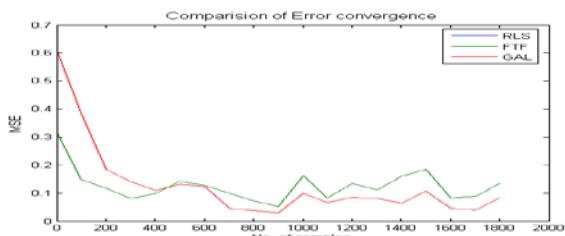


Fig. 2 Comparison of noise cancellation for RLS, FTF, GAL algorithms for sinusoidal signal.

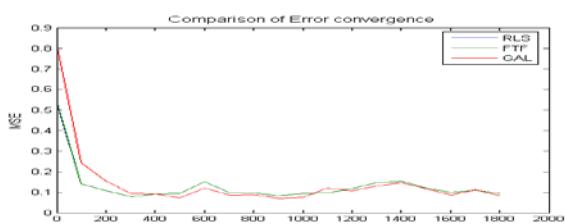


Fig. 3 Comparison of noise cancellation for RLS, FTF, GAL algorithms for saw-tooth signal.

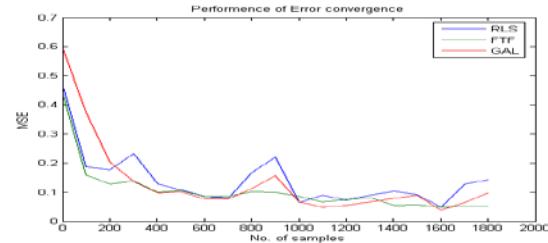


Fig. 4 Comparison of noise cancellation for RLS, FTF, GAL algorithms for Chirp signal.

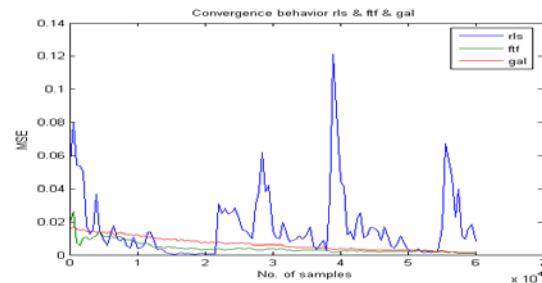


Fig. 5 Comparison of noise cancellation for RLS, FTF, GAL algorithms for audio signal.

In the case of periodic signal (mixed with white noise), RLS and FTF algorithms perform better than GAL and they almost show same convergence behavior. But, GAL's performance is not satisfactory in this case. The same thing is repeated also for the sawtooth signal when it is corrupted by white noise and the chirp signal which is distorted by pink noise. Concluding by audio noise, which is corrupted by white noise it can be told that FTF performs the best, then comes GAL and after that RLS comes.

4.2 Comparison of Correlation Coefficient

A tabular method is used to compare the correlation coefficients of the algorithms. In this comparison, the algorithms have been compared for the same signal to noise ratio combination

Table 1: Comparison of correlation coefficients of RLS, FTF, GAL for different signals mixed with white noise

| Correlation Coefficients | | | |
|--------------------------|--------|--------|--------|
| Signal Types | RLS | FTF | GAL |
| <i>Chirp</i> | 0.8401 | 0.8501 | 0.9218 |
| <i>Sinusoidal</i> | 0.9464 | 0.9459 | 0.9422 |
| <i>Saw tooth</i> | 0.8935 | 0.8909 | 0.9021 |
| <i>audio</i> | 0.9798 | 0.9988 | 0.9989 |

| Convergence Time (S) | | | |
|--------------------------|-------|-------|-------|
| Signal Type | RLS | FTF | GAL |
| <i>Chirp signal</i> | 0.453 | 1.797 | 0.672 |
| <i>Sinusoidal signal</i> | 0.641 | 2.438 | 0.844 |
| <i>Saw tooth signal</i> | 0.422 | 1.453 | 1.453 |

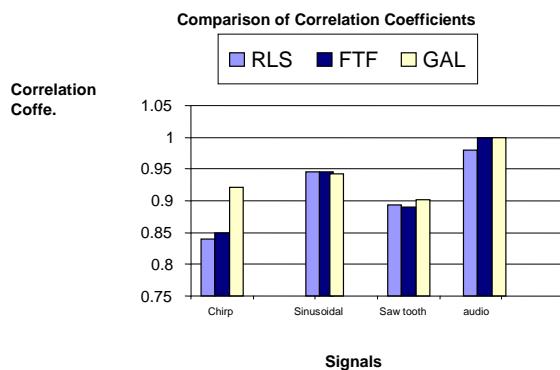


Fig. 6 Comparison of correlation coefficients of RLS. FTF, GAL for different signals mixed with white noise.

The above chart (Fig. 6) reveals the decision that Gradient adaptive lattice (GAL) has the best noise cancellation performance since the correlation coefficient for all signals except sinusoidal signal than other two algorithms. This means that GAL has achieved close approximation of the desired signal in all cases. On other side RLS and FTF show same performance in case of sinusoidal signal

4.3 Comparison of Convergence Time

Taking the same signal to noise ratio, the performance of the RLS, FTF and GAL algorithms are compared in terms of their convergence time, which is given in tabular and graphical form. Analyzing fig. 7, it is revealed that RLS takes the least convergence time than the other two. GAL takes the second position in this occurrence. Lastly comes the FTF algorithm that requires more convergence time than standard RLS and GAL algorithms

Table 2: Comparison of convergence time of RLS, FTF, GAL for different signals mixed with white noise

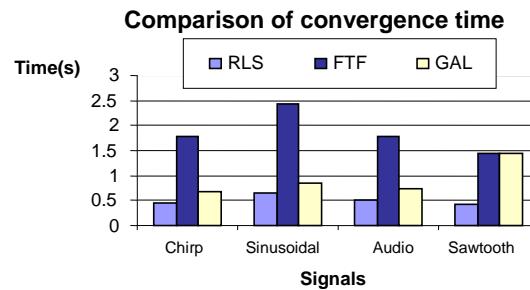


Fig. 7 Comparison of convergence time of RLS. FTF, GAL for different signals mixed with white noise.

4.4 Comparison of Signal-to-Noise Ratio.

The experiment is divided into three parts: In part 1, the input signals-to-noise ratio (SNR) is high; in part 2, it is mid and in part 3, it is low.

Part 1:

Table 3: Comparison of signal to noise ratio of RLS, FTF, GAL algorithms when given SNR=30dB (high)

| Signal to noise ratio (30dB) | | | |
|------------------------------|---------|---------|---------|
| Signal Types | RLS | FTF | GAL |
| <i>Chirp</i> | 13.9296 | 24.7287 | 68.74 |
| <i>Sinusoidal</i> | 14.5297 | 14.513 | 72.7454 |
| <i>Saw tooth</i> | 12.7979 | 12.788 | 71.474 |
| <i>Audio</i> | 13.0794 | 25.513 | 46.6549 |

Part 2:

Table 4: Comparison of signal to noise ratio of RLS, FTF, GAL algorithms when given SNR=10dB (mid)

| Signal to noise ratio (10dB) | | | |
|------------------------------|--------|--------|---------|
| Signal Types | RLS | FTF | GAL |
| <i>Chirp</i> | 9.7591 | 8.355 | 19.3497 |
| <i>Sinusoidal</i> | 7.7829 | 7.7703 | 18.1702 |
| <i>Saw tooth</i> | 7.5157 | 7.5087 | 20.4481 |
| <i>Audio</i> | 8.2704 | 9.3365 | 10.0652 |

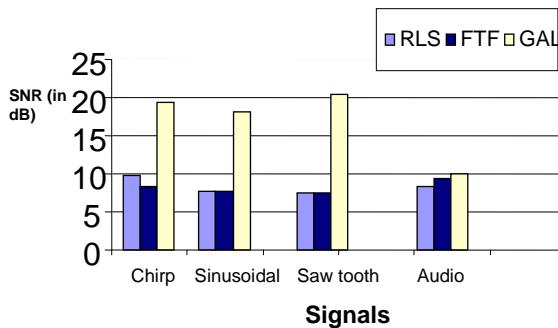


Fig. 8 Comparison of signal to noise ration of RLS, FTF, GAL algorithms when SNR=10dB (mid).

Part 3:

Table 5: Comparison of signal to noise ratio of RLS, FTF, GAL algorithms when given SNR=-10dB (low)

| Signal-to-noise ratio(-10db) | | | |
|------------------------------|---------|----------|---------|
| Signal Types | RLS | FTF | GAL |
| <i>Chirp</i> | -9.2383 | -9.3945 | -8.7799 |
| <i>Sinusoidal</i> | -9.7834 | -9.7822 | -9.4333 |
| <i>Saw tooth</i> | -9.2232 | -9.2027 | -9.0513 |
| <i>Audio</i> | -9.8538 | -10.0622 | -9.3343 |

Based On the simulated results, the following facts have been noted.

- For a fixed signal-to-noise ratio of 30 dB and considering different type of signals, we analyze the noise cancellation performance of the three algorithms. In all cases, Gradient Adaptive Algorithms (GAL) shows the best performance. In case of chirp signal, periodic signal and sawtooth signal, the output SNR of the GAL algorithms is higher than double of the input SNR value that is approximately 70dB. For audio signal, it is about 50dB.

- Almost in every case of noise cancellation, RLS and FTF algorithm show worse performance than GAL algorithms. For both of the algorithms, the output SNR values are decreased.

The same occurrence happens for both 10dB and -10dB respectively in the case of mid and low SNR. The results presented in Table 3 and 4 clearly show the superior value of the output SNR of the GAL over other two algorithms.

5. Conclusions

The components of adaptive noise canceller were generated by computer simulation using MATLAB. The analysis of the results offered useful insight into the characteristics of the algorithms. For the RLS algorithm, two parameters were varied to find their effect on the performance. One of them is the filter length, and the other is the forgetting factor. In Fast RLS algorithms, FTRLs and GAL, the performances were analyzed by varying different filter length, forgetting factor and step size parameter. The study revealed that, for the RLS, FTRLs, GAL algorithms, the increase in filter length results in increased MSE and increased convergence time. For step size, such generalization cannot be made. If the step size is increased, the algorithms converge faster. But the error tends to become unstable. Forgetting factor is other parameter which also controls the stability and the rate of convergence. Typically, it has been seen that a forgetting factor, ranges between .99 to 1, gives satisfactory results. When it comes to convergence time, the length of the filter is a big factor. It takes the RLS and FTRLs significantly longer time to compute of coefficient increases. To draw a comparison among three algorithms, the main factors that should be kept in mind are noise cancellation performance, convergence time and making the signal to noise ratio high. It is found in all cases that RLS has performed as medium level in canceling noise. Fast RLS algorithms have achieved more effective noise cancellation. In some cases FTRLs may have taken slightly more time to converge, but its error has always dipped down below that of the RLS algorithms. In the case of convergence time, GAL algorithm shows the best performance among three algorithms. The situations in which the amplitude or

frequency in signal encounters abrupt changes, the RLS and FTRLs algorithms show poor performance. In these cases, RLS and FTRLs graphs show sudden rise of error whereas the GAL remains stable to zero. Signal-to-noise ratio can be increased by canceling noise from signal and providing more strength to signal. In this case GAL always shows better performance and enhances SNR value in all types of signal either the SNR is high, low or mid. In the end, it can be stated that Fast RLS algorithms especially GAL should be preferred over the standard RLS for noise cancellation unless error convergence time, output SNR is a matter of great concern.

References

- [1] Bernard Widrow,et. al. "Adaptive Noise canceling: Principles and Applications", Proceedings of the IEE,1975,Vol.63,No.12,Page(s):1692-1716
- [2] Saymon Haykins, and Thomas Kailath, Adaptive Filter Theory, Fourth Edition, Pearson Education.
- [3] B.widrow and S.D.Stearns, adaptive Signal Processing. Englewood Cliffs, NJ:Prentice-Hall. 1985,p.474
- [4] Yuu-Seng Lau, Zahir M. Hossain, and Richard Harris, "Performance of Adaptive Filtering Algorithms", Proceedings of the Australian Telecommunications, Networks and Applications Conference(ATNAC), Melbourne, 2003.
- [5] Hyun-Chool Shin, Ali H. Sayed, and Woo-Jin Song, "Mean Square Performance of Adaptive Filters using Averaging Theory", IEEE Signal Processing Letters, May 1999, Vol. 6, pp. 106-108.
- [6] R. C. North, J. R. Zeidler, T. R. Albert, and W. H. Ku, "Comparison of Adaptive Lattice Filters to LMS Transversal filters for Sinusoidal Cancellation", Acoustics, Speech, and Signal Processing, 1992.ICASSP-92, March 1992, Vol. 4, pp. 33-36.
- [7] Syed .A.Hadei and M.loftizad, "A Family of Adaptive Filter Algorithms in Noise Cancellation for Speech Enhancement", International Journal of Computer and Electrical Engineering, April 2010, Vol. 2, No. 2, pp. 307-315.
- [8] V.R Vijaykumar,P.T. Vanathi,P.Kanagasabapathy, "Modified Adaptive Filtering Algorithm for Noise Cancellation in Speech Signals",Electronics and Electrical Engineering,Kanuša:Technologija,2007.No.2(74). P.17-20.
- [9] Ying He,et. al. "The Applications and Simulation of Adaptive Filter in Noise Canceling", 2008 International Conference on Computer Science and Software Engineering,2008,Vol.4,Page(s):1-4.
- [10] Sanaullah khan,MArif and T.Majeed, "Comparison of LMS,RLS and Notch Based Adaptive Algorithms for Noise Cancellation of a Typical Industrial Workroom",8th International Mutitopic Conference,2004,Page(s):169-173.
- [11] Yuu-Seng Lau, Zahir M.Hussian and Richard Harris, "Performance of Adaptive Filtering Algorithms:A Comparative Study", AustralanTelecommunications,networks and Applications Conference(ATNAC),Melbourne,2003.
- Lilatul Ferdouse** studied Computer Science and Engineering at University of Dhaka, Bangladesh from 1999 to 2003. In 2003 she received the Bachelor degree. She received the M. S. degree in Computer Science and Engineering in 2004 from University of Dhaka. She is working as a Senior Lecturer in Department of Computer Science, Stamford University Bangladesh. Her area of interest includes Digital Signal Processing, Data Mining, and Wireless Communication.
- Nasrin Akhter** studied Computer Science and Engineering at University of Dhaka, Bangladesh from 1999 to 2003. In 2003 she received the Bachelor degree. She received the M. S. degree in Computer Science and Engineering in 2004 from University of Dhaka. She is working as an Assistant Professor in Department of Computer Science, Stamford University Bangladesh. Her area of interest includes Digital Signal Processing, Data Mining, Cryptography and Security.
- Tamanna Haque Nipa** studied Computer Science and Engineering at Stamford University Bangladesh 1999 to 2003. In 2003 she received the Bachelor degree. Currently she is studying M.S. at Bangladesh University of Engineering and Technology. She is working as a Lecturer in Department of Computer Science, Stamford University Bangladesh. Her area of interest includes Digital Signal Processing and Wireless Communication.
- Fariha Tasmin Jaigirdar** studied Computer Science and Engineering at Chittagong University of Engineering and Technology, Bangladesh from 2001 to 2005. In 2005 she received the Bachelor degree. Currently she is studying M.S. at Bangladesh University of Engineering and Technology. She is working as a Lecturer in Department of Computer Science, Stamford University Bangladesh. Her area of interest includes Digital Signal Processing and Wireless Networking.

A Novel Architecture for Real Time Implementation of Edge Detectors on FPGA

Sudeep K C¹ and Dr. Jharna Majumdar²

¹Research Associate, Dept. of Computer Science, Nitte Meenakshi Institute of Technology,
Bangalore 560 063, Karnataka, India

²Prof. and Dean, Dept. of Computer Science, Nitte Meenakshi Institute of Technology,
Bangalore 560 063, Karnataka, India

Abstract

With the introduction of reconfigurable platform such as Field Programmable Gate Arrays (FPGA) and advent of new high level tools to configure them, image processing on FPGA has emerged as a practical solutions for most of computer vision and image processing problems. This paper briefly explains the implementation of several edge detection algorithms like Sobel, Prewitt, Robert and Compass edge detectors on FPGA and makes a comparative study of their performance. The design utilizes powerful design tool System Generator (SysGen) and Embedded Development Kit (EDK) for hardware-software codesign and integrates the edge detection hardware as a peripheral to the Microblaze 32 bit soft RISC processor with an input from a CMOS camera and output to a DVI display and verified the results video in real time.

Keywords: Field Programmable Gate Arrays (FPGA), Image processing algorithms, Edge detection, System Generator (SysGen), Embedded Development Kit (EDK), Real Time

1. Introduction

An edge in an image is a contour across which the brightness of the image changes abruptly. However, image data is discrete, so edges in an image often are defined as the local maxima of the gradient. An edge detector is basically a high pass filter that can be applied to extract the edge points in an image. The goal of edge detection is to mark the points in a digital image at which the luminance changes abruptly. The mathematical representation for the same is a convolution sequence given as:

$$g(x,y) = \sum_{k=-L}^{L} \sum_{l=-L}^{L} f(x-k, y-l) h(k,l)$$

Where, $f(x,y)$ is the input image, $h(x,y)$ is filter mask and $g(x,y)$ is the resultant image.

General purpose microcontrollers are proved to be less useful when it comes to the implementation of image processing algorithms on embedded scale. In certain instances, image processing algorithms are implemented on a dedicated Application Specific Integrated Circuits (ASIC) and more commonly on Digital Signal Processors (DSP). With the advent of Field Programmable Gate Arrays (FPGA), it has become an alternative for the implementation of Image processing algorithms on ASIC as it provides speed comparable to an ASIC and is easily reconfigurable.

The objective of this work is to develop a real-time edge detection system where the input comes from a live video acquired from a CMOS camera and the detected edges to be displayed on a DVI display screen.

2. The Setup

The setup for implementation consists of Video Starter Kit (VSK) consisting of Spartan 3A DSP XCSD3400A FPGA connected to a Micron CMOS camera of resolution 720 x 480 pixels delivering frames at 60 fps through a FPGA Mezzanine Card (FMC) Daughter card used for decoding the data arriving through the serial LVDS camera interface. The de-serialized input consists of V-Sync, H-Sync and 8 line data bus which serves as the input for the Edge detection model. The edge filter is applied in the Camera Processing block on the input signal arriving from the Camera In block. The output signal is Gamma corrected for the output DVI monitor and is driven by Display controller to the DVI output monitor. Video to VFBC and MPMC core helps us to store the image data and buffer them to the output screen.

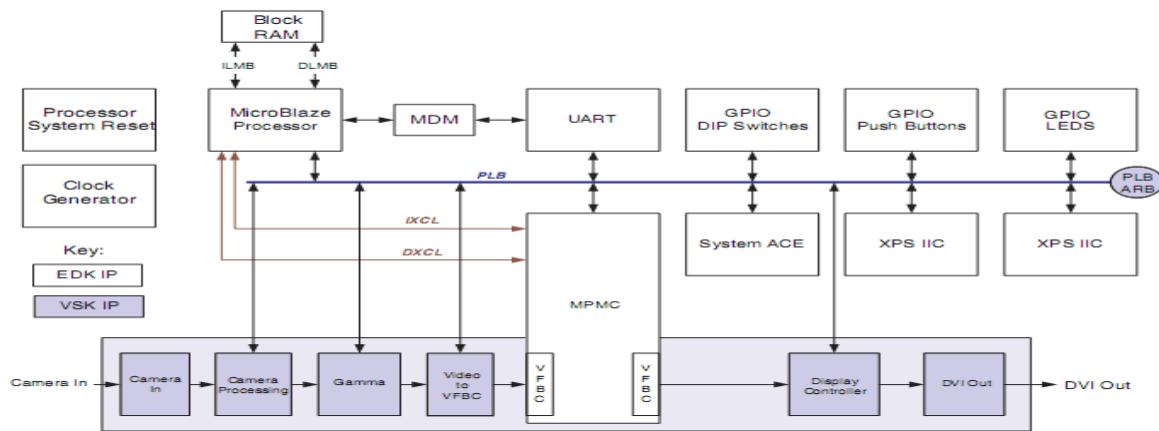


Fig. 1 Block Diagram of the complete setup

The status reporting and controlling of these blocks are carried out by Micro Blaze processor through Processor Local Bus (PLB). The block diagram of the setup is shown in Fig 1.

In the above setup a DVI display shows the output edge from the camera. However, for sake of convenience the edge detectors output is verified on few standard still images. The input is fed from a DVI video source and the edge is observed on the output display screen. The screen shot from both input and output is stored and is presented here.

3. Implementation

3.1 Sobel Edge Detector

The Sobel edge detection uses two masks, one for detecting image derivatives in horizontal direction and the other for detecting image derivatives in vertical direction. One mask is simply the other rotated by 90°. The Sobel kernels can also be thought of as 3×3 approximations to first-derivative-of-Gaussian kernels.

| | | |
|----|----|----|
| -1 | -2 | -1 |
| 0 | 0 | 0 |
| 1 | 2 | 1 |

| | | |
|---|---|----|
| 1 | 0 | -1 |
| 2 | 0 | -2 |
| 1 | 0 | -1 |

(a) (b)

Fig. 2 Sobel Edge detector – (a) Horizontal and (b) Vertical Kernel

Gradient magnitude is given by:

$$G = \sqrt{G_x^2 + G_y^2} \approx |G_x| + |G_y|$$

Gradient direction is given as,

$$\theta = \tan^{-1} \left(\frac{G_y}{G_x} \right)$$

Where,

$$G_x = \frac{\partial f}{\partial x} = (-1) * f(x-1,y-1) + (-2) * f(x-1,y) + (-1) * f(x-1,y+1) + (0) * f(x,y-1) + (0) * f(x,y) + (0) * f(x,y+1) + (1) * f(x+1,y-1) + (2) * f(x+1,y) + (1) * f(x+1,y+1)$$

$$G_y = \frac{\partial f}{\partial y} = (-1) * f(x-1,y-1) + (0) * f(x-1,y) + (1) * f(x-1,y+1) + (-2) * f(x,y-1) + (0) * f(x,y) + (2) * f(x,y+1) + (-1) * f(x+1,y-1) + (0) * f(x+1,y) + (1) * f(x+1,y+1)$$

The implementation of the Edge detector consists of RGB to grayscale color space conversion and line buffers to synchronize H-sync, V-sync and Data-enable signals. The filter and buffer block consists of line buffer to hold the corresponding rows and edge detector block. The edge detector block consists of two convolution block which performs the operation specified by the kernel. Individual rows of each kernel are multiplied by the delayed elements which are stored in the line buffers to yield vertical and horizontal edges. These are added and gradient magnitude is found out. The gradient magnitude is thresholded using a manual threshold to obtain proper edges.

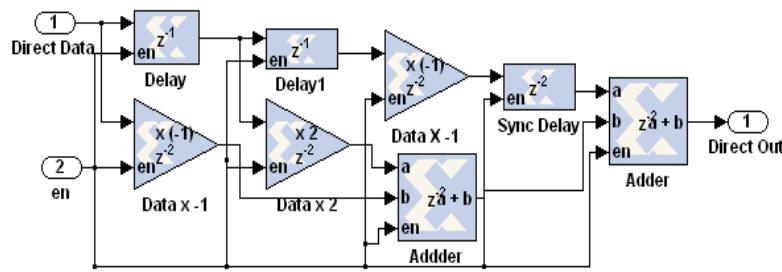


Fig. 3 Convolution block.

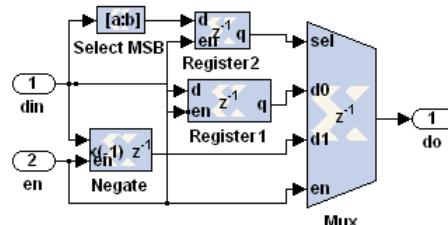


Fig. 4 Calculation of magnitude.

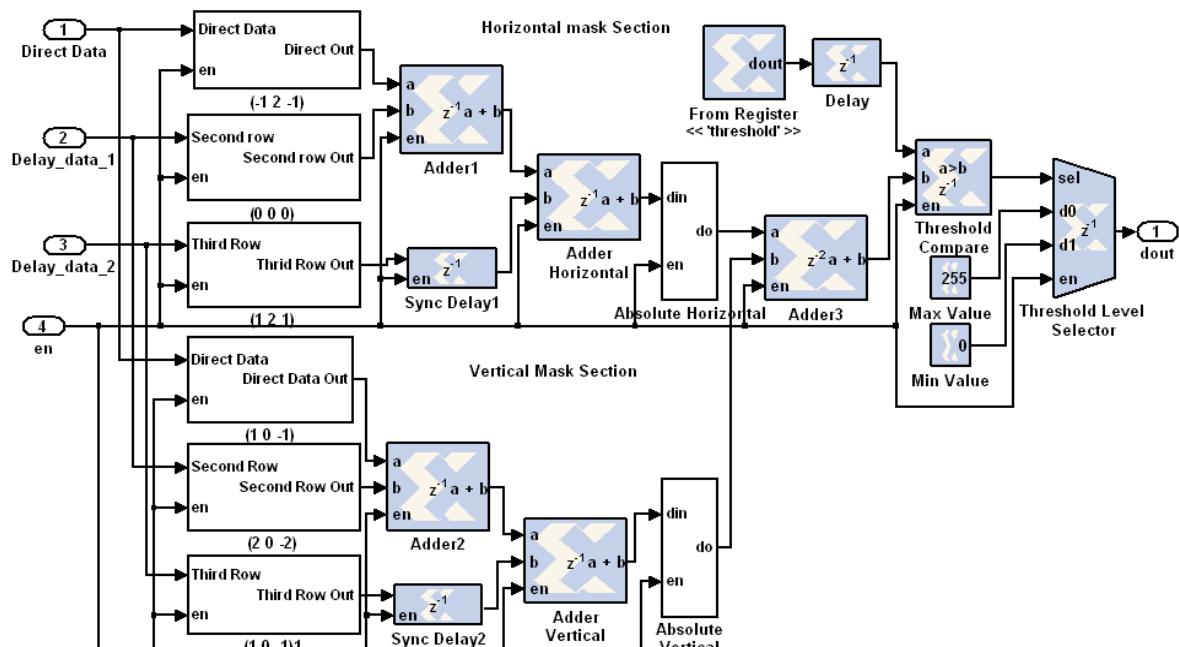


Fig. 5 Mask Implementation Block.

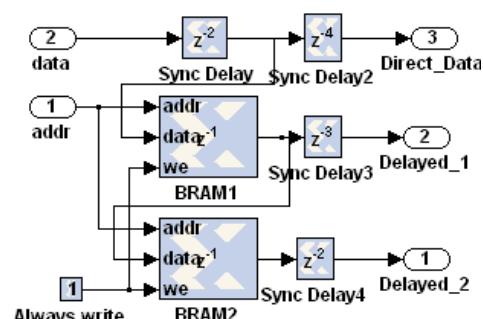


Fig. 6 Line Buffer.

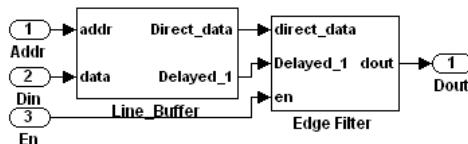


Fig. 7 Edge Detection Block.

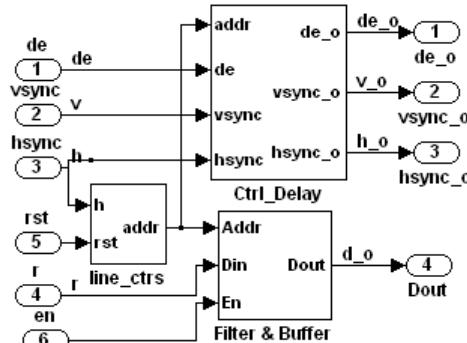


Fig. 8 Synchronization Block.

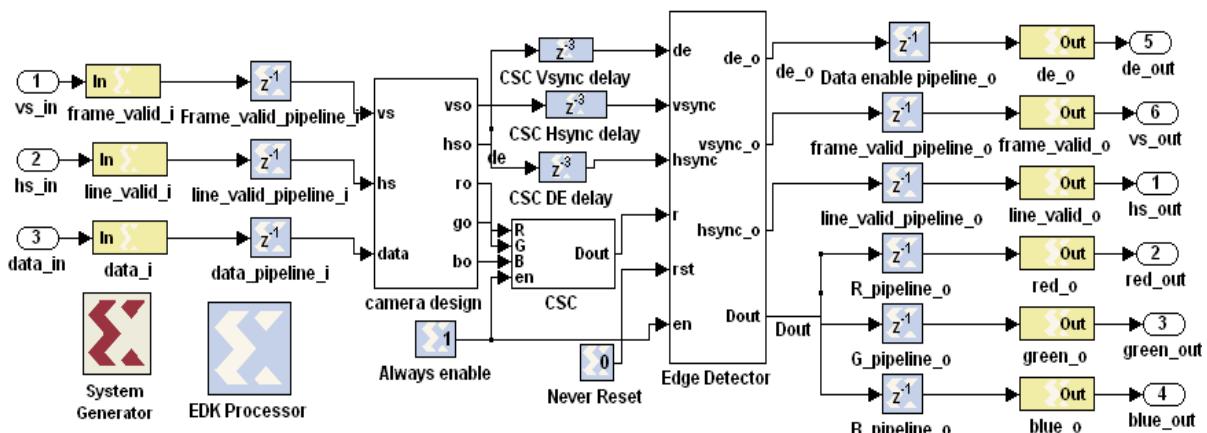


Fig. 9 Color Space Converter and Edge Detection Block.

3.2 Prewitt Edge Detection

The Prewitt edge detection is based on the idea of the central differences and give equal weightage to all pixels when averaging. Prewitt Cross operator performs a simple, quickly computable, 2-D spatial gradient measurement on the image.

These kernels are, however, sensitive to noise. We can reduce some of the effects of noise by averaging. This is done in the Prewitt kernels by averaging in y when calculating $\frac{\partial f}{\partial x}$ and by averaging in x when calculating $\frac{\partial f}{\partial y}$. Together, these kernels give us the components of the gradient vector.

| | | |
|----|----|----|
| -1 | -1 | -1 |
| 0 | 0 | 0 |
| 1 | 1 | 1 |

| | | |
|----|---|---|
| -1 | 0 | 1 |
| -1 | 0 | 1 |
| -1 | 0 | 1 |

Fig. 10 Prewitt Edge Detector – (a) Horizontal & (b)Vertical kernel

The FPGA implementation is same as Sobel's edge detector, with the change in the mask values.

3.3 Robert Edge Detector

Roberts's operator consists of a pair of 2×2 convolution kernels which are orthogonal to each other. These masks are designed to respond maximally to edges running at 45° to the pixel grid, one mask for each of the two perpendicular orientations. The masks are shown below.

| | |
|---|----|
| 1 | 0 |
| 0 | -1 |

| | |
|----|---|
| 0 | 1 |
| -1 | 0 |

Fig 11. Robert Edge Detector (a) Horizontal & (b) Vertical

Kernel.

$$G = \sqrt{|G_x|^2 + |G_y|^2} \approx |G_x| + |G_y|$$

Gradient direction is given as

$$\theta = \tan^{-1} \left(\frac{G_y}{G_x} \right)$$

Where,

$$G_x = \frac{\partial f}{\partial x} = 1 * f(x-1, y-1) + 0 * f(x-1, y) + 0 * f(x, y-1) + -1 * f(x, y)$$

$$G_y = \frac{\partial f}{\partial y} = 1 * f(x-1, y-1) + 0 * f(x-1, y) + 0 * f(x, y-1) + -1 * f(x, y)$$

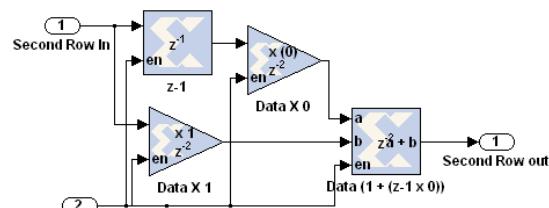


Fig. 12 Convolution Block for Robert Edge Mask.

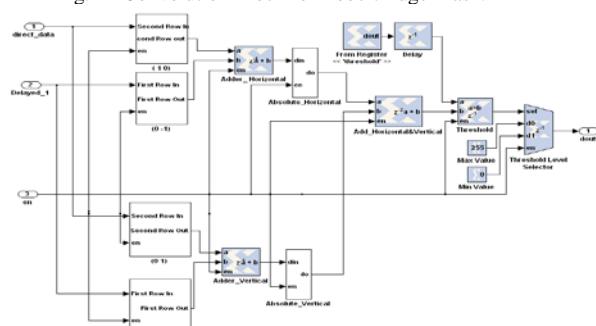


Fig. 13 Mask implementation Block.

3.4 Compass Edge Detector

Compass Edge Detectors are an alternative approach to the differential gradient edge detection. The edge magnitude and orientation of a pixel is determined by the template that matches the local area of the pixel the best. These operators comprise of eight kernels each. The whole set of 8 kernels of each operator are produced by taking one kernel of the respective type and by rotating its coefficients circularly. Each of the resultant kernels is sensitive to an edge orientation ranging from 0° to 315° in steps of 45° , where 0° corresponds to a vertical edge. In this paper, four types of compass edge detectors namely,

1. Compass,
2. Prewitt Compass,
3. Kirsch Compass,
4. Robinson Compass

are implemented

These kernels are convolved with the input image to obtain the gradients along the corresponding directions . G_S , G_{SE} , G_E , G_{NE} , G_N , G_{NW} , G_W , and G_{SW} represents gradients along South, South-East, East, North-East, North, North-West, West and South-West directions respectively.

The *edge magnitude* is given by the maximum of calculated gradients. Mathematically it can be expressed as,

$$G_{max} = \max \{G_S, G_{SE}, G_E, G_{NE}, G_N, G_{NW}, G_W, G_{SW}\}$$

where, G_{max} is the gradient corresponding to maximum pixel value.

The local *edge orientation* is estimated with the orientation of the kernel that yields the maximum response. The values for the output orientation image lie between 0 and 7, depending on which of the 8 kernels produced the maximum response. These can be quantized in terms of angle ranging from 0 to $(7 \cdot \pi/4)$ radians differing by $\pi/4$ in anticlockwise direction. Direction is perpendicular to the edge. Generally, direction points to the brighter side of the edge. Mathematically it can be expressed as,

$$\theta = \tan^{-1} \left(\frac{G_{max}}{G_z} \right)$$

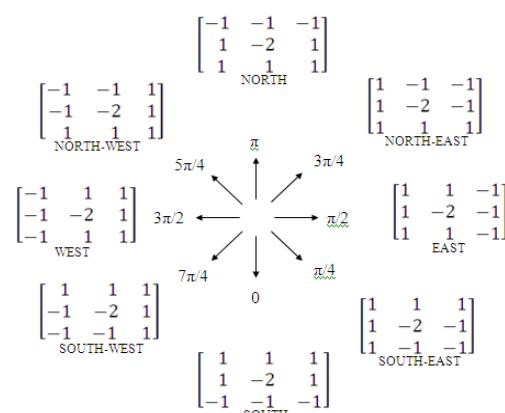


Fig. 14 Compass Masks and their orientation

The south-east direction masks for each type methods are shown below.





Fig. 15 Compass Edge Detector – South East Directional kernel

The FPGA implementation involved implementation of 8 Kernel operators which are similar to the Sobel kernel discussed above and a

maximum block, to find the maximum of the calculated edge gradients.

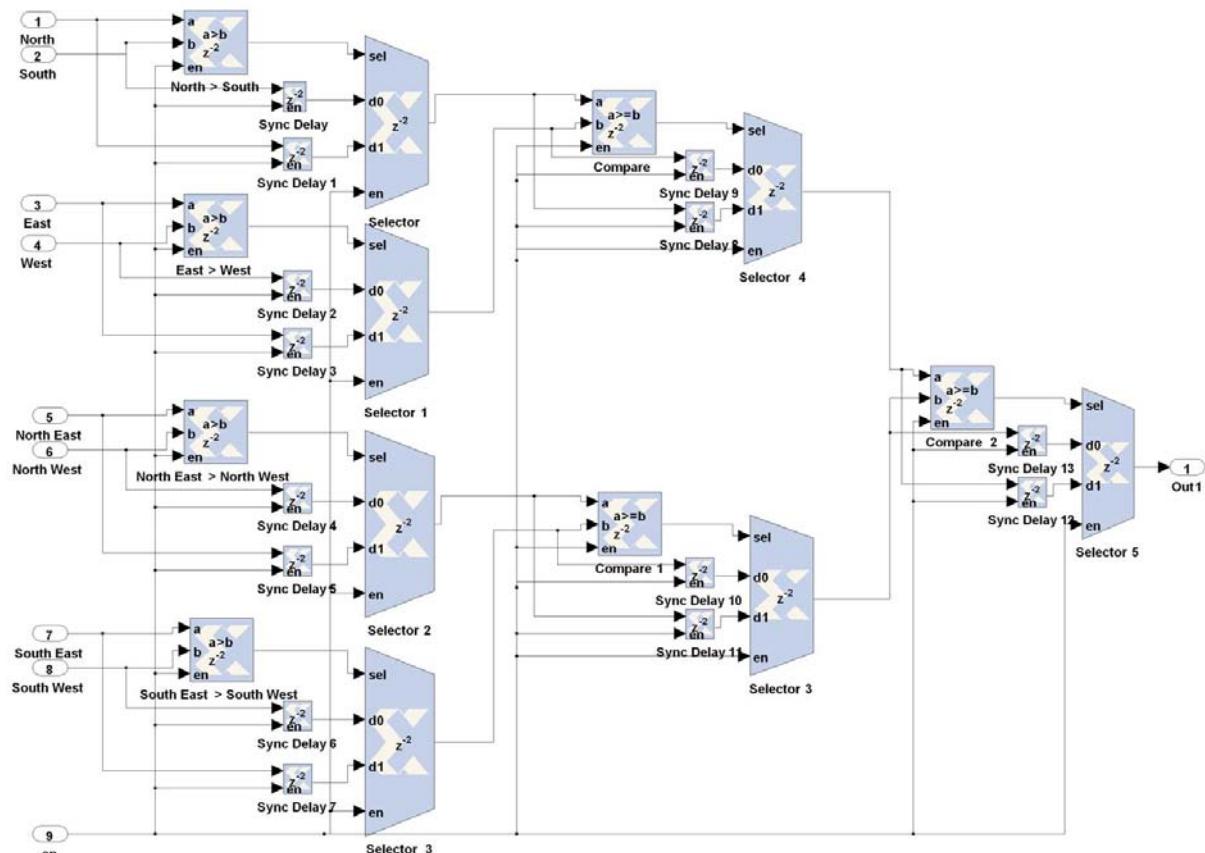


Fig. 16 Edge Maximum Block.

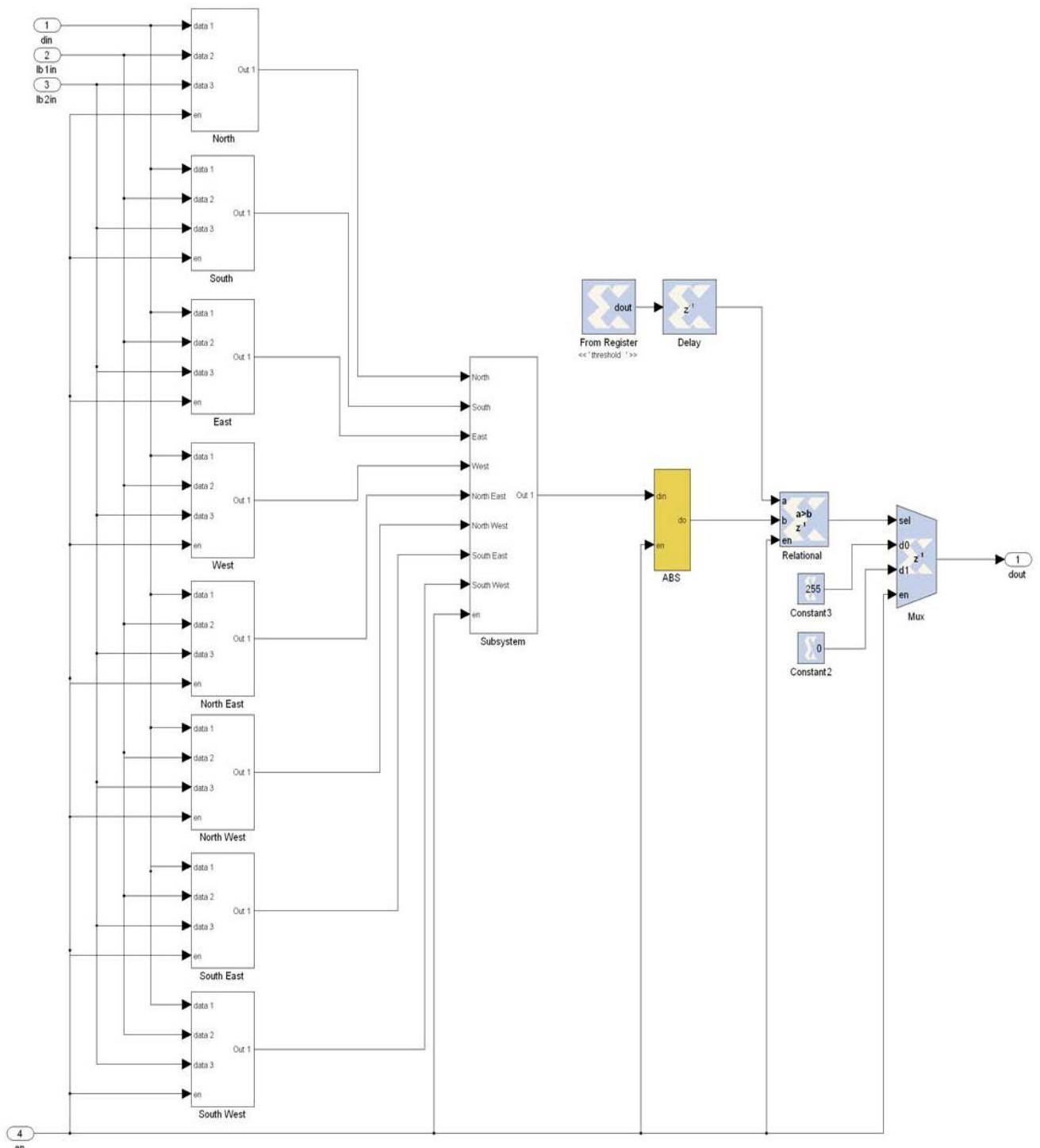


Fig. 17 System Generator Block for Compass Edge Detectors.

4. Results

The output image for input images of resolution 800 x 600 discussed above are shown below.

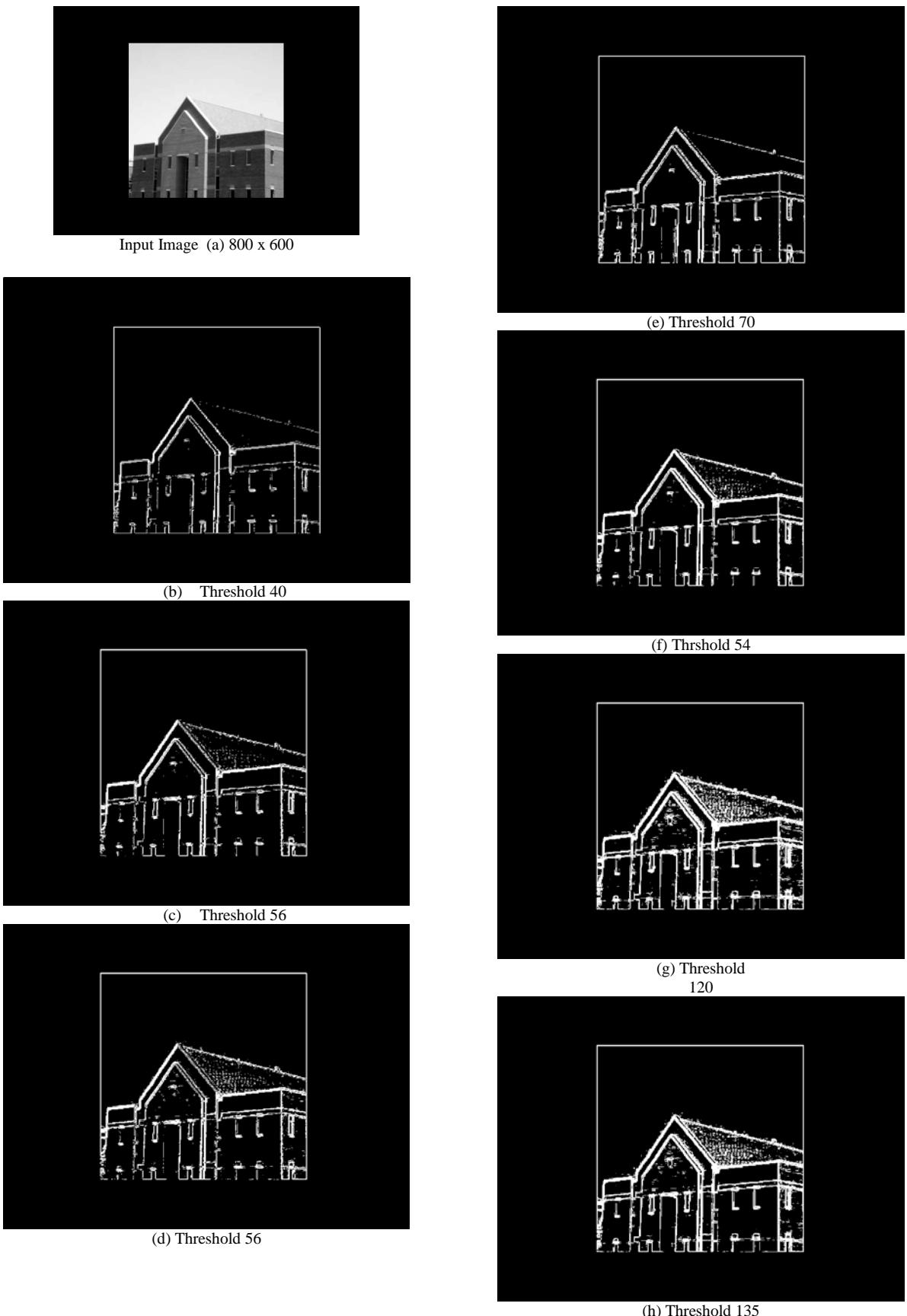


Fig. 18 (a) Input image of resolution 800 x 600 . Edge output using (b) Robert (c) Prewitt (d) Sobel (e) Robinson Compass (f) Prewitt Compass (g) Kirsch Compass Masks (h) Sobel Compass Masks.

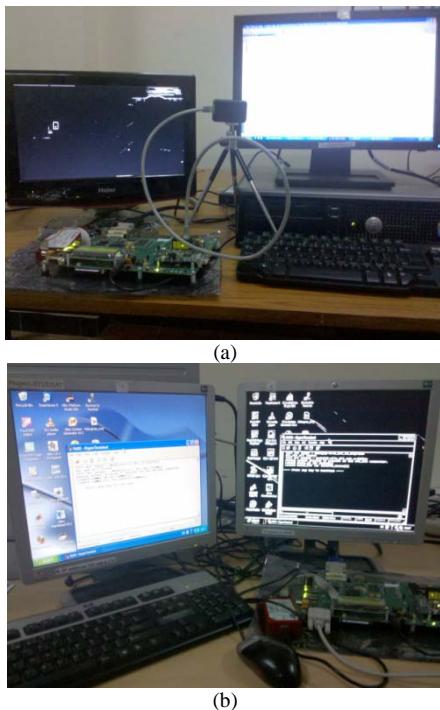


Fig. 20 (a) Experimental setup for implementation of edge detection. Input is from CMOS camera and the output is on a DVI display (b) Setup for verification. Input is obtained from a DVI Video source and output is on a DVI display.

Edge detection methods investigated so far are further assessed by quality measures that give reliable statistical evidence to distinguish among the edge maps obtained. The absence of the ground truth edge map reveals the search for an alternative approach to assess and compare the quality of the edge maps resulted from the detectors exploited so far. The evidence for the best detector type is judged by studying the edge maps relative to each other through statistical evaluation. Upon this evaluation, an edge detection method can be employed to characterize edges to represent the image for further analysis and implementation.

Table 1: Relative frequencies of edge appearance for various edge detectors on edge output image

| Operators | Sobel | Roberts | Prewitt |
|-------------------------|--------|---------|---------|
| Sobel | 1 | 0.7232 | 1.0014 |
| Roberts | 1.3827 | 1 | 1.3847 |
| Prewitt | 0.9986 | 0.7222 | 1 |
| Robinson Compass | 0.9976 | 0.7215 | 0.999 |
| Prewitt Compass | 1.0128 | 0.7325 | 1.0142 |
| Compass Sobel | 0.9977 | 0.7215 | 0.9991 |
| Compass Kirsch | 0.9928 | 0.781 | 0.9942 |

| Operators | Robinson Compas | Prewitt Comp | Sobel Compas | Kirsch Comp |
|-------------------------|-----------------|--------------|--------------|-------------|
| Sobel | 1.0024 | 0.9874 | 1.0023 | 1.0072 |
| Roberts | 1.3861 | 1.3652 | 1.386 | 1.3927 |
| Prewitts | 1.001 | 0.986 | 1.0009 | 1.0058 |
| Compass | 1 | 0.985 | 0.9999 | 1.0048 |
| Prewitt Compass | 1.0152 | 1 | 1.0152 | 1.0201 |
| Robinson Compass | 1.0001 | 0.9851 | 1 | 1.0048 |
| Kirsch Compass | 0.9952 | 0.9803 | 0.9952 | 1 |

The above table summarizes the relative frequencies of edge pixel occurrence for various edge detectors. For each edge map, max (ndf) where ndf is the frequency f of occurrence for the filter is reported, and the ratio with respect to each other gives comparative statistics for the occurrence of edges. The Compass filter with Kirsch mask reports higher detected edge pixels.

Table 2: Resource estimation for various edge detectors on Spartan 3A DSP XCSD3400A FPGA

| Edge Operator | Slices | Flip Flop s | LUT | IOBs | BRAMs | Max. Freq (Mhz) |
|------------------|--------|-------------|------|------|-------|-----------------|
| Sobel | 4% | 4% | 3 % | 29% | 4% | 154.918 |
| Prewitt | 4% | 4% | 3 % | 29% | 4% | 154.918 |
| Robert | 2% | 1% | 1 % | 29% | 3% | 154.959 |
| Compass operator | 19% | 16% | 12 % | 29% | 4% | 137.646 |

5. Conclusion

Various edge detectors were implemented on FPGA at a rate of 60 fps for an input image of resolution 720x480. The design was tested for a resolution upto 1024 x 768 on a DVI input and output. The implementation can be extended to color video images as well for much faster frame rate.

Acknowledgments

The authors express their sincere gratitude to the Director and the Principal of Nitte Meenakshi Institute of Technology for providing Spartan 3ADSP Video Starter Kit to carry out the research

work. This work is inspired by the work done by Sneha et al[2].

References

- [1] R. Gonzalaz, R. Woods, "Digital Image Processing". New Jersey: Prentice-Hall 2002, ch 10.
- [2] Sneha H L, Rashmi R, Sushma R P, Rajesh N, Dr. Jharna Majumdar, "FPGA Implementation of Real – Time Edge Detectors using Xilinx System Generator", National conference for Emerging trends in Control, Communication, Signal Processing & VLSI Techniques
- [3] Ownby, M.; Mahmoud, W.H.; , "A design methodology for implementing DSP with Xilinx® System Generator for Matlab®, " System Theory, 2003. Proceedings of the 35th Southeastern Symposium on , vol., no., pp. 404- 408, 16-18 March 2003
- [4] Hong Shan Neoh, Asher Hazanchuk, "Adaptive Edge Detection for Real-Time Video Processing using FPGAs"
- [5] Zhang Zongcheng; Wang Hongyan; Liu Weixiao; Yu Ming; , "Real-time image video edge enhancement using hardware convolver," TENCON '93. Proceedings. Computer, Communication, Control and Power Engineering.1993 IEEE Region 10 Conference on , vol., no.0, pp.398-401 vol.3, 19-21 Oct 1993
- [6] Mamta Juneja, Parvinder Singh Sandhu, "Performance Evaluation of Edge Detection Techniques for Images in Spatial Domain", International Journal of Computer Theory and Engineering, Vol. 1, No. 5, December, 2009
- [7] Xilinx, Spartan 3A DSP FPGA Family Datasheet



Sudeep K C graduated from Nitte Meenakshi Institute of Technology in 2009 from Dept. of Electronics and Communication. He has been a part of On

Board Computer subsystem of the STUDSAT – a student pico satellite program in 2009 and since then has been a part of Industrial Robot Automation program. His research areas include implementation of image and computer vision algorithms on embedded and reconfigurable hardware. He has been working in Dept. of Computer Science in Nitte Meenakshi Institute of Technology since 2009.

Dr. Jharna Majumdar Dr. Jharna Majumdar is currently working as Dean R & D and Professor and Head of Computer Science and Engineering (PG) at the NITTE Meenakshi Institute of Technology, Bangalore. Prior to this Dr. Majumdar served Aeronautical Development Establishment, Defence

Research and Development Organization (DRDO), Ministry of Defence, Govt. of India from 1990 to 2007 as Research Scientist and Head of Aerial Image Exploitation Division, Bangalore. Dr. Majumdar has 37 years. of experience in R & D and Academics in the country and abroad. She has published large number of papers in National, International Conferences and Journals. Her research areas include Image and Video Processing for defense and non defense application, Robot Vision, Vision based autonomous guided systems, development of Computer Vision Algorithms in FPGA etc.. Recently as the Project Coordinator, she has provided leadership to a team of undergraduate students from seven engineering for the development of STUDSAT, the smallest STUDent SATellite of PICO Category (Less than 1 Kg), first time developed in the country.



FARS: Fuzzy Ant based Recommender System for Web Users

Shiva Nadi¹, Mohammad. H. Saraei², Mohammad Davarpanah Jazi³ and Ayoub Bagheri⁴

¹ Department of Computer Engineering, Islamic Azad University of Najafabad
Isfahan, Iran

² Department of Electrical and Computer Engineering, Isfahan University of Technology
Isfahan, Iran

³ Department of Computer Engineering, Foulad Institue of Technology
Fouladshahr, Isfahan, Iran

³ Department of Electrical and Computer Engineering, Isfahan University of Technology
Isfahan, Iran

Abstract

Recommender systems are useful tools which provide an adaptive web environment for web users. Nowadays, having a user friendly website is a big challenge in e-commerce technology. In this paper, applying the benefits of both collaborative and content based filtering techniques is proposed by presenting a fuzzy recommender system based on collaborative behavior of ants (FARS). FARS works in two phases: modeling and recommendation. First, user's behaviors are modeled offline and the results are used in second phase for online recommendation. Fuzzy techniques provide the possibility of capturing uncertainty among user interests and ant based algorithms provides us with optimal solutions. The performance of FARS is evaluated using log files of "Information and Communication Technology Center" of Isfahan municipality in Iran and compared with ant based recommender system (ARS). The results shown are promising and proved that integrating fuzzy Ant approach provides us with more functional and robust recommendations.

Keywords: *Web personalization, Recommender Systems, Ant colony optimization, Fuzzy set.*

1. Introduction

Recommender systems (RS) are useful tools which guarantees that right information are accessible for right users at right time [1]. RSs are useful in different domains, such as web personalization, information filtering, e-commerce, providing recommendations of books, movies and music. One of the most popular applications of recommender systems is web environment personalizing by providing a list of items related to user's interests. This paper proposes a fuzzy-ant based Recommender system (FARS) which provides a list of recommendations for currently online user by comparing active user's navigational behavior with data collected from other users.

Ant colony optimization (ACO) is a computational algorithm that mimics the behavior of ants and is proposed by Dorigo in 1996. As the uncertainty is the nature of user's behavior the dynamic intelligent FARS proposed in this paper, uses ACO and fuzzy logic to prepare the high potential and suitable promoting recommendations for active user.

Usually demographic, content based and collaborative based filtering techniques are employed to generate recommendations. In demographic filtering technique (DMF), users are categorized based on their personal attributes such as their age range and provide recommendation based on these demographic categories. Recommendations produced by these systems are too general and not adaptive with changes in user preferences over time. Content based filtering (CBF) provides users with similar conceptual recommendations based on previously evaluated items. Content-based filtering methods usually utilize text extraction techniques for building user profiles. These methods have some disadvantages such as mismatch between user profile terms and item profile terms, that leads to decreasing the performance [2]. Collaborative based filtering technique (CF) suggests items based on similar users preferences. This method provides poor prediction when the number of similar users is small. Overcoming disadvantages of these models, in this paper we employ an integrated CF and CBF based hybrid model. Applying this hybrid model on the web user's behaviors obtained from the analysis of log files, we propose a fuzzy-ant based recommender system. The rest of the paper is organized as following. In section 2 the background of this research is explained. The purposed model is described theoretically in section 3. Experimental results about the implemented system and a

brief discussion about model is represented in section 4 and finally in section 5 the conclusion and future research directions are presented.

2. Research Background

During the recent years, many researches are done for personalizing websites using Variety of techniques. In the field of swarm intelligence algorithms, Ujjin and Bently have presented a recommender system based on particle swarm optimization algorithm [3]. They proved that the results obtained from their PSO recommender system are more accurate than the genetic and Pearson algorithm. In another work a fuzzy genetic recommender system with the accuracy of memory based CF and the scalability of model based CF. their novel user model helps achieving complexity and sparsity reduction in system. The performance of their method is proved by comparing the results with Pearson and fuzzy recommender system [4]. Sobecki used ant colony metaphor for selecting optimal solutions in his hybrid recommendation method and Bedi also, presented a recommender system based on collaborative behavior of ants. He used collaborative filtering approach and generated recommendations for Jester dataset [5]. Clustering is an important step in all recommendation systems, choosing an appropriate clustering algorithm leads to producing more qualified recommendations. The c-means and k-means algorithms are most well known clustering algorithms [6]. Fuzzy c-means is a method of clustering which allows one piece of data to belong to two or more clusters. This method developed by Dunn in 1973 and improved by Bezdek in 1981. Swarm algorithms are also used for clustering items. Ant based clustering has been introduced by Deneuborg, in this algorithm ants discriminate between different kinds of items and spatially arrange them according to their properties. This algorithm is modeled of the real behavior of ants in nature. The proposed approach by Kanade and Hall (2003), presents the combination of ant based clustering and FCM [7]. Their model is employed in this study for clustering web users based on their accesses to web pages.

3. Proposed fuzzy-ant recommender system

In this section, a dynamic intelligent recommender system for providing high potential recommendations is proposed, which uses fuzzy logic in accordance with ant colony optimization for increasing the accuracy and relevancy of predictions. The system analyzes a user's navigational behavior during a period of time and find out the most ideal recommendations for him. Figure 1 show the schema of the proposed system, which consists of modeling and recommendation phases. In first step, user preferences are

identified using web access log data called web usage data. In the next step, the knowledge which is achieved through the previous step is used to identify the possibly interested URLs and provide recommendations to the users. This recommendation can be done in different manners such as adding related hyperlinks to the web page requested by the user.

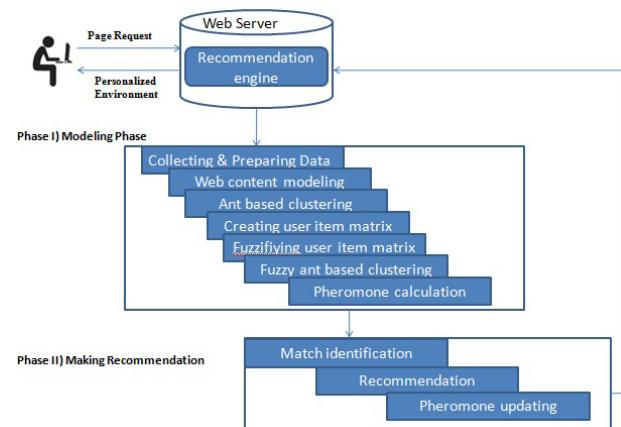


Fig. 1 Proposed work schema

In this paper, Represented method dynamically recommends the highest match score URLs to the users with similar interests, the recommended URLs are also related conceptually to each other. In this way, we integrated fuzzy logic to ant colony semaphore to provide a recommender system which produce optimum recommendations for users, considering the uncertainty among user's interests.

3.1 Data Modeling Phase

In this phase through several sub phases, pure data extracted from logs files become prepare for grouping users in appropriate clusters based on their interests' similarity.

Step1) Data Collection

Web servers provide log files which have useful information about access of all users to a specific website. Extracting these information, some preprocesses should be done on log files. The result will be a reformed log file which contains useful information about the accessed URLs and IP addresses.

Step2) Web Content Mining

Various numbers of clustering techniques are used for clustering documents. In this work, following algorithm is used for document clustering. Assume $P = \{p_1, p_2, \dots, p_k\}$ is the set of k website's pages that will be grouped in content based clusters using following steps:

Step 1. Assign each page to a single cluster

Step 2. Merge clusters based on Jaccard coefficient similarity measure by Eq. 1.

$$sim(p_x, p_y) = \frac{|p_x \cap p_y|}{|p_x \cup p_y|} \quad (1)$$

Where $|p_x \cap p_y|$ is the number of common words between two basic clusters and $|p_x \cup p_y|$ is total number of words in both clusters.

Step 3. Repeat step 2 until all documents being clustered.

The result is the set, $DC = \{DC_1, DC_2, \dots, DC_n\}$ and DC_i represents a set of URLs with similar content.

Step3) Creating User-Item Matrix

Most of recommender systems which are CF based use user-item matrix for identifying similar users and recommending items which are highly ranked by those users. User-item matrix is a matrix of size $m \times n$ where each entry represents the interest degree of i-th user to j-th document cluster. The interest degree is the value of each document cluster for each user. We use the interest degree which proposed by Castellano (2007), they defined it as the ratio of the number of accesses to each document cluster to the total number of accesses to all document clusters for each user [8].

$$Val(DC_i, U_i) = \frac{|\{DC_j | DC_j \in A_i\}|}{|A_i|} \quad (2)$$

$A = \{A_1, A_2, \dots, A_m\}$ is a set of user's accesses to document clusters. For example A_k indicates the access list of K -th user to a subset of document clusters ($A_k \in DC$, for $K=1$ to n)

Step4) Normalizing and fuzzifying user-item matrix

Fuzzy sets theory which model vagueness was introduced by Lotfi A. Zadeh in 1965. Fuzzy sets support a flexible membership of elements to the sets. While in crisp set theory, an element absolutely belongs to a set or doesn't but in fuzzy sets theory variety of membership degrees in the range of 0 to 1 can be allocated to items in classifying interest domains. We use triangular fuzzy numbers to characterize user interests to document features. A triangular number has a triangle shaped membership function, which can be viewed as possibility distribution. To express user interests, the linguistic terms are used to linguistically evaluate the importance of user interests. Five linguistic sets can be used to describe the diversity of user's interests: VS (Very Small), S (Small), M (Medium), H (High), VH (Very High)

The membership function of user's interests is shown in Figure 2.

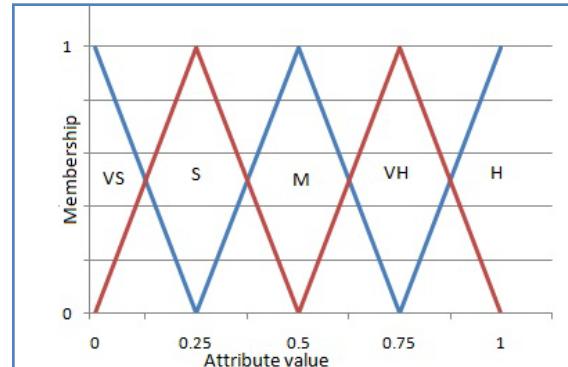


Fig. 2 membership function

The distance between two users, fuzzy sets illustrates the degree in which these users are the same. The global distance is an aggregation of distances between any two partial feature distances. The aggregation operator may be the average of the local fuzzy distances [9].

$$fd(x, y) = \frac{1}{N} \sum_{i=1}^N fd(x_i, y_i) \quad (3)$$

Step 5) Fuzzy ant based clustering

The ant based algorithm provides acceptable clusters of data without any knowledge of the initial clusters. In the ant based algorithm if an object is assigned to an inappropriate heap then it takes long time to be transported to a better heap. Overcoming this problem, a synthetic algorithm proposed by Kanade and Hall (2003) that uses the integrated FCM and ant based clustering algorithms [7]. In the web recommendation field we supposed that objects are users associated with a vector of fuzzy numbers which indicates their interests to document clusters.

The fuzzy C-means algorithm requires good initialization. These initial values are provided by the ant based algorithm. The result will be small homogenous heaps that will be merged by repeating the steps. By increasing the number of iterations the number of heaps decreases. The following algorithm is used in this study to cluster users in appropriate groups: [7]

- Scatter the users randomly on the board
- Initialize the ants with random position, and random direction
- For N iterations do
 - For each ant do
 - Move the ant
 - If the ant is carrying an user U then possibly drop the user U else
 - Possibly pick up a user

- Use the cluster centers obtained in step 3 to initialize cluster centers for the fuzzy C-means algorithm
- Cluster the data using the fuzzy C-means algorithm
- Harden the data obtained from the Fuzzy C-means algorithm, using the maximum membership criterion, to form new heaps
- Repeat step 1-6 by considering each heap as a single object

The advantage of this algorithm lies within the calculation of initial clusters' centers which is based on Ant colony algorithm. Table 1 illustrates a cluster containing 3 users. The center of the cluster is a vector (table 2), which is computed as the mean of user preferences in the user cluster.

Table 1: k-th user cluster

| UC _k | DC1 | DC2 | DC3 | ... | DC8 |
|-----------------|-----|------|------|-----|-----|
| U1 | 0.3 | 0.0 | 0.5 | | 0.2 |
| U6 | 0.2 | 0.1 | 0.7 | | 0.0 |
| U9 | 0.0 | 0.06 | 0.64 | | 0.3 |

Table 2: Center of cluster

| UC _k | DC1 | DC2 | DC3 | ... | DC8 |
|-----------------|-------------|------------|------------|-----|--------------|
| U1 | 0.080200 | 0.0000 | 0.0100 | | 0.208000 |
| U6 | 0.208000 | 0.604000 | 0.002080 | | 0.0000 |
| U9 | 0.0000 | 0.76024000 | 0.00440560 | | 0.080200 |
| Center | 0.060500600 | 0.45021000 | 0.00540450 | | 0.0605300600 |

Step 6) Calculating pheromone for each cluster

The pheromone associated with each cluster is calculated based on the density of each cluster using Eq. 4.

$$P_{\text{pheromone}}(t) = \frac{\text{Number of users in cluster } i}{\text{total Number of users}} \quad (4)$$

3.2. Recommendation process for active user

When a new user starts a transaction, our model matches the new user with the most similar user clusters and provides suitable recommendations for him/her. It will be done by computing the similarity of the active user profile with center of the user cluster. By combining the similarity and pheromone value allocated to each cluster the active user can be matched with existing clusters.

Step1) Matching active user with best user clusters

Similarity of fuzzy sets is based on their distance which is calculated using Eq. 5 [10].

$$SM(U_{\text{new}}, U_{\text{center}}) = \frac{1}{1 + DM(U_{\text{new}}, U_{\text{center}})} \quad (5)$$

Where SM is the similarity measure and DM is the distance measure of two U_{new} and U_{center} fuzzy sets. U_{new} is the active user fuzzified profile and U_{center} is center of i -th cluster.

The distance between two fuzzy sets of the active user profile and the center of the cluster can be computed using Eq.6 [10].

$$DM(U_{\text{new}}, U_{\text{center}}) = \sum_{i=1}^l w_i \cdot DM_k(U_{\text{new}}, U_{\text{center}}) \quad (6)$$

Where l is the total number of fuzzy sets, w_i is the weighting for i -th linguistic variable, here we got it as 1, and DM_k is the distance measure along the fuzzified features variables.

$$DM = \sqrt{\sum_{i=1}^l (U_{\text{new}}_i - U_{\text{center}}_i)^2} \quad (7)$$

Where, l is the total number of fuzzy sets.

Using E.q. 5 the similarity of active user is calculated with other clusters and clusters which have users with similar preferences are distinguished.

The clusters which have more pheromone and more similarity to active user are chosen as clusters with most match score with active user. The match score at time t is calculated using Eq. 8 [11].

$$MatchScore_i(t) = \frac{pheromone_i(t) \cdot SM(U_{\text{new}}, U_{\text{center}})}{\sum_{i=1}^k pheromone_i(t) \cdot SM(U_{\text{new}}, U_{\text{center}})} \quad (8)$$

The clusters in the range of [highest MatchScore – α , highest MatchScore] are chosen for recommendation.

Table 3 illustrates an active user profile applying fuzzy ant based clustering algorithm described in section 3.1.5, 4 user clusters are obtained. The match score for the active user and these 4 clusters are illustrated in table 4.

Table 3: Normalized active user profile

| | DC1 | DC2 | DC3 | DC4 | DC5 | DC6 | DC7 | DC8 |
|---------------------|------|-----|------|-----|-----|-----|-----|------|
| active user profile | 0.12 | 0 | 0.24 | 0 | 0.3 | 0 | 0 | 0.64 |

Table 4: match score between active user and user clusters

| | Pheromone | Similarity | MatchScore |
|-----|-----------|------------|------------|
| UC1 | 0.4 | 0.237 | 0.364 |
| UC2 | 0.2 | 0.153 | 0.117 |
| UC3 | 0.3 | 0.370 | 0.426 |
| UC4 | 0.1 | 0.240 | 0.092 |

By choosing $\alpha = 0.1$, the clusters between the range of $0.426 - 0.1 \leq \text{Match Score} \leq 0.426$ are chosen for extracting suitable recommendations. Here UC1 and UC3 have more compatibility with active user.

Step 2) Recommendation

The values of documents which have been accessed in chosen clusters and not by active user are candidates to be presented for active user as recommended items. Table 5 illustrates the centers of clusters U_{C_1} and U_{C_3} which has the most match score with the active user. As presented in active user profile DC₁, DC₃, DC₅ and DC₈ are accessed before and are not considered for recommendation.

Table 5: the value of candidate document clusters in chosen user clusters

| | DC2 | DC4 | DC6 | DC7 |
|---------------------|-------|-------|-------|-------|
| Center of U_{C_1} | 0.222 | 0.208 | 0.358 | 0.095 |
| Center of U_{C_3} | 0.193 | 0.064 | 0.194 | 0.111 |

DC2, DC4 and DC6 are chosen from user clusters 1, and DC7 is chosen from user cluster 3.

Step 3) Pheromone Updating

The pheromone associated with good solutions must be increased and the pheromone values associated with bad ones must be decreased. This is possible by decreasing the value of pheromone through pheromone evaporation and increasing the pheromone levels associated with a suitable solution. The amount of pheromone associated with each cluster is decreased by small value and pheromone associated to the cluster from which recommendation is generated is increased according to formula. 9 [11].

$$\text{pheromone}_i(t) = (1 - \rho) \times \text{pheromone}_i(t-1) + \Delta Q \times \text{pheromone}_i(t-1) \quad (9)$$

Where, $\rho = 0.01$ is pheromone evaporation rate. ΔQ is calculated using Eq. 10.

$$\Delta Q = \frac{Q_{cc}}{Q_{cc} + 1} \quad (10)$$

Where, Q_{cc} is the value of item in selected user cluster. The result of pheromone updating process is shown in Table 6.

Table 6: Pheromone updating

| | UC1 | UC2 | UC3 | UC4 |
|----------------------------------|-------|-------|-------|-------|
| Initial Pheromone | 0.400 | 0.200 | 0.300 | 0.100 |
| Pheromone after recommending DC2 | 0.468 | 0.190 | 0.290 | 0.090 |
| Pheromone after recommending DC4 | 0.543 | 0.180 | 0.280 | 0.080 |
| Pheromone after recommending DC6 | 0.680 | 0.170 | 0.270 | 0.070 |
| Pheromone after recommending DC7 | 0.670 | 0.160 | 0.260 | 0.076 |

The pheromone assigned to each cluster is updated based on recommendations made for active user and will be used for future recommendations.

4. Discussion and experimental results

This paper improved web user recommendation quality by considering the fuzzy behavior of users for modeling them in suitable clusters and supports them by a range of relevant recommendations. The fuzzy ant based recommender system proposed in this paper has been

applied on the access log file of “Information and Communication Technology Center” of Isfahan municipality in Iran for IP address 80.191.136.6. We collected log file during a period of one week. After data cleaning in preprocessing step, the number of requests was 5232. The number of accessed URLs in this website was 200 pages. Employing content based document clustering algorithm the URLs were grouped to 15 clusters. In next step, user’s behaviors have been modeled as access matrix. Using equation 2, interest degrees are calculated as a 5232×15 matrix. A set with 4000 rows and 5 columns is considered as the training set and is used as the input of user clustering algorithm.

In clustering step, the combination of ant based clustering and fuzzy C-means was used for clustering web site’s users. The number of algorithm iterations was set to 1000 and the values of T_{create} , P_{drop} , P_{destroy} , P_{load} , T_{remove} was set to 0.5, 0.2, 0.3, 0.3, 1.5, respectively and the number of ants was set to $n/3$ where n is the total number of objects to be clustered [10]. Here we set the number of ants to be 1000 and the value of 0.01 is taken for pheromone evaporation rate. Following the steps described in section 3 a set of documents are derived to be recommended for active user.

The proposed model predicts user preferences not only based on the interests of other users but also the conceptual similarity between website pages is intended. The combination of ant clustering and FCM used in this model locates users in appropriate classes and provide perfect clusters of users with similar interests. As providing recommendations is based on the interest of users in the same clusters, improving quality of clusters increases the completeness and exactness of recommendations. In clustering process, the fuzzy C-means algorithm requires some initializations that feed by ant based algorithm without any initial knowledge of the center of clusters. For calculating the exactness and completeness of produced recommendations, precision and recall measures are computed. In this way, a user profile is chosen and divided into a training and test set. According to the knowledge obtained from analyzing the training set, some predictions are produced. The ratio of the number of correctly predicted items over the number of predicted items indicates the recall measure. Precision is defined as the ratio of the number of correctly predicted items over the size of top n set. Recall and precision are computed for the proposed method and were compared with the results of ant based recommendation method in which user clustering method was singly ant based. The results are shown in Figure 3 and 4. N is the number of topmost recommendations.

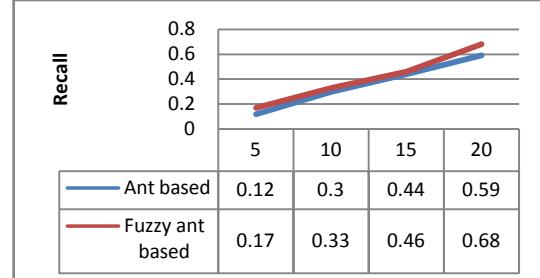


Fig. 3 Recall measures for N top recommendations

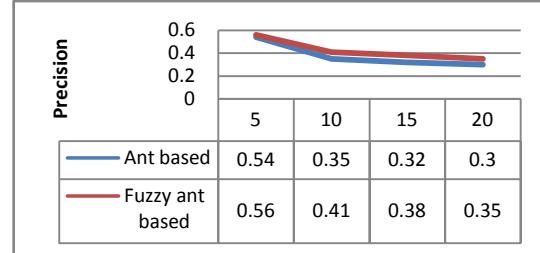


Fig. 4 Precision measures for N top recommendations

The Figure 3 and 4 indicates the proposed approach improves the performance of Ant based method and modifies the results with respect to the user demand and she/he interests and therefore our approach shows more qualified recommendations.

5. Conclusions

Recommender systems are tools which provide a personalized environment for the users of a web site by investigating their navigational behavior in a period of time. In this paper a fuzzy ant based recommender system (FARS) was proposed. In this method, the user’s interests to the web pages are extracted from web server log files. By applying FCM and ant based clustering algorithms users are grouped in appropriate clusters. Ant based algorithm helps in providing optimal solutions and FCM considers the uncertainty in user’s interests. Then the match score of currently online user with existing clusters is calculated based on their similarity and the amount of pheromone on each cluster. The pheromone assigned to each cluster is updated at end of recommendation process for future uses. Finally the precision and recall are computed to expressing the exactness and completeness of produced recommendations. The results indicate that using the proposed method for clustering users will lead to offering more qualified recommendations. Our work can be extended by assigning an appropriate weight to the documents in each document cluster for increasing their effectiveness in match score identification.

References

- [1] Andomavicius, G. and A. Tuzhilin," Toward the next generation of recommender system: A survey of the state-of-the-art and possible extensions". IEEE Trans, Knowledge Data Eng., 17: 734-749, 2005.
- [2] pertz shoval, Veronica Maidel, Brancha shapira, "international journal of information theories and applications", Vol.15,pp. 303-314,2008.
- [3] Ujjin, S. and P.J. Bentley. "Particle swarm optimization recommender system". Proceedings of the 2003 IEEE Swarm Intelligence Sysposium-SIS '03, April 24-26, London, UK. PP: 124-131, 2003.
- [4] Mohammad Yahya H., Al-Shamiri, Kamal K., Bharadwaj, "Fuzzy Genetic Approach to Recommender Systems based on a Novel Hybrid User Model",Expert systems with applications,Elsevier,pp.1386-1399, 2007.
- [5] Sobecki, J., "Web-Based System User Interface Hybrid Recommendation Using Ant Colony Metaphor. In: Knowledge-Based Intelligent Information and Engineering Systems", Apolloni, B. et al. (Eds). LNCS 4694, Springer-Verlag, Berlin, Heidelberg, ISBN: 978-3-540-74828-1, pp: 1033-1040, 2008.
- [6] Vathatis,M.N., B. Boutsinas, P. Alevizos and G. Pavlides, "The new K-windows algorithm for improving the K-means clustering algorithm". J.Complexity, 18: 375-391, 2002.
- [7] Kanade P. M. and L. O. Hall, "Fuzzy ants as a clustering concept". North American Fuzzy Information Processing Society, NAFIPS 2003,22nd International Conference of the, pp. 227–232, 2003.
- [8] Castellano, G. and Fanelli, A. M. and Mencar, C. and Alessandra Torsello, M., "Similarity-based Fuzzy clustering for user profiling", IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, pp. 75-78, 2007.
- [9] Gadi, T., Daoudi, R. B. M., & Matusiak, S., "Fuzzy similarity measure for shape retrieval", In Vision Interface '99, Trois-Rivie`res,Canada (pp. 19–21), 1999.
- [10] Koczy T. Laszlo, Tikk Domonkos: Fuzzy rendszerek, Typotex, 2000.
- [11] Bedi P., Sharma R. and kaur H. , "Recommender System based on collaborative behavior of ants", Journal of artificial intelligence,ISSN 1994-5450, PP. 40-55.

Shiva Nadi received her BS and MS degrees in computer engineering from Islamic Azad University of Najafabad, Iran, in 2010. Her research interests are in several areas of artificial intelligence, data mining, web mining, evolutionary algorithms and software engineering.

Mohamad H Sarae received his PhD from University of Manchester in Computation, MSc from University of Wyoming, USA in Software Engineering and BSc from Shahid Beheshti University, Iran. His main areas of research are intelligent databases, Mining advanced and complex data including medical and Bio, Text Mining and E-Commerce. He has published extensively in each of these areas and served on scientific and organizing committee on number of journals and conferences.

Mohamad Davarpanah Jazi received his PhD from University of Manchester in Information Systems, MSc from University of Caltech, USA in Software Engineering and BSc from Shahid Beheshti University, Iran. He is now a member of computer department of Foulad Institute of Technology. His main areas of

research are Information Systems, DataBases, Programming Languages and Operating Systems.

Ayoub Bagheri received his BS degree in computer engineering from Ferdowsi University of Mashhad, Iran, in 2007, and the MS degree in computer engineering from Isfahan University of Technology, Iran, in 2009. Now he is PhD student in computer engineering in Isfahan University of Technology. His research interests are in several areas of artificial intelligence, data mining, machine learning and evolutionary algorithms.

Fusion Of Facial Parts And Lip For Recognition Using Modular Neural Network

Anupam Tarsauliya¹, Shoureya Kant², Saurabh Tripathi³, Ritu Tiwari⁴ and Anupam Shukla⁵

¹Department of Information Technology, IIITM
Gwalior: 474010, India

²Department of Information Technology, IIITM
Gwalior: 474010, India

³Department of Information Technology, IIITM
Gwalior: 474010, India

⁴Department of Information Technology, IIITM
Gwalior: 474010, India

⁵Department of Information Technology, IIITM
Gwalior: 474010, India

Abstract

Face and Lip recognition has been benchmark problems in the field of biometrics and image processing. Various Artificial neural networks have been used for recognition purpose. This paper attempts at improving the recognition result for individuals by fusion of facial parts and lip using modular artificial neural network which employs parallel local experts with combinatory recognition techniques. Principal Component Analysis (PCA) and Regularized-Linear Discriminant Analysis (R-LDA) algorithm are used to extract low dimensional feature vector of images to drive neural networks effectively. Backpropagation Neural Network (BPNN) and Radial Basis Function Neural Network (RBFNN) are used as training algorithm for the database. Grimace database is used in this paper for carrying out the proposed methodology. Each facial image is divided into three sub image and a lip image. The three facial parts and the lip part are trained and tested individually. The fusion technique is applied using modular neural network by grouping sub images and the lip image in different network modules. Separate results obtained from each module are integrated to get the final result from the methodology used. This result set is compared with the result set obtained by training the sub images and the lip image individually. From the empirical and results finding it can be seen that the proposed methodology performs out better result.

Keywords: Face, Lip Recognition, PCA, R-LDA, BPANN, RBFNN, Modular

1. Introduction

Biometrics parameters like face, lips, iris and other various parameters are being used for individual identification and verification, recent works in field has enabled similar recognition for various states or expressions automatically. A major technical

advancement in past few years has propelled face recognition technology into spotlight [1]. Face recognition is natural and passive and hence has clear advantage over other biometrics like fingerprints, iris etc. One of the major challenges involved in face recognition technique is the handling of various postures. Appearance based approaches for face recognition are being successfully developed and tested now a day [2]. These approaches are based on pixel intensity or intensity derived features. The performance of these methods depends upon types of training phase data, variations in pose, lighting and expressions [3].

Lip is one of the most important benchmark and advance parameter used for individual recognition of its varying postures [4]. Both Face and Lip can be represented as an image of size p x q pixels and further can be represented by a vector in p,q dimension space. In practical applications face and lip as parameters for identifying individuals outperform other biometrics [5]. Need for a automatic lip reading system is in spotlight because of its complication that stands for various moods and postures for facial expression.

Using Lip as modality for human identification has many advantages such as Lips biometrics is passive biometric i.e. individual interaction is needed. Images may be acquired from the distance without the knowledge of examined person. Lips biometrics is anatomical i.e. better results are expected than in behavioral biometrics. They are usually visible, may be implemented in hybrid face or voice recognition system [6].

Often a single biometric feature fails to provide sufficient evidences for verifying the identity of an individual. By fusion of multiple modalities pertaining to the field of biometric, the performance reliability of identification system can be improved. Due to its promising application, the fusion of multimodal biometric aspect is drawing more attention in recent years [7]. Face and lip multimodal biometrics are advantageous due to the use of non encroaching low cost image acquisition. Fusion of multimodal systems makes it difficult for intruders to trespass multiple biometric traits simultaneously.

Usually facial and lip images have their own representation of basis vectors of a high dimensional face vector space [8]. This large dimension is reduced for projecting the face vector to the basis vector by using various techniques such as Principal Component Analysis and Regularized- Linear Discriminant Analysis by approximating the original data with lower dimensional feature vector. PCA provides an effective technique for dimensionality reduction which involves computation of eigen values and eigen vectors by using the covariance matrix of original input data vector [9]. The orthonormal vectors are computed from it which are the basis of the computed data. R-LDA reduces high variance of the eigen value estimates of the within class scatter matrix at expense of potentially increased bias [10].

In this paper, we attempt at presenting a novel fusion strategy for personal identification using facial parts and lip biometrics. The proposed paper shows that integration of face parts and lip biometrics can outperform single biometric indicators. We present a architecture based on modular neural network for integrating facial sub parts and lip based on PCA and R-LDA.

2. Methodology

We performed face recognition task on Grimace database (shown in Fig. 1), which contains 360 colored face images of 18 individuals forming 18 classes while there are 20 images present for each subject. Database images vary in expression & position. Figure 1 shows examples from Grimace database.



Fig 1. Grimace database of 18 individuals

We divide each individual image of 180 x 200 pixels into three subimages and a lip image. The three subimages which are left half, right half and lower half have dimension of 90 x 130 pixels, 90 x 130 pixels and 180 x 60 respectively. Lip image which corresponds to the lip section of image face has dimension of 70 x 32 pixels. Then the colored images are converted into gray scale & processed using histogram equalization. Figure 2 below shows the partitioning of the facial image.



Fig 2. Sub Images and Lip Image used

We then process the image database using Principal Component Analysis and Regularized-Linear Discriminant Analysis for normalization and dimensionality reduction i.e. we extract low dimensional feature vector of the images.

After the pre-processing stage Modular Neural Network (MNN) is used for the classification purpose. Backpropagation and radial basis algorithms are used for training the neural network on training set obtained from processed image database. 70% of each sub image including the lip image is used for training the ANN and 30% are used for testing purpose. Figure 3 below shows the training and testing methodology.

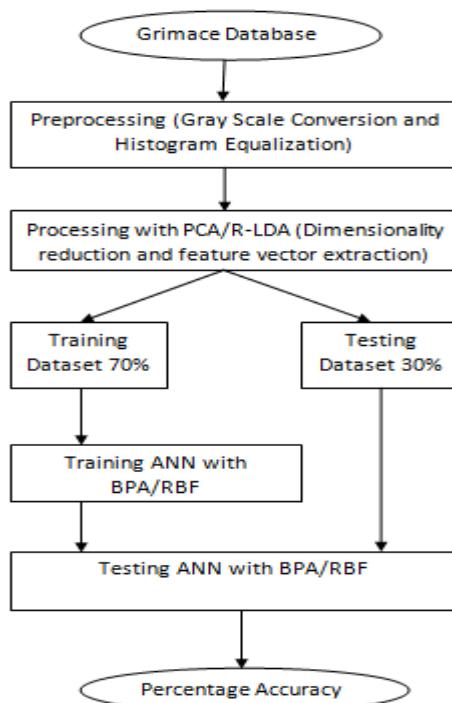


Fig 3. Flow Diagram for Methodology

The architecture consists of 3 sub-images and a lip image forming the reason to create 4 major sub-image modules. Each of these modules consists of 2 network modules, which are made to implement different combinations of principal component analysis, regularized linear discriminant analysis along with back propagation neural network and radial basis function neural networks. After learning the type 1 network module of each sub-images are integrated using in dedicated integration unit and type 2 network modules are integrated in integration unit dedicated for them. Both of these units are further integrated in module integration unit. The architecture is shown in figure 4.

The facial input image is sub divided into three sub facial image modules represented as Sub1, Sub2, Sub3 and a lip module as Lip respectively. Further each of these modules is divided into two separate network modules as N1 and N2 respectively. Sub1, Sub2, Sub3 are the upper left, upper right and lower half parts of the facial image respectively. N1 and N2 module of Sub1 is processed and trained using PCA with BPNN and R-LDA with RBFNN. N1 and N2 module of Sub2 is processed and trained using PCA with RBFNN and R-LDA with BPNN. N1 and N2 module of Sub3 is processed and trained using R-LDA with RBFNN and PCA with BPNN. N1 and N2 module of Lip is processed and trained using R-LDA with BPNN and PCA with RBFNN respectively. Network Module 1 – N1 of each of the sub image module and the lip module is integrated using various integration techniques stated below to form integrated result module M1. Similarly, Network Module 2 – N2 of each of the sub image module and

the lip module is integrated using various integration techniques stated below to form integrated result module M2. M1 and M2 thus obtained are integrated using the same integration techniques to yield the final result.

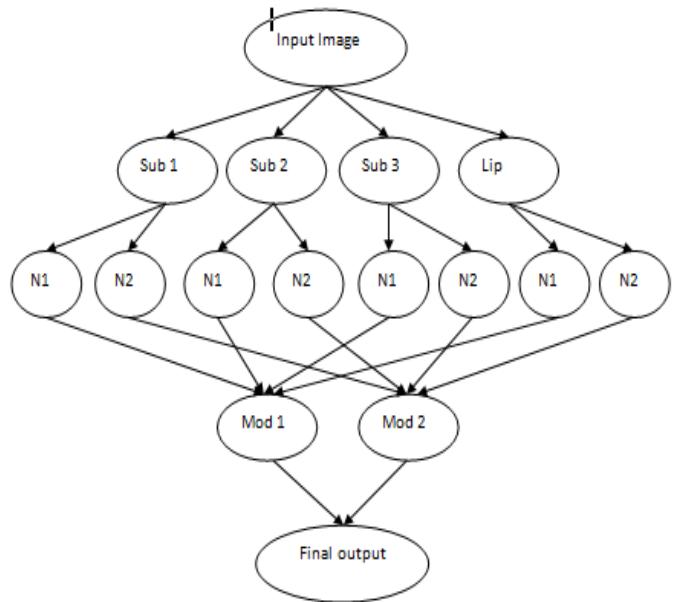


Fig 4. Used Modular Neural Network Architecture

Following strategies are used for integration: Probabilistic Sum Integration, Product Integration, Max integration, Min Integration and Polling Integration.

2.1 Back Propagation Algorithm

Backpropagation is the generalization of the Widrow-Hoff learning rule to multiple-layer networks and nonlinear differentiable transfer functions. Input vectors and the corresponding target vectors are used to train a network until it can approximate a function, associate input vectors with specific output vectors, or classify input vectors in an appropriate way as defined by you [11]. Networks with biases, a sigmoid layer, and a linear output layer are capable of approximating any function with a finite number of discontinuities.

Standard backpropagation is a gradient descent algorithm, as is the Widrow-Hoff learning rule, in which the network weights are moved along the negative of the gradient of the performance function. The term backpropagation refers to the manner in which the gradient is computed for nonlinear multilayer networks. There are a number of variations on the basic algorithm that are based on other standard optimization techniques, such as conjugate gradient and Newton methods [12].

2.2 Radial Basis Network

Radial basis function neural networks (RBFNNs) share features of the back propagation neural networks (BPNNs) for

pattern recognition. They are being extensively used for on- and off-linear adaptive modeling and control applications. RBFNNs store information locally whereas the conventional BPNNs store the information globally.

RBF nets belong to the group of kernel function nets that utilize simple kernel functions, distributed in different neighborhoods of the input space, whose responses are essentially local in nature. The architecture consists of one hidden and one output layer. This shallow architecture has great advantage in terms of computing speed compared to multiple hidden layer nets.

The function newrb iteratively creates a radial basis network one neuron at a time. Neurons are added to the network until the sum-squared error falls beneath an error goal or a maximum number of neurons has been reached. The function newrb takes matrices of input and target vectors P and T, and design parameters goal and spread, and returns the desired network [13].

2.3 Principal Component Analysis

Principal component analysis (PCA) is a classical statistical method. It is a variable reduction procedure and it is appropriate to use, when obtained a number of observed variable s and wish to develop a smaller number of artificial variable. Principal component analysis (PCA) is a mathematical method that transforms a number of correlated variables into a number of uncorrelated variables called principal components. Hence it is used to approximate the original data with lower dimensional feature vectors.

Principal Component analysis is one of the most successful techniques used in image recognition. The principal component can be used for prediction, redundancy removal, feature extraction, data compression [14].

Let $X = (x^1, x^2, \dots, x^i, \dots, x^n)$ represents the $n \times N$ data matrix, where each x_i is a lip vector of dimension n , concatenated from a $p \times q$ lip image.

$$Y = W^T q \quad (1)$$

Where Y is the $m \times N$ feature vector matrix, m is the dimension of the feature vector and transformation matrix W is an $n \times m$ transformation matrix whose columns are the eigenvectors corresponding to the m largest eigen values.

The principle component w_1 of the data set X is as follows:

$$W^i = \arg \max_{\|w\|=1} \text{var}\{W^T x\} \quad (2)$$

2.4 Regularized - Linear Discriminant Analysis

The R-LDA method presented here is based on a novel regularized Fisher's discriminant criterion, which is particularly robust against the SSS problem compared to the traditional one used in LDA. The purpose of regularization is to reduce the high variance related to the eigen value estimates of the within-class scatter matrix at the expense of potentially increased bias. The trade-off between the variance and the bias, depending on the severity of the SSS problem, is controlled by the strength of regularization [15].

Unlike the PCA method that extracts features to best represent face images; R-LDA method tries to find the sub space that best discriminates different face classes. By applying this method one can find the projection direction that on one hand maximizes the distance between the face images of different classes. On the other hand minimizes the distance between the face images of the same class.

2.5 Modular Neural Network

A modular architecture allows decomposition and assignment of task to several modules. Therefore, separate architecture can be developed to each solve a sub-task with the best possible architecture and individual modules or building blocks may be combined to form a comprehensive system. The module decomposes the problem into two or more subsystem that operates on inputs without communicating to each other. The input units are mediated by an integrating unit that is not permitted to feed information back to modules. The modular architecture combines two learning schemes supervised and competitive [16] . The supervised learning scheme is used to train the different module of networks and a gating netwok operates in a competitive mode to assign different patterns of the task to a module through a mechanism that acts as a mediator. Figure 5 given below illustrates the architecture of a typical MNN.

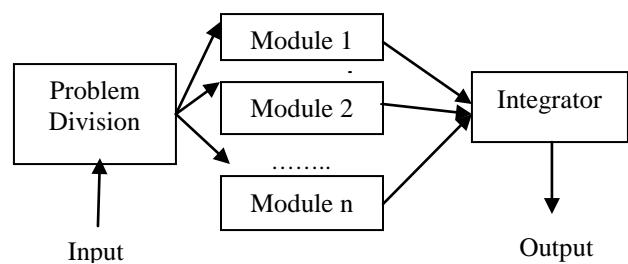


Fig 5. General Modular Neural Network Architecture

The integration of the result from various modules is performed using Probabilistic Sum Integration, Product Integration, Max integration, Min Integration and Polling Integration techniques. On the basis of the outputs that the various modules or ANNs generate, the decision regarding the final

system output is made. Integration is the mechanism to combine the various outputs into a single output of the system. The integration again depends upon the design of the MNN and the manner in which the division of the problem has taken place.

The mechanism to compute the final output by these outputs is done by the integrator using a variety of methods.

Sum rule is applicable when a high level of noise and/or high ambiguity in the classification problem cause the posterior estimated by a classifier not to deviate much from the prior. Max rule approximates the mean by the maximum of the posteriors. Min rule approximates the mean by the minimum of the posteriors [17].

In polling [18], each of the modules gives as its output the class to which the input may belong as per its knowledge and methodology. The integrator gets the various classes that are potentially the output of the system by the various modules. The task of the integrator is to decide the system output as a means of voting between the various modules. Each module casts one vote in favor of the class which is its output. The votes for the various classes are collected and the class getting the largest vote is regarded as the final system output by the integrator. In case of a tie between the modules, any one of the classes may be randomly chosen out of the classes involved in a tie.

In probabilistic sum, every module gives as many outputs as there are classes in the system. Each output measures the probability of the occurrence of any class as the final output class. The probability lies in between 0 to 1. An output of 1 by any module for any class means that as per the recordings of the module that particular class is surely the class to which the input perfectly maps to. An output of 0 by any module for any class means that as per the recordings of the module that particular class is surely not the class to which the input perfectly maps to. Every module hence computes the probability for each of the classes in the system. This probability vector comprising of all the probabilities is passed to the integrator for computing the final output.

3. Results

Table 1 and Table 2 below depict the network parameters for BPNN and RBFNN respectively.

Table 1. Network Parameters for BPNN

| Learning Rate | Momentum | Epochs | Goal |
|---------------|----------|--------|--------|
| 0.001 | 0.5 | 10,000 | 1.0e-3 |

Table 2. Network Parameters for RBFNN

| Spread | Epochs | Goal |
|--------|--------|--------|
| 0.4 | 10,000 | 1.0e-3 |

Subimage 1 which is upper left of the face having dimension of 90 x 130 pixels has been taken as separate module and is

further divided into two network modules as PCA with BPNN and PCA with RBFNN respectively. The individual network module are trained and tested with their corresponding ANN's respectively. The result thus obtained is shown in table1. It can be seen from the table that module trained with RBFNN gives better result than module trained with BPNN.

Table 3. Individual results for Sub Image 1

| SubImage 1 Network Module | Feature Extraction Classification Techniques | Matching Score |
|------------------------------|--|----------------|
| Network Module 1 | PCA and BPNN | 159/180 |
| Network Module 2 | PCA and RBFNN | 163/180 |

Subimage 2 which is upper right of the face having dimension of 90 x 130 pixels has been taken as separate module and is further divided into two network modules as R-LDA with BPNN and R-LDA with RBFNN respectively. The individual network module are trained and tested with their corresponding ANN's respectively. The result thus obtained is shown in table2. It can be seen from the table that module trained with BPNN gives better result than module trained with RBFNN.

Table 4. Individual results for Sub Image 2

| SubImage 2 Network Module | Feature Extraction Classification Techniques | Matching Score |
|------------------------------|--|----------------|
| Network Module 1 | R-LDA and RBFNN | 160/180 |
| Network Module 2 | R-LDA and BPNN | 163/180 |

Subimage 3 which is upper left of the face having dimension of 180 x 60 pixels has been taken as separate module and is further divided into two network modules as R-LDA with RBFNN and PCA with BPANN respectively. The individual network module are trained and tested with their corresponding ANN's respectively. The result thus obtained is shown in table3. It can be seen from the table that module trained with RBFNN gives better result than module trained with BPNN.

Table 5. Table Individual results for Sub Image 3

| SubImage 3 Network Module | Feature Extraction Classification Techniques | Matching Score |
|------------------------------|--|----------------|
| Network Module 1 | R-LDA and RBFNN | 163/180 |
| Network Module 2 | PCA and BPNN | 157/180 |

Lip Image which is lip part of the face having dimension of 70 x 32 pixels has been taken as separate module and is further divided into two network modules as R-LDA with BPNN and PCA with RBFNN respectively. The individual network module are trained and tested with their corresponding ANN's respectively. The result thus obtained is shown in table4. It can be seen from the table that module

trained with RBFNN gives better result than module trained with BPNN.

Table 6. Individual results for Lip Image

| Lip Image Network Module | Feature Extraction Classification Techniques | Matching Score |
|--------------------------|--|----------------|
| Network Module 1 | R-LDA and BPNN | 161/180 |
| Network Module 2 | PCA and RBFNN | 165/180 |

Table 5 below shows the integrated results obtained from five various integrated techniques namely Sum, Max, Min, Product and Polling for network module 1 and network module 2 of sub images (1 – 3) and the lip image modules.

Table 7. Integrated results for network modules

| Integration Technique | Network Module 1 | Network Module 2 |
|-------------------------------|------------------|------------------|
| Probabilistic Sum Integration | 164/180 | 166/180 |
| Max Integration | 163/180 | 165/180 |
| Min Integration | 162/180 | 166/180 |
| Product Integration | 167/180 | 165/180 |
| Polling Integration | 169/180 | 170/180 |

Table 6 below shows the final integrated result obtained from module 1 and module 2 using various integrated techniques namely Sum, Max, Min, Product and Polling for network module 1 and network module 2 of subimages (1 – 3) and the lip image modules.

Table 8. Final result obtained

| Integration Technique | Final Network Module Output |
|-------------------------------|-----------------------------|
| Probabilistic Sum Integration | 165/180 |
| Max Integration | 164/180 |
| Min Integration | 168/180 |
| Product Integration | 167/180 |
| Polling Integration | 172/180 |

From the table obtained we can see that the most optimized result is obtained from polling integration technique with recognition of 95.56%.

5. Conclusion

This paper attempts at bringing forth the advantage of fusion of multiple biometrics modalities over single biometric feature identification mechanism. The facial parts along with the lip image are fused together using modular neural network architecture and this system is then analyzed and compared with individual performance. The preprocessing stage involves the image preprocessing of gray scale conversion and histogram equalization, processing with PCA/R-LDA for dimensionality reduction and feature vector extraction. The preprocessing stage is followed by the training the

classification mechanism using BPA/RBF algorithms. From the table 1-4, it can be seen that training of module with RBFNN gives better result for three of the image module which is also the case for the lip image. It is also noticed that lip image module gave better matching score than other three sub images taken for the experiment. Different integration techniques are used for integrating the results of the modules, integrated results are found to better than the individual scoring results obtained of a module; it can be seen from table 5 and table 6 that polling integration technique gives the best result in comparison to rest of the techniques used for the integration purpose. The best matching score obtained from the three facial sub image comes out to be 90.55% with RBFNN as training algorithm. Lip image shows the matching score of 91.67% using RBFNN as training algorithm. Results obtained from the proposed fusion techniques stands better than individual matching scores at 95.56% with polling as the integration technique followed by min integration at 93.34%.

6. Future scope

Further research work can be done in this field, by fusion of other biometric features. Neural network parameter can be optimized using evolutionary algorithms for getting better results. Other training algorithm in place of BPA and RBF may give better performance than used algorithms. Modular architecture i.e. the number of modules used for an image can be varied to get better result. Combinations of other different processing, training and integration technique may be used.

7. Acknowledgement

We greatly acknowledge and admire the co-operation extended to us by Mr. Rahul Kala, Ph.d. Scholar, School of Cybernetics, University Of Reading. His sincere efforts and guidance paved the way towards the completion of this paper.

8. References

- [1] Szewczyk, R.; "New Features Extraction Method for People Recognition on the Basis of the Iris Pattern," Mixed Design of Integrated Circuits and Systems, 2007. MIXDES'07. 14th International Conference on , vol., no., pp.645-650, 21-23 June 2007
- [2] Kresimir Delac, Mislav Grgic and Sonja Grgic, "Generalization Abilities of Appearance-Based Subspace Face Recognition Algorithms", 12th Int. Workshop on Systems, Signals & Image Processing, 22-24 September 2005.
- [3] Sheetal Chaudhary, Rajender Nath, "Hybrid Approach for Template Protection in Face Recognition System", Global Journal of Computer Science and Technology, Page 34 Vol. 10 Issue 5 Ver. 1.0 July 2010
- [4] Mok, L.L.; Lau, W.H.; Leung, S.H.; Wang, S.L.; Yan, H.; , "Person authentication using ASM based lip shape and intensity information," Image Processing, 2004. ICIP '04. 2004 International Conference on , vol.1, no., pp. 561- 564 Vol. 1, 24-27 Oct. 2004

- [5] Lievin, M.; Luthon, F.; , "A hierarchical segmentation algorithm for face analysis. Application to lipreading," Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on , vol.2, no., pp.1085-1088 vol.2, 2000
- [6] Michal Choras, "The lip as a biometric", Theoretical Advances, Pattern Anal Applic (2010) 13:105–112 8 January 2009
- [7] Nageshkumar.M, Mahesh.PK and M.N. Shanmukha Swamy," An Efficient Secure Multimodal Biometric Fusion Using Palmprint and Face Image," IJCSI International Journal of Computer Science Issues, Vol. 2, 2009
- [8] Nor'aini , A.J.; Raveendran, P.; Selvanathan, N. "A comparative analysis of feature extraction methods for face recognition system" Sensors and the International Conference on new Techniques in Pharmaceutical and Biomedical Research, 2005 Asian Conference on 5-7 Sept. 2005.
- [9] M.H. Tan, J.K. Hammond." A non-parametric approach for linear system identification using principal component analysis", Mechanical Systems and Signal Processing 21 (2007) 1576–1600, 7 September 2006
- [10] Juwei Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, "Face Recognition Using LDA Based Algorithms", IEEE Transactions on Neural Networks, Vol. 14, No.1, Page: 195-200, January 2003.
- [11] K.Rama Linga Reddy, G.R Babu, Lal Kishore, M.Maanasa," Multiscale Feature And Single Neural Network Based Face Recognition", Journal of Theoretical and Applied Information Technology,2005
- [12] Saeid Iranmanesh, M. Amin Mahdavi," A Differential Adaptive Learning Rate Method for Back-Propagation Neural Networks", World Academy of Science, Engineering and Technology 50,2009.
- [13] Ch. Satyananda Reddy, P. Sankara Rao, KVSVN Raju, V. Valli Kumari," A New Approach For Estimating Software Effort Using RBFN Network", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.7, July 2008
- [14] Khoukhi, A.; Ahmed, S.F.; , "Fuzzy LDA for face recognition with GA based optimization," Fuzzy Information Processing Society (NAFIPS), 2010 Annual Meeting of the North American , vol., no., pp.1-6, 12-14 July 2010
- [15] Salehi, N.B.; Kasaei, S.; Alizadeh, S.; , "Face Recognition Using Boosted Regularized Linear Discriminant Analysis," Computer Modeling and Simulation, 2010. ICCMS '10. Second International Conference on , vol.2, no., pp.89-93, 22-24 Jan. 2010
- [16] Sarosh I. Khan , Stephen G. Ritchie," Statistical and neural classifiers to detect traffic operational problems on urban arterials", Transportation Research Part C 6 (1998) 291-314,1998
- [17] Shakhnarovich, G.; Darrell, T.; , "On probabilistic combination of face and gait cues for identification," Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on , vol., no., pp.169-174, 20-21 May 2002
- [18] Rahul Kala, Ritu Tiwari, Anupam Shukla, "Real Life Applications of Soft Computing", CRC Press, May 21, 2010

Health Care Implementation by Means of Smart Cards

Dr. Magdy E. Elhennawy
High Institute of Computers and Information Technology,
Computer Dept., El-Shorouk Academy,
Family Card Project Consultant,
Ministry of State for Administrative Development,
Cairo, Egypt,

Dr. M. Amer
Higher Institute for Specialized Technological Studies, Future Academy,
Information System Dept., El-Future Academy,
Cairo, Egypt,

A. Abdelhafeez
Ministry of State for Administrative Development, Project Manager,
Cairo, Egypt,

Abstract

Smart card technology is a reliable and proven solution that has had decades of use in various industries and is now making its mark on healthcare. It is portable, secure, and can hold health information. In this research, the smart card has been introduced in a new usage; it is used to control the referral process. It holds the control parameters that link between the referring entity and the referred entities. This leads to control the management of the claims in a timely and correct basis. The smart cards technology, in this research, have been used for: verifying citizen eligibility, registering recent citizen medical visits to family medical units as well as the corresponding visits to contracted providers besides the link between them, reporting, and facilitating claim management. Accordingly, the usage of smart cards in this way will add. The research allows remarkable improvements in the healthcare service provision.

Keywords: Smart card technology, Healthcare using smart cards, smart healthcare cards.

1. Introduction

Healthcare is seeing a steady and increasing dependence on information technology that is rapidly transforming the practice of medicine and the delivery of care. Technology is an ever-changing and evolving aspect of modern business. In healthcare, most agree that the use of information technology is essential to achieving many of the milestones critical needs of the healthcare reform. Three primary drivers are increasing the use of technology in healthcare, namely: 1) The need to lower costs and allows administrative efficiencies, 2) The need to improve patient outcomes and enhance physician and patient relations, and 3) The need to meet increasing privacy, security and identity concerns.

Smart card technology can provide innovative, practical and cost-effective solutions in healthcare domain. In this research, we will introduce the smart card technology as

a tool to link between cycles of health care, including the on-line delivery of claims for appropriate and timely claims management.

This research introduces the subject and objective of the research, this section. In Section 2, smart card technology standards are introduced. In section 3, various smart cards applications are surveyed, in Section 4, how smart cards can improve healthcare is discussed, in section 5, the research contribution for how we can adapt smart cards to facilitate claims management is introduced, in section 6, we analyze our proposal and a sample of assumed results are presented and commented, and finally, in section 7 the conclusions and future work are stated.

2. Smart Card Technology Standards

Smart cards first emerged onto the worldwide stage in 1974, when Frenchman Roland Moreno patented a smart card to use as a payment source for a telephone call. Since then, smart cards have found their way into virtually every industry, including healthcare [1].

Smart cards are essentially miniature computers without display screens or keyboards containing an embedded integrated circuit (or chip) that can be either a secure microcontroller with internal memory or simply a memory chip. Usually, microcontrollers contain a microprocessor and different kinds of memories: RAM, ROM, and EEPROM. The chip is a powerful minicomputer that can be programmed in different ways. It contains some sensors (like light sensors, heat sensors, voltage sensors, etc.), which are used to deactivate the card when it is somehow physically attacked [2]. Some smart cards hold various types of information in

electronic form and protect the information with sophisticated security mechanisms; others provide a key that unlocks a particular database on a particular server. Because of their portability, inhibited security features and size, smart cards provide an ideal solution for secure data exchange.

Smart cards can perform many functions, such as storing data, making calculations, processing data, managing files, and executing encryption algorithms. Smart cards make possible sophisticated and portable data processing applications and are far more secure and reliable than passwords or magnetic stripe ID cards [3][4].

Standards help ensure smart cards can be read by any retailer equipped with a smart card reader [1]. Smart objects, in general, need to fulfill a set of requirements such as controllability, maintainability, scalability, interoperability, security, and reliability [5]. Smart card technology conforms to international standards ISO/IEC 7816 and ISO/IEC 14443.

3. Smart cards Applications

Because of their size, flexible form factors and relatively low cost, smart cards are ideal for applications in markets where personal identity, privacy, security, convenience, and mobility are key factors [6].

In 2008, 5.045 billion smart cards were shipped worldwide—an impressive 13.2% increase over the 2007 figure of 4.455 billion—with about 15% of these cards entering the U.S. market [7]. Smart cards are deployed all over the world for personal identification, playing a critical role in logical and physical access control systems. Across the globe, public and private sector organizations are recognizing the security and efficiency of using smart card in identification applications. Smart cards are used in a wide variety of payment applications, and extensively in the telecommunications industry all over the world.

One of the today's applications is the health care. Within the U.S. healthcare industry, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) is driving the use of smart cards for both patients and providers to improve the security of healthcare IT systems and protect the privacy of patient information [8]. Healthcare organizations worldwide are implementing smart healthcare cards to support a variety of features and applications.

4. How Smart Cards Can Improve Healthcare

Over the past few years, smart card use in the U.S. healthcare sector has grown significantly and recently has used smart cards to upgrade level of significance. Current programs focus on patient identification: streamlining admissions, managing payments, and

moving patient data from point to point. Four needs have driven smart card use in health care to date. They are identification and patient authentication, matching patients to their particular data, synchronizing data from disparate sources, and security and access control.

Numerous benefits devolve to different healthcare stakeholders from using smart cards. Employing smart card can help in healthcare cost reduction, help in improving stockholders revenue, control fraud and misuse, and other features.

4.1 Cost Reductions

A major advantage of using smart cards in healthcare is the reduction in costs that results from improving the efficiency of handling medical and administrative information, which, on the other hands, increases the quality of service.

Smart cards can be integrated with healthcare systems in such a way that it can reduce cost such as: reduced administrative time and cost by automating patient identification, reduced duplication of records, fewer errors and adverse events through the use of accurate and timely information, reduced number of rejected claims and faster payments, by using accurate patient, reduced claims processing costs through real-time adjudication of claims and insurance coverage verification information.

4.2 User Identification, Authentication, and Authorization

Identification, authentication, and authorization are the pillars of security in the electronic world. Recently, based on various techniques, many password authentication schemes using smart cards have been proposed by some researchers. These schemes can allow a legal user to login to remote server and access its facilities [9]. Accordingly, entity authentication is one of the most important security services that can be applied by smart cards. It is necessary to verify the identities of the communicating parties when they start a connection [10].

In addition to the financial loss incurred by healthcare fraud, fraud poses tangible health risks for patients whose records are compromised. With the creation of large clinical data exchanges and the ready availability of information on the Internet, all system users need to be accurately identified, and properly authenticated before being allowed to access information. And finally, all individuals must have the appropriate authorization to access medical data and initiate particular transactions.

Smart cards can provide positive identification of the patient at the registration desk, by allowing personnel to be verified, however, facilitate rapid identification of a patient arriving at an emergency room and rapid retrieval

of lifesaving information about medical history, recent tests, treatments, and medications. This critical information can be stored on the smart card chip or the smart card can provide secure access to data stored elsewhere.

Using a smart card to verify patient identity can offer healthcare providers the following benefits: 1) Make it easy to link patients to the correct medical records, 2) Reduce the creation of duplicate records, 3) Reduce the potential for medical identity theft and fraud, 4) Improve the efficiency of the registration process and the accuracy of data, and 5) Improve the revenue cycle and reduce the number of denied claims.

Studies have found that on average, 5%–15% of a hospital's medical records are duplicated or overlaid [11]. The more duplicates there are in a system, the higher the rate of new duplicates. The growth rate becomes exponential with the size of the patient database [12].

So, by implementing smart card technology as part of the admission and registration process, an institution can reliably identify its patients, increase the accuracy of data capture, optimize patient throughput, accurately link patients to their medical records, verify eligibility, and ultimately improve patient experience and satisfaction.

4.3 Claims Denial and Revenue Capture

Two of the most common reasons for claims denials are incomplete demographic information and incomplete insurance information, which can cost a healthcare institution millions of dollars in lost or delayed revenue. Most healthcare CFOs are acutely aware of the high cost of reviewing and resubmitting old claims and the revenue lost because of cumbersome claims processing, including detailed chart reviews and outreach to patients and physicians for additional information.

The healthcare revenue cycle is highly dependent on the front-end registration process, which drives much of the downstream claims process. Studies estimate that 50%–90% of claim denials could be prevented by securing accurate patient information at the front desk [13][14]. According to a study by PNC Financial Services, one out of five claims submitted is delayed or denied by insurers. Smart card technology can greatly improve the accuracy of routine data capture. Instead of transcribing information from paper forms and increasing the risk of human error, smart cards can access or provide insurance information, demographics and other patient information, reducing claim denials and increasing cash flow.

4.4 Immediate Access to Lifesaving Information

Everyone in the continuum of healthcare, from ambulance crews to emergency room personnel to physicians and nurses, needs immediate access to accurate medical information such as a patient's medical history, allergies, prescriptions, and over-the-counter drugs. According to a recent study conducted by the Boston Consulting Group, as much as 40% of patient information is missing when needed by a medical professional for proper care [15]. A report published in the *Journal of the American Medical Association* found that adverse drug interactions and medical errors result in an estimated 225,000 deaths per year [16].

Smart cards carried by patients allow immediate access to vital information and information from other points of care that otherwise might not be available. Even when hospital records are not available, information stored on a smart card or accessed from the smart card with a portable reader provides an easy way to triage patients in emergency and disaster situations. Such information can be accessed from an ambulance en route to a hospital or in the field as part of disaster response. Medical information stored on a smart card can be accessed even when computer networks and power lines are inoperable [17].

4.5 Healthcare Fraud, Abuse, and Misuse

The National Health Care Anti-Fraud Association (NHCAA) estimates that 3% of USA annual healthcare pending (\$68 billion in 2007) is lost due to healthcare fraud [18]. Other estimates by Government and law enforcement agencies place the loss as high as 10% of USA annual expenditure, or \$200 billion, and growing [19].

The impact of healthcare fraud and abuse reaches far beyond cost; quality of care is compromised by false or inflated claims. The health and well-being of a patient are jeopardized when the patient is exposed to unnecessary and dangerous tests and procedures. Some patients have become "paper pawns" when fabricated histories add erroneous information to their medical records. Fraud can also threaten patients' future insurability. Smart cards can be used to secure access to electronic medical records. Implementing strong authentication within a medical facility will not eliminate but will certainly reduce the risks that personal health information is compromised.

4.6 Support for a National Health Network

Federal- and private-sector initiatives have established a framework for the creation of the Nationwide Health Information Network (NHIN). The main goal of the NHIN is to develop a scalable and secure system for exchanging healthcare information on a national level in USA.

A highly reliable identity management infrastructure is critical to the success and viability of a national network. Smart card technology can play a critical role in this infrastructure. Smart cards can be used to positively identify patients at the point of care and securely track their access to care across multiple providers. The card can be used to aggregate all medical record numbers for a patient as the patient receives care. This can greatly facilitate linkages with local data exchanges and regional health information organizations (RHIOs). Using the smart card in this way greatly improves the fidelity of the linked medical records and reduces reliance on statistical methods for matching patients to medical records, which can propagate errors. Smart cards would also provide access control for those viewing the medical records on the network.

Other advantages to using smart cards as part of a national network are that fraud and abuse can be greatly curtailed, and medical identity theft would be more difficult if an identity credential were part of the process [20]. Additionally, the smart card can be used as a security token for patients to access their personal health records online and can promote greater patient involvement in health and care management.

5. Adapting Smart Cards to Our Solution

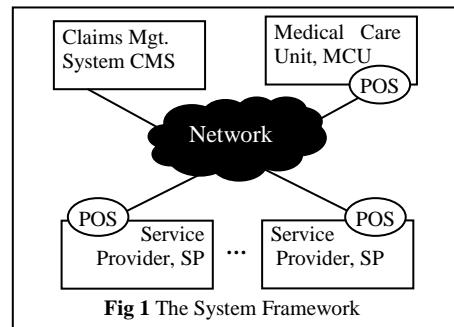
Ministry of State for Administrative Development, MSAD, has proposed, designed, and implemented a pilot for the achievement of the health care systems using smart cards, with the cooperation with Health Insurance Organization. The following is the approach of this pilot.

5.1 System Framework

Smart card has been used as an efficient tool to deliver services. In our research the use of smart cards has extended to cover the eligibility verification starting from check-in in the admission desk, the control of the referral process, as well as facilitating claims management. The idea of the research is based on using the smart cards to hold the control parameters that link between the referring entity and the referred entities, which leads to control and manage the claims management in a timely basis more correctly.

Our solution is based on the fact that the patient starts his medical course by visiting his Medical Care Unit, MCU. The population has been divided geographically into groups, each has assigned to a specific MCU. The MCU is typically a medical unit, prepared with medical specialties' doctors, labs, radiologists, and a pharmacy, each with limited capabilities. It has an automated Clinical Information System, CIS. In this unit, the patient would be investigated, and his prescription can be defined. If his status needs, the prescription may define other Service Provider, SP to be referred. SP may be other consultant doctor, test lab, radiologist, for further

investigation or pharmacy to receive medicine. However, patient should visit the referred SP and may come back to the MCU for finalization of his case. The patient medical course starts in MCU, then various SPs, and come back to MCU.



Each MCU is equipped with point of sale, POS and have a smart card. Each SP is equipped with POS and also has a smart card. The CMS, referred to sometimes as Payer, is equipped with a server with a proper specifications. The system is supported a medical database, called the Family Medical Database, FMDB. The eligibility status of the population is registered on FMDB, managed by CMS, and continually updated. The service providers, various system codes, any other needed data, are stored on FMDB.

Each eligible citizen have a smart card (basically, one for the family holder and may be for family individuals if requested). Eligible citizen smart card will be used to verify the citizen eligibility as well as to register the citizen medical visits, then to control eligibility of such visits with the CMS. Smart card, issued for eligible citizens, will contain, for healthcare management, the needed data items such as: citizen national number, NID, service eligibility status, related health insurance law, list of medical visits together with the associated data of the service provided, and claims data needed to CMS. Smart card, issued for MCUs and SPs, will be used to start his POS daily work, and do the needed system administrative work. Initially, the eligibility status will be stored on the FMDB, as well as patient's smart card. Any changes to this status will be sent from FMDB by means of CMS to related MCUs' and SP POSs and accordingly installed on the patient's card automatically whenever he is going to receive the service.

When the patient visits the MCU, his eligibility will be verified against the value stored on the smart card by comparing the MCU ID registered on the smart card with that registered on the POS of that MCU. In MCU, if the patient has been referred to other SPs, their IDs will be stored on the patient smart card. This will allow the SPs' system to verify the eligibility of this patient by

comparing the SPs IDs registered on the smart card with that registered on the POS of that SP.

In a daily basis, MCU¹ and SP² POSs will send the list of medical visits and their associated data to CMS to allow claims management. The CMS will communicate with FMDB for the updating of related data and the service provision status. **Figure 2** shows the proposed system interactions.

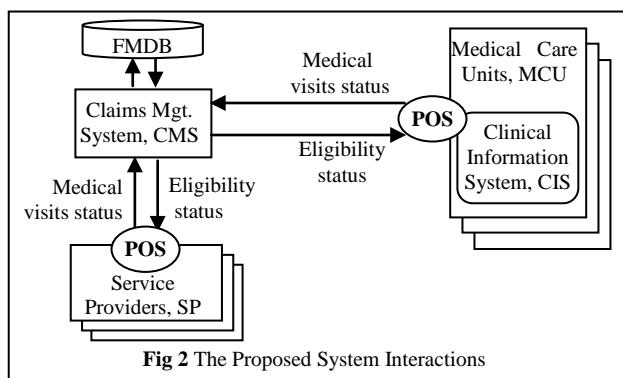


Fig 2 The Proposed System Interactions

5.2 System Preliminaries

Health care packages provided to patient are defined. Each package contains a set of elementary medical services. Each of them is distinctly coded and priced. Elementary medical services cover all medical possible provided cases. Sometimes, a patient may need medical services that are not defined inside his eligible package. SP might be obliged to provide such service, in some circumstances, to save patient life. The rules governing these exceptions will, also, be pre-defined properly.

All elementary medical services are priced, eventually, each total package is priced. However, the provided elementary services are identified and can be measurable. Accordingly, the medical services provided to a patient can be priced and measurable.

Health care packages and their internal elementary services provided to patients can be defined as follows:

Package #1: code: 000000

Elementary medical service #1: . code 000, price: 000,
 Elementary medical service #2... code 000, price: 000,
 ...
 Elementary medical service #n: .. code 000, price: 000,

Package #m: code: 000000

Elementary medical service #1: . code 000, price: 000,
 Elementary medical service #2... code 000, price: 000,
 ...
 Elementary medical service #v: .. code 000, price: 000,

Where n, .. v are the number of elementary services in individual packages, and m is the number of packages provided by the system. International codes can be used.

For other cases that are not covered by international codes, local codes can be designed and used. The elementary service can be identified by composite key constituting package code and service code.

When a patient receives his services, a transaction record is formed by the SP's electronic system, and sent electronically in a daily basis to the CMS. The transaction is composed of the following data items: date/time, SP's identification, and provided services data items. Provided services data items are of two categories, services inside the package and services outside the package code. Services outside the package will be provided to the patient by SP in emergency cases only and according the pre-defined rules. The classification into these two categories is done by the system according to eligibility status. When CMS receives a patient transaction from SP, the services inside eligible package can be differentiated from services outside the package. However, inside services can be calculated and outside services can be controlled and calculated.

The smart card data can be: Patient NID, Patient eligibility, IDs of allowed packages, Referred visits (Consultants, Test lab, Radiology, Physical treatment, and Pharmacy). Referred visits will contain its ID, status, and date/time, where: status means whether the visit and the service provision have been done or not yet, date/time is the date and time of performing the visit. Figure 3 the proposed system interactions

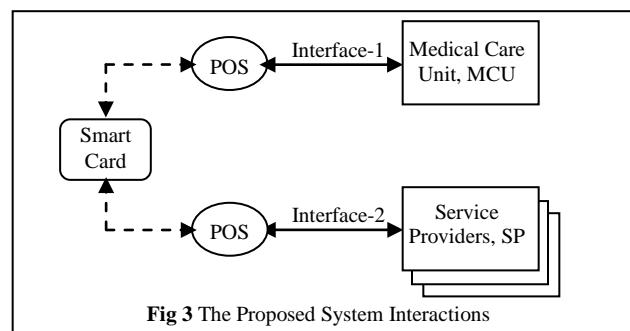


Fig 3 The Proposed System Interactions

Interface-1 is responsible for: verifying the patient eligibility in MCU, storing the type of referrals visits, and identifying and registering the referred entity. The identifying and registering the referred entity is done by the Case Management Office (CMO). Interface-2 is responsible for verifying the patient eligibility in SPs, and storing the status of referrals visits together with the date/time. Enough data about the referrals will be stored.

All referred visits have been stored in batches inside the service providers' POS's and batches are sent to CMS, electronically, in a daily basis.

5.3 The Proposed Blocks of the Healthcare

Accordingly, the proposed healthcare system comprises three basic components, namely: 1) Clinical Information System, CIS, and POS application in MCU 2) POS applications in Service Provider's premises, and 3) Claims Management system CMS. The three components are linked together through network. The smart card, as will be described below, is used to link patient's visits to various referrals with initial visit to family medical unit.

The CIS is an electronic application that can manage the registration of the patient, follow-up the patient medical visit in MCU. At the end of the day, the patient record may be of one of two cases: 1) Closed case: in which he receive medicine from MCU's pharmacy and close his case, or 2) Open case: in which types of other consultants for further medical investigation, labs, radiologists, or pharmacies to receive other medicines have been referred. The referrals will be registered on the patient smart card by means of the POS application.

Accordingly, patient will visit the referred SP. Eligibility will, first, be verified using patient smart card and his POS application, and if valid he will receive the appropriate service. The status of service provision will be stored on the card and sent to CMS, in a daily bases.

The claims management system basic role is to manage the periodical settlements among claims received from various system service providers. To be appropriate, it should receive accurate claims from various service providers. CMS receives from SPs and MCU, electronically, the claims data which are: Patient NID, Patient eligibility, IDs of allowed packages, Referred visits (Consultants, test lab, radiology, physical treatment, and pharmacy). Referred visits will contain its ID, status, and date/time, besides the SP IDs. In case provided service code indicates a medicine delivery from outside pharmacy, the medicine's codes are provided. Such information allows CMS to manage efficiently, in daily bases, the claims. The daily claims management allows accurate computation, minimize expected errors, prevent claims management delay, and allow timely stockholders revenue.

The use of smart cards will facilitate the roles of the above components, and makes them work properly. Smart cards will store appropriate data about the patient eligibility, case status, defined referrals, result of each referral visit, and last status after finally closing the case in the MCU. On the other hand, a periodical exchange of the referral claims, by means of distributed POSs, will be sent to CMS for appropriate settlements. The usage of the smart cards can, in our proposal, control the patient

eligibility, control the referral visits, and allow accurate data exchange by the cards.

5.4 Benefits of the proposed System

The combination between smart cards and information technology to manage the claims in healthcare, according to our research, leads to a set of benefits. First: since the eligibility and related package services are stored on the card and periodically/electronically updated; the correct verification of the patient service eligibility can be reached. We are sure the patient receives correct and eligible services. Second: since health care packages and their internal services are defined clearly, and coded correctly, the correct and precise pricing of individual health care services can be defined. This leads to the fact that the pricing of the individual claims are verifiable and can help the correct claim management. Third: since the exceptions that may happen during the service provision, and their application rules are clearly defined, when such exceptions happen during service provision adding additional medical services to the approved packages, it will be legal to add such exceptions. This will facilitate the claims management. Fourth: under the above circumstances, the electronic exchange of claims between service providers and claims management system will facilitates the correct, speedy, and verifiable claims settlement. This will compensate a loss and denial in claims for service providers about 20% as previously stated (refer to Section 4.3 out of five claims submitted is delayed or denied). Fifth: since health care services are timely registered in the system database, a transactional database will be built. This leads to the fact that all provided health care services and their related data (date, time, service provider ID, ...) will be available for further analysis, investigation, and inspection. A lot of conclusions can be reached and forecasting can be very useful benefits to system.

6. System Analysis and Simulation Results

A comparison between the manual and expected electronic systems has been assumed. A random number generator has been used to assume claims data received from system SPs to simulate the system operation. The applied time frame was 6 months from January to June, 2010. The assumed results have been generated under the following assumption: 1) the number of denied and delayed claims cannot exceed 20% of the total claims according the figures stated in Section 4.3, 2) the number of delayed claims will be greater than denied one by 3% as the delay will be more due to the manual intervention in the above system, 3) delayed claims in the above system will last not more than 3 months, 4) the delayed claims will be 100% for the first month, 30% for the second month, and 10% for the third month. **Annex 1** shows the assumed data together with the calculated results.

It is clear that the difference between the actual received claims differ from the valid claims because of the paper wise management of the claims, and the manual exchange of claims before applying the proposed system.

As our research will control the claims management processing, the stated above denied and delayed claims will be prevented. **Figure 3** shows the total number of received claims against total number of managed claims in previous system, noting that in our proposed system the managed claims will be the received claims. **Figure 4** shows the total number of received claims against total number cumulative number of denied in addition to delayed claims previous system.

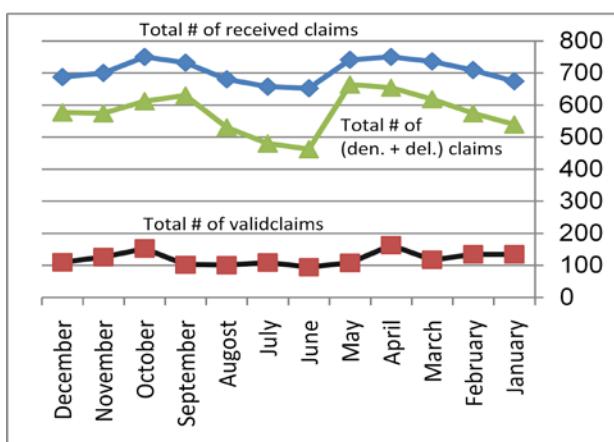


Fig 3 Total # of Received Versus Managed Claims in Previous System

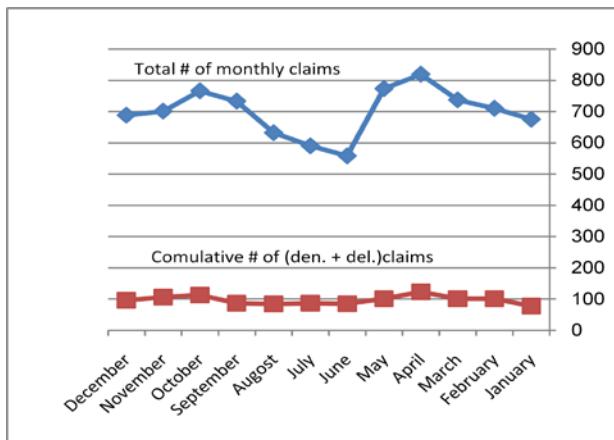


Fig 4 Total # of Received Claims Versus Cumulative # of Den. + Del. claims

According to our assumption, 1% of the electronic claims in proposed system will be lost; however the proposed system still saves in claims management.

Conclusions and Future Work

Smart cards provide a valuable and trusted tool for identification and for the privacy and security of electronic information. In this context, smart cards offer extraordinary value to the healthcare sector. They provide a means to using technology to save time, effort, resources, and, most importantly, live. Moreover, in our research, it is used for claims management. There are clearly demonstrable returns on investment gained by leveraging smart card technology.

In the research, the use of smart cards proof facilitating the following: 1) Verifying patient eligibility in both MCU and SPs, 2) allowing the patient eligibility continual updating on his smart card remotely, and 3) facilitating the claims management. The usage of smart cards technology in our research for health care application, for claims management has proofed the following: 1) proof the citizen eligibility, 2) registering the citizen medical visits to family medical units as well as the visits to contracted providers, 3) reporting, and 4) Facilitate conducting claim management. Accordingly, we feel the usage of smart cards in this tender will add.

The topic has not yet closed, a lot of future work need to complete the view. One of the important future work is to unify health record, from point of view of contents and storage location. Second, we should unify medical codes, covering the doctor's specialties', surgery operations, and other codes that need to be applied. Third, it is needed to create coding system to medicines. It will be a big contribution if the medicines have been coded as well. A simulated results has been proposed, analyzed and proofed the objectives of the proposed solution.

Appendix: System Analysis and Simulation Results

The data stated in **Table 1** contains the assumed numbers of received claims during year 2010 together with the percent of delayed and denied claims, according to our assumption. The corresponding denied and delayed claims have calculated according to above assumptions. In **Table 2** the corresponding delays of claims over three months have been calculated showing the impact of the number of claims on the following months.

Table 1

| Month | Tot # of claims | % monthly of Den.+ Del. | Total # of Den.+ Del. | Total # of denied claims |
|----------|-----------------|-------------------------|-----------------------|--------------------------|
| January | 675 | 20 | 135 | 57 |
| February | 710 | 19 | 135 | 57 |
| March | 737 | 16 | 118 | 48 |
| April | 819 | 20 | 164 | 70 |

| | | | | |
|-----------|-----|----|-----|----|
| May | 773 | 14 | 108 | 43 |
| June | 558 | 17 | 95 | 39 |
| July | 590 | 19 | 109 | 46 |
| August | 632 | 16 | 101 | 41 |
| September | 733 | 14 | 103 | 40 |
| October | 766 | 20 | 153 | 65 |
| November | 701 | 18 | 126 | 53 |
| December | 688 | 16 | 110 | 45 |

| Month | Total # of delayed claims | # of Delayed in 1st month | # of Delayed in 2nd month | # of Delayed in 3rd month |
|-----------|---------------------------|---------------------------|---------------------------|---------------------------|
| January | 78 | 78 | 23 | 8 |
| February | 78 | 78 | 23 | 8 |
| March | 70 | 70 | 21 | 7 |
| April | 94 | 94 | 28 | 9 |
| May | 66 | 66 | 20 | 7 |
| June | 56 | 56 | 17 | 6 |
| July | 63 | 63 | 19 | 6 |
| August | 60 | 60 | 18 | 6 |
| September | 62 | 62 | 19 | 6 |
| October | 88 | 88 | 26 | 9 |
| November | 74 | 74 | 22 | 7 |
| December | 65 | 65 | 20 | 7 |

Table 2

| | Jan. | Feb. | March | April | May | June |
|------------------------------------|------|--------|-------|-------|------|------|
| Total # of Monthly Claims | 675 | 710 | 737 | 819 | 773 | 558 |
| Cumulative # of Den. + Del. Claims | 78 | 101 | 101 | 123 | 101 | 85 |
| | July | August | Sept. | Oct. | Nov. | Dec. |
| Total # of Monthly Claims | 590 | 632 | 733 | 766 | 701 | 688 |
| Cumulative # of Den. + Del. Claims | 87 | 85 | 87 | 113 | 106 | 96 |

Acknowledgment

This research was developed, in close cooperation with the Ministry of State for Administrative Development, MSAD, to show how smart cards is used in healthcare industry to facilitate claims management. I would like to thank the MSAD for the support in development of the pilot, sponsoring, and financial support.

I would like to thank also the Ministry of Health (MoH) and Health Insurance Organization for their contribution in implementation of the pilot and the medical technical support.

References

- [1] Katherine M. Shuler and J. Drew Procaccino, "Smart Card Evolution", Communications of ACM, Volume 45, Issue 7, ACM Press, July 2002.
- [2] Gilles Grimaud, Jean-Louis Lanet, and Jean-Jacques Vandewalle, @ACM SIGSOFT Software Engineering Notes, Proceedings of the 17th European Software Engineering Conference held jointly with the 17th ACM SIGSOFT International Symposium on Foundations of Software Engineering ESEC/FSE-7, Volume 24, Issue 6, Springer-Verlage, ACM Press, October, 1999.
- [3] EMVCO, EMV2000 Integrated Circuit Card Specification for Payment Systems, BOOK 1 - Application Independent ICC to Terminal Interface Requirements, Version 4.0, December, 2000, <http://www.emvco.com/Specifications.cfm>.
- [4] EMVCO , EMV2000 Integrated Circuit Card Specifications for Payment Systems, Book 4 - Cardholder, Attendant, and Acquirer Interface Requirements, Version 4.0, December, 2000, <http://www.emvco.com/Specifications.cfm>.
- [5] Craig W. Thompson, @Smart Devices and Softcontrollers@, IEEE Internet Computing Journal, P.P. 82-85, January/February 2005.
- [6] "Additional information on smart card use in different vertical markets can be found on the Smart Card Alliance web site," <http://www.smartcardalliance.org>.
- [7] "Worldwide Smart Card Shipments 2008," <http://www.eurosmart.com/index.php/publications/market-overview.html>.
- [8] Smart Card Alliance, "HIPAA Compliance and Smart Cards: Solutions to Privacy and Security Requirements," September 2003, <http://www.smartcardalliance.org/pages/publications-hipaa-report>.
- [9] Cheng-Chi Lee, Min-Shiang Hwang, and Wei-Peng Yang, "A Flexible Remote User Authentication Scheme Using Smart Cards", ACM SIGOPS Operating Systems Review, Volume 36, Issue 3, ACM Press, July 2002.
- [10] B . Schneier, Applied cryptography, John Wiley & Sons, Inc., 1996.
- [11] Madison Information Technologies, Inc., "Medical Record Number Errors: A Cost of Doing Business?", April 2001.
- [12] Mays, Susan; Swetnich, Donna; and Gorken, Lynda, "Toward a Unique Patient Identifier," Health Management Technology, March 2002.
- [13] Pesce, Jim, "Staunching Hospitals' Financial Hemorrhage with Information Technology," Health Management Technology, August 2003, <http://archive.healthmgttech.com/archives/h0803stanching.htm>.
- [14] Atchison, Kara, "Surefire strategies to reduce claim denials," Healthcare Financial Management, May 2001, 2003.
- [15] Von Knoop C; Lovich D; Silverstein MB; Tutty M, "Vital Signs: E-Health in the United States," Boston: Boston Consulting Group, 2003. www.bcg.com/publications/files/Vital_Signs_Rpt_Jan03.pdf.

- [16] "Is US Health Really the Best in the World?",
JAMA 2000; 284: 483-485,
<http://jama.ama-assn.org/cgi/content/full/284/4/483>.
- [17] "A Healthcare CFO's Guide to Smart Card Technology and Applications," A Smart Card Alliance Healthcare Council Publication, February, 2009.
- [18] The National Health Care Anti-Fraud Association (NHCAA), "The Problem with Health Care Fraud,"
<http://www.nhcaa.org/eweb/DynamicPage.aspx>
- [19] "Financial crimes report to the public—Fiscal year 2006," U.S. Department of Justice, Federal Bureau of Investigation, September 2006,
http://www.fbi.gov/publications/financial/fcs_report_2006/financial_crime_2006.htm
- [20] Booz Allen Hamilton, "Medical Identity Theft Environmental Scan," October 15, 2008,
<http://www.hhs.gov/healthit/documents/IDTheftEnvScan.pdf>.



shraf Abd El-Hafeez is working as health project manager in MSAD, and has worked in the domain of hospital management information system, information technology, and education with experience about 22 years in IT industry (business & technical).



Magdy El-Hennawy is a lecturer in the Higher Institute of Computer Science & Information Technology, El-Shorouk Academy, and in the same time working as a consultant in the family card system in MSAD. Before and since 1978 working as manager of SW development and maintenance center specialized in mission critical SW systems development.

Before he was working in that place as deputy manager, chief of the system engineering team. Working as a team manager in the system analysis, design and implementation of SW systems, at the same time. He has researches and has taught courses in various subjects.



Mohamed Amer is lecturer in the Higher Institute for Specialized Technological Studies, Future Academy. He has worked in several positions such as maintenance & quality, crises management, info. system development. He has some researches and has taught a sort of courses in various places.

He is interested in Computer networks including Wireless Sensor Networks, Security in Computer Networks and Satellite Communications.

Enhancing Decision Making Using Intelligent System Solution

Sushanta Kumar Panigrahi¹, Amaresh Sahu² and Sabyasachi Pattnaik³

¹ Information Communication & Technology, Fakir Mohan University
Balasore, Orissa 756019, India

² I Computer Science Department, Siksha O Anusandhan University
Bhubaneswar, Orissa 751030, India

³ Information Communication & Technology, Fakir Mohan University
Balasore, Orissa 756019, India

Abstract

The development and deployment of managerial decision support system represents an emerging trend in the business and organizational field in which the increased application of Decision Support Systems (DSS) can be compiling by Intelligent Systems (IS). Decision Support Systems (DSS) are a specific class of computerized information system that supports business and organizational decision-making activities. A properly designed DSS is an interactive software-based system intended to help decision makers compile useful information from raw data, documents, personal knowledge, and/or business models to identify and solve problems and make decisions. Competitive business pressures and a desire to leverage existing information technology investments have led many firms to explore the benefits of intelligent data management solutions such as Particle Swarm Optimization (PSO). This technology is designed to help businesses to finding multi objective functions, which can help to understand the purchasing behavior of their key customers, detect likely credit card or insurance claim fraud, predict probable changes in financial markets, etc.

Keywords: Linear problem, Intelligent System, particle swarm optimization, simplex method

1. Introduction

Organizations generate and collect large volumes of data, which they use in daily operations. Yet despite this wealth of data, many organizations have been unable to fully capitalize on its value because information implicit in the data is not easy to distinguish. However, to compete effectively today, taking advantage of high-return opportunities in a timely fashion, decision-makers must be able to identify and utilize the information. These requirements imply that an intelligent system must interact with a data warehouse and must interface with decision support systems (DSS), which are used by decision-makers in their daily activities [1].

There is a substantial amount of empirical evidence that human intuitive judgment and decision-making can be far from optimal, and it deteriorates even further with complexity and stress. Because in many situations the quality of decisions is important, aiding the deficiencies of human judgment and decision-making has been a major focus of science throughout history. Disciplines such as statistics, economics, and operations research developed various methods for making rational choices. More recently, these methods, often enhanced by a variety of techniques originating from information science, cognitive psychology, and artificial intelligence, have been implemented in the form of computer programs as integrated computing environments for complex decision making. Such environments are often given the common name of decision support systems (DSS). An other name sometimes used as a synonym for DSS is knowledge-based systems, which refers to their attempt to formalize domain knowledge so that it is amenable to mechanized reasoning [5] [6].

An intelligent technology is the duplication of human thought process by machine. It learning from experience, interpreting ambiguities, rapid response to varying situations, applying reasoning to problem-solving and manipulating by applying knowledge, thinking and reasoning [1]. Different from traditional optimization technique, evolutionary computation techniques work on a population of potential solutions (points) of the search space. The most commonly used population-based evolutionary computation techniques is **Particle Swarm Optimization (PSO)**.

The success of management depends on execution of managerial functions and all managerial functions revolve around decision-making and the manager is a decision maker. Financial decision of a company is very complex and risk problem. Due to the constrained nature of the problem, this paper is looking for a new solution that improves the robustness against existing decision with high effectiveness [1]. In this paper we presents the comparison and the relative performance of Traditional Method with intelligent computing techniques like **Particle Swarm Optimization** (PSO) through which a decision maker can enhance decision making, and asses the benefits of variety of intelligent computing techniques. The objective of this paper is to determine the efficiency and accuracy of PSO method for the financial decision of any company.

2. Particle Swarm Optimization

A Swarm can be defined as population of interacting elements (particles) that are able to optimize some global objective through collaborative search of space. It is initialized with a group of random particles and then searches for optima by updating generations. At each step, each particle keeps track of the best solution that it has achieved so far and keeps also track of the overall best value that is obtained thus far by all particles in the population. The nature of interactive elements depends on the problem domain. If the search space is an n-dimensional space, the i^{th} particle of the swarm may be represented by an n-dimensional vector $X_i = (x_{i1}, x_{i2}, \dots, x_{in})$. The velocity of this particle can be represented by another n-dimensional vector $V_i = (v_{i1}, v_{i2}, \dots, v_{in})$. The fitness of each particle can be evaluated according to the objective function of optimization problem. The best previously visited position of the particle i is noted as its individual best position $pbest_i = (p_{i1}, p_{i2}, \dots, p_{in})$. The best position of the swarm is noted as the global best position $gbest_i = (g_1, g_2, \dots, g_n)$. At each step, the velocity of each particle and its new position will be re-estimated according to the following two equations:

$$V_i^{k+1} = \omega V_i^k + c_1 r_1 (pbest_i^k - X_i^k) + c_2 r_2 (gbest^k - X_i^k) \quad (1)$$

$$X_i^{k+1} = X_i^k + V_i^k \quad (2)$$

where, ω is called the inertia weight that controls the impact of previous velocity of particle on its current one. r_1 and r_2 are independently uniformly distributed random variables in the range [0,1]. C_1 and C_2 are positive constant parameters called acceleration coefficients which control the maximum step size and K denotes evolutionary iterations. In PSO, equation (1) is used to calculate the new velocity according to its previous velocity and to the distance of its current position from both its own best

historical position and the best position of the entire population. The particle flies toward a new position according to equation (2). The PSO algorithm is terminated with a maximal number of generations or the best particle position of the entire swarm cannot be improved further after a sufficiently large number of generations. Figure 1 shows the concept of modification of searching points in PSO [12], [13], [14], [15], [16], [19].

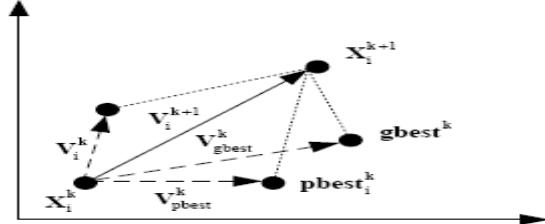


Fig. 1 The concept of modifications of Searching points.

A pseudocode of PSO algorithm is given below,

```
// Initialization
For each particle i
Randomly initialize  $X_i$ ,  $V_i$  for particle i
End For
// Optimization
Do
```

```
    For each particle i
        Call calculate_fitness_value
        If current_fitness_value is better than
            previous_best_fitness_value ( $p_i$ )
        Then
            Current_fitness_value of particle i becomes  $p_i$ 
        End If
    End For
    Call find_global_best_fitness
    For each particle i
```

```
        Call calculate_V_i based on eq. (2)
        Call calculate_X_i based on eq. (3)
    End For
    While MAX_iterations or min_error_criteria is not attained
```

The flow diagram of PSO algorithm is presented in figure 2.

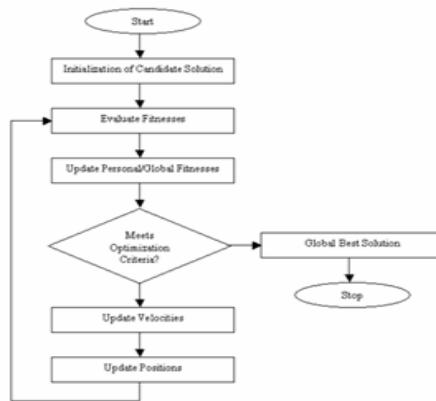


Fig. 2 A simple flow diagram of PSO algorithm.

3. Traditional Method

Applying some well-defined mathematical algorithm known as optimization technique in which the decision theory is based on the assumptions of rational decision makers, whose objective is to optimize the attainment of goals? A well-known Optimization method is linear programming [3] [4].

3.1 Linear Programming

A linear programming is the most commonly applied form of constrained optimization. It may be defined as the problem of maximizing or minimizing a linear function subject to linear constraints. The constraints may be equalities or inequalities. The main components of linear programming problem are decision variable, variable bounds, constraints and objective functions [2] [3] [4].

Example: Product Mix Linear Programming Model [2].

Goal: Maximize Total Profit / Month

Decision variables: X_1 and X_2

Uncontrollable variables and parameters:

Market requirements: $X_1 \geq 0$; $X_2 \geq 0$

Profit contribution of each X_1 is 3 and X_2 is 2

Result variable: Profit = $3X_1 + 2X_2$

Constraints:

$X_1 + X_2 \leq 4$

$X_1 - X_2 \leq 2$

4. Analysis & Discussion

The key element of an optimization problem is the definition of a profit and cost function. This function is a mathematical function which represents the objectives of the expected solution. The goal of the optimization is usually to find the minima or the maxima of this function.

Sometimes, the relationships among the objectives of the optimization problem are so complex that the profit and cost function cannot be defined, or even there is no point in defining a quantitative function (e.g. when the goal is to optimize the quality of a product when the quality is determined by human taste). In this kind of situation, it is very difficult to apply traditional optimization algorithms.

In this section a number of experiments are carried out which outlines the effectiveness of the algorithm described above. The purpose of these experiments is to compare the performance of Simplex Method approach with Particle Swarm Optimization approach for the Product Mix Linear Programming Model. The experiments were conducted on 'Mat lab' and 'c' programming tool. Experimental results obtained from these algorithms were generated with 500 iteration per data point e.g. 40 different populations were created for all the algorithms and each algorithm was run 30 independent runs per data. The best result for each data was produced data point. For each algorithm there are number of different parameters, which need to be varied to "fine-tune" the optimization process. Below we have given two comparison graphs for objective values and fitness values for the respective table 1 and table 2.

4.1 Traditional Procedure

It is a scientific approach to automate managerial decision making and it consists of steps i.e. Define the problem, Classify the problem into a standard category, Construct a mathematical model, Find and evaluate potential solutions to model, Choose and recommend a solution to problem [3] [4].

There are several types of traditional methods, i.e. Simplex Method, Dual Method, etc. We follow the simplex method for the above product mix model and the Solution is found as $X_1 = 3$ and $X_2 = 1$, Profit=Rs 11 after 10 to 12 generations.

Table 1: Objective values after 120 generation

| Generations | Traditional LP | |
|-------------|----------------|-------|
| | X_1 | X_2 |
| 10 | 3 | 1 |
| 20 | 3 | 1 |
| 30 | 3 | 1 |
| 40 | 3 | 1 |
| 50 | 3 | 1 |

| | | |
|-----|---|---|
| 60 | 3 | 1 |
| 70 | 3 | 1 |
| 80 | 3 | 1 |
| 90 | 3 | 1 |
| 100 | 3 | 1 |
| 110 | 3 | 1 |
| 120 | 3 | 1 |

| | | |
|-----|---|---|
| 110 | 3 | 1 |
| 120 | 3 | 1 |

4.3 Result Analysis

Table 3 and Table 4 summarize the empirical results of the LP Model and Proposed PSO Model on optimization of the Product Mix Problem for fitness value and maximization of profit respectively. The result by the test dataset show that the accuracy and multi-objective resultant of the PSO model is much better than obtained from the LP Simplex model and figure 4 and 5 are the graphically representation of fitness value and optimization value respectively.

4.2 Linear Programming Model Using PSO

For the above linear programming model the Particle swarm optimization was set to,

Population size = 40 Maximum iteration = 500
 Max Weight = 0.4 Min Weight = 0.9(Decreasing order)
 $C_1 & C_2 = 1.4$ Dimension = 2
 Velocity = 0 to 10(increasing order)
 Agent initialization between 0 & 1
 Fitness Function is, $3X_1 + 2X_2$ in maximization,
 $X_1 + X_2 \leq 4$
 $X_1 - X_2 \leq 2, X_1, X_2 \geq 0$
 Weight = $W_{max} - ((W_{max} - W_{min}) / \text{max. iter}) \times \text{iter}$
 Velocity = $V_{min} + (V_{max} - V_{min}) \times \text{Random (pop, dim)}$
 where $V_{min}=0$ & $V_{max}=10$

Table 2: Objective values after 120 generation

| Generations | PSO | |
|-------------|--------|--------|
| | X_1 | X_2 |
| 10 | 2.9593 | 1.0149 |
| 20 | 2.9975 | 1.0025 |
| 30 | 2.9999 | 1 |
| 40 | 3 | 1 |
| 50 | 3 | 1 |
| 60 | 3 | 1 |
| 70 | 3 | 1 |
| 80 | 3 | 1 |
| 90 | 3 | 1 |
| 100 | 3 | 1 |
| 110 | 3 | 1 |
| 120 | 3 | 1 |

Table 3: Fitness values after 120 generation

| Generations | Traditional LP | | PSO | |
|-------------|----------------|-------|--------|--------|
| | X_1 | X_2 | X_1 | X_2 |
| 10 | 3 | 1 | 2.9593 | 1.0149 |
| 20 | 3 | 1 | 2.9975 | 1.0025 |
| 30 | 3 | 1 | 2.9999 | 1 |
| 40 | 3 | 1 | 3 | 1 |
| 50 | 3 | 1 | 3 | 1 |
| 60 | 3 | 1 | 3 | 1 |
| 70 | 3 | 1 | 3 | 1 |
| 80 | 3 | 1 | 3 | 1 |
| 90 | 3 | 1 | 3 | 1 |
| 100 | 3 | 1 | 3 | 1 |
| 110 | 3 | 1 | 3 | 1 |
| 120 | 3 | 1 | 3 | 1 |

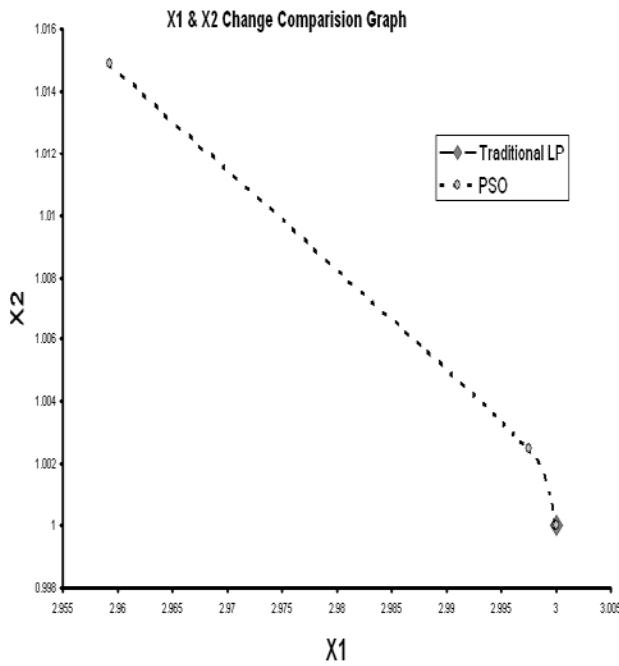


Fig. 3 Fitness comparison graph.

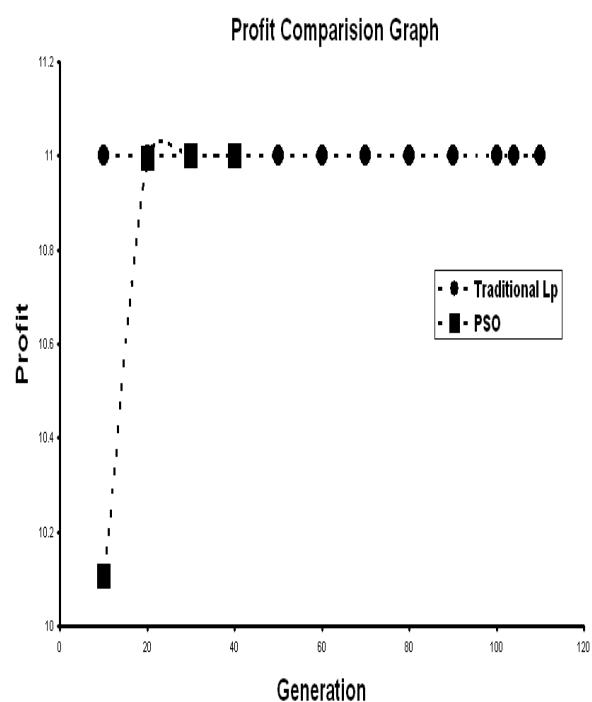


Fig. 4 Objective comparison graph.

Table 4: Objective values after 120 generation

| Generations | Traditional LP | PSO |
|-------------|----------------|---------|
| 10 | 11 | 10.1077 |
| 20 | 11 | 10.9965 |
| 30 | 11 | 10.9997 |
| 40 | 11 | 11 |
| 50 | 11 | 11 |
| 60 | 11 | 11 |
| 70 | 11 | 11 |
| 80 | 11 | 11 |
| 90 | 11 | 11 |
| 100 | 11 | 11 |
| 110 | 11 | 11 |
| 120 | 11 | 11 |

5. Conclusion

In some cases, achievement of optimization problems can not be defined in quantitative way. In this kind of situation, it is very difficult to apply traditional and common optimization methods. But PSO may be a good approach. This paper presented a new approach for the product mix linear programming model with simplified & standard algorithm to optimize combinatorial problem. All the algorithms are based on search technique to further improve individual's fitness that may keep high population, diversity and reduce the likelihood premature convergence. Our objective is to determine the performance of particle swarm optimization algorithm in comparison with simplex method for the financial decisions. It seems that the proposed new comprehensive optimization algorithm may be an efficient system in financial analysis.

References

- [1] "Decision Support Systems and Intelligent Systems", E. Turban and J. Aronson, Prentice Hall.
- [2] "Operation Research", S. D. Sharma, Kedarnath, Ramnath & Co., 2000.
- [3] "Linear Programming", Thomas S. Ferguson.
- [4] James K. Strayer, Linear Programming and Applications, (1989) Springer-Verlag.

- [5] "Decision Support Systems", Marek J. Druzdzel and Roger R. Flynn, University of Pittsburgh, Pittsburgh, PA 15260
- [6] Marek J. Druzdzel. Probabilistic Reasoning in Decision Support Systems: From Computation to Common Sense. PhD thesis, Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA, December 1992.
- [7] L.A. Zadeh, Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems Soft Computing 2 (1998) (Springer-Verlag 1998).
- [8] Kathryn, A.D., "Genetic Algorithms a Tool for OR?", Journal of the operational Research Society 47, pp. 550-561, (1996).
- [9] A Directed Genetic Algorithm for Treating Linear Programming Problems, Dr. Bayoumi M.A.H. & El-Feky E.Z., Cairo University, Egypt.
- [10] "Genetic algorithms and financial applications", Davis L., Deoock, 1994
- [11] "Evolutionary Module Acquisition", Angeline, P. J. and Pollack, J. B., In Proceedings of the Second Annual Conference on Evolutionary Programming, D.B. Fogel and W. Atmar (eds.), La Jolla, CA: Evolutionary
- [12] "Swarm Intelligence", J. Kennedy and R. Eberhart, Morgan Kaufmann Publishers, San Francisco, CA, 2001.
- [13]. "An Introduction to Particle Swarm Optimization", Matthew Settles Programming Society, 1993.
- [14] J. Kennedy and R. Eberhart (1995). "Particle swarm optimization," in *Proc. IEEE Int. Conf. Neural Networks*, vol. 4, pp. 1942–1947.
- [15] Y. Shi and R. C. Eberhart, "Parameter selection in particle swarm optimization," in *Proc. Evolutionary Programming VII*, vol. 1447, 1998, pp. 591–600.
- [16] M. Clerc, J. Kennedy (2002). "The particle swarm: explosion, stability and convergence in a multi-dimensional complex space", In IEEE Transaction on Evolutionary Computation Vol 6, pp. 58–73.
- [17] Kou, G., Liu, X., Peng, Y., Shi, Y., Wise, M., & Xu, W. (2003). Multiple criteria linear programming approach to data mining: Models, algorithm designs and software development. *Optimization Methods and Software*, 18, 453-473.
- [18] "Research and Trends in Data Mining Technologies and Applications", David Taniar, Monash University, Australia, Idea Group Publishing, Melbourne.
- [19] Saeed Rategh, Farbod Razzazi, Amir Masoud Rahmani, Shayan Oveis Gharan "A Time Warping Speech Recognition System Based on Particle Swarm Optimization" in Second Asia International Conference on Modelling & Simulation-2008.

Sushanta Kumar Panigrahi: Mr.Sushanta Kumar Panigrahi received a MCA from IGNOU in 2002.Currently he is Head in IT at Interscience Institute of Management & Technology, Bhubaneswar, India. He is performing research in Cluster Analysis, Neural Network, Optimization Technique and Soft Computing in ICT at Fakir Mohan University, Balasore. He is served more than 8 years in different Colleges in the state of Orissa. He is published 2 research paper in national journals and conferences.

Amaresh Sahu: Mr.Amaresh Sahu received a M. Tech. (CS) from Utkal University, Bhubaneswar 2005.Currently he is HOD in MCA at Ajay Binay Institute of Technology, Cuttack, India. He is performing research in Cluster Analysis, Neural Network and Soft Computing in Computer Science at SOA University, Bhubaneswar. He is served more than 10 years in different Colleges in the state of Orissa. He is published 4 research paper in national journals and conferences.

Dr.Sabyasachi Pattnaik: Dr Sabyasachi Pattnaik has done his B.E in Computer Science, M Tech.from IIT Delhi. He has received his PhD degree in Computer Science in the year 2003, now working as Reader in the Department of Information and Communication Technology, in Fakir Mohan University, Vyasavihar, Balasore, Orissa, India. He has got 15 years of teaching and research experience in the field of neural networks, soft computing techniques. He has got 22 publications in national & international journals and conferences. He has published three books in office automation, object oriented programming using C++ and artificial intelligence. At present he is involved in guiding 6 scholars in the field of neural networks in cluster analysis, bio-informatics, computer vision & stock market applications. He has received the best paper award & gold medal from Orissa Engineering congress in 1992 and institution of Engineers in 2009.

Teaching Software Engineering: Problems and Suggestions

Osama Shata

Department of Computer Science and Engineering, Qatar University
Doha, Qatar

Abstract

Teaching Software Engineering is a challenging task. This paper presents some problems encountered during teaching the course of software engineering to computer science and computer engineering students for few offerings. We present problems encountered and which are related to its title and contents and present suggested solutions.

Keywords: *Software Engineering, Development Cycle, Object-Oriented.*

1. Introduction

This paper presents some problems encountered during teaching the course of software engineering to computer science and computer engineering students for few offerings. We present problems encountered as well as suggested solutions.

I teach Software Engineering, which is a common compulsory course in many Computer Science and Computer Engineering curriculums. Probably because programming courses are part of those curriculums and software engineering is being defined, in general, as concerned with developing quality software. However, I found that in many cases, and during my discussions with colleagues on how to improve the course, the course is being looked at as an intruder to both curriculums. Computer scientists do not feel that it is a real computer science course. While the word “engineering” in its title may be contributing to this feeling, the course is also different in its nature from real computer science courses such as Computer Architecture, Operating Systems, Algorithms ... etc. Computer engineers also do not feel that it is a real Computer Engineering course, probably the word “software” is contributing to this feeling, but more importantly, it lacks hardware components and lacks real design experiences. This paper begins with a brief introduction to the origin of the Software Engineering discipline. Next the paper will discuss contents that are usually being taught in a typical Software Engineering

course and highlights problems faced and offer suggestions. The paper concludes with a summary.

2. Problems

The term Software Engineering (SE) was first introduced in 1968 in a NATO conference to address software crisis which came to surface in that period, when many large software projects faced great difficulties such as unexpected delay in delivery, and exceeding estimated costs [1]. Some of the problems encountered during teaching the SE course are related to its title while others are related to its contents. We begin with those related to its title.

2.1 Course Title

One of the first problems faced during teaching the course was to explain its title and why the word “engineering” was in its title. The IEEE Computer Society’s Software Engineering Body of Knowledge defines Software Engineering as the: “application of a systematic, disciplined, quantifiable approach to the development, operation and maintenance of software, and the study of these approaches; that is the application of engineering to software” [2]. This means that “engineering” is the application of a “systematic, disciplined, quantifiable approach”. However, according to The American Engineer’s Council for Professional Development, “engineering” is: “the creative application of scientific principles to design or develop structures, machines, apparatus, or manufacturing processes, or works utilizing them singly or in combination; or to construct or operate the same with full cognizance of their design; or to forecast their behavior under specific operating conditions; all as respects an intended function, economics of operation and safety to life and property” [3].

The relationship between the two definitions is not, and cannot be, tight. Software development is very different

from engineering, for example: software is intangible meanwhile engineering applications are tangible, the existence of many programming languages with many features makes it possible to have many solutions to the same problem, and the reusability and reproduction of a solution to a problem in many other problems makes it hard to assess the effort involved. Other resources [4] point out that software development should follow an engineering paradigm. This means that there is a standalone “engineering paradigm” which has well defined steps. However, top ranked results returned from searching the Internet with various search engines for that term returned resources having “software engineering paradigm”. That was really confusing to students. The software development process must follow the engineering paradigm which itself does not have a clear definition. A good solution for this problem is to accept Alistair’s claim that [5]: “The phrase ‘software engineering’ was deliberately chosen as being provocative, in implying the need for software manufacture to be based on the types of theoretical foundations and practical disciplines, that are traditional in the established branches of engineering”. Specially that the term “software engineering” first appeared in the 1968 NATO Conference on Software Engineering, and which was aimed to stimulate software professionals and researchers to respond to software crisis at that time [1].

A second justification for using the word “engineering” in the title is to relate it to “Systems Engineering”. According to The International Council on Systems Engineering (INCOSE), Systems Engineering is: “An interdisciplinary approach and means to enable the realization of successful systems” [6]. An expanded definition for systems engineering is given by the National Aeronautics and Space Administration (NASA) [7]: “System Engineering is a robust approach to the design, creation, and operation of systems. In simple terms, the approach consists of identification and quantification of system goals, creation of alternative system design concepts, performance of design trades, selection and implementation of the best design, verification that the design is properly built and integrated, and post-implementation assessment of how well the system meets (or met) the goals” . So depending on the previous definitions one may justify that the term “engineering” in software engineering was either borrowed from system engineering to mean “an interdisciplinary approach and means to enable the realization of successful software systems”, or to denote that if an engineering system has a software component, and most probably it would, then Software Engineering “is a robust approach to the design, creation, and operation of software systems. In simple terms, the approach consists of identification and quantification of software system goals, creation of

alternative software system design concepts, performance of design trades, selection and implementation of the best design, verification that the design is properly built and integrated, and post-implementation assessment of how well the system meets (or met) the goals”. In all cases, in our opinion, this does not make Software Engineering an engineering discipline.

The some may agree or disagree with the above trials to explain the title of the course. However, we believe that there has been much room for trials because the 1968 NATO Software Engineering Conference did not give an explanation. according to Alistair [3]: “despite having the term as a focal point for the conference, the participants showed little understanding of either the term “Software Engineering” or engineering in general, and provide little guidance as to just what readers are supposed to infer from the term “Software Engineering.” Alan Perlis’ keynote speech contains the following: this is the first conference ever held on Software Engineering and it behooves us to take this conference quite seriously since it will likely set the tone of future work in this field in much the same way that Algol did. We should take quite seriously both the scientific and engineering components of software, but our concentration must be on the latter. Unfortunately, that is all he offers on the intention of the term.”

Questioning whether software engineering is an engineering discipline at all is not new [7, 8, and 9].

Also, the teaching of Software Engineering as a subject is in continuous debate [10, 11]. It is not the goal of this paper to add to the doubts about the Software Engineering as a discipline or its education, but rather to find solutions to problems encountered during teaching the course. We find that Alistair’s justification that the term was deliberately chosen as being provocative is an acceptable solution to the confusing title of the course.

2.1 Course Contents

Other problems encountered in the course were related to the course contents. Browsing syllabi of many software engineering courses, including ours, would lead to the conclusion that most of them have the following contents in common:

- a- Introduction to software engineering
- b- The software development process and software life cycle
- c- Requirements specifications
- d- Analysis and design (structured, object oriented approaches and UML)
- e- Implementations, testing, maintenance and reliability
- f- CASE tools
- g- Other topics (e.g. project managements)

Problems related to (a) above would mainly involve the title and this has been dealt with earlier.

The software development process and software life cycle usually introduces students to the waterfall model, iterative models (e.g. spiral model), agile model, extreme model and rational united process. The waterfall model is well defined and the differences between this model and other models are clear. However, the differences between the other models are not that clear and could be confusing. For example, it is difficult to explain and highlight rigid differences between the spiral and agile models. Both are incremental and iterative. Both work in order of risk. The difference may be in the scope. While the spiral focuses on big design from the beginning and is recommended for large projects, the agile focuses on one increment at a time and may work for small projects. That difference is not really sharp to require two names for almost the same model. It was going to be easier if agile was considered a special case of the spiral model. Also, it is not clear what is meant with big and small projects, this is proportional. If the course delves into the discussion of the Extreme Programming (XP) model / technique then more confusion is added to the course as follows: The PC magazine says about XP that “it is based on a formal set of rules about how one develops functionality such as defining a test before writing the code and never designing more than is needed to support the code that is written” and “XP is designed to steer the project correctly rather than concentrating on meeting target dates, which are often unrealistic in this business” [12]. But is not that what software developers need? Just to design what is needed for coding and to steer the project correctly? If so, then why the need for other models? Even more, TechTarget [13] claims that: “Kent Beck, author of Extreme Programming Explained: Embrace Change, developed the XP concept. According to Beck, code comes first in XP”. But this contradicts what we have been teaching students that software engineering is concerned with the careful analysis and design so that the coding phase goes smoothly. Now, we teach them that code comes first. Furthermore, according to Don Wells [14], XP “has already been proven to be very successful at many companies of all different sizes and industries worldwide”. Again, if XP is the perfect model for all different sizes and industries then why trying other models? On the other hand, some suggest that XP is waning [15]. While most literature suggests that XP is a special case of agile, Extreme Programming (XP) happens to be the most well-known of agile methodologies [16]; others suggest that agile itself is only an implementation of the spiral model [17]. The point here is that there is no consensus on the relationship between the different models and there is no

clear recommendation on when to use each. We transfer this confusion to students in our teaching. Now, how about adding the Rational Unified Process (RUP) to the picture? We suggest not to overwhelm students with many techniques and models, but rather to introduce them to the waterfall model and the spiral model and we list the agile, pair programming, and Extreme programming as different implementation of the spiral model and focus on the XP since it seems to be working and we believe that students actually have been following this technique in their programming courses without actually realizing that it has the name XP. Once students learn a programming language they become enthusiastic to using it and start coding quickly. So, they actually design little and later and code first. Of course, we have to shape their skills in using these techniques, but it is the closer to them.

A third source for problems encountered was the topic of “Analysis and design (structured, object oriented (OO) approaches and UML)”. This is due to the similarity between some of the tools used in the structured and object-oriented approaches. A student once asked why I should use use-cases, sequence diagrams and class diagrams when I can use the entity-relationship diagram I learned in the database course and the data flow diagram and process flow diagram which I have learned in other courses. The structured approach mainly uses the entity-relationship diagram (E-R) and the data flow diagram (DFD), whereas the object-oriented approach may use the UML including:

- Use case diagrams
- Class diagrams
- Sequence diagram
- Object diagram
- Package diagram
- Deployment diagrams
- State machine diagram
- Activity diagram
- Communication diagram
- Component diagrams
- Interaction overview diagrams
- Timing diagrams

Although there are differences between the structured (functional decomposition) approach and the OO approach, but there are also big similarities between some of their tools (e.g. the E-R diagram and the class diagram). This makes students ask why the E-R diagram is not part of the UML. An entity in the E-R diagram corresponds to the class in the class diagram. Attributes in the E-R diagram corresponds to attributes in the class diagram. Relationships between entities correspond to relationships between classes. Of course the latter have methods and

operations as well. But students wonder if the structured approach is supposed to be considered a completely different approach from the object-oriented approach whereas major tools are almost the same (or very similar) in both. The instructor focuses on the fact that a class diagram represents the behavior features of a system through the operations. A similar argument can be said about the similarity between the DFD and the sequence diagram or the activity diagram. Since the UML with its various diagrams are more comprehensive then we believe that it should be used directly without actually considering the E-R diagram and the DFD. A brief introduction to the structured approach may be considered but without delving into the tools.

4. Conclusions

We have identified and presented some problems encountered during teaching the course of software engineering with some brief and quick suggestions. We believe that most of these problems encountered are due following a traditional course syllabus that addresses both the structured and OO approaches in detail, and also for considering many diagrams that are part of UML. We believe that the course must involve extensive programming and many case studies to be interesting to students and to clarify to students that this course does not provide one proper solution to developing software, but rather various approaches could be adopted and that there is room for creativity. We are currently working on developing a new syllabus which addresses the contents problems raised in this paper in more details and which we expect to make the course interesting and more applied.

References

- [1] P. Naur and B. Randell, Eds. Software Engineering. Report on a Conference held in Garmisch, Oct. 1968, sponsored by NATO
- [2] SWEBOK executive editors, Alain Abran, James W. Moore; editors, Pierre Bourque, Robert Dupuis. (2004). Pierre Bourque and Robert Dupuis. Ed. Guide to the Software Engineering Body of Knowledge - 2004 Version. IEEE Computer Society. pp. 1–1. ISBN 0-7695-2330-7
- [3] Science, Volume 94, Issue 2446, pp. 456: Engineers' Council for Professional Development
- [4] A Brief History of Software Engineering. Online resource: http://www.comphist.org/computing_history/new_page_13.htm Retrieved October 17, 2010.
- [5] Alistair. The end of software engineering and the start of economic-cooperative gaming. Online Resource: <http://alistair.cockburn.us/The+end+of+software+engineering+and+the+start+of+economic+cooperative+gaming> Retrieved Oct 1, 2010.
- [6] Systems Engineering Handbook, version 2a. INCOSE. 2004.
- [7] NASA Systems Engineering Handbook. NASA. 1995. SP-610S.
- [8] Mahoney, Michael. 2004. Finding a History for Software Engineering, Online resource <http://www.princeton.edu/~hos/Mahoney/articles/finding/finding.html> Retrieved September 15, 2010.
- [9] M. Shaw, "Prospects for an Engineering Discipline of Software", IEEE Software vol. 7, no. 6, Nov. 1990, p. 15.
- [10] Parnas, D. L., Software Engineering Programs are not Computer Science Programs, IEEE Software, November/December, 1999, Vol. 16, No. 6, pp. 19-30.
- [11] Demarco, T. Point-Counter Point: It Ain't Broke, So Don't Fix It, IEEE Software, November/December, 1999, Vol. 16, No. 6, pp. 67-69.
- [12] PCMag. Extreme Programming. Online resource: http://www.pcmag.com/encyclopedia_term/0,2542,t=XP&i=55075.00.asp Retrieved Sep18, 2010.
- [13] Techtarget. Extreme Programming. Online resource http://searchsoftwarequality.techtarget.com/sDefinition/0,,sid92_gci214366.00.html Retrieved sep15, 2010.
- [14] Extreme Programming. : Extreme Programming: A gentle introduction. Online resource <http://www.extremeprogramming.org/> Retrieved Sep13, 2010.
- [15] Smith, Steve. Is Extreme Programming Dying? Is Agile Growing in Popularity? Online resource <http://stevesmithblog.com/blog/is-extreme-programming-dying-is-agile-growing-in-popularity/> Retrieved Oct 1, 2010.
- [16] Hutagalung, Wilfrid. 2006. Extreme Programming. Online resource <http://www.umsl.edu/~sauterv/analysis/f06Papers/Hutagalung/#xp> Retrieved Oct 18, 2010.
- [17] Stackoverflow. Online resource <http://stackoverflow.com/questions/253789/agile-vs-spiral-model-for-sdlc> Retrieved Oct 1, 2010.

Osama Shata is an Associate Professor in the department of Computer Science and Engineering at Qatar University, Qatar. He has an extensive industrial, academic and administrative experience at both the postgraduate and undergraduate levels. His research interests began with intelligent database systems and knowledge base systems. As information technology became an integral component of any successful education process, his research interests focused on e-learning / distance education, electronic course delivery, curriculum development and integration, multimedia, HCI, curriculum design and development, and Accreditation.

Software Vulnerabilities, Banking Threats, Botnets and Malware Self-Protection Technologies

Wajeb Gharibi¹, Abdulrahman Mirza²

¹ Computer Networks Department, Computer Science & Information Systems College, Jazan University
Jazan 82822-6694, Saudi Arabia

² Information Systems Department, King Saud University, Center of Excellence in Information Assurance
Riyadh 11482, Saudi Arabia

Abstract

Information security is the protection of information from a wide range of threats in order to ensure success business continuity by minimizing risks and maximizing the return of investments and business opportunities. In this paper, we study and discuss the software vulnerabilities, banking threats, botnets and propose the malware self-protection technologies.

Keywords: *Informatics, Information Security, Cyber Threats, Malware Self-Protection Technologies.*

1. Introduction

Nowadays, there is a huge variety of cyber threats that can be quite dangerous not only for big companies but also for an ordinary user, who can be a potential victim for cybercriminals when using unsafe system for entering confidential data, such as login, password, credit card numbers, etc. Among popular computer threats it is possible to distinguish several types depending on the means and ways they are realized. They are: malicious software (malware), DDoS attacks (Distributed Denial-of-Service), phishing, banking, exploiting vulnerabilities, botnets, threats for mobile phones, IP-communication threats, social networking threats and even spam. All of these threats try to violate one of the following criteria: confidentiality, integrity and accessibility.

Obviously that hackers use the malicious programs to gain control of targeted computer in order to use it further for other types of cyber attacks. As a result, malicious software has turned into big business and cyber criminals became profitable organizations and able to perform any type of attack. An understanding of today's cyber threats is a vital part for safe computing and ability to counteract the cyber invaders.

Our paper is organized as follows: Section 2 demonstrates the software vulnerabilities. Section 3 proposes banking threats. Section 4 defines botnets. Conclusions have been made in Section 5.

2. Software Vulnerabilities

The term 'vulnerability' is often mentioned in connection with computer security, in many different contexts. It is associated with some violation of a security policy. This may be due to weak security rules, or it may be that there is a problem within the software itself. In theory, all computer systems have vulnerabilities [1-5].

MITRE, a US federally funded research and development group, focuses on analyzing and solving critical security issues. The group has defined the followings:

Definition 2.1 A universal vulnerability is a state in a computing system (or set of systems) which either allows an attacker to execute commands as another user, or to access data that is contrary to the specified access restrictions, or to pose as another entity to conduct a denial of service.

Definition 2.2 An exposure is a state in a computing system (or set of systems) which is not a universal vulnerability, but either allows an attacker to conduct information gathering activities or hide activities or includes a capability that behaves as expected, but can be easily compromised.

It is a primary point of entry that an attacker may attempt to use to gain access to the system or data is considered a problem according to some reasonable security policy.

Microsoft Windows, the operating system most commonly used on systems connected to the Internet, contains multiple, severe vulnerabilities. The most commonly exploited are in IIS, MS-SQL,

Internet Explorer, the file serving and message processing services of the operating system itself [6, 7].

A vulnerability in IIS, detailed in Microsoft Security Bulletin MS01-033, is one of the most exploited Windows vulnerabilities ever. A large number of network worms have been written over the years to exploit this vulnerability, including 'CodeRed' which was first detected on July 17th 2001 and is believed to have infected over 300000 targets. Still some versions of CodeRed worm are spreading throughout the Internet [8].

Spida Network Worm, detected almost a year after CodeRed appeared, relied on an exposure in MS-SQL server software package to spread.

Slammer Network Worm, detected in late January 2003, used an even more direct method to infect Windows systems running MS-SQL server: a buffer overflow vulnerability in one of the UDP packet handling subroutines. As it was relatively small - 376 bytes - and used UDP, a communication protocol designed for the quick transmission of data, Slammer spread at an almost incredible rate. Some estimate the time taken for Slammer to spread across the world at as low as 15 minutes, infecting around 75000 hosts [9].

However, Lovesan Worm, detected on 11th August 2003, used a much more severe buffer overflow in a core component of Windows itself to spread. This vulnerability is detailed in Microsoft Security Bulletin MS03-026.

Sasser Worm was first appeared at the beginning of May 2003, exploited another core component vulnerability, this time in the Local Security Authority Subsystem Service (LSASS). Sasser spread rapidly and infected millions of computers world-wide, at an enormous cost to business [10].

From last incidents also it is possible to note that epidemic of Worm Kido/Conficker/Downadup which as one of distribution methods used vulnerability MS08-067 in service "Server" (<http://www.microsoft.com/technet/security/Bulletin/MS08-067.mspx>).

Inevitably, all operating systems contain vulnerabilities and exposures which can be targeted by hackers and virus writers. Although Windows vulnerabilities receive the most publicity due to the number of machines running Windows, Unix and MacOS have also their own weak spots.

3. Banking Threats

Definition 3.1 Banking - one of the remote bank service kinds at which management is made through the Internet.

In 2007, antivirus vendors saw a huge increase in

the number of malicious programs targeting banks (financial malware) according to Kaspersky Lab stats (Figure 1). In spite of the lack of clear information from the financial sector, this indicates a corresponding increase in the number of banks attacks.

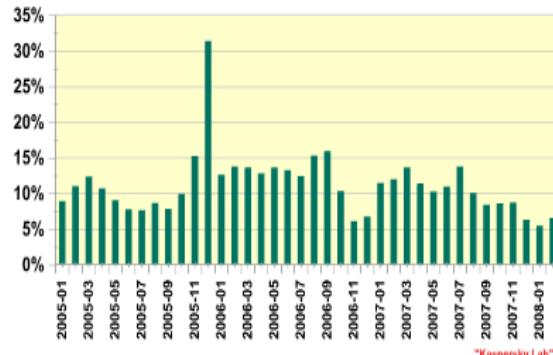


Figure1. Percentage of financial malware among all malicious programs detected

Notwithstanding an increased number of attacks, as the graph above shows, the percentage of financial malware detected each month is dropping. The reasons for this are detailed below:

- Malware authors constantly change their programs in order to evade detection by antivirus solutions. However, if the changes made are minor, AV vendors will still be able to detect new malware samples using signatures created for previous variants.
- The graph above covers only financial malware. However, banking attacks are usually a multi-step process: social engineering, phishing, and the use of Trojan-Downloaders which then download the financial malware. It's easier for the criminals to modify the Trojan-Downloader programs (which are usually smaller in size, and generally less complex) than the financial malware itself.

In 2007, there was an upsurge in the number of password stealing Trojans designed to steal all data entered into web forms. These target the most popular browsers i.e. Internet Explorer, Opera and Firefox. Such Trojans can obviously be used to steal credit cards, and using such malware may be enough to breach a bank's defenses – it all depends on the sophistication of the security measures employed [11, 12].

Actually, malicious programs delivered via email are more likely to attract the attention of antivirus vendors and financial institutions, not to mention the media and end users. Stealth is the key factor in the success of attacks on financial institutions, so conducting a drive-by download using exploits is obviously an attractive method. Moreover, it is a significant factor in terms of evading quick

detection by antivirus solutions – malicious programs which infect victim systems via the web are hosted on a web server. This means that the cyber criminals using these programs to conduct attacks can modify the malicious files very easily using automated tools – a method known as server-side polymorphism. In contrast to regular host polymorphism (where the algorithm used to modify the code is contained in the body of the malicious program) it's impossible for antivirus researchers to analyze the algorithm used to modify the malware, as it's located on the remote server.

In addition, some of the more sophisticated Trojan-Downloaders used to deliver financial malware to its eventual destination are designed to self-destruct (or 'melt') once they have successfully or unsuccessfully downloaded the financial malware.

The use of Transaction Authorization Numbers (TAN) for signing transactions makes gaining access to accounts somewhat more complex. The TAN may come from a physical list issued to the account holder by the financial organization or it may be sent via SMS.

Another method used by cyber criminals is to redirect traffic. There are several ways of doing this, and the easiest of these is to modify the Windows "hosts" file which is located in the %windows%\system32\drivers\etc directory, can be used to bypass DNS (Domain Name Server) lookups. DNS is used to translate domain names, such as www.kaspersky.com, into an IP address. Domain names are used purely for convenience; it's the IP addresses which are used by computers. If the host files are modified to point a specific domain name to the IP address of a fake site, the computer will be directed to that site.

Another method for redirecting traffic is to modify the DNS server settings. Instead of trying to bypass DNS lookups, the settings are changed in such a way that the machine uses a different, malicious, DNS server for the lookups. Most people surfing from home use the DNS server belonging to their ISP for lookups. As a result, the vast majority of this type of attack has been directed at workstations. However, when a router is used to access the internet, by default it's the router performing DNS lookups and passing them on to the workstations.

Yet another method which can be used to redirect traffic is to place a Trojan on the victim machine which monitors the sites visited. As soon as the user connects to a banking site (or that of another financial organization) the Trojan will redirect the traffic to a fake website. The traffic may be redirected from an HTTPS site to an HTTP (potentially insecure) site. In such cases, the Trojan is usually able to suppress any warning message

issued by the browser.

4. BOTNETS

Botnets have been in existence for about 10 years; experts have been warning the public about the threat posed by botnets for more or less the same period.

Definition 4.1 A botnet is a network of computers made up of machines infected with a malicious backdoor program. The backdoor enables cybercriminals to remotely control the infected computers (which may mean controlling an individual machine, some of the computers making up the network or the entire network).

Malicious backdoor programs that are specifically designed for the use of creating botnets are called bots. Botnets have vast computing power. They are used as a powerful cyber weapon and are an effective tool for making money illegally. The owner of a botnet can control the computers which form the network from anywhere in the world – from another city, country or even another continent. Importantly, the Internet is structured in such a way that a botnet can be controlled anonymously [13].

Botnets can be used by cybercriminals to conduct a wide range of criminal activities, from sending spam to attacking government networks:

- Sending spam - the most common use for botnets (over 80% of spam is sent from zombie computers).
- DDoS attacks - using tens or even hundreds of thousands of computers to conduct DDoS (Distributed Denial of Service) attacks.
- Anonymous Internet access; cybercriminals can access web servers using zombie machines and commit cybercrimes such as hacking websites or transferring stolen money.
- Selling and leasing botnets. One option for making money illegally using botnets is based on leasing them or selling entire networks. Creating botnets for sale is also a lucrative criminal business.
- Phishing; a botnet allows phishers to change the addresses of phishing pages frequently, using infected computers as proxy servers. This helps conceal the real address of the phishers' web server.
- Theft of confidential data; botnets help to increase the haul of passwords (passwords to email and ICQ accounts, FTP resources, web services etc.)
- There are currently only two known types of

botnet architecture:

- a) Centralized botnets; in this type of botnet, all computers are connected to a single command-and-control center or C&C. The C&C waits for new bots to connect, registers them in its database, tracks their status and sends them commands selected by the botnet owner from a list of bot commands. All zombie computers in the botnet are visible to the C&C. The zombie network owner needs access to the command and control center to be able to manage the centralized botnet.

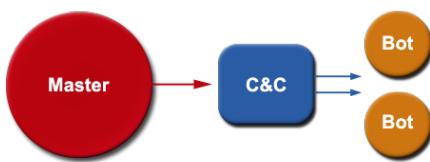


Figure 2 Centralized topology (C&C)

Centralized botnets are the most widespread type of zombie network. Such botnets are easier to create, easier to manage and they respond to commands faster. However, it is also easier to combat centralized botnets, since the entire zombie network is neutralized if the C&C is put out of commission.

- b) Decentralized or P2P (peer-to-peer) botnets; in a decentralized botnet, bots connect to several infected machines on a bot network rather than to a command and control center. Commands are transferred from bot to bot: each bot has a list of several 'neighbors', and any command received by a bot from one of its neighbors will be sent on to the others, further distributing it across the zombie network. In this case, a cybercriminal needs to have access to at least one computer on the zombie network to be able to control the entire botnet.

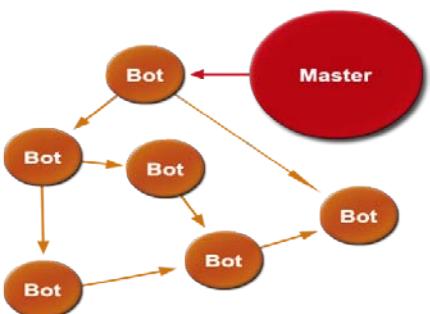


Figure 3 Decentralized topology (P2P)

In practice, building decentralized botnets is not easy task, since each newly infected computer

needs to be provided with a list of bots to which it will connect on the zombie network.

Today, botnets are among the main sources of illegal income on the Internet and they are powerful weapons in the hands of cybercriminals [14].

What makes botnets increasingly dangerous is that they are becoming easier and easier to use. In the near future, even children will be able to manage them. The ability to gain access to a network of infected computers is determined by the amount of money cybercriminals have at their disposal rather than whether they have specialized knowledge. Additionally, the prices in the well-developed and structured botnet market are relatively low.

5. Malware Self-Protection Technologies

Trying to hide the presence of a malicious component in binary code of the program or in script of the web-page, hackers use various techniques, such as: encryption, or polymorphism, or obfuscation, or packing. Thus the malicious program complicates the process of signature detection of the code by the anti-virus scanner. Such methods of protection have received the name passive.

Encryption is the universal mechanism which can be applied for protection of code as well, as for ciphering the data of the user and demanding the payment for their decoding, as it has been realized in one of the first viruses - Cascade which contained polymorphic encryptor and ciphered each new copy of a virus a unique key, and as dangerous malicious functional, for example, as it was implemented with virus GpCode [15].

Definition 5.1 Polymorphism — a technology that allows a self-replicating program to fully or partially modify its outward appearance and/or the structure of its code during the replication process.

Definition 5.2 Obfuscation — a combination of approaches used to obscure the source code of a program.

This is designed to make the code as difficult as possible to read and analyze it while retaining full functionality. Obfuscation technologies can be applied at the level of any programming language (including high level, script and assembler languages). Examples of very simple obfuscation include adding neutral instructions (which do not alter program functionality) to the code or making the code harder to read by using an excessive number of unconditional skips (or unconditional changeovers disguised as conditional skips).

There are many approaches that can be applied for these purposes (dynamic code generators,

polymorphism, etc.), but in most cases, the authors of malicious programs don't spend much time or effort on developing these types of mechanisms. They use a much simpler solution in order to achieve their goals: so-called packers. These are utilities that use dedicated algorithms to encode the target executable program while retaining its functionality. The use of packers makes a malicious user's task much easier: in order to prevent an antivirus program from detecting an already known malicious program, the author no longer has to rewrite it from scratch - all he has to do is re-pack it with a packer that is not known to the antivirus program. The result is the same, and the costs are much lower.

The continued acceleration of the increase in new malicious programs (Figure 4) is now accompanied by an increase in the total number of malicious programs which actively combat antivirus solutions. First and foremost, this involves virus writers using rootkit technologies in order to increase the lifespan of a virus in the infected system: if a malicious program has stealth capabilities, then it is less likely to end up in an antivirus company's database [16, 17].



Figure 4 Increase in the number of new modifications of malicious programs that actively combat security solutions.
Source: Kaspersky Lab

There have always been malicious programs that have actively defended themselves. Self-defense mechanisms include:

- Performing a targeted search of the system for an antivirus product, firewall or other security utility, followed by disrupting the functioning of that utility. An example might be a malicious program that searches for a specific antivirus product in the process list and subsequently attempts to disrupt the functioning of that antivirus.
- Blocking files and opening them with exclusive access as a counter measure against file scanning by the antivirus.
- Modifying the hosts file in order to block access to antivirus update sites.
- Detecting query messages sent by the security system (for example, a firewall window with

an inquiry such as "Allow this connection?") and imitating a click on the "Allow" button.

Rootkit, or as it is usually referred to, bootkit because it runs during the boot sequence, is based on the eEye Bootroot code. Essentially, it's not so much a separate piece of malware as a tool to hide Trojans...any Trojans. Consequently, it seems a reasonable conclusion that Sinowal is being shared (possibly for a fee) in certain circles and that we haven't seen the last of it by any means [18-21].

6. Conclusions

The most common types of contemporary threats are considered as well, as mechanism of malware self-protection aimed to counteract against antivirus. New areas of hackers' attacks were highlighted in this paper. Thus, it gives us clear understandings what is going on in the world of cyber security and helps to protect our computer systems from undesirable intruders and confidential data theft.

Nowadays, we use more and more online services in Internet which can be threat of personal information stealing by third party. It is getting more important to keep our data in secure place protected by antivirus and DLP (Data Leakage Protection) systems.

References

- [1] Alexander Adamov, «Computer Threats: Methods of Detection and Analysis», Kaspersky Lab, Moscow 2009.
- [2] www.securelist.com
- [3] Infosecurity Magazine: Phishing and the economics of e-crime, Sep 2007.
- [4] Z. Chen and C. Ji, "A self-learning worm using importance scanning," in ACM CCS Workshop on Rapid Malcode (WORM'05), 2005.
- [5] C. C. Zou, W. Gong, and D. Towsley, "Code red worm propagation modeling and analysis," in 9th ACM Conference on Computer and Communication Security (CCS'02), 2002.
- [6] C. Shannon and D. Moore, "The spread of the witty worm," IEEE Security and Privacy Magazine, 2004.
- [7] C. Zou, L. Gao, W. Gong, and D. Towsley, "Monitoring and early warning of internet worms," in ACM Conference on Computer and Communications Security (CCS'03), 2003.
- [8] M. Rajab, F. Monroe, and A. Terzis, "Fast and evasive attacks: Highlighting the

- challenges ahead,” in 9th International Symposium on Recent Advances in Intrusion Detection (RAID’04), 2006.
- [9] «Computer Threats: Methods of Detection and Analysis», Kaspersky Lab, Moscow 2010.
- [10] www.securelist.com, Kaspersky Security Bulletin, Malware evolution 2008.
- [11] www.securelist.com, «Software Vulnerability», Encyclopaedia.
- [12] www.securelist.com, «Examples and Descriptions of Various Common Vulnerabilities», Encyclopaedia.
- [13] www.securelist.com, «Attacks on banks», Roel Schouwenberg, 23.10.2008.
- [14] www.securelist.com, «The botnet business», Vitaliy Kamluk, 13.05.2008.
- [15] www.securelist.com, «”Instant” threats», Denis Maslennikov, Boris Yampolskiy, 27.05.2008.
- [16] www.securelist.com, «Skype and Corporate Network Security», Infowatch, SecurityLab.ru, 4.04.2007.
- [17] www.securelist.com, «The evolution of self-defense technologies in malware», Alisa Shevchenko, 28.06.2007,
- [18] www.securelist.com, «Rootkits evolution», Alisa Shevchenko, 28.08.2008
- [19] www.dnt-lab.com, «Rootkits and Antirootkits», Vitaliy Kiktenko, 2009
- [20] www.wikipedia.com
- [21] www.alexa.com

A Remote Robotic Laboratory Experiment Platform with Error Management

Chadi Riman¹

¹Computer Engineering, Fahad Bin Sultan University, Tabuk, KSA

Abstract

Remote control of experiments is gaining more importance in training and education. However, remote real-time training on instruments programming still have some unresolved problems such as error management. In this paper, a platform for training students on system's control by Tele-Programming is presented. Programming sessions can be done by the trainee at many levels of control with built-in error management in order to avoid system freezing or malfunction. We showed an illustrative application: programming navigation control of a mobile robot in the presence of obstacles using fuzzy control.

Keywords: *Remote lab experimentation, HCI, Robotics, Computer Simulation.*

1. Introduction

Remote Experimentation is a distant control of an experimental setup accessed from different places and by different users (figure 1). The application carried out in Remote Experimentation can vary from a simple demonstration where interaction between the student and the experiment is on a simple level (view only), to a complex application where the student has more control over the experiment. The first case is safe with limited teaching possibilities. The complex case has more teaching advantages but it also has malfunction risks due to a higher probability in committing errors by students. This type of training can be improved if it is accompanied with a tutorial and simulation software to be used before the real experiment.

Tele-Programming is a remote laboratory platform in which control is done using program files exchange. These files are usually text files of small size which requires very low bandwidth. This type of remote experimentation is therefore suitable to low speed networks. This study evaluates some existing major platforms in tele-programming and suggests an improved low-cost platform with three programming levels based on student and course levels. For illustration purposes, this platform will be used for remote training on a mobile robot in a fuzzy logic environment.

The main problem in self programming is the need of a tutor either in the local place or in the remote

lab, which can be replaced by a tele-tutorial system [1-3]. This is also achieved in our platform by an error management module to identify errors, notify the student, and prevent system malfunction. Our idea is to support the training of the students by allowing failures in the experimentation. The system can manage different kind of failures and then send feedback to student. Moreover, all processes are built using free software.

In this paper, section 2 presents an analysis of some existing tele-experimentation training platforms used in robotics. The need of different programming levels in training is discussed in section 3. The suggested platform architecture is given in section 4. Section 5 presents the robot and its fuzzy controller module which is used for training in our suggested platform. Error managements and Simulation modules are respectively described in sections 6 and 7. A case study is given in section 8 and concluding remarks are given in section 9.

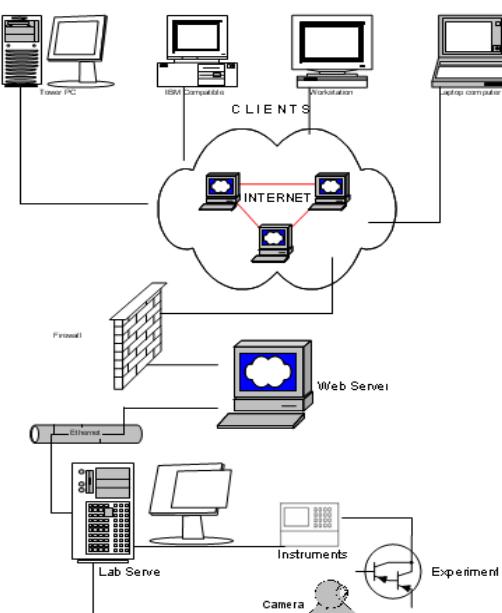


Figure 1. Remote Experimentation

2. Existing Robotics Platforms

Distance Learning platforms on robotics (DLR) are mainly concerned with learning, evaluation and security. In this section, the characteristics of some Existing DLR platforms [4-7] will be studied and used in the analysis of our suggested one.

The Open Learning platform presented in [4] uses MS NetMeeting and Matlab real-time tools for control purposes. It has the following characteristics:

- Software development is simple, reducing expenses and minimizing faculty work.
- Existing off-the-shelves freeware software, are used.
- No need for Web-enabled Interfaces

This platform uses the *Learning-by-Doing* methodology but it doesn't present any solution related to safety problems.

The platform described in [5] uses the *Active-Learning* methodology. The design provides the remote user with the perception of reality which is due to the use of Learning Objects. Safety is provided by running a VRML (Virtual Reality Module Language) simulation before executing the control program, which may reduce any damage due to some manipulation errors. Control learning is limited to changing pre-defined controllers with their parameters.

In the Platform described in [6] uses the *Learning by Tele presence*, the *Learning-by-Doing* or *Active-Learning*. The remote user indicates obstacles to be avoided and the target to be reached. A simulation module based on a potential field algorithm draws the path and a control program runs in order to follow it. Infrared sensors are installed on the robot to provide security and increase autonomy. This system has some deficiencies due to the limited interaction with the user and limited experience of the student.

The platform developed in [7] uses all learning methodologies previously mentioned for remote control of Lego mobile robots. The student uses Matlab/Simulink in order to design a controller to track a user defined trajectory. The control program is next transmitted to the server and executed. Three lights have been placed on the top of the robot in order to detect position and direction by means of a camera. A safety mechanism stops the experiment whenever the robot reaches a forbidden region. The control accuracy is based on a predefined model of the robot dynamics and needs to be changed with the physical environment.

The platform suggested in our work focuses on displacement control of a mobile robot. It benefits from useful techniques and methodologies

developed in previous studies and add improvements to them.

3. Programming Protocol

The suggested protocol for training on programming shows that there are three levels of tele-programming (Figure 2):

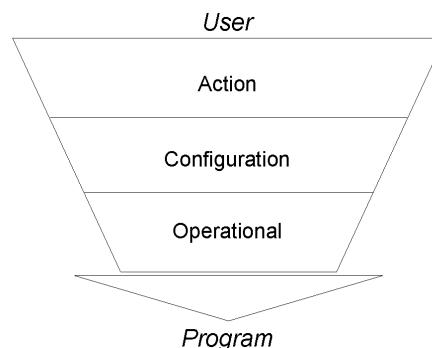


Figure 2. The three programming levels

- 1) **The direct action level** is for introductory courses where a program corresponds to a sequence of instructions (advance, turn...). The first set of experiments is made at this level where the student will be learning the system specifications (functionalities and workspace) and the basic programming structures (loops, iterations...).
- 2) **The configuration level** is used for more specialized courses where programming integrates internal specifications. The student can modify internal parameters. This facilitates the understanding of control principles.
- 3) **At the operational level** the student learns how to control system by advanced programming. The program file is transmitted for execution on the server. The program can be directly executed either on the robot or its virtual simulated model. Virtual reality, for design engineer workshop, allows transmitting control instructions without syntax constraints. Contrarily, a textual way for transmission of control instructions needs more abstraction in the programming phase.

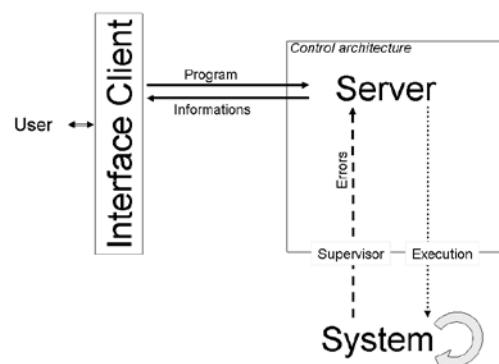


Figure 3. Process of failure supervision

An error management module (Figure 3) is included to deal with failures produced at any of the described levels. This module prevents deadlocks [8] and system freezing, and transmits feedback information which improves the student learning. It runs on the server in order to validate syntax and semantics of programs. It also supervises program execution like in a watchdog concept. In addition, it controls the execution time and odometrical task limits. Therefore, students can test their programs without human tutor because of the feedback provided due to an abnormal response of the robot.

4. Architecture of Suggested Platform

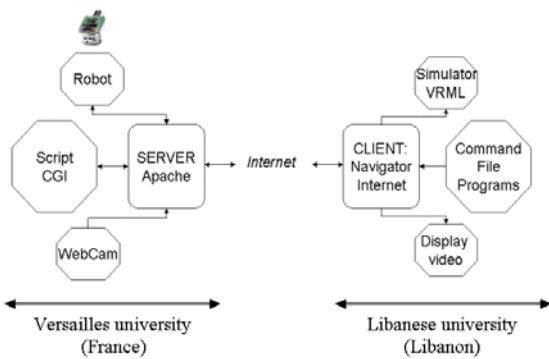


Figure 4. The platform architecture

The suggested platform uses client-server architecture (figure 4). The server provides training information such as description of programming methods, programming steps, and programmable devices. It is also responsible of communication with the remote client and with the robot.

Interoperability among users is achieved by using an Apache server and an HTTP browser. Apache server is very stable and widely available. The server software includes the modes of displacements, the reachable workspace of the robot, the response times of the actuators and the sensors as well as the sequence of procedures required for a given task. Interactions between the user interface and the platform is based on sending predefined instructions to the robot, and receiving its status (Figure 5).

This interaction is performed in order to explore the robot parameters and them to program its operation. Exploring parameters allows the understanding and the comparison between structures and sensitivity ranges of various controllers. Programming provides the ability of integrating the controller in a programming language. The program has to follow a predefined structure in order to be compatible with the error management module. In case of errors, the system sends an error report and reinitializes the platform for restarting the exercise.

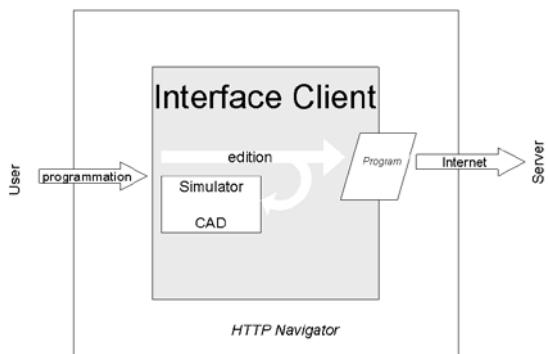


Figure 5. Interface process

The Apache server communicates with the robot through a Common Gateway Interface (CGI) using C language. The user program is transmitted by the server to the CGI which transmits information to the robot according to the programming level. The robot is connected to the server through an RS232 HF serial port. Next, the CGI Program returns results to the client. A Webcam connected to the site returns feedback information on the framework environment.

5. The Robot and its Fuzzy Controller

The Khepera robot (Figure 6) has a diameter of 55 mm and a height of 30 mm. It is controlled by a Motorola processor 68331 with 256 KB of RAM and 18 KB of ROM. Its motion is due to two DC motors with encoders. It is also provided with 8 infra-red sensors. Because of its modularity and its important number of options, this product is widely used by researchers and teachers. It can be programmed in GNU C with LabView or Matlab.



Figure 6. Khepera mobile robot

At the first level of programming, operations are based on actions available in the robot integrated libraries. These actions are simple: advance, turn, pause, avoid, measure sensors values... The user is therefore able to carry out a simple task in a complex environment.

At the configuration level, uploaded programs use fuzzy controllers that are structured with heuristic features close to human actions. Configuration is done on parameters relative to a classification of

entries (data generated by sensors or robot status) and on fuzzy rules.

At the operational level, the transmitted source code implements the algorithmic structure and the robot control.

6. Error Management

For a navigation task towards a goal in a complex space, two situations are considered as [8]:

- Blocking (figure 7): a bad choice of control parameters (control law or range of operation of proximity sensors) may cause a freezing in operation during obstacles crossing.
- Vagrancy (figure 8): the error causes divergence from the goal.

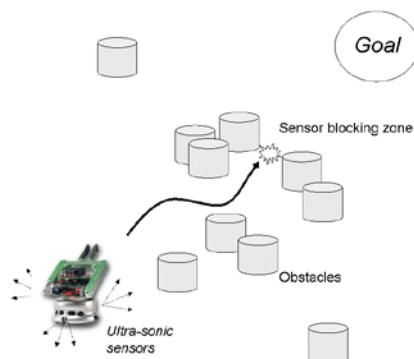


Figure 7: blocking situation

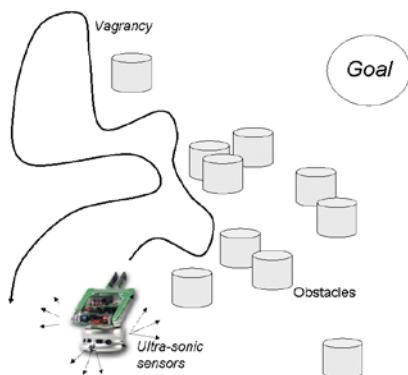


Figure 8. Vagrancy situation

A first type of errors is due to a bad parameters configuration. These fuzzification parameters on inputs and on selected rules of inference, lead to freezing during a simple navigation task in presence of obstacles. This allows students to understand the structural and logical definition of the configured fuzzy subsets. A simple modification of the number of subsets or the use of a symmetrical or asymmetrical triangular structure makes a navigation task successful or not.

The second type of errors occurs because of a programming problem (infinite loops, interruption,

exception...). For example, infinite loops are suspected when blocking occurs without being associated with sensors configuration. Memory allocation problem is detected by the mechanism of integrated watchdog: the robot must send periodically information on the execution status. The training at this stage is concerned with the algorithm, its implementation, and its execution. Tools for coding and reliability analysis could be used during this learning stage [9].

7. Simulation Module

The aim of the simulation module is to perform a local test on the student program before being uploaded. The simulation phase has the following objectives:

- 1- Understanding the robot specifications and functions.
- 2- Testing programs without risks or without using the robot.

Simulation is done in a VRML (Virtual Reality Modeling Language) environment, which is a 3-dimensional scene description language installed on the client station. First, the server transmits various documents to the client browser which interprets these documents with a possible help of plug-ins. In case this interpretation fails, these documents are transmitted to the VRML.

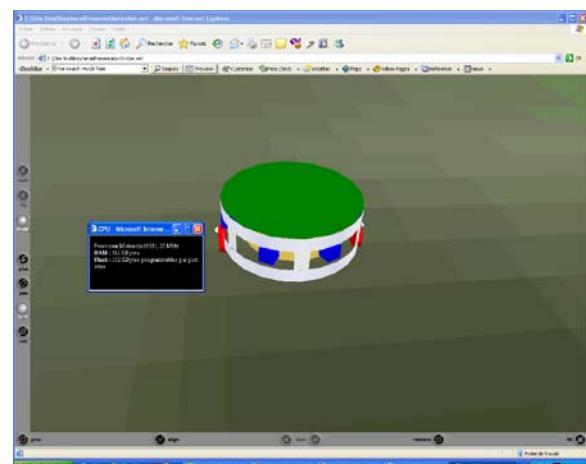


Figure 9. VRML Simulator

In a first step, the simulator visualizes the robot in order to present its functionalities (figure 9): motion, perception and processing devices. Students, at this stage of training, can study these functions relative to mechanical, electronic and data processing concepts:

- Perception: Allows studying the principles, types, ranges, and positioning measurements of infra red sensors.
- Motion: Allows to study actuators specifications and types, as well as kinematics and power module devices.

- Processing: Allows studying processor characteristics (memory, temporal diagrams...).

In a second step, VRML simulator serves as a trial stage for programming. Students can study the behavior of the simulated robot in the action programming level.

8. Case Study

Experiments were carried out between Lebanon and France for various levels of programming. This allows evaluating the risks of operation due to the reduced quality of connection. The system was installed on a computer server at the LISV lab (Versailles University). The server was connected by wireless link to the robot. A visual feedback was used during the experiment in order to simplify operations and to improve learning process.

The test task was a simple navigation with obstacle avoidance. The Khepera mobile robot was installed in a limited enclosure for a safe displacement. Examples are given for various types of users (specialized engineers, students without or with limited experience) operating at different programming levels.

a. Direct Action level

This level is tested with high school students. They have to program the robot in order to test sensors and control strategies based on program files composed of predefined orders. A list of commands accessible from the client site in action mode is given in Appendix.

Table 1 summarizes the common errors which are mainly due to bad values of PID parameters or other parameters relative to uploaded actions. Errors are always detected by means of mobility status.

| Cause | Effect |
|---|---|
| Bad choice of PID terms, Often Integral term too high | Instability and Vagrancy situation |
| Displacement without sensors feedback | Blocking situation because of security process occurrence |

Table 1. Common Students errors in the Task level

b. Configuration level

At this level, a student tests his/her knowledge of control based on fuzzy logic. Thus, to simplify illustration of obstacle avoidance management, outputs of sensors will be fuzzified in terms of detected distance and orientation. An example in figure 10 can be adapted by modifying the structure of the fuzzy subsets: i.e. by modifying each subset limits FSij. This information will modify the mobile robot behavior by affecting its sensitivity to sensors

values. In the same way, it is possible to modify the rules of fuzzy inferences. This technique is based on the Sugeno-takagi approach [10], in which conclusions of the rules are singletons Si (Figure 11). The file of the program on the configuration level contains the parameters FSij and Si. Table 2 summarizes common errors that may occur at this level.

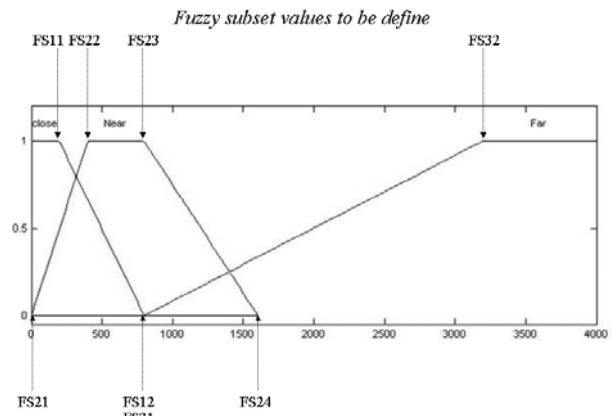


Figure 10. Example of fuzzy sub sets

| | close | near | far |
|----------|-------|------|-----|
| negative | S1 | S2 | S3 |
| zero | S4 | S5 | S6 |
| positive | S7 | S8 | S9 |

Figure 11. Inference rule table with singletons

| Cause | Effect |
|---|--|
| FSij terms shift on the left space (obstacle detection too close) | Blocking situation |
| FSij terms shift on the right space (far obstacle detected) | Vagrancy situation |
| Si terms in the inference rule table are symmetric | Blocking situation by opposition between two obstacles |

Table 2. Common Students errors in the Configuration level

c. Operational level

At this level, students must carry out the complete compilation in order to upload the program file. To simplify this phase, students may use a library of programs made of classical Khepera instructions and stored in the server. It is possible to proceed to simple actions like flickering LEDs of the robot, reading values of the infra-red sensors or writing a program containing navigation instructions with

obstacle avoidance such as the one based on the principle of Braitenberg [11]. Uploaded files in this level must respect the S37 format [12]. The robot being provided with a microcontroller Motorola 68331, programs written in C language, must be compiled using an appropriate cross-compiler like KTproject compiler under Windows. In this mode, the program is automatically executed after it is uploaded on the robot.

Table 3 contains common errors that may occur at this level.

| Cause | Effect |
|----------------------------|--|
| Braitenberg simple control | Blocking situation by antagonism between two obstacles |
| Bad choice of control law | Vagrancy or blocking situations |
| Infinite loop problem | Hazardous direction vagrancy situation |
| Memory allocation problem | On place blocking situation |

Table 3. Common Students errors in the Operational level

9. Conclusion

In this study, a platform for training on systems control by tele-programming was presented. According to the student academic profile, and in order to improve self-learning process, a protocol including various training levels and error management was validated.

This approach was evaluated based on experiments carried out through Internet connection between Lebanon and France. The experiment for the student learning phase was a mobile robot programming using different levels: from predefined order to fuzzy logic programming. During this stage, we tested the error management to improve knowledge feedback for students.

The realized platform uses free software and works with low bandwidth Internet connection. Our idea is to answer typical student constraints in term of flexibility and cost.

After this first trial supported by CEDRE program¹, different modes of operation will be introduced in a future work. A suggested mode of operation is to configure the system to be used by handicapped persons. For this purpose, the learning platform would integrate an adaptable and easy configurable man machine interface.

¹ "Coopération pour l'Évaluation et le Développement de la Recherche", Cooperative program between Lebanese and French Governments.

Appendix

List of commands accessible from the client site in action mode

A: for parameters configuration of the PID velocity controller: proportional (K_p), integral (K_i) and derivative (K_d).

Default values of these parameters are: $K_p=3800$, $K_i=800$, $K_d=100$.

C: Indicates to the position controller the absolute position to reach. The robot trajectory will produce three phases: acceleration, constant velocity and braking.

D: Configures the velocity of both wheels. The unit is pulse/10ms which corresponds to a velocity of 8mm/s, its maximum value is 1m/s.

E: Reads the instantaneous wheels velocity.

H: Reads the position 32 bits counter of each wheel.

I: Reads on 10 bits the value of the analog input relative to the selected channel. Its maximum digital value corresponds to an analog input of 4.09 Volts.

Channel 0: Detects battery status.

Channel 1: Measures instantaneously the intensity of the reflected light.

Channel 2: Measures instantaneously the intensity of the ambient light.

Channels 3, 4, 5: Free channels to be used by analog inputs: 36, 37 and 38 of the KBus.

Channel 6: Reads the Khepera current consumption in mA.

J: Configures velocity profiles using motion parameters (Maximum velocities and accelerations).

N: Reads on 10 bits each value of all eight proximity sensors.

P: Configuration of the desired amplitude of the PWM relative to each wheel. The modulation factor varies between 100% lagging and 100% leading with 0% as middle range. These values correspond respectively to +255, -255 and 0 as binary reading.

Acknowledgments

The author thanks the Assistance and Handicap team headed by Dr. Eric Monacelli in LISV research laboratory of Versailles University (France) for their help in applying this work.

References

- [1] A. Böhn, K. Rütter, B. Wagner, “Evaluation of tele-tutorial in a remote programming laboratory”, American society for engineering education annual conference, 2004
- [2] C. Riman, A. El Hajj and I. Mougharbel. “A Remote Lab Experiments Improved Model”, International Journal of Online Engineering, Volume 7 Number 1, 2011.

- [3] I. Mougharbel, A. El Hajj, H. Artail, and C. Riman, *Remote Lab Experiments Models: A Comparative Study*, International Journal of Engineering Education, Volume 22 Number 4, 2006.
- [4] N. Swamy, O. Kuljaca, F.L. Lewis, “*Internet-Based Educational Control Systems Lab Using NetMeeting*”, IEEE Transactions on Education, VOL. 45, N°. 2, May 2002.
- [5] D. Fabri, C. Falsetti, S. Ramazzotti, T. Leo, “*Robot Control Designer Education on the Web*”, Proceedings of the 2004 IEEE International Conference on Robotics and Automation, New Orleans, April 2004.
- [6] F.D. Von Borstel, B.A.. Ponce, J.L. Gordillo, “*Mobile Robotics Virtual Laboratory Over the Internet*”, Proceedings of the Fourth Mexican International Conference on Computer Science (ENC’03), 2003.
- [7] F. Carusi, M. Casini, D. Hattichizzo, A. Vicino, “*Distance Learning in Robotics and Automation by Remote Control of Lego Mobile Robots*”, Proceedings of the 2004 IEEE International Conference on Robotics and Automation, New Orleans, April 2004.
- [8] A. Smirnov, E. Monacelli, S. Delaplace “*Single adaptation mechanism for collaborating multi-robots*”, ICINS’2002 conference, St Petersburg, Russia, 2002.
- [9] M. H. Klein, and al., “*A Practitioners' Handbook for Real-Time Analysis: Guide for Real-Time Systems*”, Boston, Kluwer Academic Publishers, 1993.
- [10] F. Abdesselmed, K. Benmahammed, E. Monacelli, “*A fuzzy based reactive controller for a non holonomic mobile robot*”, Robotic and autonomous systems journal, pp 31-46, 2004
- [11] V. Braitenberg, “*Vehicles: Experiments in Synthetic Psychology*”, MIT Press, 1984
- [12] K-Team, “*User's Guide for Khepera mobile robot*”, <http://www.k-team.com/download/khepera.html>

Chadi Rimani received his Bachelor of Engineering and Masters of Engineering degrees both in Computer & Communication Engineering from the American University of Beirut (AUB), Lebanon in 1994 and 2004, respectively. He finished his PhD degree at the University of Versailles (UVSQ), France in January 2008. He worked from 1994 to 1998 in the software engineering domain, and was the IT manager in AinWazein Hospital, Lebanon from 1999 to 2008. He is currently Assistant Professor at Fahad Bin Sultan University, Tabuk, KSA. His research interests include software systems for handicap rehabilitation and remote engineering education.

Performance of Distributed System

Abdellah Ezzati¹, Abderrahim Beni hssane² and Moulay Lahcen Hasnaoui³

¹LAVETE laboratory, Mathematics and Computer Science Department, Sciences and Technics Faculty Settat, Morocco.

^{2,3}MATIC laboratory, Mathematics and Computer Science Department, Sciences Faculty, University Chouaïb Doukkali University, El Jadida, 24000, Morocco.

Abstract

Many distributed systems are still too large to be handled. Thus, it's important to find techniques that can be used to extend the size of the systems that can be verified and analyzed. In this paper, we study the qualitative and quantitative performance of the distributed systems that can be interacting with each other by using Temporized Stochastic Petri Net (TSPN) [5]. We consider then the composition asynchronous operation for deducing properties of a global distributed system from the properties of its components [2, 9]. Introduction of a structured interface net allows us to preserve properties of components in the global system.

Keywords: *Distributed Systems, Temporized Stochastic Petri Net, Liveness, Boundedness and Interface Net.*

1. Introduction

Using a standard web browser, the user can access information stored on Web servers situated anywhere on the globe. This gives the illusion that all this information is situated locally on the user's computer. In reality, the Web represents a huge Distributed System (DS) that appears as a single resource to the user available at the click of a button. According to Leslie Lamport [10], a distributed system is defined as "one on which I cannot get any work done because some machine I have never heard of has crashed". This reflects the huge number of challenges faced by distributed system designers. Despite these challenges, the benefits of distributed systems and applications are many, making it worthwhile to pursue. Performance modeling and evaluation constitute an important aspect of the design of distributed systems. Performance models are mainly of two types : simulation and analytical. In this paper, we propose to use Temporized Stochastic Petri Nets (TSPNs) as an analytical models to study the conception and evaluation of performance of DS.

The paper is organized as follows : Section 2 outlines an introduction to TSPNs with an illustrative example. In section 3, we propose a composition of components via a

structure of interface in order to facilitate the conception and performance evaluation of DS. Section 4 presents the preservation of qualitative and quantitative properties of the global system. Finally, Section 5 concludes the paper.

2. Temporized Stochastic Petri Net (TSPN)

Petri nets have emerged as a prominent modeling tool of concurrent systems. A class of timed Petri nets called Temporized stochastic Petri nets (TSPN) are well suited for performance modeling. In the framework of TSPN and Extended stochastic Petri nets, several features of distributed system such as concurrency, non-determinism and synchronization can be captured in an elegant way. Informally, any TSPN comprises a set of places, a set of transitions, a set of arcs, an initial marking; a random variable, and a time interval are assigned to transitions.

In the TSPN representation of a distributed system, places represent logical conditions or resources in the system; transitions represent events or activities; arcs represent interdependencies among places and transitions; initial marking refers to the initial state of the system; and the random variables model the durations of various activities in the system. The evolution of a TSPN in time constitutes a stochastic process called the marking process. The use of TSPN as an analytical model is based on a set state analysis of the marking process.

Example of TSPN:

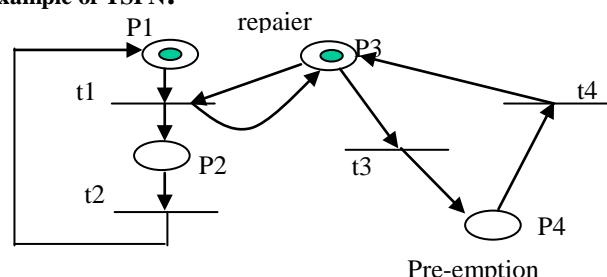


Fig1. Breakdown and pre-emption process.

Table 1: Distribution functions of example above

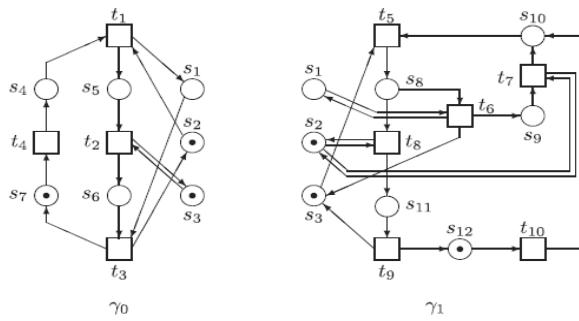
| Transition | Density |
|------------|------------------------|
| t1 | $\delta_{(x-10)}$ |
| t2 | $(1/100)\exp(-1/100)x$ |
| t3 | $\delta_{(x-4)}$ |
| t4 | $\delta_{(x-5)}$ |

A Temporized Stochastic Petri Net is called event graph [4] if and only if each place has exactly one transition in input and one transition in output.

3. Structure of interface between the components of DS Equations

3.1 Asynchronous composition of TSPN

Generally speaking, a composition operation of distributed systems combines two models into a single one whose behavior captures, in some sense, the interaction between that two models. There are two major ways of forming the composition of two models, synchronous and asynchronous, and for each of them different variants are known. In synchronous composition, the models run and synchronize on actions from a given set of actions. The main use of such an operation is for coupling a system with a tester which tests for the satisfaction of a given property. Opposite to the synchronous composition is the asynchronous composition, which does not assume any action synchronization but the systems may communicate via a set of shared variables (locations). The execution of such a system can be viewed as the interleaved execution of the components. For example of asynchronous composition of Temporized Stochastic Petri Net, we present Owicki-Lamport's Mutex algorithm [8] of composition of two models TSPN (reader and writer).



```

s1 = writer involved
s2 = writer detached
s3 = reader detached
s4 = prep1
s5 = prep2
s6 = writing
s7 = producing
s8 = pend2
s9 = failed
s10 = pend1
s11 = reading
s12 = using

```

Fig 2. Tow components of TSPN.

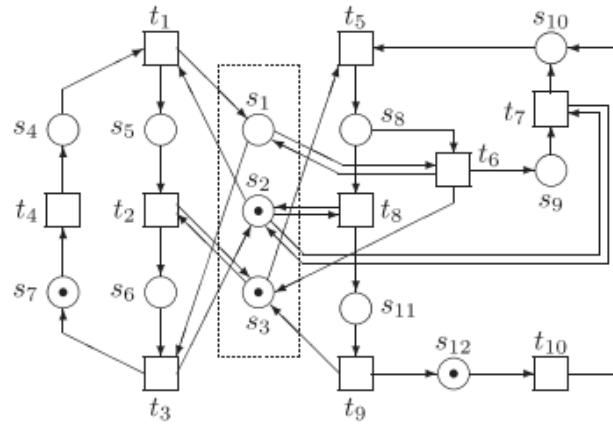


Fig 3. Composition of the two components above

It consists essentially of two sites: the writer and reader site, the first one to the left, and the second one to the right, of the dash box in figure. The net uses three flags: the flag writer detached (s_2) signals to the reader that the writer is presently not striving to become writing, the flag reader detached (s_3) likewise signals to the writer that the reader is presently not striving to become reading, and the flag writer involved (s_1) is just the complement of writer detached (for a detailed discussion about this net model the reader is referred to [11]). These two sites of the net in Figure 3 are connected each other by means of s_1 , s_2 and s_3 .

3.2 Structure of interface

In the case of asynchronous composition, the interface places are used by a component to interact with an environment [9]. During an execution, their content is updated by the system (component) or by the environment. The content of the internal places can be updated only by the component itself.

The aim of this section is to propose a new structure of interface between components in a distributed system that preserve some qualitative and quantitative properties in the global system.

The interface is defined as follow:

Let N_0 be a Temporized Stochastic Petri Net, with P_0 and T_0 the set of places and set of transitions respectively. T_0 is the union of the sets of T_{01} , T_{00} and T_{02} . We say that N_0 is a structure of interface S_0 if and only if

- a) (P_0, T_0) generate an event graph GE [4, 6].
- b) $\Gamma(T_0) \subseteq P_0$; that means, P_0 is the set of input and output places of T_0 .
- c) For all Ω_i , elementary circuit, of GE , the set of places of Ω_i is S -Invariant (i.e the total of the tokens in Ω_i remain invariant during the execution).

d) $\forall t \in T_0 / t^* \cap P_0 \neq \emptyset \Rightarrow f_t(x) = \delta(x)$, with t^* is the set of input places of the transition t , and f_t distribution fonction .

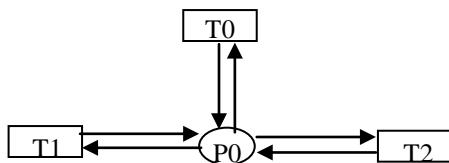


Fig 4. Interface S_0

In figure 5 below, the channel represents the interface S_0 between the producer (P) and consumer(C) with T_0 null.

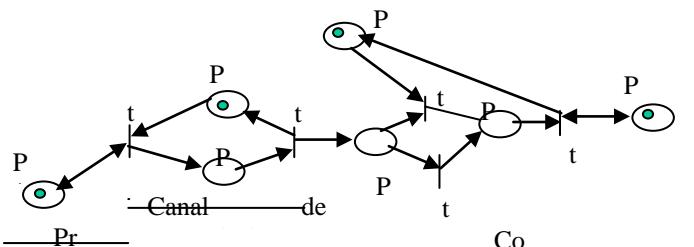


Fig 5. Canal de transmission as S_1

4. Preservation of properties by the composition of components via S_0

The goal of this paragraph is to deduce the properties of the global system from those of its components.

4.1 Qualitative properties

We are interested here by the liveness and boundedness properties.

The first property liveness permits to control the dysfunction of the distributed system. We say that a transition t is live if from any reachable marking, there is a reachable marking enabling t . If we expect that the activity modeled by t can always take place from any state, t should be live.

We take into account that we consider in our study of Temporized Stochastic Petri Net the case of the memory of trajectory [5].

Proposition

Let N_1 and N_2 two Temporized Stochastic Petri Net. The composition of N_1 and N_2 via the interface S_0 generates a Temporized Stochastic Petri Net N . Then, N_1 and N_2 are live $\Rightarrow N$ is live. In otherwise, the composition via the interface S_0 preserves the property liveness. This means that associate many components of the distributed system which are live via interface S_0 guarantees that the whole system is also live.

Proof: For each marking M in the N then M/P_i is a marking of N_i ; where M/P_i is a projection of M to N_i and P_i the set of places of N_i . Also, for each marking M_i of N_i there is a marking M in the N such that $M/P_i = M_i$. This result is deduced from 3.2 c). This remark is the main idea for proving the result looked for.

The second qualitative property is the boundedness which ensures that each place of the net is bounded (for example, there is no overflow in storage areas).

Proposition

Let N_1 and N_2 two Temporized Stochastic Petri Net. The composition of N_1 and N_2 via the interface S_0 generates a Temporized Stochastic Petri Net N . Then, N_1 and N_2 are bounded $\Rightarrow N$ is bounded.

Proof: Let p be a place of N , M_i a marking of N_i and k an integer such that $M_i(p) \leq k$. For each marking M in the N , $M/P_i(p) \leq k$ since the number of tokens is not changed in the N_i , then $M(p) \leq k$. This is deduced from 3.2 c).

4.2 Quantitative property

Execution path is a quantitative property which can be defined as (E_i, σ, E_f) ; where E_i is the initial state and E_f is a final, and σ a transition sequence from E_i to E_f .

Proposition: Let N_1 and N_2 two Temporized Stochastic Petri Net. The composition of N_1 and N_2 via the interface S_0 generates a Temporized Stochastic Petri Net N . Then, if (E_i, σ, E_f) is a path execution of N_1 (or N_2), then (E_i, σ, E_f) is an execution path of N .

Proof: The proof is deduced from the fact that, for any transitions' sequence σ in N , the projection of this transition' sequence σ/T_i is a transitions' sequence of N_i .

5. Conclusions

In this paper we've shown that the complexity of distributed system can be dressed by studing its components. By using the composition of Temporized Stochastic Petri Net, we have introduced a new structure of interface which preserve the qualitative and quantitative

properties of components in the global system. In the perspective, we shall present new structures of interface that will preserve other properties.

References

- [1] Y. Atamna Thèse “Réseaux de Petri Temporisé Stochastiques Classiques et bien formés” LAAS TOULOUSE 1994.
- [2] Y. Souissi “composition des réseaux de Petri et validation modulaire des systèmes distribués”, Revue Réseaux et Informatique répartie, vol 3 n°2 Hermes 1993.
- [3] L. Gallon Thèse “Le modèle réseau de petri temporisé stochastique” LAAS TOULOUSE 1997
- [4] G. Memmi: “Méthodes d’analyse de réseaux de Petri, Réseaux à file, et application aux systèmes réels ” Thèse de Doctorat Paris 6,1983.
- [5] Y. ATAMNA, R.CARMO, G.JUANOLE, F.VASQUES “Le modèle réseau de petri temporisé stochastique pour l’analyse quantitative et qualitative des systèmes distribués ” Rapport LAAS 92438, 1994.
- [6] Y. Souissi “Composition des réseaux de Petri” Thèse de Doctorat, Paris 6, 1990.
- [7] Sheng-Uei Guan, Sok-Seng Lim “ Modeling with enhanced prioritized Petri Net” Computer Communications, Volume 25, Issue 8, 15 May 2002.
- [8] F.L Tiplea, A. Tiplea “Petri net reactive modules” Theoretical Computer Science, Volume 359, Issues 1-3, 14 August 2006, Pages 77-100 Ferucio Laurențiu Tiplea, Aurora Tiplea
- [9] A. Ezzati “Etude des performances des systems distributes” 1er Work Shop On NEXT GENERATION NETWORK p 117-121 WNGN 2008
- [10] M. A. Marsan, A. Bobbio, S. Denatelli “Petri Net in performance analysis: An introduction” Journal and circuits, systems and computers 8(1): 119-158 (1998)
- [11] W. Reisig, Elements of Distributed Algorithms. Modeling and Analysis with Petri Nets, Springer, Berlin, 1998

A Generalized Framework for Energy Conservation in Wireless Sensor Network

V.S.Anita Sofia¹ , Dr.S.Arockiasamy²

¹Research Scholar, Department of Computer Applications, Karunya University,
Coimbatore – 641 114, Tamilnadu, INDIA

²Head, Information System, University of Nizwa

Abstract

A Wireless Sensor Networks (WSN) consists of spatially distributed autonomous sensors to cooperatively monitor physical or environmental conditions, such as temperature, sound, vibration, pressure, motion or pollutants. WSN contains a large number of nodes with a limited energy supply. A wireless sensor network consists of nodes that can communicate with each other via wireless links. Sensors are be remotely deployed in large numbers and operates autonomously in unattended environments. One way to support efficient communication between sensors is to organize the network into several groups, called clusters, with each cluster electing one node as the head of cluster To support scalability, nodes are often grouped into disjoint and mostly non-overlapping clusters. This paper deals about the frame work for energy conservation of a Wireless sensor network. The frame work is developed such a way that the nodes are to be clustered, electing the cluster head, performing intra cluster transmission and from the cluster head the information is transmitted to the base station.

Keywords: Wireless Sensor network, clustering, energy, cluster head.

1. Introduction

A wireless sensor network (WSN) consists of largely deployed sensor nodes which has limited battery power. Sensor nodes of WSN have the capability of self organizing the network. The transmission between the sensor nodes are done through wireless medium. WSN is

used to sense the physical or environmental conditions such as temperature, sound, vibration, pressure, motion or pollutants. Unique characteristics of a WSN include:

- Limited power they can harvest or store
- Ability to withstand harsh environmental conditions
- Ability to cope with node failures
- Mobility of nodes
- Dynamic network topology
- Communication failures
- Heterogeneity of nodes
- Large scale of deployment
- Unattended operation
- Node capacity is scalable, only limited by bandwidth of gateway node.

A sensor network is composed of a large number of sensor nodes that are densely deployed either inside the environment or close to it. The position of sensor nodes need not be engineered or predetermined. This allows random deployment in inaccessible terrains or hazardous environments. Some of the most important application areas of sensor networks include military, natural calamities, health, and home. When compared to traditional ad hoc networks, the most noticeable point about sensor networks is that, they are limited in power, computational capacities, and memory. Hence optimizing the energy consumption in wireless sensor networks has recently become the most important performance objective.

The main task of a sensor node in a sensor network is to monitor events, i.e., collect data, perform quick local data aggregation, and then transmit the data. Power consumption can hence be divided into three domains: sensing, aggregation, and communication. This paper proposes a new frame work to conserve energy of WSN, thereby the lifetime of the network is increased.

2. Motivation

The wireless network topology must be approached from a point of view different from that of a wired technology. In wireless sensor network (WSN), the definition of network technology is derived from the physical neighborhood and transmission power. Much of the related research in WSN is in the area of being mobile and battery powered.

Many literatures are concentrated on finding solution at various levels of the communication protocol, including being extremely energy efficient. Energy efficiency is often gained by accepting a reduction in network performance [1]. Low-energy adaptive clustering hierarchy (LEACH) [2][3] is a new communication protocol that tries to distribute the energy load evenly among the network nodes by randomly rotating the cluster head among the sensors. Sensor protocols for information via negotiation (SPIN) [4][5] is a unique set of protocols for energy efficient communication among wireless sensors.

Pottie has studied design issues and trade-offs that need to be considered for power-constrained WSNs with low data-rate links [6] and advocates “aggressive power management at all levels”, noting that the communication protocol is more helpful in reducing the power consumption than is optimizing the hardware.

3. Framework Design

Figure 1 shows the framework to conserve energy of the sensor node in the WSN. The principle of the framework is as follows

- (i) To identify the changing pattern of the sensor reading of the sensor node in the network.
- (ii) To identify and eliminate the redundancy of information in the base station
- (iii) To identify the failed nodes and assign their duties to some other nodes
- (iv) To combine the residual energy of the sensor nodes.

Sensor node is a device that receives and responds to a stimulus or signal. Sensors measure real-world conditions, such as heat or light, and then convert this condition into an analogue or digital representation.

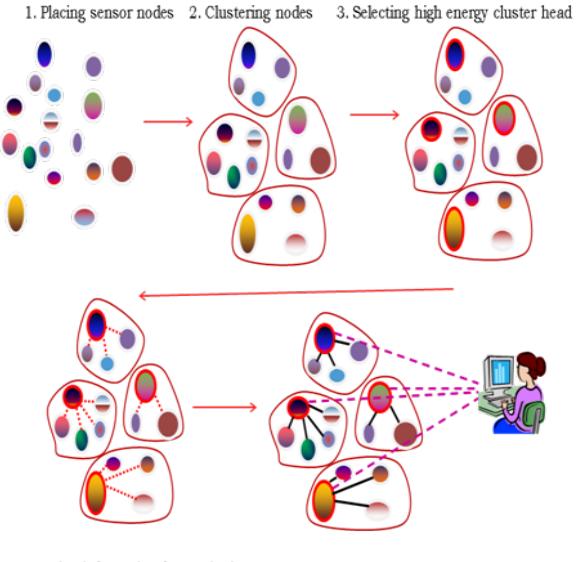


Fig. 1 Framework to conserve energy of sensor node 3.2 Equations

The steps are as follows

- A Placing sensor node
- B Clustering nodes
- C Selecting high energy cluster head
- D Sensing information from sub cluster
- E Transferring information to base station

A Placing the sensor node

The architecture of a sensor node is shown in Fig 3. The sensor node consists of the sensor head, ADC, Transmitter/Receiver and limited processor, memory and power source. The sensor nodes are deployed in a random manner.



Fig 2. Sensor Node

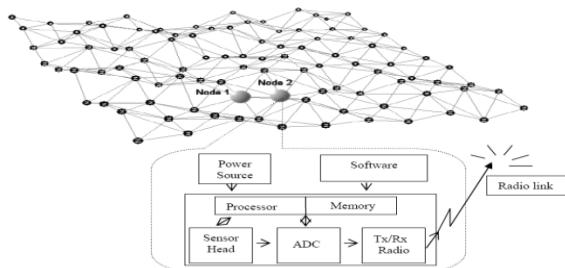


Fig 3 – Sensor Node Architecture

B Clustering Nodes

Clustering is a process by which the nodes are combined together to form a group.

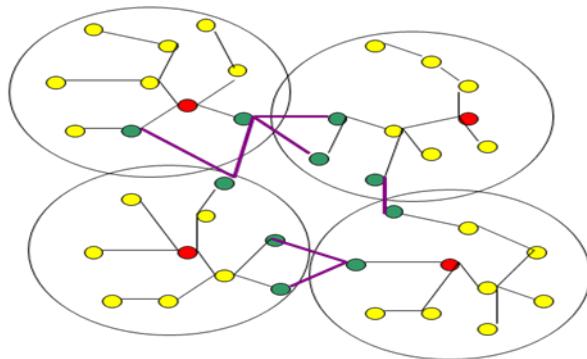


Fig 4. Clustering of sensor nodes

- Cluster member
- Cluster head
- Gateway node
- Intra-Cluster link
- Cross-cluster link

Fig 4 shows the clustering of sensor nodes into four groups. The cluster members interact with each other with the help of cluster head. Interaction is made through intra cluster link. The inter cluster communication is done using gateway node which acts as a mediator between two cluster heads with the help of a cross cluster link.

C Selecting high energy cluster head

Each cluster elects cluster head based on the energy level. The cluster head will not be stable and another cluster head is elected when the energy drains out.

D Sensing information from sub Clusters

Cluster head receives the sensed information from cluster members. The cluster head check for (i) the changing pattern of sensor reading of the cluster members and (ii) the redundancy of information obtained and eliminates the information in the cluster head.

E Transferring information to Base Station

In the network which is clustered, the cluster head transfers the data to the base station. The base station collects the data from all cluster head and checks for data redundancy and identifies the changing pattern of the sensor reading.

Psuedocode for Framework

1. $S = \{u_1, u_2, \dots, u_n\}$
2. Compute the distance d_j from u_i to u_j
3. If ($N(u_i)$) and distance d_j from u_i
 $C_i = \{u_{i0}, u_{i1}, \dots, u_{in}\}$
4. do
{

 Compute $E(u_{iI})$ where $I = 0$ to n
 CH_i is the cluster head of C_i if $E(U_{iI})$ is high
 U_{in} senses the information and passes the message to CH_i
} while ($E(u_{iI}) > E(u_{iJ})$)
5. CH_i aggregates the information and passes it to base station

The proposed frame work focuses on energy conservation of the sensor node in WSN. Clustering technique is used to combine the sensor node which is within a short distance into a group. A cluster head is selected for each cluster based on the energy level of that node. The main objective is to make only the cluster head communicate with the base station so that the remaining node can be put to a sleep state. This saves the energy of each node in the cluster. When the energy of the sensor node is drops, another node with high energy in the same cluster can be selected. Clustering technique also prevents dumping the same information into the base station.

4. Conclusions

This paper mainly deals about the frame work for energy conservation of Wireless Sensor Network. The frame work involves the steps which represents the different activities that are performed to conserve the energy of wireless sensor networks. Different protocols and algorithms are to be proposed to optimize the energy in the sensor network. The protocols and algorithm has to be

identified for the steps A to E which will optimize the energy in an efficient manner.

References

- [1] C. Patel, S.M. Chai, S. Yalamanchili, and D.E. Schimmel. Power/Performance trade offs for direct networks. In Parallel Computer Routing Commun. Workshop, 193 – 206, July 1997.
- [2] W.R. Heinzelman, A. Chandrakasan, and H. Balakrishnan. Energy- efficient communication protocols for wireless microsensor networks. In 33rd Ann. Hawaii Int. Conf. Syst. SCI., 2000.
- [3] A. Wang, W.R. Heinzelman, A. Chandrakasan. Energy – scalable protocols for battery – operated microsensor networks. In IEEE Workshop Signal Process. Syst., 483 – 492, Oct 1999.
- [4] W.R. Heinzelman, J. Kulik, and H. Balakrishnan. Adaptive protocols for information dissemination in wireless sensor networks. In Proc. 5th Annu. ACM/IEEE Int. Conf. Mobile Computing Networking (MobiCom'99), 174 – 185, August 1999.
- [5] J. Kulik, W.R. Heinzelman, H. Balakrishnan. Negotiation- based protocols for disseminating information in wireless sensor networks. In ACM MOBICOM, 99.
- [6] Rajesh Krishnan, David Starobinski, “Efficient Clustering Algorithms for Self-Organizing Wireless Sensor Networks”, Ad Hoc Networks, Elsevier Science Publishers, Volume 4, Issue 1 (January 2006) .



Mrs.V.S.Anita Sofia received the Master of Computer Application in 2001. She is currently working towards the Ph.D Degree in the Department of Computer Application, Karunya University, Coimbatore. Her research interest includes Computer Networks and Wireless Sensor Networks. She is life member of CSI.



Dr.S.Arockiasamy, Head, Information System, University of Nizwa. He received M.Sc, M.Phil and P.hD in Computer Science. He is specialized in Applications of Image processing. He has published a considerable number of Research papers and articles in various leading International journals and International Conferences. He is also Chief Editor of a Quarterly Computer Magazine CSI TIMES (Computer Society of India) and Voice of IT.

Platform for Assessing Strategic Alignment Using Enterprise Architecture: Application to E-Government Process Assessment

Kaoutar Elhari and Bouchaib Bounabat

Al-Qalsadi Research & Development Team, National Higher School for Computer Science and System analysis (ENSIAS),
Mohammed Vth University-Souissi
Rabat, Mohammed Ben Abdallah Regragui avenue, Madinat Al Irfane, BP 713, Agdal, Morocco

Abstract

This paper presents an overview of S2AEA (v2) (Strategic Alignment Assessment based on Enterprise Architecture (version2)), a platform for modelling enterprise architecture and for assessing strategic alignment based on internal enterprise architecture metrics. The idea of the platform is based on the fact that enterprise architecture provides a structure for business processes and information systems that supports them. This structure can be used to measure the degree of consistency between business strategies and information systems. In that sense, this paper presents a platform illustrating the role of enterprise architecture in the strategic alignment assessment. This assessment can be used in auditing information systems. The platform is applied to assess an e-government process.

Keywords: Strategic Alignment, Enterprise Architecture, Platform, Information System, Assessment Metrics.

1. Introduction

The information technology investment impacts positively on business performance. In order to reach a good impact, IT must constantly be appropriated to the business strategy. The strategic alignment (SA) has been studied since 1993 [1] how to coordinate the company's strategy with the information system strategy in order to improve the efficiency of information systems which support the company's business. Indeed, misaligned solutions have negative effects on the business level and, in turn, can reduce the value of services provided by the company.

On the other hand, the concept of enterprise architecture has come, more than twenty years ago, to address two problems: systems complexity and poor strategic alignment [2]. The enterprise architecture is the best way of representing information as a model illustrating the links between strategy, business and information systems [3].

Thus, this article presents a platform which assesses SA using the enterprise architecture. It is based on a set of metrics collected from several researches, classified according to the links between the layered structures proposed by enterprise architecture. The platform helps architects to improve the SA maturity

level by (a) analyzing the structure of enterprise architecture and (b) suggesting the effort to do in order to reach a better level.

This article uses many concepts of [4]. It is recommended to read it before [4].

The layout of this paper is as follows. The second section is devoted to EA and SA concepts; the third section presents an e-government process which will be used as an example to illustrate the platform functionalities. Finally, the fourth section presents the platform developed to support SA assessment by comparing the two versions of the platform. The conclusion and future work are presented in Section 5.

2. Strategic Alignment Evaluation

Many terms are used in the literature to refer to the SA [5]. Thus, a lot of synonymous of alignment are proposed: congruence, harmony, correspondence, coherence, and so on. The diversity of terms used involves the diversity of meaning given to the SA concept. [5] defines it as the correspondence between a set of components (e.g. between business process and system that supports them). [6] sees it as the act of applying information technology in harmony with the strategies, needs and objectives of the business. Some others study it as the harmony between architecture and software architecture of business processes [7]. Others consider the alignment between information systems and its environment [8]. And yet others are interested in aligning business processes and systems supporting these processes [9], [10].

In this article, we study the SA as harmony or correspondence between the company strategy represented by business processes and the systems supporting them.

2.1 Strategic alignment evaluation

Luftman proposes a framework for measuring the alignment between a company's strategies and the information technology strategies [11]. This framework is based on the foundations of CMM (Capability Maturity Model). He proposed five levels

of maturity from 1 (not alignment) to 5 (strong line). To evaluate SA in [11], six criteria were studied: communication, competency, governance, partnership, scope, architecture and skills.

[9] suggests an alignment strategy corresponding to a sequence of activities (represented by UML activity diagram). One of these activities is the evaluation of alignment. He proposed two metrics for this evaluation: Technological Coverage and Technological Adequacy. These two metrics are insufficient to assess the SA.

[12] proposes a framework for measuring alignment using a set of metrics classifying them according to four categories: intentional alignment, information alignment, functional alignment and dynamic alignment.

The purpose of this article is to assess the strategic alignment based on enterprise architecture

2.2 Strategic alignment evaluation based on enterprise architecture concepts

Enterprise architecture describes the enterprise structure. It represents all aggregate artifacts that are relevant to a company. There are many frameworks used to describe enterprise architecture such as [13], [14], [15] etc. But, it is often modelled as a layered organisation. The layers that are usually recognised in this context are the business layer, the application layer, the information layer and the technology layer.

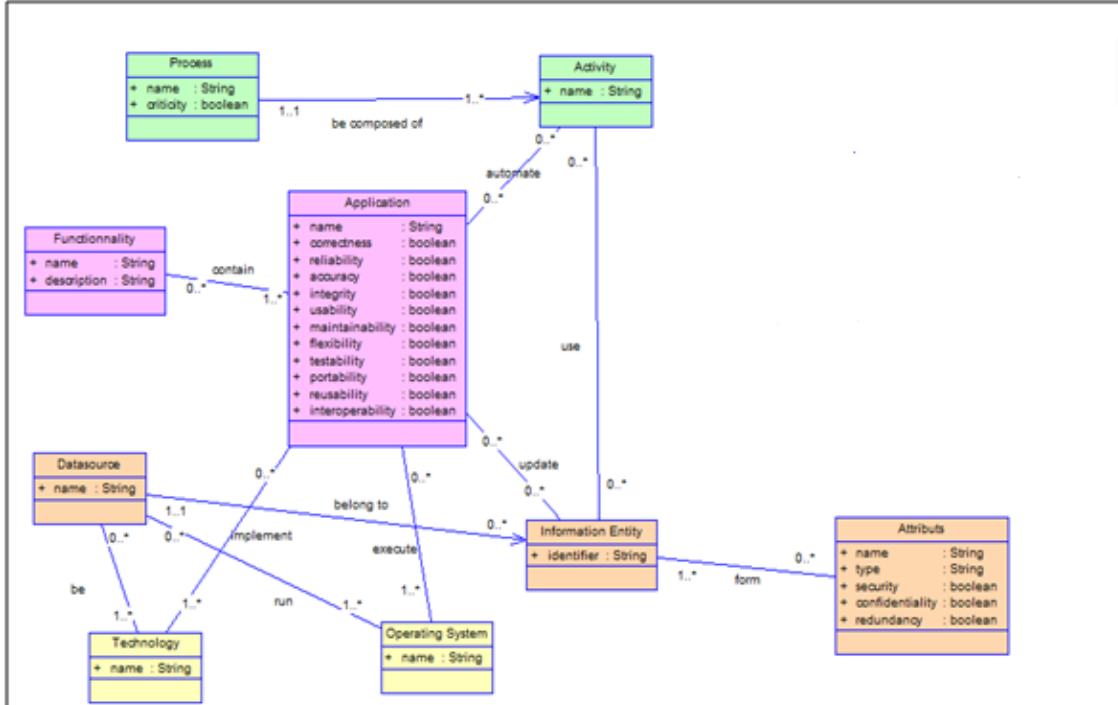


Fig1: Enterprise architecture metamodel

Many authors such as [17], [18], [19], [20], [21], [22], have associated the EA and SA concepts.

The definitions given to different layers in this paper are:

- The business layer represents the business of the company which is represented by a set of processes. Each process may consist of several activities (or sub processes). The processes or activities are supported by applications and use information entities. A process is characterized by its criticity.
- The application layer represents the application layer that automates the processes and activities. Each application has functionalities that meet the needs of the business processes. An application is described by a set of quality factors defined by [16]:
- The information layer is the data layer which is represented by information entities that can be found in data sources and which are formed by attributes. An attribute can be described by several qualifiers: secure, confidential, redundant.
- The technology layer is the layer of technical infrastructure including operating systems and technologies.

Figure 1 presents the metamodel used in this paper using a UML class diagram.

In this article, we are interested in detailed assessment of SA by examining the links between the various enterprise architecture layers.

Thus [19] develops an assessment of the SA from the links between the different EA layers, especially the business-application link and the business-data link. The metrics used in [19] use the quality criteria proposed by [16] on software quality and the notion of critical business processes that means a business process priority, which contributes to specific goals within the company and which is not superfluous [19]. Furthermore, studies such as [20], [21] present metrics for assessing the information system architecture.

In the same sense, [22] proposes a model of business of non-alignment with the information system by comparing it to medical science approaches. Thus, the authors suggest a set of cases where the business is not aligned with the information system. Then they present for everyone the organ system of the non alignment, symptoms, signs, syndromes and their etiologies. Then they suggest a diagnosis, therapy and prophylaxis.

The authors in [4] propose a strategic alignment maturity model based on enterprise architecture. The authors collected a set of metrics from several researches for each enterprise architecture internal link. They use the enterprise architecture metamodel presented in Figure1 and develop an evaluation tool for strategic alignment maturity which calculates metrics values and infers the maturity level for each layer's link. They propose five levels (chaotic, poor, average, good, very good). Their approach is represented on the diagram of the figure 2.

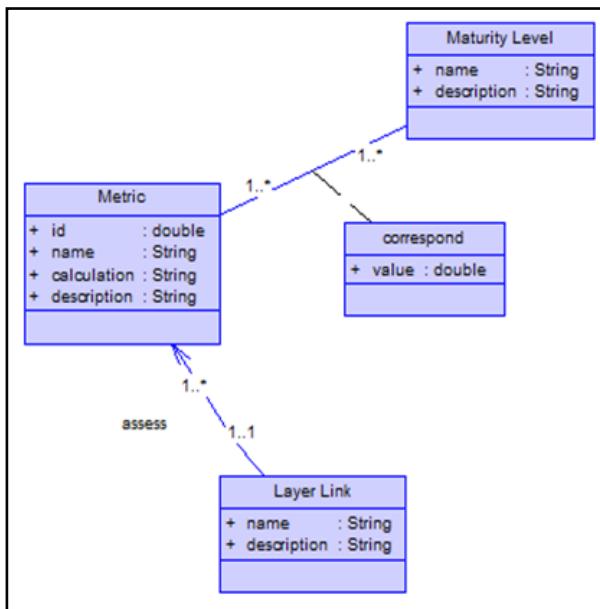


Fig2: maturity model diagram

This paper presents a platform supporting enterprise architecture modelling, calculating metrics values and

proposing where architectures have to change in order to enhance strategic alignment level. The platform is applied to an e-government process.

3. An E-Government case study

The process that will illustrate the assessment of strategic alignment in this article is surveys data production by using the characters automatic recognition. It was used in Morocco for the first time in the Census of Population and Housing 2004 [23]. Now, several surveys use the same process.

The process of data production using automatic characters recognition consists of several activities: We are going to illustrate the assessment of strategic alignment by the process: data capture which is used to produce data from questionnaires of the surveys.

Data Capture process contains 7 activities:

Activity 1: Receiving of questionnaires - The first step is to receive batches of questionnaires with an electronic file that indicates the identification number of each batch.

Activity 2: Scanning - It consists of scanning documents. Its aim is to computerize paper documents to enable and prepare the automatic optical recognition.

Activity 3: Character recognizing - It translates a group of points of a scanned image into characters readable by computer programs. It uses OCR (optical character recognition) technology.

Activity 4: Key correction and coding -The objective of this activity is to monitor, validate or correct the fields that were not recognized by the OCR with a sufficient confidence level or which have a coherence formula that indicates a suspicion of error.

Activity 5: Inter-questionnaires control and correction - This process was undertaken for each batch to verify that all questionnaires within a statistical area had been processed.

Activity 6: Quality control - The objective is to verify if the number of fields misread or misinterpreted in a document's batch is not above the targets set for production. The quality control method used was implemented to produce data with a minimum accepted error rate.

Activity 7: Data export - Data was exported in a text file format with a dictionary for further processing. This was the last step in the data processing system. The results were also exported in text files and their corresponding images of questionnaires to DVDs for backup and storage.

4. Platform Presentation

S2AEA is a Java platform dedicated to assessing strategic alignment using the concept of enterprise architecture. It contains two parts. The first part concerns the modelling of enterprise architecture and the second is dedicated to the strategic alignment evaluation.

The platform presented in this paper is the second version of S2AEA. The first version was presented in [4].

| Name | Criticity |
|------|-----------|
| ea1 | true |
| ea1 | true |

| Name |
|------|
| act1 |
| act2 |

Fig3: Description of enterprise architecture using S2AEA v1

Strategic alignment maturity is calculated based on this description. Maturity tables are generated by corresponding to each layer's link, a level of maturity.

The approach here is interested globally in an alignment overview between layers. The figure 4 is an illustration of this approach.

| Layers Links/Levels | Level1:Chaotic | Level2:poor | Level3:Average | Level4:Good | Level5:VeryGood |
|-------------------------|----------------|-------------|----------------|-------------|-----------------|
| Business-Application | | | | | |
| Business-Information | | | | | |
| Application-Information | | | | | |
| Application-Technology | | | | | |
| Information-Technology | | | | | |

Fig4: Strategic alignment Maturity level using S2AEA v1

4.2 S2AEA v2

The version which is proposed in this paper offers the opportunity to shape the enterprise architecture

graphically offering better ergonomics. The graphics incorporate the metamodel elements presented in fig3. The figure 5 illustrates how S2AEA (v2) models some activities of the process cited in section 3.

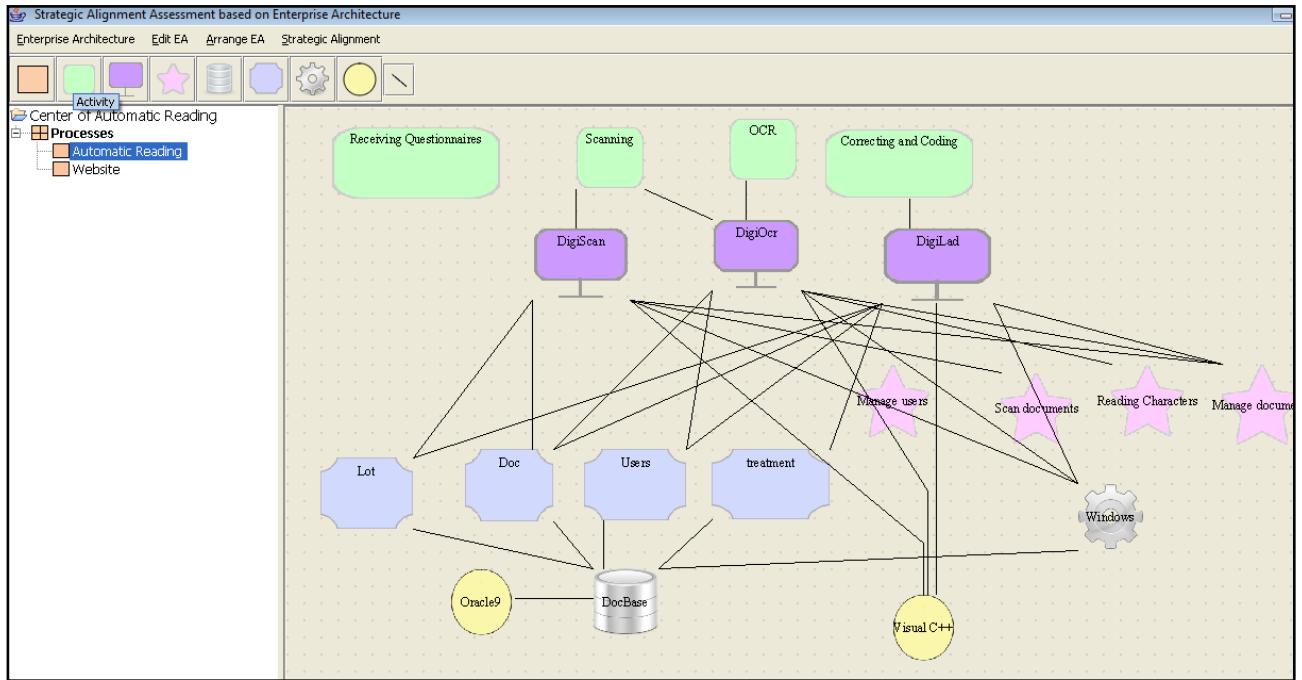


Fig5: Description of enterprise architecture using S2AEA (v2)

The table 1 contains elements constituting the figure5.

Table 1: S2AEA symbols

| Symbol | Name |
|--------|--------------------|
| | Process |
| | Activity |
| | Application |
| | Functionality |
| | Data source |
| | Information entity |
| | Operating system |
| | Technology |

The first version intends simply to calculate metrics and to infer the maturity level of each layer's link. It allows companies to locate their strategic alignment. S2AEA (v2) looks the alignment in more detail. It specifies information systems elements that affect the strategic alignment. This idea is based on 21 metrics collected in [4]. The v1 metrics targeted the whole layer while v2 metrics study case by case.

To illustrate an example of the use of S2AEA platform, we apply some metrics (M1 and M2) to the information system described in the figure5.

- M1: Number of activities not automated [4]
Indeed, each activity must be supported by an application in order to enhance alignment.
- M2: Number of applications supporting the same business process activity. [16], [18]
In fact, if a business process activity is supported by different applications; many problems can emerge:
- inserting the same data multiple times in different applications [21];
- Logging in multiple times, once for each application they need to access [21];
- etc

The figure 6 shows an example of two activities belonging to the process of automatic reading. It illustrates the role of the metrics M1 and M2.

After calculating metrics, the platform specifies the architecture elements that must be changed to reach a

higher alignment level (activities red colored in figure 6).

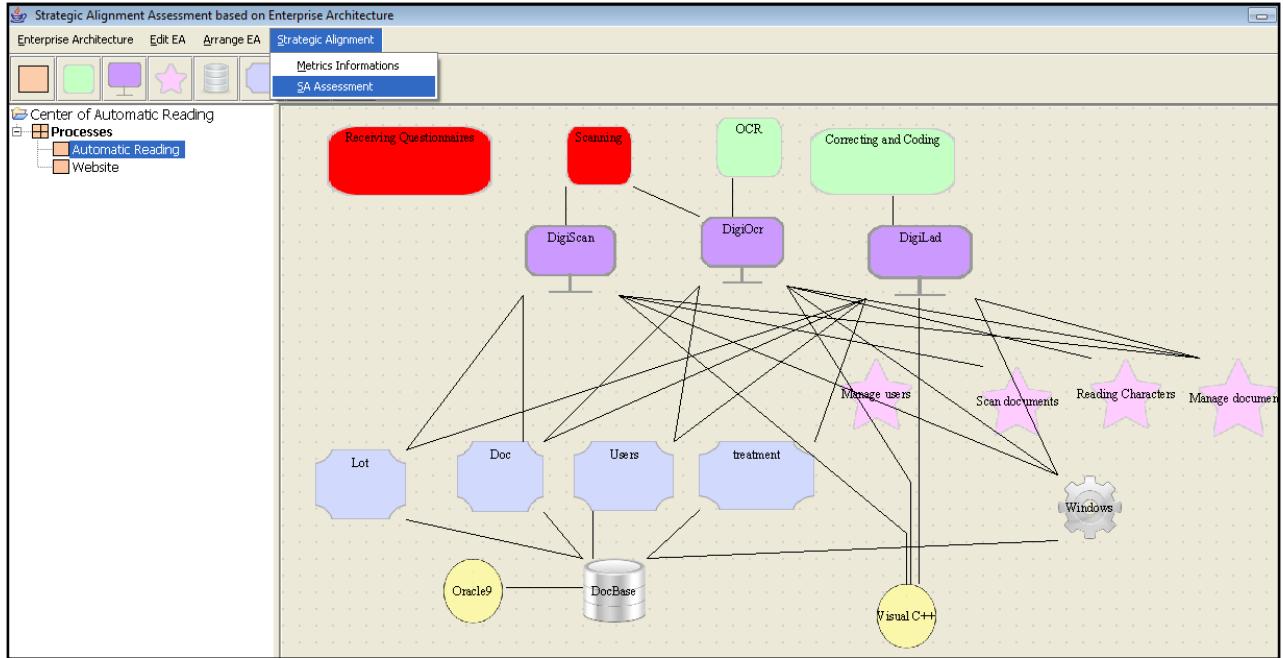


Fig 6: Strategic alignment assessment using S2AEA v2

The activity "Receiving questionnaires" harms the alignment in the sense that it is not automated (metric: M1). Architects should take it into account because it can be a real deficiency to deal with in order to reach

alignment. Indeed, non automated activities require more human resources and more time. Figure 7 shows the message given by the platform concerning the activity "Receiving questionnaires".

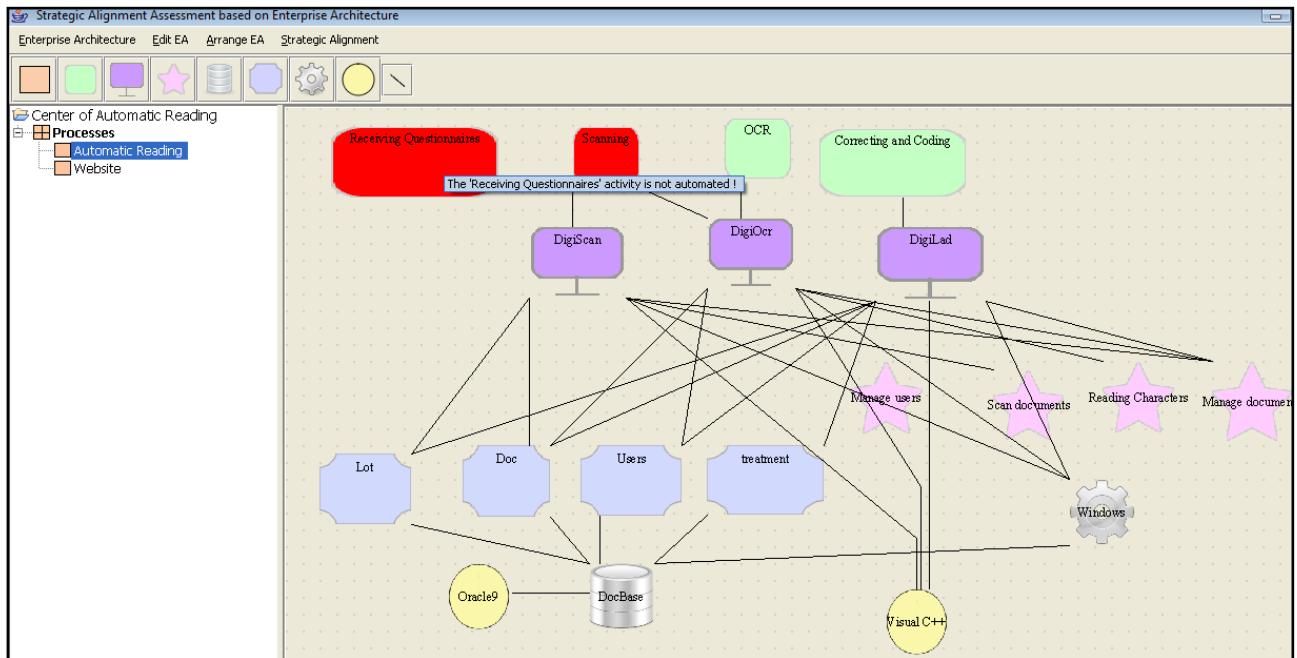


Fig7: Example of misaligned activity: activity not automated

On the other hand, figure 8 shows the problem raised by S2AEA concerning the “Scanning” activity. It takes into account the metric M2. “Scanning” activity harms the alignment because it is supported by three different applications (DigiScan, DigiOcr and

DigiLad). Indeed, an activity must be supported by a minimum number of applications: This can facilitate modification when the business process activity changes [19] and can reduce the need for distributed transactions across applications [20] [21].

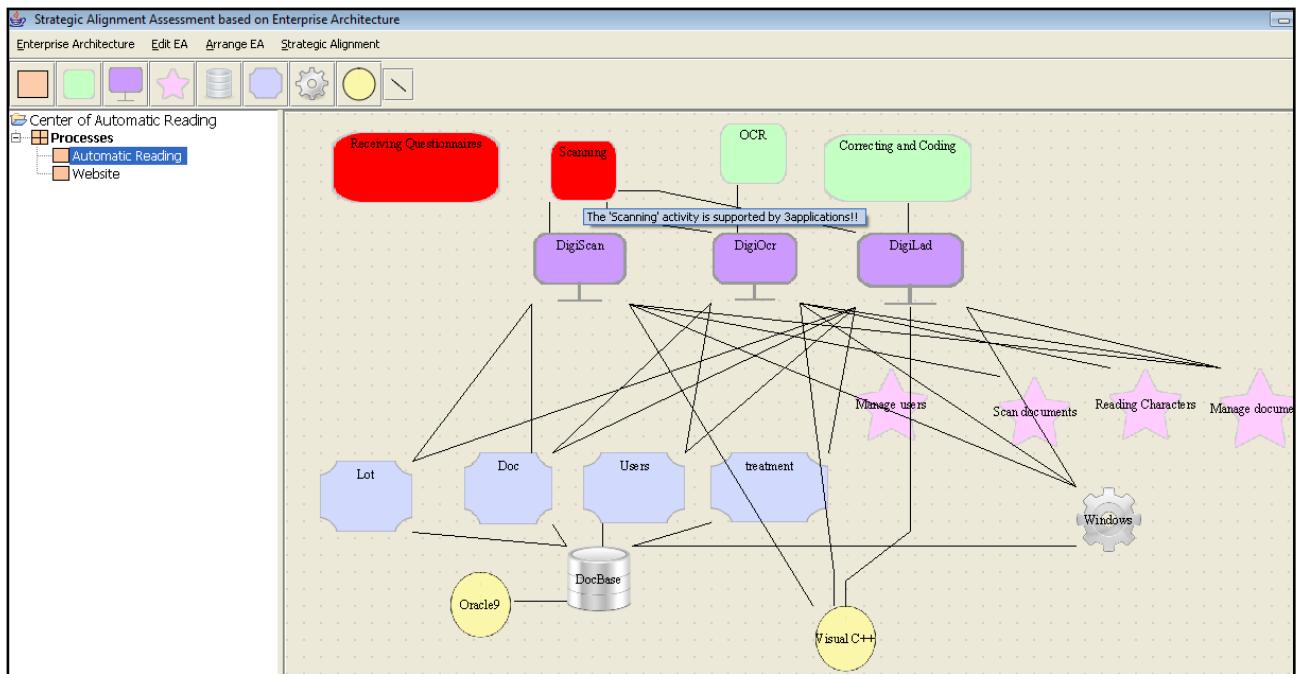


Fig8: Example of misaligned activity: activity supported by many applications

4. Conclusion

The article presents a platform S2AEA for assessing companies' strategic alignment.

The platform approach consists of using enterprise architecture concepts and its capacity to structure information system into layers. It is based on a set of metrics selected, studied and interpreted.

The platform proposed in this paper:

- graphically models enterprise architecture;
- calculates the corresponding metrics values;
- shows the information system elements harming strategic alignment;
- suggests the effort to do to reach a better strategic alignment level.

The very next steps in this research would be to improve S2AEA by adding more assessment metrics and by developing other platform functionalities.

References

- [1] J.C. Henderson, N. Venkatraman. "Strategic alignment: Leveraging information technology for transforming organizations", IBM Systems Journal, 1993, Vol. 32, No 1, 1993, pp.4-16.
- [2] R. Sessions, "A Comparison of the Top Four Enterprise-Architecture Methodologies". Microsoft Developer Network Architecture Center, 2007.

[3] R. Whittle, C. Myrick, "Enterprise Business Architecture: The Formal Link between strategy and Results", 2004, CRC Press, 2004.

[4] K. Elhari, B. Bounabat, "Strategic Alignment Assessment Based on Enterprise Architecture", Proceedings of the International Conference on Information Management and Evaluation (ICIME), 2010, pp.179-187.

[5] G. Regev, A. Wegmann, "Remaining Fit: On the Creation and Maintenance of Fit", Proceedings of BPMDS Workshop on Creating and Maintaining the Fit between Business Processes and Support Systems, 2004, Vol., pp.131-137.

[6] R. Papp, "Alignment of Business and Information Technology Strategy: How and Why?" Information Management, 1998, Vol. 11, pp. 6-11.

[7] R.J. Wieringa, H.M. Blanken, M.M. Fokkinga, P.W.P.J. Grefen, "Aligning application architecture to the business context." Conference on Advanced Information System Engineering (CAiSE 03), 2003, Vol., pp. 209–225.

[8] G. Camponovo, Y. Pigneur, "Information Systems Alignment in Uncertain Environments", IFIP International Conference on Decision Support Systems, 2004.

[9] T. Bodhuin, R. Esposito, C. Pacelli, M. Tortorella, "Impact Analysis for Supporting the Co-Evolution of Business Processes and Supporting Software Systems", in CAiSE Workshops on Creating and Maintaining the Fit between Business Processes and Support Systems, 2004, Vol. 2, pp. 146-150.

- [10] P. Soffer, "Fit Measurement: How to Distinguish Between Fit and Misfit", CAISE'04 Workshops, 2004, Vol. 2.
- [11] J. Luftman, "Assessing business-IT alignment maturity" in Communications of the Association for Information Systems, 2000, Vol. 4, pp. 1-50.
- [12] A. Etien, "Ingénierie de l'alignement :Concepts, Modèles et Processus", Ph.D thesis, Department Computer Science, University of Sorbonne, Paris, France, 2006.
- [13] J. Zachman, "A framework for Information Architecture", IBM Systems Journal, 1987, Vol. 38, pp.2-3.
- [14] The Open Group, "The Open Group Architecture Framework (TOGAF)", (2002).
- [15] Department of Defense Architecture Framework Working Group, "DoD Architecture Framework", 2003, Vol. I-II, Deskbook.
- [16] A. McCall, P.K. Richards, G.F Walters. "Factors in software quality", Rome Air Development Centre, 1977, Vols I-III.
- [17] L. Plazaola, M. Enrique, N. Vargas, J. Flores, M. Ekstedt, "A metamodel for strategic business and it alignment assessment" the International Conference on System Sciences, Vol. 21, isbn. 0-9787122-1-8.
- [18] A. Wegmann, P. Balabko, G. Regev, I. Rychkova, "A Method and Tool for Business-IT Alignment in Enterprise Architecture", Proceedings of CAiSE'05, 2005, pp: 113-118.
- [19] B. Bounabat, "Enterprise Architecture Based Metrics for Assessing IT Strategic Alignment", The European Conference On Information Technology Evaluation, 2006, Vol. 13, pp. 83-90.
- [20] A. Vasconcelos, P. Sousa, J. Tribolet J, "Information System Architecture Metrics: an Enterprise Engineering Evaluation Approach", the Electronic Journal Information Systems Evaluation, 2007, Vol. 10, pp. 91-122.
- [21] P. Sousa, C.M. Pereira, J.A. Marques, "Enterprise Architecture Alignment Heuristic", Microsoft Architects Journal, 2005, Vol. 4, pp. 34-39.
- [22] G. Carvalho, P. Sousa, "Business and Information Systems MisAlignment Model (BISMAM): an Holistic Model Leveraged on MisAlignment and Medical Sciences Approaches", International Workshop in Business IT Alignment and interoperability BUISTAL, 2008, Vol. 3, pp. 104-119.
- [23] Haut Commissariat au Plan, "Recensement Général de la population et de l'Habitat Maroc", 2004. Web site : www.hcp.ma.

K. Elhari: PhD candidate at the National High School for Computer Science and Systems Analysis (ENSIAS). She held an Extended Higher Studies Diploma from Mohammadia School of Engineers (EMI) on 2006 by working on the use of multi-agent systems in the development of Amine platform. She is a software Engineer graduated on 2003 from National Institute of Statistics and Applied Economics (INSEA) and she is working in the High Commission for Planning in the kingdom of Morocco.

B. Bounabat: PhD in Computer Sciences. Professor in ENSIAS, (National Higher School for Computer Science and System analysis), Rabat, Morocco. Responsible of "Computer Engineering" Formation and Research Unit in ENSIAS, Regional Editor of Journal of Computing and Applications, International Expert in ICT Strategies and E-Government to several international organizations, Member of the board of Internet Society - Moroccan Chapter.

Towards an Adaptive competency-based learning System using assessment

Noureddine El Faddouli¹, Brahim El Falaki², Mohammed Khalidi Idrissi³ and Samir Bennani⁴

Computer Science Department, RIME Team, Mohammadia School of Engineers (EMI), Mohammed Vth University Agdal
BP. 765 AV. Ibn Sina Agdal, Rabat, Morocco

Abstract

E-learning is not restricted to publishing online content. The challenge is to apply pedagogical models using new information and communication technologies in order to adjust the learning process. This adaptation will take place by proposing an adaptive learning system. There are several e-learning environment adaptation approaches such as adaptive hypermedia and semantic web. In our proposed system, we focus on evaluation that has not been, up till now, given its real value in the e-learning environment. The purpose is to implement an adaptive learning system for individualized pedagogical paths through a personalized diagnosis of learners' performances. This proposed system relies mainly on assessment and the competency-based approach as a framework. To achieve this goal, we adopt an enhanced cycle of formative assessment while adhering to a service-oriented architecture.

The system will be implemented as an activity in a pedagogical scenario defined responding to the learner's needs, while aligning with norms and standards

Keywords: E-Learning, adaptive learning system, Service Oriented Architecture, formative assessment, learner model IMS-LIP, IMS-LD, IMS-QTI, IMS-RDCEO, LOM

1. Introduction and context

Online education does not substitute the traditional mode and is not restricted to the publication of online content. E-Learning aims to improve the quality of education by providing an interactive environment integrating pedagogical goals and benefiting from advances in ICT. Our proposal seeks to implement an adaptive learning system based on a new approach, namely the assessment that has not been, up till now, given its real value in the e-learning environment. The proposed system relies mainly on assessment and the Competency-based learning. The purpose is to individualize the learning path through a personalized diagnosis of the learner.

Currently, competency-based approach (CBA) is at the heart of the curricula of most educational systems. This approach rests on the notion of competency, which has been formalized for implementation in e-Learning environment

Learners have different objectives and predispositions. Thus, an optimal learning path for one learner is not necessarily the same for the other [1]. Consequently, adapting learning paths is crucial to manage learner differences. However, to achieve this adaptation, many approaches can be considered. In our previous work [2] we were distinguished by proposing the implementation of formative assessment as a means to adaptive learning system.

To implement the proposed system in its environment, several constraints are to face. First, we model the competency and the learner according to CBA. Then, we consider assessment as an activity to incorporate into a learning unit. Finally, we design a bank of items (questions).

In our proposal, the enhanced formative assessment [3] is the approach that will lead to a personalized diagnosis. Thus, the proposed system offers a series of consecutively selected items. The correctness of the answer to an item determines the selection of the next one taking into account the previous responses and performances recorded in the learner model. As far as the technical context is concerned, we adhere to our research team's vision [4] making the service oriented architecture (SOA) the approach of integrating the e-learning framework. Consequently, the evaluation service will be implemented as a composed service.

To enable reuse and operability, the environment will be designed according to standards, such as IMS-LD, IMS-QTI and IMS-LIP.

The next section deals with the pedagogical scenario, complying with IMS-LD standards, which describes a learning unit nurturing our proposed service. Section 3

concerns CBA learner modeling that will interact with the proposed system in alignment with IMS LIP specifications. The two ensuing sections (4 and 5) tackle the personalization of E-learning path, taking into account the learner's performances and relying mainly on formative assessment. SOA, the choice of which is justified in section 6, will lay the ground for the developing of the proposed system (section 7), and we terminate with a conclusion and perspectives

2. The pedagogical scenario

The advent of new information and communication technologies (ICT) in education calls upon the various actors involved in the education system to rethink the instructional design to keep up with this technological revolution. This design was focused on the content and the learning object was the main element. For a long time, almost all research centred on standardizing descriptions of learning objects enabling their reuse and interoperability. But this approach is insufficient to meet the different learning situations. To complement these shortcomings and enable effective integration of technology in education, research has focused on modeling the teaching/learning situations highlighting activities rather than content. Pedagogical scripting is a central process in instructional design, which aims to define the organization of learning activities [5]. The scripting begins with content design, resource organization, activity planning and orchestration in an environment. The notion of learning activity is central to the scenario. According to Paquette (2007) a lesson plan is an ordered set of activities governed by the actors who use resources and produce results [5]. The scenario creates the conditions for an activity once it is operationalized using a variety of services and content. This script is possible by using an Educational Modeling Language (EML)

Educational Modeling Language: IMS LD standards

According to the European Committee for Standardization, an EML is "information model and aggregation semantics, describing the content and processes involved in a learning unit in a pedagogical perspective and in order to ensure reusability and interoperability" [6]. The EMLs propose formalisms for describing situations of learning mediated by ICT. Several proposals exist; that of Koper [7] sets the goal to describe learning situations with EML to define the relationship between the competence or educational objectives, activities and actors of learning. This proposal was formalized by the implementation of an EML which first inspired the IMS Learning Design (IMS-LD).

IMS-LD specifications (Fig. 1) offer a conceptual framework to model a learning unit using the educational concepts necessary for its formal description. This specification combines, according to Anne Lejeune [8], genericity of the implementation of different pedagogical approaches and a precise description power, while ensuring the exchange and interoperability of learning materials in the learning units.

The terminology used to describe a learning unit is borrowed from the theatre area, particularly the concepts of play, act, role and role-part. The involvement of different actors in a learning unit is described and organized according to a scenario, using an environment.

In our proposed adaptive learning system, formative assessment is an activity (according to IMS LD) integrated in a learning unit. In alignment with IMS LD standards, the



Fig. 1 The activity according to IMS-LD specifications [9]

assessment activity concerns actors and uses an environment composed of resources and services, to produce results. IMS LD does not impose a pedagogical model, it is a pedagogical meta-model. The choice of IMS LD is motivated by references to this specification by valuable works on the e-learning environment [10, 11] as well as recent efforts to implementations in many instructional design, and the developing of several tools for modeling and implementing learning scenarios such as Reload, CopperCore and CopperAuthor.

3. Modeling the learner in CBA

3.1 Learner modeling

In CBA, the learner is central and the learning environment must take into account his needs and expectations for the acquisition of a competency.

Learner modeling is a representation of the state of competencies. It focuses on learner characteristics and activities. It should represent information characterizing the learner at the static level (profile) as well as at the dynamic level (progression). In our contribution, the learner model is solicited in different stages of the

proposed system. It will help to highlight the root causes of the competency gap. This is possible by providing the items suited for a relevant diagnosis. The learner model can be implemented using standard templates. In our proposal, we adopt the IMS learner Information Package (IMS-LIP) specifications, which toes our vision that consider CBA as reference.

3.2 Modeling learner using IMS-LIP

To ensure the provision of competence-based learning services and facilitate the interaction with the learner, it is necessary to record his individual competencies in a persistent and standard way. Thus, the learner can find learning activities that meets his needs to achieve the desired competencies.

In our proposed system, we adopt the IMS-LIP specifications, which are “based on a data model that describes those characteristics of a learner needed for recording and managing learning-related history, goals and accomplishments” [9]. That model defines an XML structure (Fig.2) for data exchange between different learning systems involved in the learning process.

The IMS-LIP model offers the opportunity to refer to a competency described in an external source using the tag <exrefrecord>. This competency description must allow performance measurement. Thus, we expand the definition with the HR-XML model.

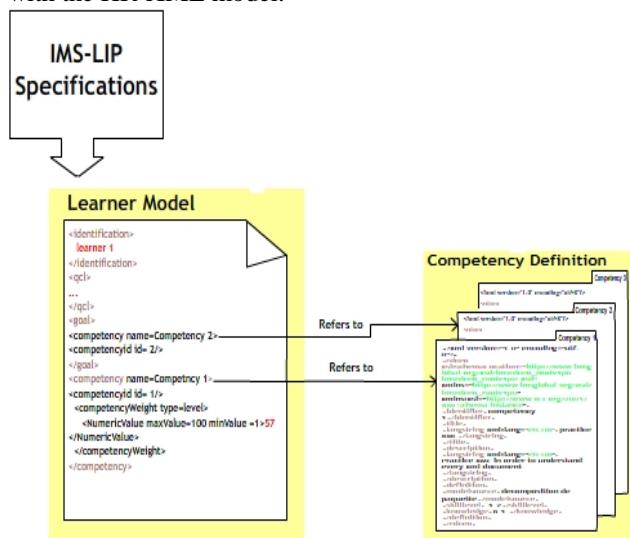


Fig. .2 Learner model using competency definition [6]

4. Personalizing learning: A content perspective

In aboard manner, personalization aims to adapt contents and services offered to the user to promote the quality of his interactions with the system [12]. In the education field, adapting is providing each learner with the feeling that the training is designed specifically to meet their expectations taking into account their capacities. According to Bellier [13], personalization targets the provision of learners with a training course perfectly suited to their level and needs. However, adaptation is based on identification of the learner, his ability, prior knowledge and current performance for the acquisition of competency. To do this, we stipulate that two ingredients are essential, namely learner modelling and a relevant diagnosis vis-a-vis the current activity. In this perspective, we modelled [14] formative assessment to offer to the learning system a relevant diagnosis customized to regulate the learning taking into account the characteristics and progression of learner. In our proposal, it would be prominent to operationalize formative assessment by describing the interfaces of its various functional components and their choreography to enable individualized diagnosis.

4.1 Referencing educational resources using LOM

The dynamic composition of learning paths (including assessment) is assembling a set of activities combining learning objects and services). The purpose is to draw for each student, the optimal path to acquire a competency (Fig.3) responding to the specific needs and the output profile desired by the educational system.

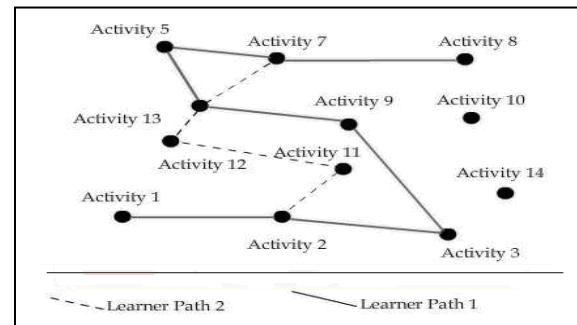


Fig. 3 The dynamic composition of learning paths using the proposed system

The production of learning objects is time-consuming and costly and can be made profitable when reused as long as possible. For this, it is necessary to specify the structure and index them [15]. The referencing of learning objects is a necessity if we want to integrate them in a course while maintaining its coherence and relevance. It is very useful for designers of educational content to adapt the choice of

educational resources based on the specified requirements [16]. In this connection, they must add semantic information which should be structured, usable and descriptive of the resource and its use. The latter is metadata: data describing data [17], that Bernes-Lee, T [18] defined as "data about data" and considers that "Metadata is machine understandable information about web resources or other things". To enable operability and reuse, a standard must exist so that the educational content developers and users use the same repository. This repository can take various forms: Metadata standard (IEEE LOM (Learning Object Metadata) [19], DC (Dublin Core), taxonomies, ontologies formalized in various languages (XML, RDF, OWL). In our proposal, we will opt for LOM.

4.2 Learning objects for assessment: Question & Test Interoperability specifications (IMS-QTI)

On the occasion of an evaluation activity proposed as part of a pedagogical scenario, items will be administrated to the learner. This uses an item bank described in a formal and standard way. Each item corresponds to a competency and will be integrated into the environment of an activity. There

may be several items assessing the same competency.

In the proposed system, we opt for the standard IMS-QTI, which allows representing the data structure of a question (item) and a test (assessment) and their corresponding results. This representation is done through an XML file (Fig.4) providing interoperability

```
<xml version="1.0" encoding="iso-8859-1">
<test>
  <question ident="1">
    <duration>20</duration>
    <objectif>comprendre les base xml</objectif>
    <condition_prestable_a_question>connaissance en xml</condition_prestable_a_question>
    <condition_denvoi_de_reponse>r1.checked or r2.checked or r3.checked or r4.checked</condition_denvoi_de_reponse>
    <presentation label="exemple de question" xmlLanguage="fr">
      <clue>
        <materiel>
          <mattext>qu'est ce que le XML?</mattext>
          <materiel>
            <response_id ident="001" cardinality="single">
              <choix_de_rende>
                <clue_label>
                  <response_label ident="1">
                    <materiel>
                      <mattext>langage de balise tout comme HTML mais avec une plus grande possibilité d'adoption</mattext>
                    </materiel>
                  </response_label>
                <response_label ident="0,5">
                  <materiel>
                    <mattext>un langage de programmation</mattext>
                  </materiel>
                </response_label>
              <response_label ident="0,75">
                <materiel>
                  <mattext>une base de donnée amovible</mattext>
                </materiel>
              </response_label>
            </response_id>
          </materiel>
        </clue>
      </presentation>
    </question>
  </test>
```

Fig. 4 Example of an item complying IMS-QTI

5. Formative assessment in CBA

5.1 Formative assessment

In an educational system, evaluation is central in the learning process. Its role is not limited to certification, which was the only outlet, but can be perceived as a process of verification to guide the teaching/learning process [20]. Therefore, it is necessary to distinguish between assessment as an integral part of the learning process (formative) and assessment for certification (summative or certification). In our proposal, we adopt the formative assessment as a component that participates into the process of learning.

The concept of formative assessment was introduced by Scriven, M [21] and supported by Bloom, B [22], when he built his model of mastery pedagogy [23]. According to Perrenoud [24], formative assessment helps the student to learn. In other words, it participates in the regulation of the learning process.

According to the review of literature, formative assessment is made up of a cycle that is built on three layers and that we enriched with pre-regulation layer [25]:

- Observation: Establish the position in relation to a repository, instead of confining the learner to be on a scale and compare him to other learners.
- Intervention: identifies symptoms to address root causes of problems. It involves analyzing metacognitive knowledge [26]. It involves identifying mental functions and highlights their weaknesses.
- Regulation: Describe the mechanisms that provide guidance, control and the adjustment of cognitive activities, emotional and social as well as their relationship with a learner [27]

5.2 Competency: object of evaluation

Competency is abstract and hypothetical [28]. Thus, how can it be the object of assessment? The evaluation is the most problematic point of CBA [29], since it is not only assessing the situation in academic knowledge, but also mobilize, in a situation sometimes close to real life, resources, skills and competencies. In our proposal, we focus on formative assessment to regulate the learning process in a competency-based approach. This requires modeling competency from an operational viewpoint. In this way, we adopt the generic skills taxonomy of Paquette [30] and his definition of competency in which competency is defined as a relation linking three areas: knowledge, skills and actors.

In order to support and integrate competency in education, there is a need to provide reusable definitions of competence, across the different systems (CEN/ISSS CWA

15455, 2005). Several models are proposed to describe the competence formally, such as the IEEE Reusable Competency Definition (IEEE RCD) [31] and the IMS Reusable Definition of Competence or Educational Objective (IMS RDCEO) [32] specification. In our system, the competency has been defined in accordance with the standard IMS RDCEO (Fig.5)

The IMS RDCEO specification defines an information model that can be used to exchange these definitions between different systems [6]. It describes the competency independently of the context and guarantees the interoperability among systems using the competency definition. The specification presents competency information in five categories: identifier, title, description, definition and Metadata.

IMS RDCEO specification permits the representation of the competency level, the success threshold of a competency and the complex competency within the title element in an unstructured format, thus, the machine can not understand, search and process this element effectively. Consequently, the scope of interoperability among different systems will be limited.

To define a common meta-model for the description of competencies, integrating; competency level, the success threshold of a competency and the complex competency .we propose possible extensions to the information model by its enrichment through HR-XML standard according to the recommendation of European commission of normalisation [6].

The HR-XML consortium was established to create a standard facilitating the exchange of competency information among different systems dealing with competency offering the organization a tool to enhance human resources activities

HR-XML describe competency in nine categories: Name, Description, Required, CompetencyId, TaxonomyId, CompetencyEvidence, CompetencyWeight, Competency and UserArea

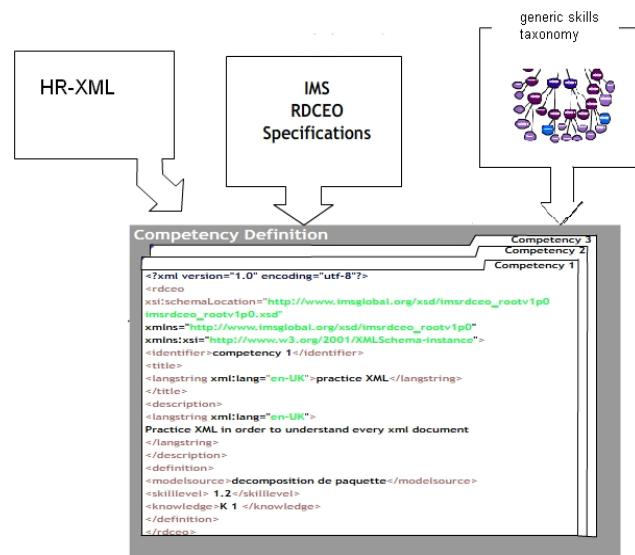


Fig 5 The proposed competency definition used in our proposed system

6. An SOA vision of e-learning platform

The implementation of an e-Learning system faces two major difficulties: (1) the modeling of pedagogical goals and operationalization of designed artifacts [4]. Thus, it must take into account the pedagogical context and the accelerated evolution of information technology. Currently, many e-Learning platforms existing in the market carry the teaching goals with features that are similar. The reuse of functionality and interoperability across different systems is neglected compared to the efforts concentrated on the structuring and reuse of learning objects (LOM, SCORM, etc) [33]. In our research team, the reuse and interoperability of components and services among different systems are in the agenda. The objective is to develop an open platform for the development, integration and management of distributed software components. In this perspective, our team adopt an approach of modeling the different components that we operationalize by enrolling in a service oriented architecture to reorganize the e-learning system (Fig.6)

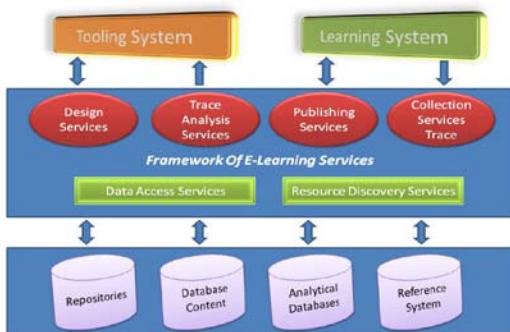


Fig. 6 The Adopted Computer System Architecture Integrating e-learning [4]

7. The proposed model

7.1 Modeling

To illustrate the progress of the evaluation process implemented in the proposed system, we propose the following flowchart (Fig.7)

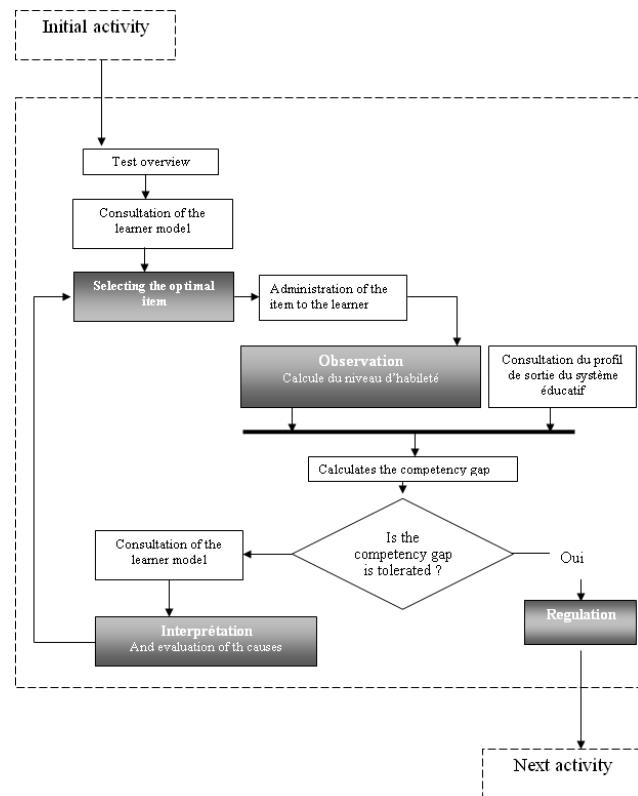


Fig. 7 Flowchart explaining the progress of the activity of evaluation implemented in the proposed system

7.1.1 Constraints of the implementation:

In the process of implementing the assessment system modelled in our previous work [25], the major difficulty is to take advantage of ICTs while faithfully reproducing didactic goals. In this process, we must consider several parameters involved in the scripting of the e-learning unit (Fig. 8)

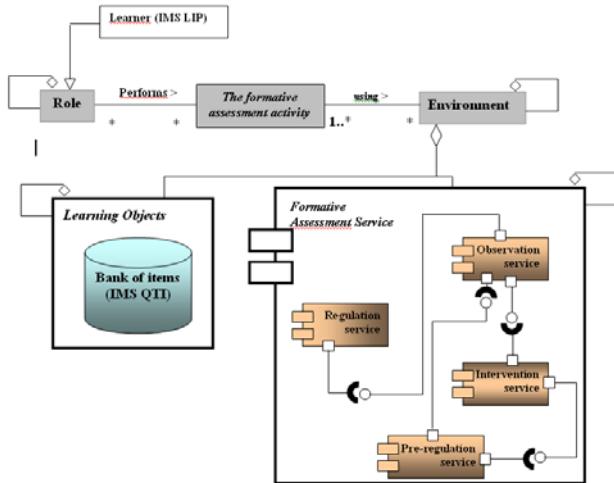


Fig.8: The environment of the formative assessment activity (according to IMS LD) in a pedagogical scenario complying with the IMS-LD specifications

The assessment activity (according to IMS LD specifications) will be operationalized in a pedagogical scenario; it uses an environment consisting of resources (according to IMS-QTI specifications) grouped in an item bank, and a composed service. The service integrates a formative assessment approach.

Several constraints must be respected when implementing the proposed services:

- 1) Reusable definition of competency to support and integrate competence across different systems. In this way we adopt the IMS RDCEO specifications.
- 2) Items must incorporate the notion of competency respecting generic skills taxonomy to construct and classify the items in terms of their difficulties, and permit reuse regardless of any platform. In this stage, we adopt the IMS-QTI standard for the implementation of criteria and performance indicators to measure the level of actual performance in the observation process. The target level is determined by the output profile
- 3) The unit of learning that will incorporate the designed activity and provide a pedagogical scenario. In this way, we adopt the IMS LD specifications.
- 4) A representation of the state of the learner's competencies as a prototypical perspective (learner profile) and a dynamic perspective (progression of the learner). For its implementation we have opted for a IMS-LIP standard.

7.1.2 Services and their interactions:

The proposed service consists of a set of services whose functions are synchronized to allow a personalized

diagnosis proposing remediation activities. Thus, we find in (Fig. 8):

1 Observation service consists of a set of activities (Fig.9). The purpose of this step is to establish the state of knowledge and skills. At this stage, we calculate the level of performance (current) using criteria and indicators of performance to identify the competency gap.

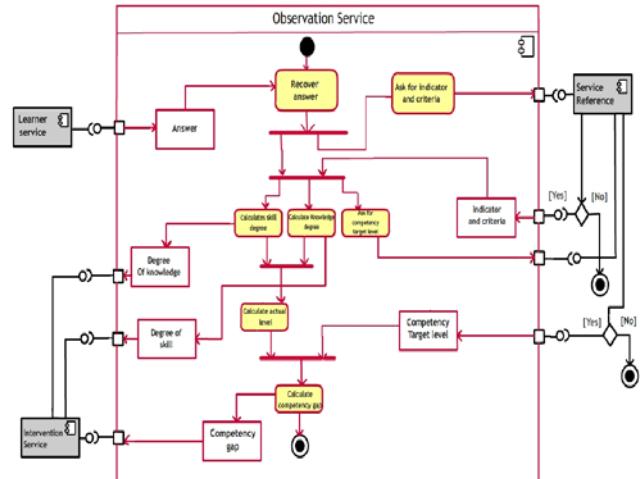


FIG.9: Observation Service activity

Operationally, through this observation stage, the following steps are crossed:

- a-The learner gives an answer to an item designed according to IMS-QTI, and the result will be forwarded to the observation service.
- b-The observation service identify the current competency level using the item and the learner answer
- c-The observation service interacts with the output profile to extract the target competency level, and accepts as inputs: the item, the response of the learner and the target competency level from the output profile. In the output, the observation service results in the competency gap (Tab.1),

Tab. 1: competency gap [31]

| | Awareness | Familiarized | Mastery | Expertise |
|------------------------------|-----------|--------------|---------|-----------|
| State | 0 | 2.5 | 5 | 7.5 |
| Competency A | Value | | | |
| Current level of performance | 0 | 2.5 | 5 | 7.5 |
| Competency gap | 2.5 | 2.5 | 2.5 | 2.5 |
| Target performance level | 10 | 10 | 10 | 10 |

2 - Intervention service: In this step, it analyzes symptoms to address root causes of the competency gap detected in the observation step. It involves analyzing metacognitive knowledge (the mental) which remains very mysterious [26]. Assessing competency based on observing only reaches limits very quickly. Say "you can do better" does

not help the learner to know how. To be useful, we must identify, isolate mental functions (generic skills), and highlight their weaknesses.

Operationally (Fig.10), (a) the process begins by intercepting the competency gap calculated in observation step and the current level performance.

b- If the competency gap is not tolerated, the intervention service consults the learner model. It compares the current level (intercepting from the first step) with the performance level carried in the learner model for the same generic skill.

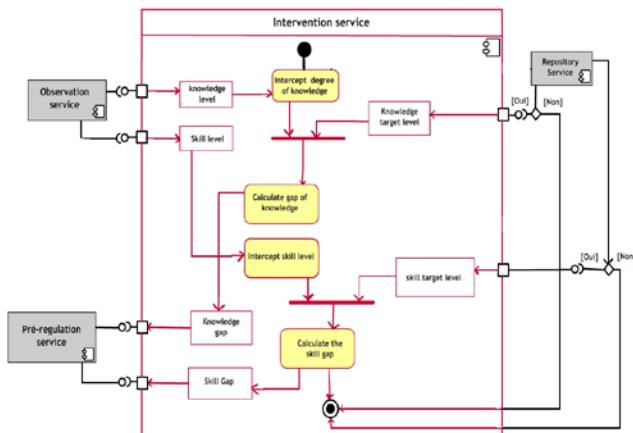


FIG.10 : Intervention service activity

3 - Pre-regulation service: in this step, we adapt the evaluation to the learner by providing items suitable for his current competency level. The purpose is to make accurate assessment with fewer items.

In this stage, the principal role is to select the optimal item based on parameters from the previous step

The pre-regulation process uses a bank of item semantically referenced and each is relative to a competency. The items are designed according to IMS QTI specifications and the evaluation is given item by item. The item is chosen in the pre-regulation stage. The bank of items will be used until the end of the assessment. Once the assessment is completed, the final competency gap will be used in the last process to choose the next learning activity (Fig.11)

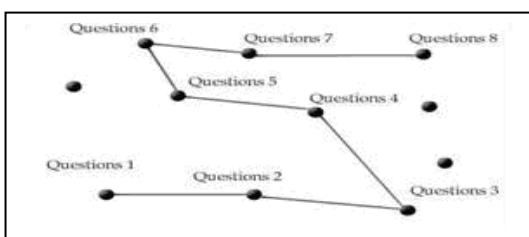


FIG.11: Individualization of formative assessment through the personalization of evaluative path for each learner

4 - Regulation Process: In this step, a mechanism that provides guidance and adjusting learning activities will be implemented and its main role would be to choose the remediation activity that is the most suited to the learner for the acquisition of competency

7.2 Implementation

Service oriented architecture

The goal is to provide "an open Platform" for the development, deployment, interaction and management of distributed e-services. [34]. The model of web services (Fig.12), is defined as an architecture calling upon a set of standardized protocols (fig 6). The Orchestration of services is carried out by IMS LD specifications.

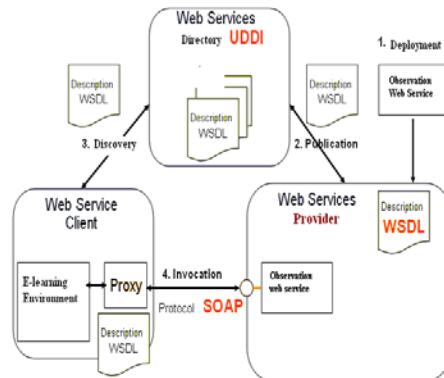


Fig. 12 SOA and observation web service

The environment:

Figure 13 illustrates the environment of our proposed system:

Output Profile: the profile desired by the educational system. In implementation, it is considered as an XML file recording for each skill level targeted.

Learner model: it is specific to each learner is an XML file that records the level of performance for each competency

Issues that will be delivered to learner described through an XML file, while respecting the standard (IMS-QTI)

Competency definition that describes the skills that will be used both by the output profile as the model of the learner

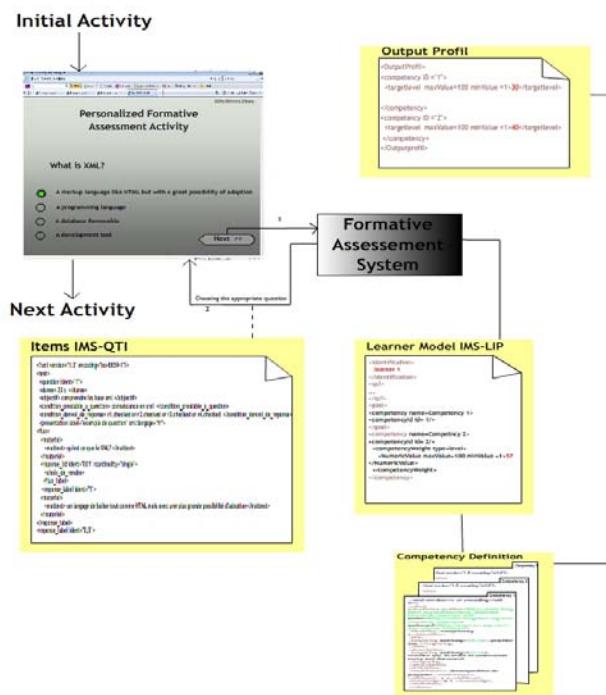


Fig. 13 the specification of environment of the proposed system

8. Conclusion & perspectives

To provide an interactive environment tailored to the learner's needs is one of the most important goals of e-learning environments. Interactivity and adaptation do not rely solely on technical artefacts, but are the result of a combination involving educational theory, and technological advances in the field of ICT. Several studies have addressed the individualization from different angles. Ours is different, both in the approach and tools; it offers a system that individualizes the evaluation process offering a personalized diagnosis to decide upon the remediation activity. In the implementation of the proposed system, interoperability and reuse justify the choice of components and the environment interacting with the system. As far as the technical architecture is concerned, we adhere to our research team's global vision. In this vision, the e-learning platform should be composed of a set of reusable, interoperable and interacting services. The composition and the orchestration of these services will be allotted to the ILMSD standards in the learning unit framework. Formative assessment is undoubtedly a central concept in the process of teaching/learning. Its implementation in the context of competency-based approach in the E-learning systems allows the individualization of learning. Several perspectives are considered, and can be summarized in:

- 1) The developing of the services which make up the system.
- 2) The deployment and testing in a learning unit.
- 3) The collecting and analysis of formative assessment activity traces.

References

- [1] Ph. Perrenoud, *Construire des compétences dès l'école*, Paris: ESF, 2000
- [2] B. El Falaki, M. Khalidi Idrissi, and S. Bennani, "Formative assessment in e-learning: an approach to personalize diagnosis and adapt learning path", in IADIS e-Society 2010 (ES 2010) à Porto, Portugal, 2010, ISBN: 978-972-8939-07-6 pp : 391.395
- [3] B. El Falaki, M. Khalidi Idrissi, and S. Bennani, "Modèle de l'évaluation Formative dans une approche par compétences pour un apprentissage à distance ", in Workshop sur les Technologies de l'Information et de la Communication'2009. Agadir. Maroc, 2009, ISBN :978-9981-0-2625-50
- [4] M. Khalidi Idrissi, F. Merrouch, and S. Bennani, "Analyse des situations e-learning : abstraction et modélisation" in 2nd Conférence internationale, systèmes d'information et intelligence économique, SIIE 2009. Hammamet, Tunisie, 12-14 Février 2009 ; IHE edition ISBN: 9978-9973-868-21-3. pp. 153-164.
- [5] G. Paquette, "An Ontology and a Software Framework for Competency Modeling and Management.", in Educational Technology and Society, Special Issue on Advanced Technologies for Life-Long Learning, Volume 10, Issue 3, 2007, pp. 1-21
- [6] CEN/ISSS cwa 15455, "A European Model for Learner Competencies", ICS 03.180; 35.240.99, November 2005
- [7] R. Van, R. Koper, "Testing the pedagogical expressiveness of IMS LD". Educational Technology & Society, 2006, Issue 9 (1), pp. 229-249.
- [8] L. Lejeune, "IMS Learning Design", in Distances et savoirs, Lavoisier 2004/4, Volume 2, ISSN 1765-0887, pp, 409-450
- [9] IMS Global Learning Consortium, Inc (2006a), IMS Learning Design Information Model,
- [10] N. Taurisson, P. Tchounikine, "Mixing Human and Software Agents: a Case Study", In IEEE International Conference on Advanced Learning Technologies, 9-11 juillet 2003, Athènes (Grèce), pp. 239-243
- [11] G. Paquette, "L'ingénierie pédagogique à base d'objets et le référencement par les compétences", International Journal of technologies in higher education, 2004, pp. 45-47.
- [12] A. Stewart, C. niederee, and B. Mehta, "State of the art in user modeling for personalization in content, service and interaction", DELOS Report on Personalization, NSF, 2004
- [13] S. Bellier, Le e-learning, Paris: Edition Liaison, 2001.
- [14] M. Khalidi Idrissi, B. El Falaki, and S. Bennani, "Implementing the formative assessment within competency-based-approach applied in e-learning", 3d International conference on SIIE;; Sousse, TUNISIA; 18-20, 2010 . IHE edition ISBN: 978-9973-868-24-4, pp. 362-368;
- [15] Y. Bourda, M. Hélier, "What Medata and XML can do for learning Objects", Webnet Journal:Internet Technologies, Applications 2001, Issues 2(1), pp. 24-31.

- [16] V. Devedic, Semantic web and education. New York: Springer, berlin Heidelberg, 2006.
- [17] O. Lassila, "Web Metadata: A matter of semantics ", In IEEE Internet Computing, July –August 1998, pp.30-37.
- [18] T. Berners-Lee, "Metadata Architecture", ([Http://www.w3.org/DesignIssues / Metadata.html](http://www.w3.org/DesignIssues / Metadata.html)), Jan 1997.
- [19] IEEE, "Draft Standard for Learning Object Metadata", Institute of Electrical and Electronics Engineers, 2002.
- [20] G. Scallon, L'évaluation des apprentissages dans une approche par compétences, Bruxelles: De boeck, 2007.
- [21] M. Scriven, "The methodology of evaluation". In Gredler, M. E. Program Evaluation. New Jersey: Prentice Hall, 1996. pp. 16. 1967
- [22] B. Bloom, Handbook on formative and summative evaluation of student learning, New York: McGraw-Hill, 1971
- [23] L. Allal, et al, L'évaluation formative dans un enseignement différencié, Berne: Lang, 1989
- [24] Ph. Perrenoud, L'évaluation des élèves. De la fabrication de l'excellence à la régulation des apprentissages, Bruxelles: De Boeck, 1998
- [25] B. El Falaki, M. Khalidi Idrissi, and S. Bennani, "A Formative Assessment Model within competency-based-approach for an Individualized E-learning Path", in world academy of science, engineering and technology. Issue 64, april 2010, ISSN. 1307-6892, pp. 208-212
- [26] Ph. Perrenoud, "Formation et Profession", Bulletin du Centre de recherche interuniversitaire sur la formation et la profession enseignante, Vol. 11, n° 1, Montréal, avril 2005.
- [27] L. Allal, L. Mottier Lopez, Régulation des apprentissages en situation scolaire et en formation, Bruxelles: De Boeck, 2007.
- [28] C. Vern, Evaluation des compétences, Paris: Liaisons, 2002.
- [29] L. Endrizzi, O. Rey, L'Évaluation au cœur des apprentissages. Dossier d'actualité de la VST 39, Lyon: INRP, 2008.
- [30] G. Paquette, Modélisation des connaissances et des compétences pour concevoir et apprendre, Sainte-Foy: PUQ, 2002.
- [31] IEEE, CEN WS-LT LTSO, (<http://www.cen-ltso.net/main.aspx?put=1054>) 05/12/2010.
- [32] IMS GLC, (<http://www.imsglobal.org/specificationdownload.cfm>), 2002.
- [33] Ivan Madjarov, "Une approche orientée services web pour l'intégration d'applications e-learning et la réutilisation des objets pédagogiques XML" Colloque international sur l'informatique et ses Applications IA'2006, ENSAO, oujda, Maroc, 2006, pp.45-51
- [34] Hai zhuge, jie liu "Flexible retrieval of web services". The journal of systems and software, 2004, ISSN. 107-116

N. El Faddouli Doctorate degree in Computer Science in 1999; Assistant Professor at the Computer Science Department at the Mohammadia School of Engineers (EMI); 3 recent publications papers between 2007 and 2010; Ongoing research interests: Web services, elearning, evaluation and information systems.

B. El Falaki Engineer degree in Computer Science in 2002; PhD Student in Computer Science; student member of IEEE education and computer science; former with the Moroccan Office of professional education (OFPPT) in the specialized institute of new information technologies and offshoring; 4 recent publications

papers between 2009 and 2010; Ongoing research interests: Web services, elearning and assessment

M. Khalidi Idrissi Doctorate degree in Computer Science in 1986, PhD in Computer Science in 2009; Former Assistant chief of the Computer Science Department at the Mohammadia School of Engineers (EMI); Professor at the Computer Science Department-EMI; 8 recent publications papers between 2008 and 2010; Ongoing research interests: SI, ontology, Web services, MDA, elearning and evalution

S. Bennani Engineer degree in Computer Science in 1982; Doctorate degree in Computer Science, PhD in Computer Science in 2005; Former chief of the Computer Science Department at the Mohammadia School of Engineers (EMI); Professor at the Computer Science Department-EMI; 10 recent publications papers between 2008 and 2010; Ongoing research interests: SI, Modeling in Software Engineering, Information System, eLearning content engineering, tutoring, assessment and tracking,

An Efficient Searching and an Optimized Cache Coherence handling Scheme on DSR Routing Protocol for MANETS

Mr. Rajneesh Kumar Gujral¹, Dr. Anil Kapil²

¹Assoc. Professor, Computer Engineering Department, M. M. Engineering College, M. M. University, Ambala, Haryana, India-133207.

²Professor M. M. Institute of Computer Technology and Business Management, M. M. University, Ambala, India-133207.

Abstract

Mobile ad hoc networks (MANETS) are self-created and self organized by a collection of mobile nodes, interconnected by multi-hop wireless paths in a strictly peer to peer fashion. DSR (Dynamic Source Routing) is an on-demand routing protocol for wireless ad hoc networks that floods route requests when the route is needed. Route caches in intermediate mobile node on DSR are used to reduce flooding of route requests. But with the increase in network size, node mobility and local cache of every mobile node cached route quickly become stale or inefficient. In this paper, for efficient searching, we have proposed a generic searching algorithm on associative cache memory organization to faster searching single/multiple paths for destination if exist in intermediate mobile node cache with a complexity $O(n)$ (Where n is number of bits required to represent the searched field). The other major problem of DSR is that the route maintenance mechanism does not locally repair a broken link and Stale cache information could also result in inconsistencies during the route discovery /reconstruction phase. So to deal this, we have proposed an optimized cache coherence handling scheme for on -demand routing protocol (DSR).

Keywords: DSR, Efficient Searching, Cache Coherence, MANETS etc.

1. Introduction

Mobile ad hoc networks (MANETS) are self-created and self organized by a collection of mobile nodes, interconnected by multi-hop wireless paths in a strictly peer to peer fashion [1]. Caching is an important part of any on-demand routing protocol for wireless ad hoc networks. In mobile ad hoc network (MANETS) [2],[3],[4] all node cooperate in order to dynamically establish and maintain routing in the network , forwarding packets for each other to allow communication between nodes not directly within wireless transmission range. Rather than using the periodic or background exchange of

routing information common in most routing protocols , an on-demand routing protocols is one that searches for the attempts to discover a route to some destination node only when a sending node originates a data packet addressed to the node. In order to avoid the need for such a route discovery to be performed before each data is sent, an on-demand routing protocol must cache routes previously discovered. Such caching then introduces the problem of proper strategies for managing the structure and contents of this cache, as nodes in the network move in and out of wireless transmission range of one another, possibly invalidating some cached routing information.

Several routing protocols for wireless ad hoc networks have used on-demand mechanisms, including temporally-ordered routing algorithm (TORA) [8], Dynamic source Routing protocols (DSR) [5]. Ad hoc on demand distance vector (AODV) [6], Zone routing protocol (ZRP) [7], and Location-Aided Routing (LAR) [9]. For example, in the Dynamic Source Routing protocol [5] in the simplest form, when some node S originates a data packet destined for a node D to which S does not currently know a route, S initiates a new route discovery by beginning a flood a request reaches either D or another node that has a cached route to D, this node then returns to S the route discovered by this request. Performing such a route discovery can be an expensive operation, since it may cause a large number of request packets to be transmitted, and also add latency to the subsequent delivery of data packet that initiated it. But this route discovery may also result in the collection of a large amount of information about the current state of network that may be useful in future routing decision. In particular, S may receive a number of route replies in response to its route discovery flood, each of which returns information about a route to D through a different portion of the network. In high-mobility environment the performance degrades rapidly of this protocol because the

route maintenance mechanism does not locally repair a broken link. Stale cache information could also result in inconsistencies during the route reconstruction phase. In this paper, for efficient searching, we have proposed a generic searching algorithm on associative cache memory organization to faster searching single/multiple paths for destination if exist in intermediate mobile node cache with a complexity $O(n)$ (Where n is number of bits required to represent the searched field). The other major problem of DSR is that the route maintenance mechanism does not locally repair a broken link and Stale cache information could also result in inconsistencies during the route discovery/reconstruction phase. So to deal this, we have proposed an optimized cache coherence handling scheme for on-demand routing protocol (DSR). In this paper, Section 2, we describes the Dynamic Source Routing Protocol (DSR), Section 3, we describe related work, Section 4, we describe associative searching Flowchart, Algorithm and their implementation with example, Section 5, we describe proposed an optimized cache handling scheme and Section 6, we had concluded the paper and future works.

2. Overview of the Dynamic Source Routing Protocols (DSR))

Dynamic source routing protocol (DSR) is an on-demand protocol designed to restrict the bandwidth consumed by control packets in ad hoc wireless networks by eliminating the periodic table-update messages required in the table-driven approach [10]. The major difference between this and other on-demand routing protocols is that it is beaconless and hence does not require periodic hello packet (beacon) transmission, which are used by a node to inform its neighbors of its presence. The basic approach of this protocol (and all other on-demand routing protocols) during the route construction phase is to establish a route by flooding Route Request packets in the network. The destination node, on receiving a RouteRequest packet, responds by sending a RouteReply packet back to the source, which carries the route traversed by the RouteRequest packet received.

Consider a Source node that does not have a route to the destination. When it has data packets to be sent to that destination, it initiates a RouteRequest packet. This RouteRequest is flooded throughout the network. Each node, upon receiving a RouteRequest packet, rebroadcasts the packet to its neighbors if it has not forwarded already or if the node is not the destination node, provided the packets time to live (TTL) counter has not exceeded. Each RouteRequest carries a sequence number generated by the source node and the path it has traversed. A node, upon receiving a RouteRequest packet, checks the sequence number on the packet before forwarding it. The packet is forwarded only if it is not a duplicate RouteRequest. The

Sequence number on the packet is used to prevent loop formations and to avoid multiple transmissions of same RouteRequest by an intermediate node that receives it through multiple paths. Thus, all nodes except the destination forward a RouteRequest packet during the route construction phase.

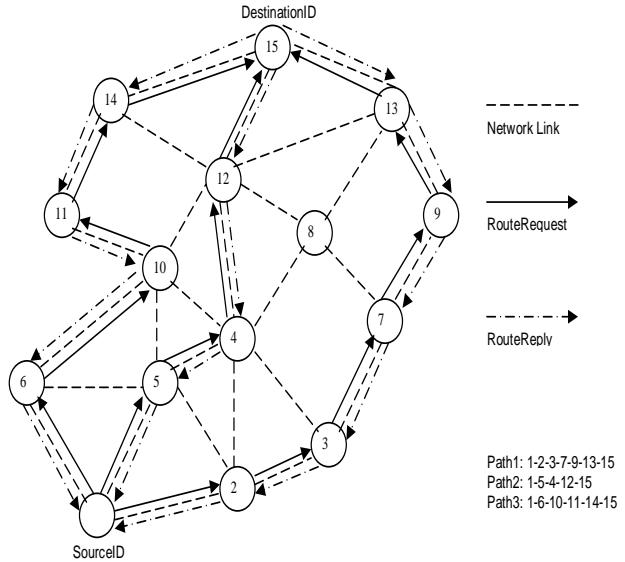


Figure 1. Route establishment in DSR

A destination node, after receiving the first RouteRequest packet, replies to the source node through the reverse path the RouteRequest packet had traversed. In Figure 1, source node 1 initiates a RouteRequest packet to obtain a path for destination node 15. This protocol uses a route cache that stores all possible information extracted from the source route contained in a data packet. Nodes can also learn about the neighboring routes traversed by data packets if operated in the promiscuous mode (the mode of operation in which a node can receive the packets that are neither broadcast nor addressed to itself). This route cache is also used during the route construction phase. If an intermediate node receiving a RouteRequest has a route to the destination node in its route cache, then it replies to the source node by sending a RouteReply with the entire route information from the source node to the destination node.

2.1 Optimizations:

Several optimization techniques have been incorporated into the basic DSR protocol to improve the performance of the protocol. DSR uses the route cache at intermediate nodes. The route cache is populated with routes that can be extracted from the information contained in the data packets that get forwarded. This cache information is used by the intermediate nodes to reply to the source when they receive a RouteRequest packet and if they have a route to the corresponding destination. By operating in the Promiscuous mode, an intermediate node learns about

route breaks. Information thus gained is used to update the route cache so that the active routes maintained in the route cache do not use such broken links. During network partitions, the effected nodes initiate RouteRequest packets. An exponential backoff algorithm is used to avoid frequent RouteRequest flooding in the network when the destination is in another disjoint set. DSR also allows piggy-backing of a data packet on the RouteRequest so that a data packet can be sent along with the RouteRequest.

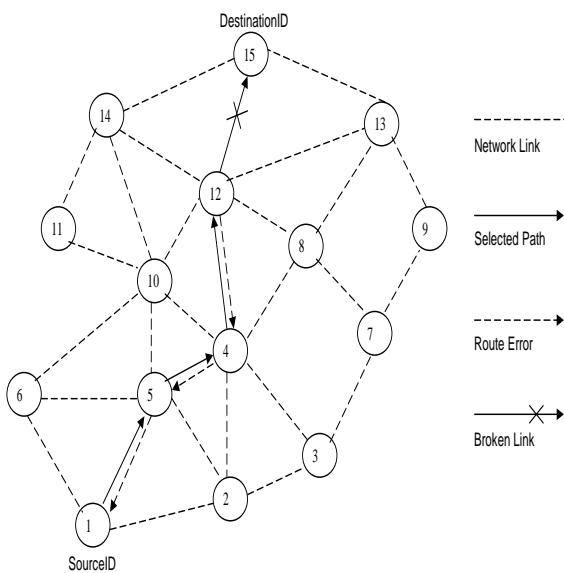


Figure 2. Route maintenance in DSR

If optimization is not allowed in the DSR protocol, the route construction phase is very simple. All the intermediate nodes flood the RouteRequest packet if it is not redundant. For example, after receiving the RouteRequest packet from node 1 from Figure 1, all its neighboring nodes, that is, nodes 2, 5, and 6, forward it. Node 4 receives the RouteRequest from both nodes 2 and 5. Node 4 forwards the first RouteRequest it receives from any one of the nodes 2 and 5 and discards the other redundant/duplicate RouteRequest packets. The RouteRequest is propagated till it reaches the destination which initiates the RouteReply. As part of optimizations, if the intermediate nodes are also allowed to originate RouteReply packets, then a source node may receive multiple replies from intermediate nodes. For example, in figure 2, if the intermediate node 10 has a route to the destination via node 14, it also sends the RouteReply to the source node. The Source node selects the latest and the best route, and uses that for sending data packets. Each node packet carries the complete path to its destination. When an intermediate node in the path moves away, causing a wireless link to break, for Example, the link between nodes 12 and 15 in Figure 2, a RouteError

message is generated from the node adjacent to the broken link to inform the source node. The source node reinitiates the route establishment procedure. The cached entries at the intermediate nodes and the source node are removed when a RouteError packet is received. If a link breaks due to the movement of the edge nodes (node 1 and node 15), the source node again initiates the route discovery.

3. Related Works.

3.1 Cache data and cache path

The cache data scheme considers the cache placement policy at intermediate nodes in the routing path between the source and destination. The node caches a passing by data item locally when it finds that the data item is popular, i.e., there were many requests for data item, or it has enough free cache space. Since cache data needs extra space to save the data, it should be used prudently. A conservative rule is proposed as follow: A node does not cache the data if all requests for the data are from the same node. However, it uses cooperative caching protocol among mobile node. Each mobile node does not independently perform the caching tasks such as placement and replacement. Cache path is also proposed for redirecting the requests to the cache node. In MANETS, the network the network topology changes fast and thus, the cached path may become invalid due to the movement of mobile nodes [11].

3.2 Neighbor Caching Technique

The concept of neighbor caching (NC) is to utilize the cache space of inactive neighbors for caching tasks. The basic operations of NC are as follow. When a node fetches a data from remote node, it puts the data in its own caching space for reuse. This operation needs to evict the least valuable data from the cache based on a replacement algorithm. With this scheme, the data that is to be evicted is stored in the idle neighbor nodes storage. In the near future if the node needs the data again, it requests the data not from the far remote source node but from the near neighbor that keeps the copy of data. The NC scheme utilizes the available cache space of neighbor to improve the caching performance. However, it lacks the efficiency of the cooperative caching protocol among the mobile nodes [12].

3.3 Node caching schemes

This is a novel approach to constrain route request broadcast based on node caching. The Intuition used is that the nodes involved in recent data packet forwarding have more reliable information about its neighbors and have better locations (e.g., on the intersection of several data routes) than other MANET nodes. The nodes which are recently involved in data packet forwarding are considered as cache nodes, and only they are used to forward route

requests. The modified route request uses a fixed threshold parameter H. The first route request is sent with the small threshold H. When a node N receives the route request, it compares the current time T with the time T(N) when the last data packet through N has been forwarded. If $T-H > T(N)$, then N does not belong to the current cache and , therefore, N will not propagate the route request. Otherwise, if $T-H \leq T(N)$, then N is in the node ache and the route request is propagated as usual[13].

3.4 Group Caching

There are some challenges and issue such as mobility of mobile nodes, power consumption in battery , and limited wireless bandwidth when caching techniques are employed in MANETs for data communication .Due to the movement of mobile nodes, MANETs may be partitioned into many independent networks. Hence, the requester cannot retrieve the desired data from the remote server (data source) in another network. The entire data accessibility will be reduced. Also, the caching node may be disconnected from the network for saving power. Thus, the cached data in a mobile node may not be retrieved by other mobile nodes and then usefulness of the cache is reduced. The mobile nodes also decide the caching policy according to the caching status of other mobile nodes. However, the existing cooperative caching in a MANET lack an efficient protocol among the mobile nodes to exchange their localized caching status for caching tasks. In this work a novel cooperative caching scheme called Group caching (GC) which maintains localized caching status of 1 hop neighbors for performing the tasks of data discovery, caching placement, and caching replacement when a data request is received in a mobile node. Each mobile node and its 1 hop neighbors form a group by using the "Hello" message mechanism. In order to utilize the cache space of each mobile node and its 1 hop neighbors form a group by using space of each mobile in a group, the mobile nodes periodically send their caching status in a group. Thus, when caching placement and replacement need to be performed, the mobile node selects the appropriate group member to execute the caching task in the group; this reduces redundancy of cached data objects [14].

Another work is intelligent caching a technique in which, a node not only saves the path discovered during route discovery for itself but also for others who are located close to it. This technique reduces the number of route request packets unnecessarily circulating in the network, when the path it requires is present in its neighborhood.[15]. Authors of [16] in order to share internet contents among mobile users by utilizing low cost wireless connectivity, a content delivery framework with a new content perfecting strategy (AGCS).Another work in which cache management, cooperative caching increase the effective capacity of cooperative caches by minimizing

duplications within the cooperation zone and accommodating more data varieties. In this work authors evaluate the performance of the neighbor Group Data caching by using NS2 and compare it with the existing schemes such as Neighbor caching and Zone Cooperative [17].In[18] Authors Proposed epoch numbers, to reduce the problem of cache staleness, by preventing the re-learning of stale knowledge of a link after having earlier heard that the link has broken. In [19] Authors discuss undesirable side effect including cache inefficiencies due to stale paths, and the use of low quality paths even when significantly shorter path become available .

4. An Efficient Associative Search Scheme

Associative memories are mainly used for the faster search and ordered retrieval of large files of records. Many researchers have suggested using associative memories for implementing relational database machines. In this paper, we have proposed a generic searching algorithm on associative cache memory organization to faster searching single/multiple paths for destination if exist in intermediate mobile node cache with a complexity $O(n)$ (Where n is number of bits required to represent the searched field).So for that tabulation of routing records of mobile nodes can be programmed into the cells of an associative memory.

Various Associative Search operations have been classified in to the following categories.

Extreme Search:

Maxima: Find the largest one among a set of records.

Minima: Find the smallest one among a set of records.

Median: Find the median according to a particular ordering.

Equivalence Search:

Equal To: Search is made for a exact match.

Not Equal To: Find all the records which are not equal to given key.

Similar To: Search is made within a masked field.

Proximate To: Find all the records which satisfy a proximate condition.

Threshold Search:

Smaller Than: Find all the records which are strictly smaller than the given key.

Greater Than: Find all the records which are strictly greater than the given key.

Not Smaller Than: Find all the records which are equal to or greater than the given key.

Not Greater Than: Find all the records which are equal to or less than the given key.

Adjacency Search:

Nearest Below: Find nearest record in which key is smaller than the given key.

Between limit Search

$[x, y]$: find all records within the closed range.

$\{Z \mid X \leq Z \leq Y\}$

(X, Y) : Find all records in the open range.

$\{Z \mid X < Z < Y\}$

$[X, Y)$: Find all records within in the range.

$\{Z \mid X \leq Z < Y\}$

$(X, Y]$: Find all records within in the range

$\{Z \mid X < Z \leq Y\}$

Ordered Search:

Ascending Sort: List all records in the ascending order.

Descending Sort: List all records in the descending order.

Table 1.List of Abbreviations:

| | |
|-------------------|--|
| C | n bit comparative register |
| M | n bit masking register |
| $I(0)$ and $T(0)$ | $I(0)$ and $T(0)$ are Index and Temporary $N * 1$ bit registers initially set. |
| N | Number to be searched |
| K | The Index within the field, where $1 \leq k \leq f$ |
| J | The index for successive bit slices, where $1 \leq j \leq m$ |
| R_i and S_i | The reset and set signals of flip flop I_i are denoted as R_i and S_i |
| S | The starting bit address of a field, where $1 \leq s \leq n$ |
| f | The field length in bits. |
| i | The index for different bit slices, where $1 \leq i \leq n$ |
| B_{ij} | Represent j th bit position of i th memory word |

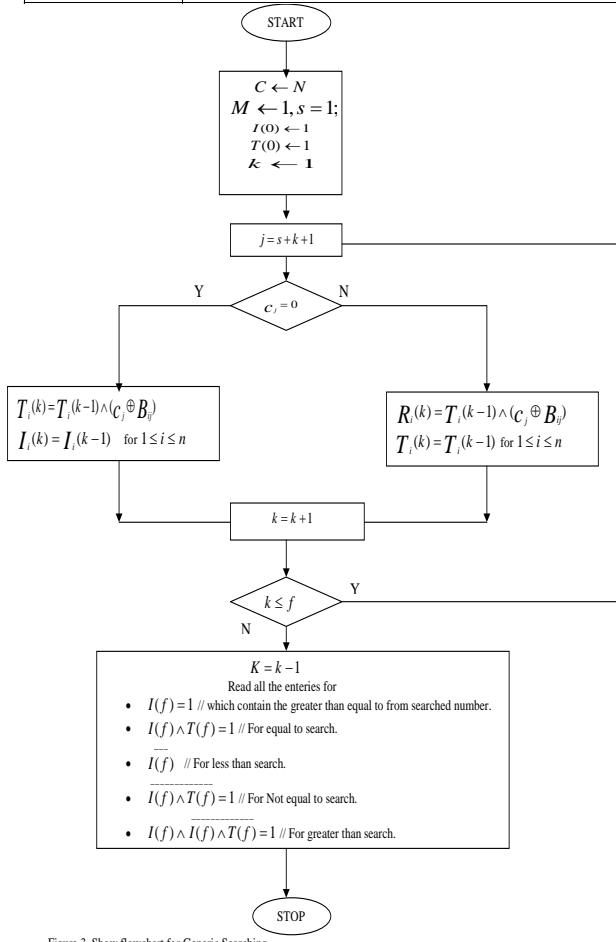


Figure 3. Show flowchart for Generic Searching

4.1 Generic search algorithm for search number from Associative Cache memory Organization

Generic-search ($C, N, n, f, s, S_i, R_i, m, k, I(0), T(0)$)

Step1 : $C \leftarrow N$

Step2 : $M \leftarrow 1, s = 1$; // M is n bit masking register & all its bits are 1 (to search all the bits of register C)

Step3 : $I(0) \leftarrow 1$ // $I(0)$ is an $N * 1$ bit Index Register (Initially set)

Step4 : $T(0) \leftarrow 1$ // $T(0)$ is an $N * 1$ bit Temporary Register (initially set)

Step5 : $k \leftarrow 1$ // k is used for two purpose
I) To point the next bit position in the field.
II) To represent stage number.

Step6 : $j = s + k + 1$

if ($C_j = 0$)

$T_i(k) = T_i(k-1) \wedge (C_j \oplus B_{ij})$

$I_i(k) = I_i(k-1)$ for $1 \leq i \leq n$

else

$R_i(k) = T_i(k-1) \wedge (C_j \oplus B_{ij})$

$T_i(k) = T_i(k-1)$ for $1 \leq i \leq n$

Step7 : $k = k + 1$

if ($k \leq f$) then go to *Step6*

else

$k = k - 1$

Read the Index register $I(f) = 1$ their set bits positions contain the number greater than equal to from searched number.

Read all the entries of $I(f) \wedge T(f) = 1$ its set bits position for equal to search.

Read all the entries of $I(f) \wedge \overline{T(f)} = 1$ its set bits position for less than search.

Read all the entries of $I(f) \wedge \overline{I(f)} \wedge T(f) = 1$ its set bits position for Not equal to search.

Read all the entries of $I(f) \wedge \overline{I(f)} \wedge \overline{T(f)} = 1$ its set bits position for greater than search.

4.2. Implementation of Generic Search Algorithm with example.

In associative memory time required to find an item stored in memory can be reduced considerably because stored data can be identified by the content of the data

itself rather than by the address. An associative memory is also known as content addressable memory (CAM).The block diagram of an associative memory is shown in figure. It consists of a memory array for 4 words with 4 bits per word. The comparative register C hold the item that you want to search and masking register M are also 4bits. The masking register provides a mask for choosing a particular field. The entire bits of register C are compared with each memory word if the masking register contains all 1's. There are also two 4*1 bit size index (I) and temporary (T) which are initially set. In this section, Implementation of Generic search Algorithm on example is shown below.

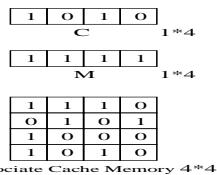
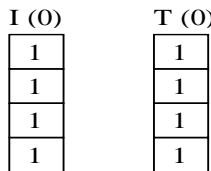


Figure 4. Shows Associate Memory



step1: $k = 1$, $j = s + k - 1 = 1$ and $f = 4$;

step2: $C_1 = 1$ \\There fore the Index Register effected

and no change in Temporary Register.

$R_1(1) = (1) \wedge (1 \oplus 1) = 0$ i.e. No Change So the value of I(1) & T(1)

$R_2(1) = (1) \wedge (1 \oplus 0) = 1$ Reset



$R_3(1) = (1) \wedge (1 \oplus 1) = 0$ i.e. No Change

$R_4(1) = (1) \wedge (1 \oplus 1) = 0$ i.e. No Change

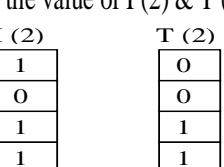
step1: $k = 2$, $j = s + k - 1 = 2$

step2: $C_2 = 0$ \\There fore the Temporary Register

effected and no change in Index Register

$T_1(2) = (1) \wedge (0 \oplus 1) = 0$ So the value of I(2) & T(2)

$T_2(2) = (1) \wedge (0 \oplus 1) = 0$



$T_3(2) = (1) \wedge (0 \oplus 0) = 1$

$T_4(2) = (1) \wedge (0 \oplus 0) = 1$

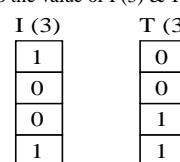
step1: $k = 3$, $j = s + k - 1 = 3$

step2: $C_3 = 1$ \\There fore the Index Register effected

and no change in Temporary Register.

$R_1(3) = (0) \wedge (1 \oplus 1) = 0$ i.e. No Change So the value of I(3) & T(3)

$R_2(3) = (0) \wedge (1 \oplus 0) = 0$ i.e. No Change



$R_3(3) = (0) \wedge (1 \oplus 0) = 1$ Reset

$R_4(3) = (0) \wedge (1 \oplus 1) = 0$ i.e. No Change.

step1: $k = 4$, $j = s + k - 1 = 4$

step2: $C_4 = 0$ \\There fore the Temporary Register

effected and no change in Index Register.

So the value of I (4) & T (4)

| $T_1(4) = (0) \wedge (0 \oplus 0) = 0$ | I (4) | T (4) |
|--|-------|-------|
| $T_2(4) = (0) \wedge (0 \oplus 1) = 0$ | 1 | 0 |
| $T_3(4) = (1) \wedge (0 \oplus 0) = 1$ | 0 | 1 |
| $T_4(4) = (1) \wedge (0 \oplus 0) = 1$ | 1 | 1 |

Read the Index register $I(4) = 1$ its set bits position contain the number greater than equal to from searched number

Read all the entries of $I(4) \wedge T(4) = 1$ its set bits position for equal to search.

Read all the entries of $I(4) = 1$ its set bits position for less than search.

Read all the entries of $I(4) \wedge T(4) = 1$ its set bits position for Not equal to search.

Read all the entries of $I(4) \wedge I(4) \wedge T(4) = 1$ its set bits position for greater than searched number.

5. Proposed an Optimized Cache Coherence Handling Scheme

Ad hoc networks (MANETS) are self-created and self organized by a collection of mobile nodes, interconnected by multi-hop wireless paths in a strictly peer to peer fashion. Caching is an important part of any on-demand routing protocol for wireless ad hoc networks. In mobile ad hoc network (MANETS) all nodes cooperate in order to dynamically establish and maintain routing in the network, forwarding packets for each other to allow communication between nodes not directly within wireless transmission range. For that Several optimization techniques have been incorporated into the basic DSR protocol to improve the performance of the protocol. DSR uses the route cache at intermediate nodes. The route cache is populated with routes that can be extracted from the information contained in the data packets that get forwarded. This cache information is used by the intermediate nodes to reply to the source when they receive a RouteRequest packet during Route Discovery Phase.

Due to presence of private cache for each mobile node in an ad hoc network necessarily introduces problems of cache coherence, which may result in data inconsistency. Clearly, the cache coherence problem cannot be solved by a memory Write-through policy. If a Write-through policy is used, the main memory location is updated, but the possible copies of the routing information in other caches are not automatically updated by the write-through mechanism. So “Write-through: is neither necessary nor sufficient for cache coherence. For that in this paper, we have proposed a dynamic coherence check scheme for

cache coherence in routing table of mobile nodes for MANETs.

In this existing scheme, called dynamic coherence check, multiple copies are allowed. However, whenever a mobile node moves and modifies routing information in its local cache, it must check the other caches to invalidate possible copies. This operation is referred to as a *cross-interrogate* (XI). In other words, when a mobile node writes into a shared block X in its cache, the node sends a signal to all the remote caches to indicate that the “data at memory address X has been modified.” At the same time, it writes through memory. Note that, to ensure correctness of execution, a mobile node which requests an XI must wait for an acknowledge signal from all other mobile nodes before it can complete the write operation. The XI invalidates the remote cache location corresponding to X if it exists in that cache. When the other mobile node references this invalid cache location, it results in a cache miss, which is serviced to retrieve the block containing the updated information. In this approach, for each write operation, $(n - 1)$ XIs result, where n is the number of mobile nodes. Note that the two sources of inefficiency for this technique are the necessity of a write-through policy, which increases the network traffic, and the redundant cache XIs which are performed. In the latter case, a cache is purged blindly whether or not it contains the data item X.

In our proposed scheme, our objective is to optimize cache coherence handling scheme. In this scheme, we focus on a more refined technique filters the *cross-interrogate* (XI) requests before they are initiated on reactive routing protocol DSR for mobile ad hoc network. For that, we have ad hoc network in which every mobile node having own local cache and there is one mobile node having the centralized shared main memory. This main memory contains the memory control element (MSC) maintains a central copy of the directories of all the caches. We will elaborate on a similar scheme called the presence flag technique, which assumes a write-back main memory update policy. There are two central tables associated with the blocks of main memory (MM) as shown in Figure 5. The first table is a two-dimensional table called the *Present table*. In this table, each entry $P[i, c] = 1$, contains a *present* flag for the i th block in MM and the c th cache. If $P[i, c] = 1$, then the c th cache has a copy of the i th block of MM, otherwise it is zero. The second table is the *Modified table* and is one-dimensional. In this table, each entry $M[i]$ contains a *modified* flag for the i th block of MM. If $M[i] = 1$, it means that there exists a cache with a copy of the i th block more recent than the corresponding copy in MM. The Present and Modified tables can be implemented in a fast random-access memory. The philosophy behind the cache coherence check is that an arbitrary number of caches can

have a copy of a block, provided that all the copies are identical. They are identical if the Mobile node associated with each of the caches has not attempted to modify its copy since the copy was loaded in its cache. We refer to such a copy as *read only* (RO) copy. In order to modify a block copy in its cache, a mobile node must own the block copy with *exclusive read only* (EX) or *exclusive read-write* (RW) access rights. A copy is held EX in a cache if the cache is the only one with the block copy and the copy has not been modified. Similarly, a copy is held RW in a cache if the cache is the only one with the block copy and the copy has been modified. Therefore, for consistency, only one mobile node can at any time own an EX or RO copy of a block.

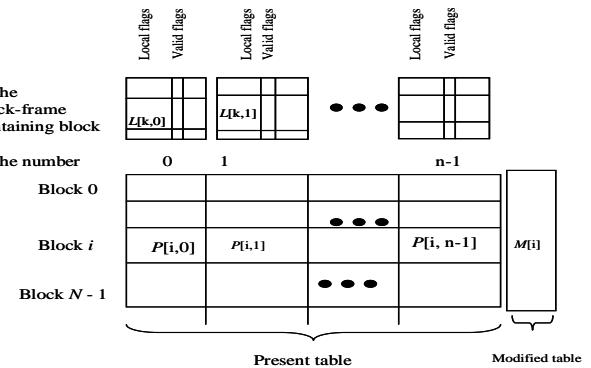


Figure 5. Organization of flags for dynamic solution to cache coherence

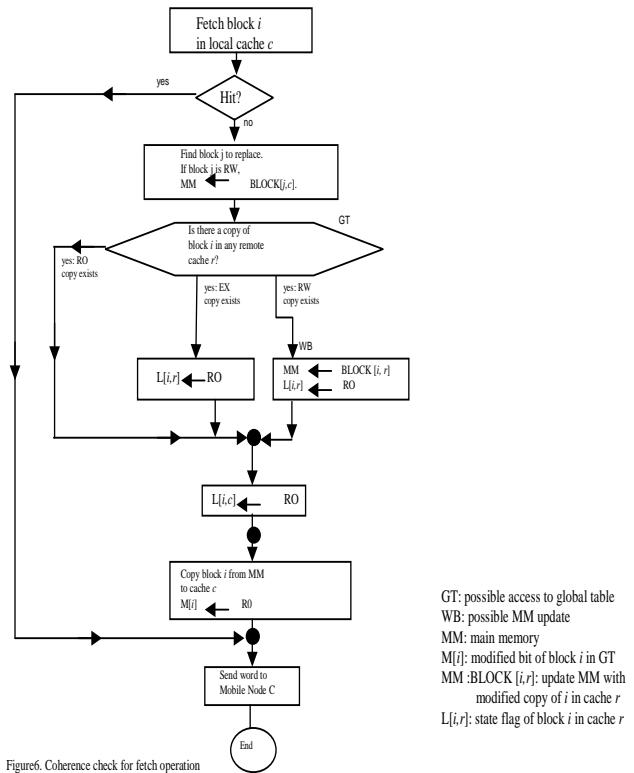


Figure 6. Coherence check for fetch operation

To enforce the cache consistency rule, local flags are provided within each cache in addition to the global tables. A local flag $L[k, c]$ is provided for each block k in cache c . This flag indicates the state of each block in the cache. A block in a cache can be in one of three states: RO, EX, or RW. Figure 6 shows the flowchart for the HIT or MISS when mobile node c fetches the block i th from their local cache.

As long as the copy of block i remains present in the cache, mobile node c can fetch it without any consistency check. If mobile node c attempts to store into its copy of block i , it must ensure that all other copies (if any) of block i are invalidated. To do this, the global table is consulted. It should indicate the mobile node caches that own a copy of block i . The modified bit for block i is updated in the global table to record the fact that mobile node c owns block i with RW access rights. Finally, the local $L[k, c]$ flag is set to RW to indicate that the block is modified.

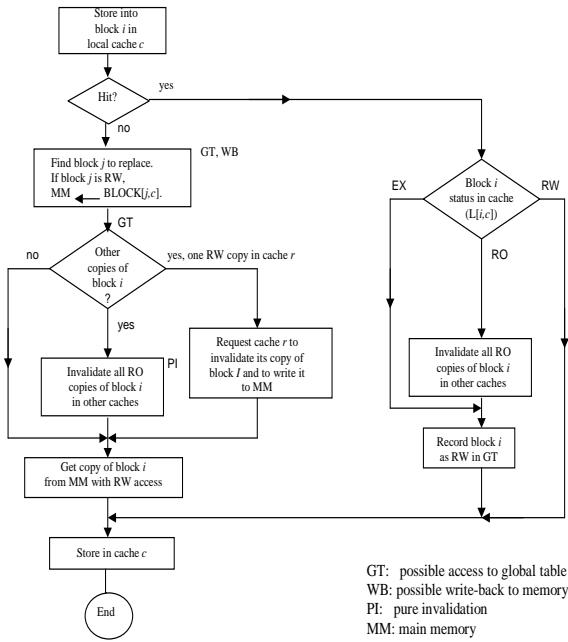


Figure 7. Coherence check for store operation

The flowchart for a store is shown in Figure 7. In this implementation, a block copy in a cache is invalidated whenever the cache receives a signal from some other mobile node attempting to store into it. Moreover, a cache which owns an RW copy may receive a signal from a remote cache requesting to own an RO copy. In this case, the RW copy's state is changed to RO.

6. Conclusions and Future work

In this paper, we have proposed a generic searching algorithm on associative cache memory organization to enhance the searching of single/multiple path for destination, if exist, in intermediate mobile node cache with a complexity $O(n)$ (Where n is number of bits required to represent the searched field). The proposed algorithm reduces the route discovery over head. We have proposed an optimized dynamic coherence check scheme to reduce the number of cross-interrogate (XI) signal sends to different mobile node caches resulting in the reduction of routing overhead on DSR protocols. In future work, the focus will be on the effect of different page replacement policies [FIFO, LRU etc.] of cache and to reduce the number of conflicts that occurs when concurrent access to the global table occurs.

References

- [1] Jeremy Pitt, Pallapa Venkataram, and Abe Mamdani, "QoS Management in MANETS Using Norm-Governed AgentSocieties" ESAW 2005, LNAI 3963, Page(s): 221- 240, 2006.
- [2] Internet Engineering Task Force MANET Working Group. Mobile Ad hoc networks (Manet) Charter Available at <http://www.ietf.org/html.charters/manet-charter.html>.
- [3] Asis Nasipuri, Mobile Adhoc networks,Department of Electrical and Computer Engineering, The university of North Carolina at charlotte.Charlotte,NC 28233-0001.
- [4] C.E. Perkins, Ad hoc networking, Addison-Wesley, 2001.
- [5] D.Johnson, Rice University; Y. Hu,UIUC and D.Maltz, Microsoft Research The Dynamic Source Routing Protocol (DSR) for mobile ad hoc networks for IPV4, February 2007 <http://www.ietf.org/rfc/rfc4728.txt>.
- [6] C. Perkins and S. Das, Ad hoc On Demand Distance Vector (AODV) Routing IETF, Internet Draft, draft-ietf-manet-aodv-13,RFC 3561, 17.February -17- 2003.
- [7] Z. Haas, M. Pearlman nad P. Smar, Zone routing protocol (ZRP), Internet Draft, Internet Engineering Task Force , Jan 2001,<http://www.ietf.org/internet-drafts-ietf-manet-zoneirp-00.txt>.
- [8] S. Bradner , temporally-ordered routing algorithm (TORA) Routing IETF, Internet draft-ietf-manet-tora-spec-04.txt,RFC 2026,July 2001.
- [9] Y. Kuo, and N.H. Vaidya, " Location –Aided Routing (LAR) Mobile Ad Hoc Networks," In proceedings of the International Conference on Mobile Computing and Networking (MobiCom'98),Oct.1998.
- [10] David .B Johnson, David. A. Maltz, and Josh Broch, " Dynamic Source Routing protocol for Multihop Wireless Ad Hoc Networks," In Ad Hoc Networking, edited by Charles E. Perkins, chapter 5, pages 139-172. Addison-Wesley, 2001.
- [11] LiangZhong Yin and Guohong Cao, " Supporting Cooperative caching in ad hoc networks ,," IEEE Transactions on Mobile computing, Vol. 5, Issue 1,pages 77-89.jan. 2006.
- [12] Joonho Cho, Seungtaek Oh, Jaemyoung Kim, Hyeong Ho Lee, and Joonwon Lee, " Neighbor caching in multi-hop

- wireless ad hoc networks," IEEE Communications Letters, Vol 7, issue Nov. , pages 525-527,2003.
- [13] Sunlook Jung,Nisar Hundewale, and Alex Zelikovsky, " Node Caching Enhancement of Reactive Ad hoc routing Protocols," IEEE Wireless Communications and Networking Conference, 2005.
- [14] Yi-Wei Ting and Yeim-Kuan Chang, " A novel Cooperative Caching Scheme for Wireless Ad hoc Networks; GroupCaching," International Conference on Networking Architecture and storage ,NAS 2007.
- [15] Shobha.K.R., and K. Rajanikanth "Intelligent caching in On-Demand Routing Protocol for Mobile Adhoc Networks" World Academy of Science Engineering and Technology 56 pages 413-420,2009.
- [16] Yaozhou Ma, M. Rubaiyat Kibria, and Abbas Jamalipour "Cache-based Content Delivery in Opportunistic Mobile Ad hoc Networks" IEEE "GLOBECOM", 2008.
- [17] Mrs. K. Shammugavadi and Dr. M. Madheswaran "CachingTechnique for Improving Data Retrieval Performance in Mobile Ad Hoc Networks" In International Journal of Computer Science and Information Technologies (IJCSIT),Vol. 1(4),pages 249-255,2010.
- [18]Yih-Chun Hu and David B.Johnson " Ensuring Cache Freshness in On-Demand Ad hoc Network Routing Protocols" In POMC'02, October 30-31, Toulouse, France ,2002.
- [19] Nikhil I. Panchal and Nael B. Abu-Ghazaleh "Active Route Cache Optimization for Ad hoc Networks" In Infocom 2002.

First Author Rajneesh Kumar Gujral is working as Assoc. Professor Department of Computer Engineering, M.M Engineering College, M.M.University Mullana, Ambala. He obtained his BE (Computers) in 1999 unit SLIET Longowal from Punjab Technical University(PTU), Jalandhar. He also obtained his MTECH (IT) in 2007 from University School of Information Technology, GGSIP University Delhi. He has about 10 publications in journals and International Conferences to his credit. His current research interest includes Wireless communications which include mobile, Adhoc and sensor based networks, Network Security and computer communication networks etc.

Second Author Dr. Anil Kumar Kapil is working as Professor. & Principal M. M. Institute of Computer Technology and Business Management, M. M. University Mullana, Ambala, India. He obtained his Ph.D. (Computer Science & Engineering) in 2007. He has about 25 publications in journals and International Conferences to his credit. His current research interest includes Wireless communications which include mobile, Adhoc and sensor based networks, computer communication networks, Distributed networks and Concurrency control etc.

Fingerprint Matching Using Hierarchical Level Features

D. Bennet¹, Dr. S. Arumuga Perumal²

¹ Assistant Professor & Scholar, Department of Computer Science, S.T. Hindu College, Nagercoil, India.

² Associate Professor & Head, Department of Computer Science, S.T. Hindu College, Nagercoil, India

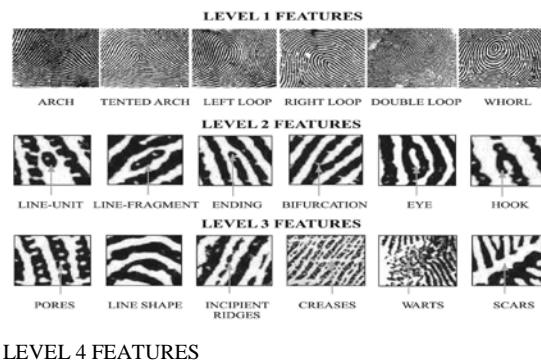
Abstract - This paper proposes a fingerprint features extraction using different levels. The hierarchical order at four different levels, namely, Level 1 (pattern), Level 2 (minutia points), Level 3 (pores and ridge contours), and Level 4 (oscillated pattern). The fingerprint feature extraction frequently take advantage of Level 4 features to assist in identification. Automated Fingerprint Identification Systems (AFIS) currently rely only on Level 1 and Level 2 features. In fact, the Federal Bureau of Investigation's (FBI) standard of fingerprint resolution for AFIS is 500 pixels per inch (ppi), which is inadequate for capturing Level 3 features, such as pores. With the advances in fingerprint sensing technology, many sensors are now equipped with dual resolution (1,000 ppi) scanning capability. However, increasing the scan resolution alone does not necessarily provide any performance improvement in fingerprint matching, unless an extended feature set is utilized. As a result, a systematic analysis to determine how much performance gain one can achieve by introducing Level 4 features in AFIS is highly desired. We propose a hierarchical matching system that utilizes features at all the four levels extracted from 1,000-ppi fingerprints scans. Level 3 features, pores and ridge contours are automatically extracted using Gabor filters and wavelet transform and are locally matched using the Iterative Closest Point (ICP) algorithm and Level 4 features, oscillated pattern including curve scanned DCT to measure the recognition rate using k -nn classifier. Our analytical study conclude Level 4 features carry significant discriminatory information. The matching system when Level 4 features are employed in combination with Level 1 Level 2 and Level 3 features. This proposed method outperforms the others, particularly in recognition rate.

Keywords - Fingerprint features, minutia, Level features, extended feature set, pores, ridge contours, oscillated pattern, curve-DCT, hierarchical matching.

I. INTRODUCTION

Fingerprint based biometric systems are rapidly gaining acceptance as one of the most effective technologies to authenticate users in a wide range of applications. In fingerprint identification is based on two properties, namely, uniqueness and permanence.

The fingerprint features are generally categorized into four levels. Level 1 features, or patterns, are the macro details of the fingerprint such as ridge flow and pattern type. Level 2 features, or minutiae, such as ridge bifurcations and endings. Level 3 features, or shape, include all dimensional attributes of the ridge such as ridge path deviation, width, shape, pores, edge contour, incipient ridges, breaks, creases, scars, and Level 4 or oscillated pattern curve scanned DCT.



LEVEL 4 FEATURES

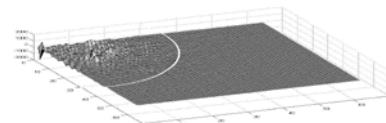


Fig.1 Fingerprint features at Level 1, Level 2, Level 3 and Level 4

Fig.1 Shown that Level 1 features, though not unique, are useful for fingerprint classification (e.g., into whorl, left loop, right loop, and arch classes), while Level 2 features have sufficient discriminating power to establish the individuality of fingerprints. Similarly, Level 3 features are also claimed to be permanent, immutable, and unique according to the forensic experts, and Level 4 features if properly utilized, can provide discriminatory information for human identification [3]. Both Level 3 and Level 4 features play important roles in providing quantitative as well as qualitative information for identification

II.HISTORY

Fingerprints as a scientific method for identification traces back to the 1880s, when Faulds suggested that latent fingerprints obtained

at crime scenes could provide knowledge about the identity of offenders [4]. In 1892, Galton introduced Level 2 features by defining minutia points as either ridge endings or ridge bifurcations on a local ridge. In 1912, Locard introduced the science of poroscopy, the comparison of sweat pores for the purpose of personal identification.

In 1962, Chatterjee discovered that some shapes on the friction ridge edges tend to reappear frequently and classified them into eight categories, namely, straight, convex, peak, table, pocket, concave, angle, and others (see Fig. 2).

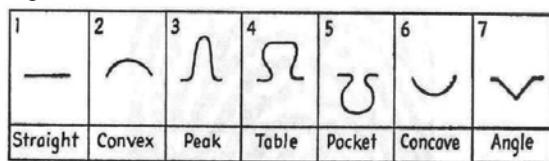


Fig. 2 Characteristic features of friction ridges

It is believed that the differences in edge shapes are caused by the effects of differential growth on the ridge itself or a pore that is located near the edge of the friction ridge.

A. Fingerprint Formation

Human fingers are known to display friction ridge skin (FRS) that consists of a series of ridges and furrows, generally referred to as fingerprints. The FRS is made of two major layers: dermis (inner layer) and epidermis (outer layer). Pores, on the other hand, penetrate into the dermis starting from the epidermis. Each ridge unit contains one sweat gland, pores are often considered evenly distributed along ridges and the spatial distance between pores frequently appears to be in proportion to the breadth of the ridge, which, on an average, is approximately 0.48 mm. A pore can be visualized as either open or closed in a fingerprint image based on its perspiration activity. A closed pore is entirely enclosed by a ridge, while an open pore intersects with the valley lying between two ridges (see Fig. 3).

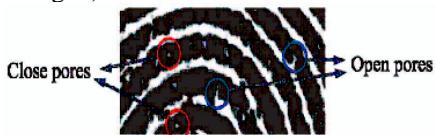


Fig. 3 Open and closed pores

One should not expect to find two separate prints of the same pore to be exactly alike, as a pore may be open in one and closed in the other print.

B. Fingerprint Sensing Technology

There are many different sensing methods to obtain the ridge-and-valley pattern of finger skin or fingerprint [10]. Historically, in law enforcement applications, fingerprints were mainly acquired offline. Nowadays, most commercial and forensic applications accept live-scan digital images acquired by directly sensing the finger surface with a fingerprint sensor based on optical, solid-state, ultrasonic, and other imaging technologies.

Direct sensing of fingerprints as electronic signals started with optical "live-scan" sensors with Frustrated Total Internal Reflection (FTIR) principle. When the finger touches the top side of a glass prism, one side of the prism is illuminated through a diffused light. While the fingerprint valleys that do not touch the glass platen reflect the light, ridges that touch the platen absorb the light. This differential property of light reflection allows the ridges (which appear dark) to be discriminated from the valleys. Solid-state fingerprint sensing technique uses silicon based, direct contact sensors to convert the physical information of a fingerprint into electrical signals.

New fingerprint sensing technologies MSI (Multispectral Fingerprint Imaging) technology appear to be of significantly better quality compared to conventional optical sensors for dry and wet fingers. One of the most essential characteristics of a digital fingerprint image is its resolution, which indicates the number of dots or pixels per inch (ppi) (see Fig. 4).



Fig. 4 Fingerprint resolution same fingerprint captured at different image resolutions
 (a) 380 ppi (b) 500 ppi and (c) 1,000 ppi

III. PREVIOUS WORKS

Only a few researchers have studied the use of Level 3 features in an automated fingerprint identification system. Skeletonization - based pore extraction and matching algorithm [6]. Specifically, the locations of all end points (with at most one neighbor) and branch points (with exactly three neighbors) in the skeleton image are extracted and each end point is used as a starting location for tracking the skeleton. The tracking algorithm advances one element at a

time until one of the following stopping criteria is encountered: 1) another end point is detected, 2) a branch point is detected, and 3) the path length exceeds a maximum allowed value. Condition 1) implies that the tracked segment is a closed pore, while Condition 2) implies an open pore. Finally, skeleton artifacts resulting from scars and wrinkles are corrected and pores from reconnected skeletons are removed. The result of pore extraction is shown in Fig. 5.



Fig.5 Pore detection based on skeletonization (a) Fingerprint detected pores. (b) The raw skeleton image

During matching, score between a given image pair is then defined as a ratio of the number of matched pores to the total number of pores extracted from template regions,

1. Skeletonization is effective for pore extraction only when the image quality is very good.
2. Comparison of small fingerprint regions based on the distribution of pores.ments.
3. The alignment of the test and the query region is established based on intensity correlation, which is computationally expensive by searching through all possible rotations and displacements.

4.

IV. LEVEL3 FEATURE EXTRACTIONS

As suggested in [7], Level 1, Level 2, and Level 3 features in a fingerprint image are mutually correlated. Based on the physiology of the fingerprint, pores are only present on the ridges, not in the valleys. During image acquisition, we observe that the ridge contour is often more reliably preserved at 1,000 ppi than the pores, especially in the presence of various skin conditions and sensor noise (see Fig. 6).



Fig. 6 Two impressions of the same finger at 1,000 ppi

In order to automatically extract Level 3 features, namely, pores and ridge contours, we

have developed feature extraction algorithms using Gabor filters and wavelet transform.

A. Pore Detection

Based on their positions on the ridges, pores can be divided into two categories: open and closed. A closed pore is entirely enclosed by a ridge, while an open pore intersects with the valley lying between the two ridges.

To enhance the ridges, we use Gabor filters, which has the form

$$G(x, y : \theta, f) = \exp\left\{-\frac{1}{2}\left[\frac{x^2}{\delta_x^2} + \frac{y^2}{\delta_y^2}\right]\cos(2\pi f x \theta)\right\}$$

Where θ and f are the orientation and frequency of the filter, respectively, δ_x and δ_y are the standard deviations of the Gaussian envelope along the x- and y-axes, respectively. Here, $(x \theta; y \theta)$ represents the position of a point (x, y) after it has undergone a clockwise rotation by an angle $(90^\circ - \theta)$. The four parameters of the Gabor filter are $(\theta, f, \delta_x, \delta_y)$ ridge frequency and orientation .

B. Ridge Contour Extraction

The ridge contour is defined as edges of a ridge. However, the flexibility of the friction skin and the presence of open pores tend to reduce the reliability of ridge edge classification. In contrast to edgeoscopy, our method utilizes the ridge contour directly as a spatial attribute of the ridge and the matching is based on the spatial distance between points on the ridge contours.



Fig. 7 Ridge contour extraction.

(a) Wavelet (b) Ridge contour (c) Gabor

Classical edge detection algorithms can be applied to fingerprint images to extract the ridge contours. However, the detected edges are often very noisy due to the sensitivity of the edge detector to the presence of creases and pores. Hence, we again use wavelets to enhance the ridge contours and linearly combine them with a Gabor enhanced image (where broken ridges are fixed) to obtain enhanced ridge contours.

V. LEVEL 4 FEATURE EXTRACTIONS

The informative features extraction method is curve-scanned DCT. The extracted features can subsequently be used for fingerprin

matching process. The curve-scanned DCT coefficients have a high matching score and also its evaluated by the k -nearest neighbor (k -NN) classifier and the time required in the processing steps are compared.

However, we found that the informative features used for matching purpose exist within the low frequency area (top-left corner) of the distribution plane only. From our observation, magnitude of most features extracted from the middle and high frequency areas (around 75%) of DCT coefficients, from different fingerprint images, are hardly changed. Hence, we bound the energy compactness area to be used for creating the fingerprint features by a white arch as shown in Fig.8. This boundary is actually defined by the oscillated pattern contained within the top-left corner of the distribution plane of the DCT coefficients, which disperses equally in all directions from the DC component. The DCT coefficients divided in this fashion are thus referred to as curve-scanned DCT coefficients. Note that the area within the boundary is approximately 25% of the distribution plane.

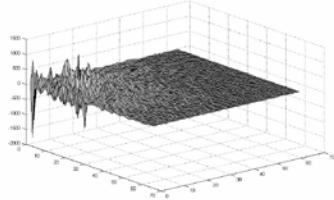


Fig. 8: DCT coefficients distributions

To generate the fingerprint features from the curve-scanned DCT coefficients, firstly we cropped a gray-scale fingerprint image to the size of 64×64 pixels, where the reference point was at the center of the cropped image. Secondly, the cropped image was quartered to obtain four non-overlapping images of size 32×32 pixels. Then, the DCT was applied to each non-overlapping image. The DCT coefficients within the boundary were divided into 5 areas in the curve-scanned fashion, where each area was 2-pixel width (see Fig. 9)



Fig.9 Curve-scanned DCT coefficients

Finally, the standard deviation of the DCT coefficients in each area from all 4 non-overlapping images was computed to create a feature vector of the length 20 (5 features from each non-overlapping image). The performance indicator we used to evaluate our proposed

method was the recognition rate obtained from the k nearest neighbors (k -NN) classifier with no rejection option. Recall that, in k -NN classifier, the database is divided into 2 data sets, namely training set and testing set. Basically, training set is a set of fingerprint images stored in a fingerprint image database, while testing set is a set of entry fingerprint images. Hence, k nearest neighbor is merely k instances in the training set, which nearest to the testing set, and nearest neighbor is evaluated by the distance between both sets. The nearest neighbors of an instance are normally defined in term of the standard Euclidean distance [1]. Given x_i and x_j are the feature vectors from the training sets and testing sets, respectively, the distance between two feature vectors is then defined to be $d(x_i, x_j)$, where

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^n (x_i - x_j)^2}$$

For the complexity indicator, we measured the processing time required in the features extraction process and the fingerprint matching process. This can be easily achieved by timing such processes.

VI. HIERARCHICAL MATCHING

Fig. 10 illustrates the architectural design of our proposed matching system. Each layer in the system utilizes features at the corresponding level.

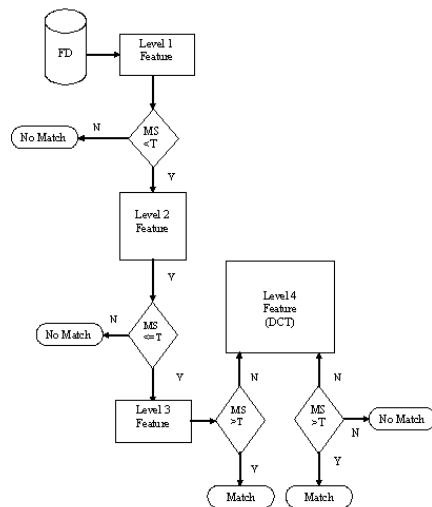


Fig.10 Hierarchical matching system flow chart

In general, the numbers of impostor minutia matches decreases after Level 3 features are used, while the number of genuine minutia

Matches remain almost unchanged. As a result, the overlap region of the genuine and impostor distributions of matched minutiae is reduced after Level 3 features were utilized. Although the latter is a more commonly used and straightforward approach, it is more time-consuming since matching at both Level 2 and Level 3 has to be performed for every query. In addition, parallel score fusion is sensitive to the selected normalization scheme and fusion rule. On the other hand, the proposed hierarchical matcher enables us to control the level of information or features to be used at different stages of fingerprint matching.

VII. EXPERIMENTAL RESULTS

In general, our experiments show significant performance improvement when we combine Level 3 and Level 4 features in a hierarchical fashion. It is demonstrated that Level 3 features do provide additional discriminative information and should be used in combination with Level 2 features. The results of this study strongly suggest that using Level 4 features in fingerprint matching at 1,000 ppi is both practical and beneficial.

VIII. SUMMARY AND CONCLUSIONS

We conclude that the fingerprint feature extraction scheme for hierachal levels the variation in matching score is presented as a Table and graph below.

Table.1 matching score

| Diff. Levels | Matching Score |
|--------------|----------------|
| Level 1 | 65 |
| Level 2 | 84 |
| Level 3 | 92 |
| Level 4 | 97 |

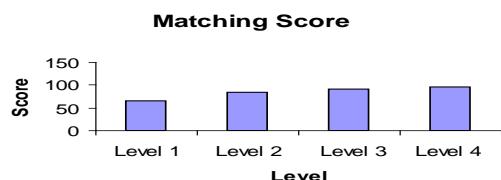


Fig.11 Matching score presentation

To obtain the discriminatory information at level 4 we introduced. The algorithm based on curve – scanned DCT to automatically extract maximum features. In our analytical results strongly proposed Level 4

feature extraction is high matching score and improve the recognition rate.

REFERENCES

- [1] S. Tachaphetpiboon, "Fingerprint Features Extraction Using Curve-scanned DCT Coefficients" 1-4244-1374-5/07@2007 IEEE
- [2] F. Galton, "Personal Identification and Description," Nature, vol. 38, pp. 201-202, 1888.
- [3] H. Cummins and M. Midlo, Finger Prints, Palms and Soles: An Introduction to Dermatoglyphics. Dover, 1961.
- [4] S. Pankanti, S. Prabhakar, and A.K. Jain, "On the Individuality of Fingerprints," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 8, pp. 1010-1025, Aug. 2002.
- [5] J.D. Stosz and L.A. Alyea, "Automated System for Fingerprint Authentication Using Pores and Ridge Structure," Proc. SPIE Conf. Automatic Systems for the Identification and Inspection of Humans, vol. 2277, pp. 210-223, 1994.
- [6] R. McCabe and M. Garris, "Summary of April 2005 ANSI/NIST Fingerprint Standard Update Workshop," <http://fingerprint.nist.gov/standard/>, 2005
- [7] Anil k. Janin, "Pores and Ridges: High Resolution Fingerprint Using Level 3 Features", IEEE vol.29 No.1 2007.
- [8] S. Meagher and A. Hicklin, "Extended Fingerprint Feature Set," Proc. ANSI/NIST ITL 1-2000 Standard Update Workshop, 2005.
- [12] A.K. Jain, S. Prabhakar, and S. Chen, "Combining Multiple Matchers for a High Security Fingerprint Verification System," Pattern Recognition Letters, vol. 20, nos. 11-13, pp. 1371-1379, 1999
- [9] P.J. Besl and N.D. McKay, "A Method for Registration of 3D Shapes," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 14, no. 2, pp. 239-256, Feb. 1992.
- [10] A. Ross, S.C. Dass, and A.K. Jain, "Fingerprint Warping Using Ridge Curve Correspondences," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 1, pp. 19-30, Jan. 2006.
- [11] A.K. Jain, Y. Chen, and M. Demirkus, "Pores and Ridges: Fingerprint Matching Using Level 3 Features," Proc. Int'l Conf. Pattern Recognition, vol. 4, pp. 477-480 Aug. 2006.
- [12] Y. Chen, S.C. Dass, and A.K. Jain, "Fingerprint Quality Indices for Predicting Authentication Performance," Proc. Audio- and Video- Based Biometric Person Authentication, pp. 160-170, 2005.

Computation of Fifth Degree of Spline Function Model by Using C++ Programming

Faraidun K. HamaSalh¹, Alan A. Abdulla² and Khanda M. Qadir³

¹ Mathematics Dept, University of Sulaimani,
Sulaimani, IRAQ

² Mathematics Dept, University of Sulaimani,
Sulaimani, IRAQ

³ Mathematics Dept, University of Sulaimani,
Sulaimani, IRAQ

Abstract

In this paper, a new quintic spline method developed for computing approximate solution of differential equations. It is shown that the present method is of the order three and four derivatives and gives approximations which are better. The numerical result obtained by the present method has been compared with the exact solution using C++ programming and also illustrate graphically the applicability of the new method. By getting the advantages of the mathematical building functions like pow (for power), exp (for exponential),...etc. are provided in C++ programming library, all processing steps are done efficiently and illustrated as Pseudocode model.

Keywords: - Quintic spline, Differential equations, Building functions, Pseudocode.

1. Introduction

A method for approximate solving initial value problems proposed for differential equations. In fact, this method is a variant of the well-known method of spline interpolation considered in [1]. A principal difference between considerations in [1] and ours is that, the new case of lacunary interpolations with others boundary conditions.

This method enables us to approximate the solution as well as its first and third derivatives at every point of the range of integration. We proved that this new method gives better numerical results than the previous known results. In recent

years, Al-Said and Noor [2, 3], Khalifa and Noor [4] and Noor and Al-Said [5, 6] have used such types of penalty function in solving a class of contact problems in elasticity in conjunction with collocation, finite difference and spline techniques.

The general fourth order initial value problem considered is of the form

$$y^{(4)} + f(x)y = g(x), \quad -\infty < x < \infty \quad (1)$$

With the boundary conditions

$$y(x_0) = y_0, \quad y'(x_0) = y'_0, \quad y''(x_0) = y''_0 \text{ and } y'''(x_0) = y'''_0. \quad (2)$$

Where $x_0 = 0$, $x_n = 1$ and that $f \in C^{n-1}([0,1] \times R^4)$, and that f is Lipschitz continuous in y, y', y'', y''' and $y^{(4)}$, similarly for the third order initial value problems.

The aim of this paper is to construct a new spline method based on a quintic spline function that has a polynomial part and to develop numerical methods for obtaining smooth approximations for the solution of the problem (1) subject to the initial conditions (2).

The existence and uniqueness for spline function of degree five which interpolate the lacunary data (1, 3) is presented and examined in Section 2, we derive the numerical method and briefly discuss its error analysis theoretically in Section 3. Convergence analysis for second order, fourth order and fifth order methods is established in Section 4. Numerical results are presented to illustrate the applicability and accuracy their practical usefulness with C++ programming in Section 5. One of the C++ programming powerful includes (cmath) header file. The cmath header file provides a collection of functions that enables programmer to perform common mathematical calculations [7]. The instructions (codes) are illustrated in Pseudocode. Pseudocode is a compact and informal high-level description of a computer programming algorithm that uses the structural conventions of a programming language, but it is intended for human reading rather than machine reading [8].

2. Explanation of the Method

We consider a mesh with nodal points the x_j on $[a, b]$ such that;

$\Delta : a = x_0 < x_1 < \dots < x_{n-1} < x_n = b$ where $h = x_j - x_{j-1}$, $j = 0, 1, 2, \dots, n$. Also we denote a quintic spline function $S_\Delta(x)$, interpolating to a function $y(x)$ defined on $[a, b]$ is such that:

$$S_i(x) = y_i + (x - x_i) y'_i + \frac{(x - x_i)^2}{2} a_{i,2} + \frac{(x - x_i)^3}{3!} y'''_i + (x - x_i)^4 a_{i,4} + (x - x_i)^5 a_{i,5} \quad (3)$$

$$S_i(x_{i+1}) = y_{i+1}, S'_i(x_{i+1}) = y'_{i+1} \text{ and } S''_i(x_{i+1}) = y''_{i+1} \quad (4)$$

On the last interval $[x_{n-1}, x_n]$ we define $S_{n-1}(x)$ as follows:

$$\begin{aligned} S_{n-1}(x) = & y_{n-1} + (x - x_{n-1}) y'_n + \frac{(x - x_{n-1})^2}{2} a_{n-1,2} + \frac{(x - x_{n-1})^3}{3!} y'''_{n-1} \\ & + (x - x_{n-1})^4 a_{n-1,4} + (x - x_{n-1})^5 a_{n-1,5} \end{aligned} \quad (5)$$

Where $a_{n-1,j}$, $j = 1, 3, 5$ and 6 , unknowns are to be determined.

Theorem 1: Existence and Uniqueness Spline Model

Gives the real numbers $y^{(r)}(x_i)$, $i=0, 1, 2, \dots, n$ and $r=0, 1, 3$ and $y'(x_0)$ and $y'(x_n)$ then they exist a unique spline function of degree six from equations (3)-(5) such that:

$$\left. \begin{array}{l} S(x_i) = y(x_i) \\ S^{(r)}(x_i) = y^{(r)}(x_i), r = 1, 3 \\ \text{and} \\ S'(x_0) = y'(x_0) \text{ and } S'(x_n) = y'(x_n) \end{array} \right\} \text{for } i = 0, 1, \dots, n \quad (6)$$

Proof: For whole interval $[x_{i-1}, x_i]$ where $i = 0, 1, 2, \dots, n$. Assuming $y(x)$ to be the exact solution of the equation (3), obtained by the spline $S_i(x)$, along with the continuity condition of the first and third derivatives at $[x_{i-1}, x_i]$ in (4), the following consistency relations are derived:

$$\begin{aligned} h^2 a_{i,2} + h^4 a_{i,4} + h^5 a_{i,5} &= y_{i+1} - y_i - h y'_i - \frac{h^3}{6} y'''_i \\ 2h a_{i,2} + 4h^3 a_{i,4} + 5h^4 a_{i,5} &= y'_{i+1} - y'_i - \frac{h^2}{2} y''_i \\ 24h a_{i,4} + 60h^2 a_{i,5} &= y'''_{i+1} - y'''_i \end{aligned}$$

Solving the above system, the coefficients of $S_i(x)$ on the interval $[x_i, x_{i+1}]$ for $i = 1, 2, 3, \dots, n-2$.

$$a_{i,2} = \frac{5}{2h^2} (y_{i+1} - y_i) - \frac{1}{4h} (3y'_{i+1} + 7y'_i) + \frac{h}{48} (y'''_{i+1} - 3y'''_i) \quad (7)$$

$$a_{i,4} = -\frac{5}{2h^4} (y_{i+1} - y_i) + \frac{5}{4h^3} (y'_{i+1} + y'_i) - \frac{1}{48h} (3y'''_{i+1} + 7y'''_i) \quad (8)$$

$$a_{i,5} = \frac{1}{h^5}(y_{i+1} - y_i) - \frac{1}{2h^4}(y'_{i+1} + y'_i) + \frac{1}{24h^2}(y'''_{i+1} + y'''_i) \quad (9)$$

By solving these equations, we see that the coefficients $a_{n-1,i}$; $i=2, 4$ and 5 are uniquely determined, since we have three equations and three unknowns, and finally, we can find the coefficients of $S_{n-1}(x)$ similarly in the interval $[x_{n-1}, x_n]$. Hence the proof is complete.

3. Convergence analysis

In this section, we investigate the convergence analysis of the quintic spline method described in Section 2. For this purpose, the error bound of the spline function $S(x)$ which is a solution of the problem (3) and (4) is obtained for the uniform partition I by the following theorem:

Theorem 2: Let $y \in C^6[0,1]$ is the exact solution of the differential equations (1) and $S(x)$ be a unique spline function of degree five which a solution of the problem (3) and (4). Then for $x \in [x_i, x_{i+1}]$; $i=0,1,2,\dots,n-1$, we have

$$\|S_i^{(r)}(x) - y^{(r)}(x)\| \leq \begin{cases} \frac{1}{120}h^{5-r}W_5(h) & \text{for } r=5 \\ \frac{1}{24}h^{5-r}W_5(h) & \text{for } r=4 \\ \frac{1}{6}h^{5-r}W_5(h) & \text{for } r=3 \\ \frac{1}{2}h^{5-r}W_5(h) & \text{for } r=2 \\ \frac{9}{4}h^{5-r}W_5(h) & \text{for } r=1 \\ \frac{7}{2}W_5(h) & \text{for } r=0 \end{cases}$$

where $W_6(h)$ denotes the modules of continuity of $y^{(5)}$, defined by $\|W_6(h)\| = \max\{|W_6(x)|; 0 \leq x \leq 1\}$

Proof:

Let $x \in [x_i, x_{i+1}]$ where $i=1, 2, \dots, n-2$.

From equation (4) and the Taylor's expansion formula, we have

$$S^{(5)}_i(x) = 120 a_{i,5}$$

$$|S_i^{(5)}(x) - y^{(5)}(x)| = \left| \frac{120}{h^5} a_{i,5} - y^{(5)}(x) \right| = \left| 120 \left[hy_i + \frac{h^2}{2} y_i'' + \frac{h^3}{6} y_i''' + \frac{h^4}{24} y_i^{(4)} + \frac{h^5}{120} y_i^{(5)} \right] \right. \\ \left. - \frac{60}{h^4} [2y_i + hy_i'' + \frac{h^2}{2} y_i''' + \frac{h^3}{6} y_i^{(4)} + \frac{h^4}{24} y_i^{(5)}] + \frac{5}{h^2} [2y_i''' + hy_i^{(4)} + \frac{h^2}{2} y_i^{(5)}] \right|$$

$$|S_i^{(5)}(x) - y^{(5)}(x)| \leq \frac{7}{2} |y_i^{(5)}(x) - y^{(5)}(x)| \leq \frac{7}{2} W(h, f^{(5)})$$

And

$$S^{(4)}_i(x) = 24 a_{i,4} + 120(x - x_i) a_{i,5}$$

$$|S_i^{(4)}(x) - y^{(4)}(x)| = |24 a_{i,4} + 120(x - x_i) a_{i,5} - y^{(4)}(x)| \\ \leq \frac{7}{2} |y_i^{(5)}(x) - y^{(5)}(x)| \leq \frac{9}{4} h W(h, f^{(5)})$$

And

$$S'''_i(x) = y_i''' + 24(x - x_i) a_{i,4} + 60(x - x_i)^2 a_{i,5}$$

$$|S_i'''(x) - y'''(x)| \leq \frac{h^2}{2} |y_i'''(x) - y'''(x)| \leq \frac{h^2}{2} W(h, f^{(5)})$$

And

$$S''_i(x) = 2a_{i,2} + (x - x_i)y_i'' + 12(x - x_i)^2 a_{i,4} \\ + 20(x - x_i)^3 a_{i,5}$$

$$|S_i''(x) - y''(x)| \leq \frac{h^3}{2} W(h, f^{(5)})$$

And

$$S'_i(x) = y_i' + 2(x - x_i)a_{i,2} + \frac{(x - x_i)^2}{2!} y_i''' \\ + 4(x - x_i)^3 a_{i,4} + 5(x - x_i)^4 a_{i,5}$$

$$|S_i'(x) - y'(x)| \leq \frac{h^4}{24} W(h, f^{(5)})$$

And

$$S_i(x) = y_i + (x - x_i)y_i' + \frac{(x - x_i)^2}{2} a_{i,2} + \frac{(x - x_i)^3}{3!} y_i''' \\ + (x - x_i)^4 a_{i,4} + (x - x_i)^5 a_{i,5}$$

$$|S_i(x) - y(x)| \leq \frac{h^5}{120} W(h, f^{(5)})$$

This proves Theorem 2 for $x \in [x_{i-1}, x_i]$, similarly we can obtain the result for $x \in [x_{n-1}, x_n]$. This completes the proof.

3. Illustration Examples:

In this section, several numerical examples are given to illustrate the properties of the method and all of them were performed on the computer using a program written in [7, 8]. The absolute errors in Tables 1–2 are the values of $|y(x) - S(x)|$ at selected points, and also the following figures are shown that if increases of the order derivatives increases the errors.

Problem 1: we consider the initial value problem $y^{(5)} = y$

where $x \in [0,1]$,

$y(0) = y'(0) = y''(0) = y'''(0) = y^{(4)}(0) = 0$, clearly that, the exact solution is

$$y(x) = e^x.$$

The Pseudocode of problem 1 is:

for (**i** = start point **to** end point, increase start point by **h**
 for each step)

{

Step 1: Find

$$y_1 = y_0 + (h * y'_0) + ((pow(h, 2) / 2) * y''_0) + ((pow(h, 3) / 6) * y'''_0) + ((pow(h, 4) / 24) * y^{(4)}_0) + ((pow(h, 5) / 120) * y^{(5)}_0)$$

$$y'_1 = y'_0 + (h * y''_0) + ((pow(h, 2) / 2) * y'''_0) + ((pow(h, 3) / 6) * y^{(4)}_0) + ((pow(h, 4) / 24) * y^{(5)}_0)$$

$$y''_1 = y'''_0 + (h * y^{(4)}_0) + ((pow(h, 2) / 2) * y^{(5)}_0)$$

Step 2: Find

$$\begin{aligned} S'' &= (5 / pow(h, 2)) * (y_1 + y_0) + (1 / (2 * h)) * (7 * y'_1 + y'_0) + (3 * y''_0) + (h / 24) * (3 * y'''_1 - 25 * y'''_0) \\ S^{(4)} &= (60 / pow(h, 4)) * (y_1 - y_0) - (30 / pow(h, 3)) * (y'_1 + y'_0) + (1 / (2 * h)) * (7 * y'''_1 + 3 * y''''_0) \\ S^{(5)} &= (120 / pow(h, 5)) * (y_1 - y_0) - (60 / pow(h, 4)) * (y'_1 + y'_0) + (5 / pow(h, 2)) * (y'''_1 + y''''_0) \end{aligned}$$

Step 3: Find

$$y'' = (0.5) * ((exp(h) + exp(-h)))$$

$$y^{(4)} = y''$$

$$y^{(5)} = (0.5) * (exp(h) - exp(-h))$$

Step 4: Find

$$\text{Error 2} = S'' - y''$$

$$\text{Error 4} = S^{(4)} - y^{(4)}$$

$$\text{Error 5} = S^{(5)} - y^{(5)}$$

Step 5: Print Error 1, Error 2, Error3 respectively.

}

Problem 2: Consider that the fifth order boundary value problem $y^{(5)} - y^{(4)} - y' + y = 0$ where $x \in [0,1]$, $y(0) = y'(0) = y'''(0) = y^{(4)}(0) = 0$ and $y''(0) = 1$ the exact solution is $y(x) = \frac{1}{4}e^{-x} + \frac{1}{4}e^x - \frac{1}{2}\cos(x)$

The Pseudocode of problem 2 is:

for (**i** = start point **to** end point, increase start point by **h**

for each step)

{

Step 1: Find

$$\begin{aligned}
 y_1 &= y_0 + (h * y'_0) + ((\text{pow}(h, 2) / 2) * y''_0) + ((\text{pow}(h, 3) / 6) * y'''(0)) + ((\text{pow}(h, 4) / 24) * y^{(4)}(0)) + \\
 &\quad ((\text{pow}(h, 5) / 120) * y^{(5)}(0)) \\
 y'_1 &= y'_0 + (h * y''_0) + ((\text{pow}(h, 2) / 2) * y'''(0)) + ((\text{pow}(h, 3) / 6) * y^{(4)}(0)) + ((\text{pow}(h, 4) / 24) * y^{(5)}(0)) \\
 y'''_1 &= y'''(0) + (h * y^{(4)}(0)) + ((\text{pow}(h, 2) / 2) * y^{(5)}(0))
 \end{aligned}$$

Step 2: Find

$$\begin{aligned}
 S'' &= (5 / \text{pow}(h, 2)) * (-y_1 + y_0) \\
 &+ (1 / (2 * h)) * (7 * y'_1 + (3 * y''_0)) + (h / 24) * (3 * y'''_1 - 25 * y'''(0)) \\
 S^{(4)} &= (60 / \text{pow}(h, 4)) * (y_1 - y_0) - \\
 &(30 / \text{pow}(h, 3)) * (y'_1 + y''_0) \\
 &+ (1 / (2 * h)) * (7 * y'''_1 + 3 * y'''(0)) \\
 S^{(5)} &= (120 / \text{pow}(h, 5)) * (y_1 - y_0) - \\
 &(60 / \text{pow}(h, 4)) * (y'_1 + y''_0) + (5 / \text{pow}(h, 2)) * (y'''_1 + y'''(0))
 \end{aligned}$$

Step 3: Find

$$y'' = \exp(-h)$$

$$y^{(4)} = y''$$

$$y^{(5)} = -(\exp(-h))$$

Step 4: Find

$$\text{Error 2} = S'' - y''$$

$$\text{Error 4} = S^{(4)} - y^{(4)}$$

$$\text{Error 5} = S^{(5)} - y^{(5)}$$

Step 5: Print Error 1, Error 2, Error3 respectively.

}

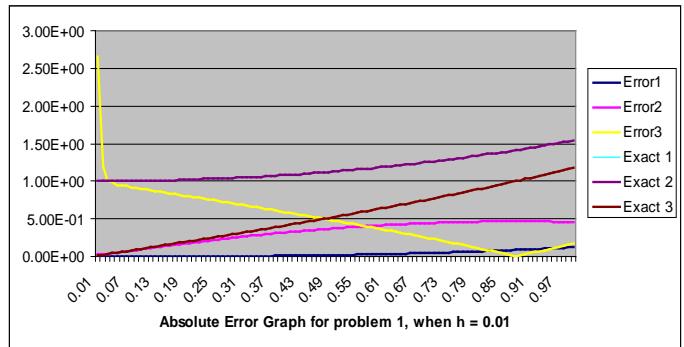
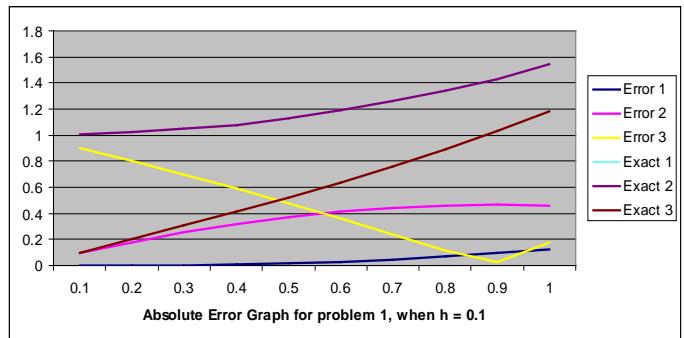
Table (1): Maximum errors in solution of problem 1

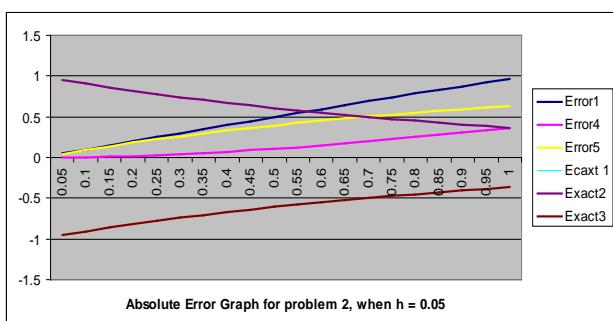
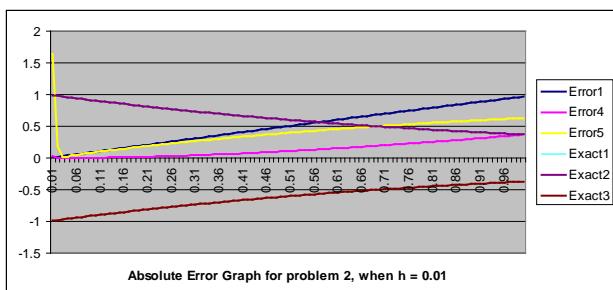
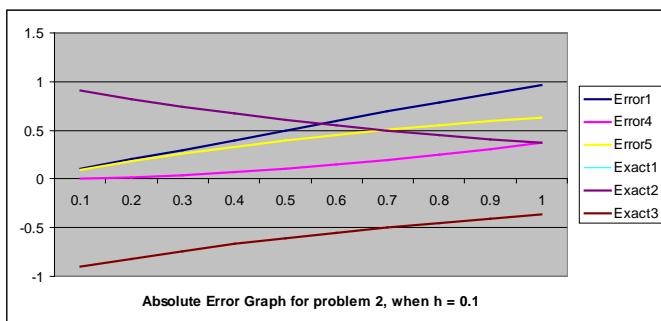
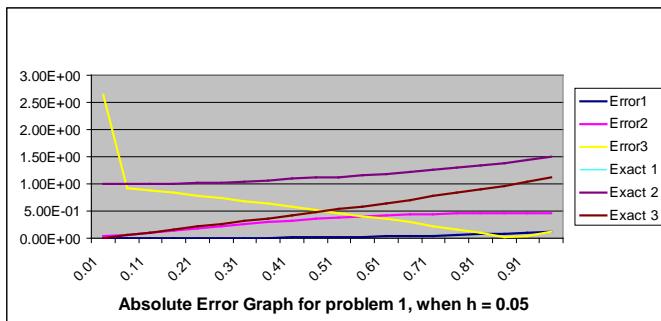
| h | E ⁽²⁾ | E ⁽⁴⁾ | E ⁽⁵⁾ |
|------|------------------|------------------|------------------|
| 0.1 | $1.62 * 10^{-4}$ | $9.49 * 10^{-2}$ | $9.01 * 10^{-1}$ |
| 0.05 | $2.06 * 10^{-5}$ | $4.85 * 10^{-2}$ | $9.59 * 10^{-1}$ |
| 0.01 | $1.58 * 10^{-7}$ | $3.69 * 10^{-2}$ | $2.65 * 10^0$ |

Table (2): Maximum errors in solution of problem 2

| h | E ⁽²⁾ | E ⁽⁴⁾ | E ⁽⁵⁾ |
|------|--------------------|--------------------|--------------------|
| 0.1 | $9.9996 * 10^{-2}$ | $4.8766 * 10^{-3}$ | $9.3849 * 10^{-2}$ |
| 0.05 | $5 * 10^{-2}$ | $1.409 * 10^{-3}$ | $3.8843 * 10^{-2}$ |
| 0.01 | $1 * 10^{-2}$ | $2.6651 * 10^{-2}$ | $16.411 * 10^{-1}$ |

The following figures observe the numerical results with respect two orders of derivative:





5. Discussion:

A new technique, using the Taylor series, to numerically solution the pantograph equations is presented.

It is observed that the method has the best advantage when the known functions in equation can be expanded to Taylor series with converge rapidly. In order to get the best approximation, we take more terms from the Taylor expansion of functions; that is, the truncation limit N must be chosen large enough.

On the other hand, from Table 1, it may be observed that the solutions found for different h show close agreement for various values of x . In particular, our results in tables are usually better than the other methods, are shown in the above figures. Another considerable advantage of the method is that Taylor coefficients of the solution are found very easily by using the computer programs.

References

- [1] Abbas Y. Al Bayati, Rostam K. Saeed and Faraidun K. Ham-Salh (2009) The Existence, Uniqueness and Error Bounds of Approximation Splines Interpolation for Solving Second-Order Initial Value Problems, Journal of Mathematics and Statistics 5 (2):123-129, , ISSN 1549-3644.
- [2] E.A. Al-Said, M.A. Noor, Computational methods for fourth-order obstacle boundary value problems, Comm. Appl. Nonlinear Anal. 2 (1995) 73–83.
- [3] E.A. Al-Said, M.A. Noor, Quartic spline method for solving fourth-order obstacle boundary value problems, J. Comput. Appl. Math. 143 (2002) 107–116.
- [4] A.K. Khalifa, M.A. Noor, Quintic splines solutions of a class of contact problems, Math. Comput. Modell. 13 (1990) 51–58.
- [5] M.A. Noor, E.A. Al-Said, Fourth-order obstacle problems, in: T.M. Rassias, H.M. Srivastava (Eds.), Analytic and Geometric Inequalities and Applications, Kluwer Academic Publishers, Dordrecht, Holland, 1999, pp. 277–300.
- [6] M.A. Noor, E.A. Al-Said, Numerical solutions of fourth-order variational inequalities, Int. J. Comput. Math. 75 (2000) 107–116.
- [7] P.J. DEITEL, H.M. DEITEL , How To Program C++, Sixth Edition,2008.
- [8] Y. Daniel Liang, Introduction to Programming With C++, 2007.

A Method for Designing an Operating System for Plug and Play Bootstrap Loader USB Drive

Dr. T. Jebarajan ¹, K. Siva Sankar ²

¹ Principal, V.V Engg College
Tamil Nadu, India.

² Lecturer, Noorul Islam University.
Tamil Nadu, India.

Abstract

This paper lays out different issues and solutions in the design of an operating system with an inbuilt kernel memory for data storage and USB (Universal Serial Bus) drive with bootstrap loader. This relates from the minimum features required for a program to become the kernel and how this kernel should be written into the boot sector of a hard disk drive depend upon the machine architecture, so that it gets loaded into the computer memory automatically and it restores the disk drive in to its original state. It highlights how this operating system can be made to support user specific authentication, keyboard, networking, peripherals, file system access etc. Most of the frequently used drivers are added to the kernel images. This also lays out the specifications for a shell to issue system commands and system utilities, interfacing FAT (File Allocation Table) file system for smooth boot of an operating system and how to communicate with another similar system using hardware device.

Keywords- *USB flash drive, plug and play, kernel memory, boot loader, shell*

1. INTRODUCTION

Any computer system can be considered to have four basic components [1] [2] - users, application programs, operating system and hardware. The hardware, comprising of central processing unit, memory and input output devices provides the basic computing resources.

The operating system provides a platform for proper use of these resources [6] [7]. It can be considered as a program that manages the computer hardware or as an intermediary between a user of a computer and the computer hardware. The application programs that run on top of the operating system provide the users with solutions they are looking for. This paper will explore the design of a small operating system for and plug and play USB device, which can be built using C and assembly language. In addition to this some modules are configured with this operating system.

2. BUILD ENVIRONMENT

In order to start building, a development box running on Windows or Linux with a C editor, C compiler,

Nasm is required. It requires a target machine for the operating system. It supports a range of object file formats, including Linux and *BSD a.out, ELF, COFF, Mach-O, Microsoft 16-bit OBJ, Win32 and Win64. It will also output plain binary files. Its syntax is designed to be simple and easy to understand, similar to Intel's but less complex for fast accessing. The design and bootstrap strategy will vary with the underlying machine architecture. For this design, consider the target machine as an Intel x86 based hardware with minimum 1MB RAM, USB disk drive, keyboard and monitor.

3. DESIGN OF SYSTEM

A Linux-based system is a modular Unix-like operating system. It derives much of its basic design from principles established in linux. Such a system uses a monolithic kernel version 2.26, the Linux kernel, which handles process control, networking, peripheral and file system access. Device drivers are integrated directly with the kernel

Separate projects that interface with the kernel provide much of the system's higher-level functionality. The user land is an important part of most Linux-based systems, providing the most common implementation of the C library, a popular shell, and many of the common Unix tools which carry out many basic operating system tasks. The graphical user interface (or GUI) used by most Linux systems is built on top of an implementation of the X Window System.

In order to design such a system it can divide it into different logical modules [11], Boot Loader, Kernel, FAT File System, Reverse Mapping, Initial Ramdisk, and Packaging.

3.1 Boot Loader

Boot Loader program loads the kernel of the operating system into the main memory for execution [1] [4]. The Boot Loader must be of size 512 bytes and should reside in the first sector of the disk drive. The conventional

MBR code expects the MBR partition table scheme to have been used, and scans the list of (primary) partition entries in its embedded partition table to find the only one that is marked with the active flag. It then loads and runs the volume boot record for that partition.

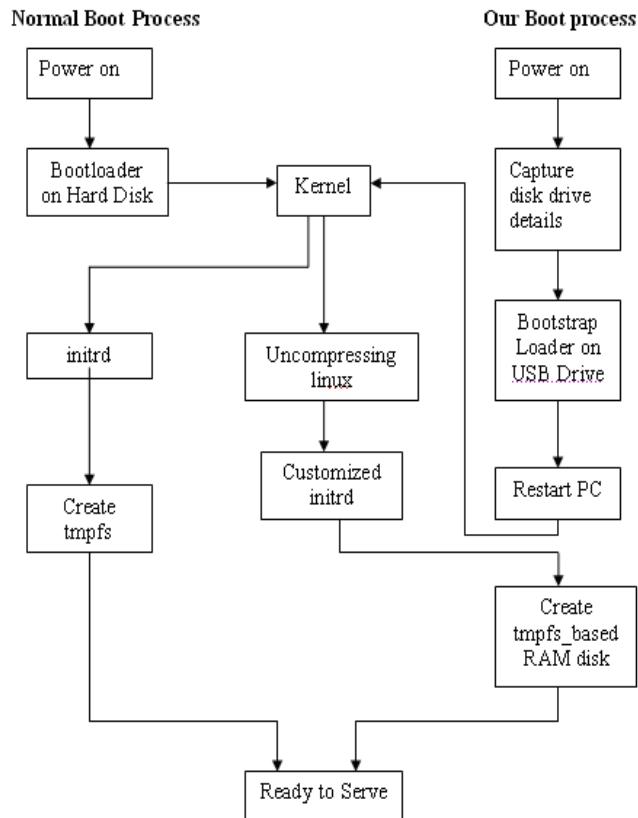


Fig.1. Boot Process comparison

The procedure of Boot Loader is as follows:

Check the boot signature 0AA55h at 510,511-th bytes of the first sector of Boot Disk. If boot signature is present, it loads the code present in the first sector (512bytes) to memory address 07c00h. Next the code at 07c00h is executed. This code then tries to find the available physical memory and divides it into 64KB pages. After that, 2 KB boot stack is allocated at (A0000-512) h and stack pointer is setup. Then Space for IVT and BIOS routines are reserved, and kernel is loaded at 00600h. The kernel that is to be loaded can be an EXE, BIN or COM file. Search for this kernel file will be conducted in the Root Directory (19th sector of the Boot Disk). On getting the file, it is allocated properly with all needed segments and memory pointers. If the kernel is in BIN or COM format it will have a single segment with all DS, CS, ES, SS integrated. If the kernel is in EXE format, it will have separate code, data, extra and stack segments. In such cases the exe header will be ripped off and proper relocation factors are added as needed. After this, the loader loads the kernel into memory.

3.2 Kernel

Kernel is nothing but a program that resides in the memory, takes in user inputs, process those user inputs and give out a suitable response. Kernel can be EXE, BIN or COM file. Though the tasks done by the kernel varies with design [5], and many operating systems delegate system functions to different layers or child programs, the absolute minimal functions common to all other subsystems should be kept in the kernel. In this design, the shell is the main component. This shell program should perform the following functions.

User Authentication – The first routine in the shell should be to authenticate the user. The username and password entered by the user is checked against stored username/password combinations in the system and if found valid, the user is given a ‘prompt’ from where he can issue system commands.

Command Interpreter - The user input parser is nothing but the first component of the Command Interpreter. Once the parser identifies a command as valid, an action has to be taken corresponding to the command. For example, if the command is copy, the shell may call the File Manager subsystem to read the source file and then issue another command to create and write the contents just read as a new file. Similarly all the system commands that the operating system supports can be implemented

IO Functions - Internal working of the Operating System can be based on the basic display and keyboard driver routines [2] [4]. These routines are implemented using BIOS Interrupts [4]. For Keyboard, the routines for reading a keystroke, converting it to number format, checking for keystroke presence etc are included. For Video, the routines for displaying a character, message, error message, printing at specific location on the screen, video settings, color settings etc are included. String manipulation routines are also implemented for simplicity of the high level operations.

Memory Location Routine

- 10h - Video Support Routines
- 12h - RAM size
- 15h - Delay Support Routines
- 16h - Keyboard Routines
- 19h - Reboot Functions
- 1Ah - CMOS Support Routines

That is to say, when a user presses a key, the keyboard support routes at 16H kicks in, based on the context, then appropriate action is taken. For example if the user is in typing on the shell, the video support routines at 10h will be called to display the character typed onto the screen. Any program that implements the above functions qualifies to be a kernel. Below is the pseudo code for the simplest of all kernels. The only function done by this operating system kernel is to display a message after

booting and it is invoked by the script called initrd (initial RAM disk).

```
int main()
{
char *vidmem = (char *) 0xb8000;
vidmem[0] = 'O';
vidmem[1] = 0x7;
vidmem[2] = 'S';
vidmem[3] = 0x7;
return 0;
}
```

Note that `char *vidmem = (char *) 0xb8000` is the memory mapped location of video memory.

3.3 FAT File System

BIOS interrupts are available for low level disk services like reading sectors from drive, writing sectors into the drive, formatting a track etc [3]. The operating system can invoke these bios services for disk activities whenever it has to read/write into the disk drive [9]. But, in order to do low-level reading and writing on a hard drive with a FAT file system, it is required that the address assigned to files/directories by the file allocation table is converted to the absolute sector address understood by BIOS. And for this conversion, a good understanding of the underlying structure of FAT and FAT chaining is required. Disk structure has got 4 logical parts: Boot Sector, File Allocation Table (FAT), Directory and Data space. Of these, the Boot Sector contains information about how the disk is organized. That is, how many sides does it contain, how many tracks are there on each side, how many sectors are there per track, how many bytes are there per sector, etc.

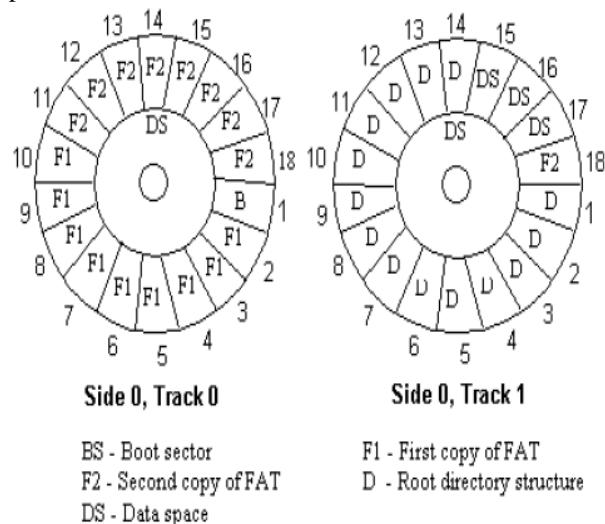


Fig.2. FAT files system architecture

The files and the directories are stored in the Data Space. The Directory contains information about the files like its attributes, name, size, etc. The FAT contains information about where the files and directories are stored in the data space. Fig.2 shows the four logical parts of a 1GB USB flash drive. The basic functions that should be supported by the operating system on the file system should be, List Files / Directories, Create Directory, Change Directory, Create File, Display File contents, Copy File, Rename File, Delete File, Modify File etc. It also makes sense to provide users with an in-built editor which can be used for creating and editing files structure.

FAT is more robust and it can relocate the root folder and use the backup copy of the file allocation table instead of the default copy. In addition, the boot record on FAT USB drives is expanded to include a backup copy of critical data structures. Therefore, FAT USB drives are less susceptible to a single point of failure than existing other file system specified drives.

3.4 Reverse Mapping

Reverse mapping, or RMAP, was implemented in the kernel to solve memory problem. Reverse mapping provides a mechanism for discovering which processes are using a given physical page of memory. Instead of traversing the page tables for every process, the memory manager now has, for each physical page, a linked list containing pointers to the page-table entries (PTEs) of every process currently mapping that page. This linked list is called a PTE chain. The PTE chain greatly increases the speed of finding those processes that are mapping a page, as shown in below

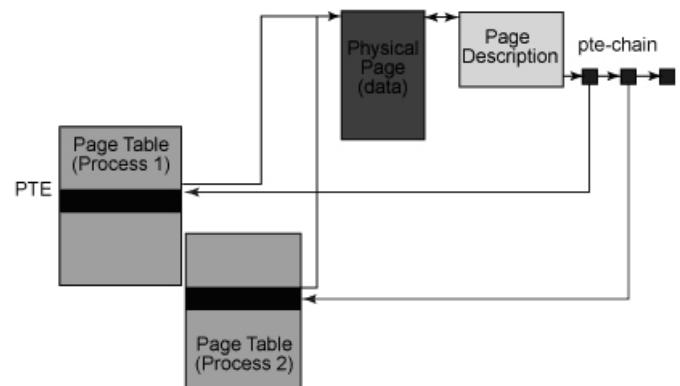


Fig. 3. Reverse mapping page description

The most notable and obvious cost of reverse mapping is that it incurs some memory overhead. Some memory has to be used to keep track of all those reverse mappings. Each entry in the PTE chain uses 4 bytes to

store a pointer to the page-table entry and an additional 4 bytes to store the pointer to the next entry on the chain. This memory must also come from low memory, which on 32-bit hardware is somewhat limited. Sometimes this can be optimized down to a single entry instead of using a linked list it should be compatible with one other.

4. IMPLEMENTATION

4.1 Initial Ramdisk

The PC has only 256MB total RAM, maybe not even any swap partition or swap file, how on earth does this operating system avoid writing to the Flash drive during a session. This is one of the key architectural points of this approach. At boot up, pup_save.2fs is mounted read-only from where it is on the Flash drive, and its contents are not copied into RAM. Instead, a tmpfs (temporary file system) in RAM holds all new and changed files [15]. This is still actually very fast, as all the "working files" are in RAM. Periodically and at end of session, those "working files" are written back to the pup_save.2fs file. This approach has a much smaller initial ramdisk file, named initrd.gz (instead of image.gz), only about 1.1MB, and these accounts for a significantly faster boot time.

This operating system tackles the problem other way round, by always booting up in ramdisk-only the first time you boot on a PC, then at shutdown you are asked if you want to create a personal storage file with different storage option. This operating system mounts the persistent storage file pup_save.sfs at the top level, that is, on root directory (""). The read only compressed file with all the Puppy files, pup_xxx.sfs, mounts on root directory. The kernel is configured with a 12288KB maximum ramdisk and this is increased to 13824K for this approach, so the size should be mentioned in boot parameter and can have built in memory in the operating system. This design uses a tiny initial ramdisk, initrd.gz that is only 0.9M compressed. Installing extra applications, such as dotPups that install into /root/my-applications, do not add to initrd.gz. The initial ramdisk file remains fixed in size, and everything under root directory ("") goes into pup_save.2fs, the squashfs file that gets attached later in the boot process. The initial ramdisk file that loads into the fixed-size-limit ramdisk. The boot sequence then creates a tmpfs ramdisk, which is variable in size, and will use as much RAM and swap space as available.

4.2. Architecture Overview

The way to understand the diagram is to view each of those layers as a complete filesystem, that is, a

complete directory hierarchy from root directory ("") down. These layers are laid one on top of the other, which is achieved by the unions filesystem. This file will be visible at the top layer. If the "off-blue" layer has the same file, it will not be visible, as it is overlaid by the same file on a higher layer. Depending on this version of Linux you are running, the method for creating the initial RAM disk can vary. The initrd is constructed using the loop device. The loop device is a device driver that allows you to mount a file as a block device and then interpret the file system it represents [18]. The loop device may not be present in this kernel, but you can enable it through the kernel's configuration tool by selecting Device Drivers > Block Devices > Loopback Device Support. The small, but necessary, set of applications are present in the ./bin directory, including nash (not a shell, a script interpreter), insmod for loading kernel modules, and lvm (logical volume manager tools).

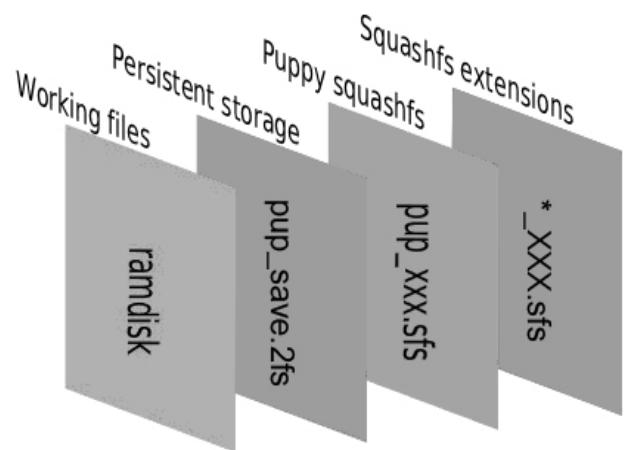


Fig. 4. Structure of Operating System

| | |
|--------------|--|
| ramdisk | This is the tmpfs filesystem running in RAM, with new and updated files. |
| pup_save.2fs | This is the persistent storage, where all your data, settings, email, installed packages, etc., get saved permanently. The ".2fs" means that the file contains a FAT or ext2 filesystem. |
| pup_xxx.sfs | The built-in applications, window manager, scripts, everything. The ".sfs" means the file contains a squashfs compressed filesystem. The "xxx" is the version number without the dots. |
| *_XXX.sfs | These are additional squashfs files. The "*" can be anything. For example, devx_xxx.sfs is the complete environment for compiling C/C++ applications |

While running this operating system, the outlook seen is one filesystem, which is the top layer. Thus you see /usr/lib/libgdkxft.so and you don't care what layer it is actually on. An exciting alternative to the squashfs extensions is to use an existing installed Linux distro as the bottom layer.

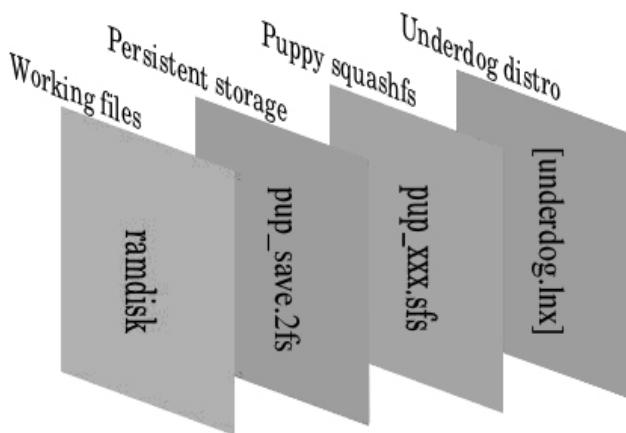


Fig. 5. Operating System Distro as the bottom layer

There are other variations, and it has a "state variable" named PUPMODE that shows what state, the operating system currently used. There is a file, etc/rc.d/PUPSTATE that has the PUPMODE variable defined in the following modes

- PUPMODE 5
- PUPMODE 12
- PUPMODE 13
- PUPMODE 2
- PUPMODE 77

4.3. PUPMODE 5

This is the configuration the very first time that operating system is booted from USB Flash drive. The first time that you plug in the USB and boot up, there is no persistent storage, and the "union" consists of only two layers, the top "working files" and the pup_xxx.sfs squashfs filesystem that has all the operating system files [18]. These two layers appear overlaid at root directory; however they can be viewed individually, at their respective mount points. So, we describe this approach but not touching the hard drive at all. You can run applications, configure, download, install packages, but it is all happening in the tmpfs ramdisk, so not getting saved. The way that we have been using pupmode is to create a "pup100" file on the USB drive, which has a FAT file system. This file is copied into RAM at boot up, if there is enough RAM, thus avoiding writes to the Flash drive

during a session. Then the files are copied back at shutdown.

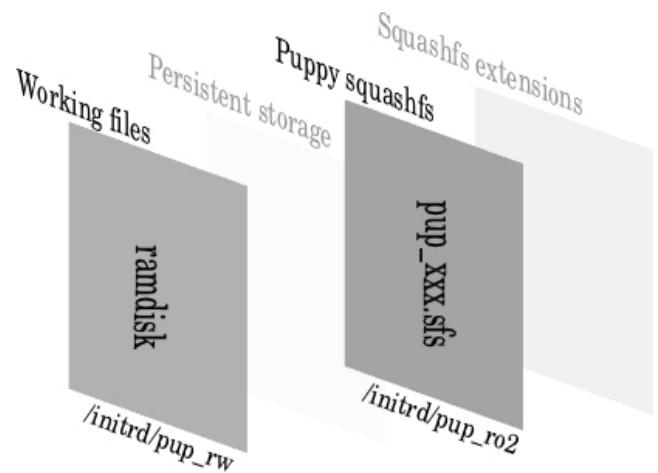


Fig. 6. First time boot configuration of USB flash drive

The amount of space you have in the ramdisk depends on how much RAM is in the PC. The really interesting part is when you decide to end the session and shutdown the PC. The shutdown script, which is actually /etc/rc.d/rc.shutdown, will execute and will bring up a dialog window asking you to save the session with different allocation of memory size. Whatever directories and files that have been created in the ramdisk can now be saved [18]. The choice of storage location depends on whether a partition is a Linux filesystem, or FAT filesystem, a file called pup_save.2fs can be created in USB flash drive and stored periodically every time.

4.4 PUPMODE 13

Configure an operating system to a USB Flash drive, perhaps by using the operating system Universal Installer program, you will have a bootable drive with the files vmlinuz (the Linux kernel), initrd.gz (the initial ramdisk), pup_xxx.sfs (squashfs filesystem with all the os files) and syslinux.cfg (Syslinux config file). The situation is just like booting from a live-CD on first boot of this operating system and it will be in PUPMODE 5, as no persistent storage has yet been created. On first shutdown, as described in the PUPMODE 5 section above, you will create a persistent storage called pup_save.2fs file [18]. On the second boot, operating system will discover the persistent storage and boots. In the case of the persistent storage on Flash memory, which is the second layer, operating system will save everything from the top layer to the second layer every 30 minutes.

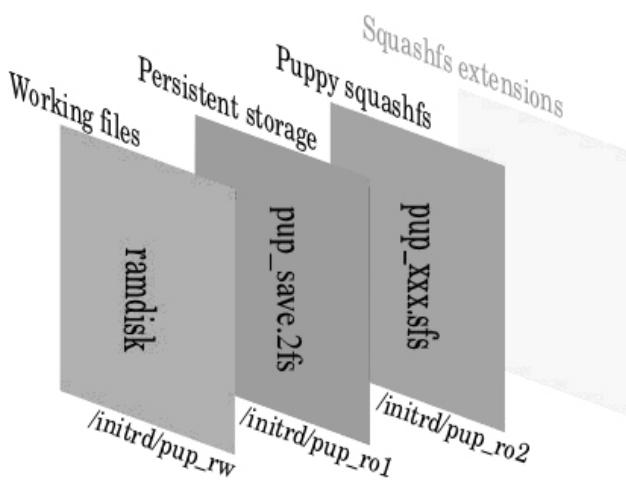


Fig. 7. Second time boot configuration of USB flash drive

From the "unions" point of view, the second layer is mounted read-only, it is only the top layer that is written to, however this design can able to "flush" the top layer down to the next layer at periodic intervals [18]. This has an option to select the user storage in and operating system, according to this user can use two storage one in operating system and another in normal disk drive. The updating made in the operating system is stored in that particular space allotted for it.

4.5 Flash technology

However, there is a downside to flash technology, and that is it is not designed for unlimited writes. That is, you can save onto it just so many times, and then it will collapse [16].

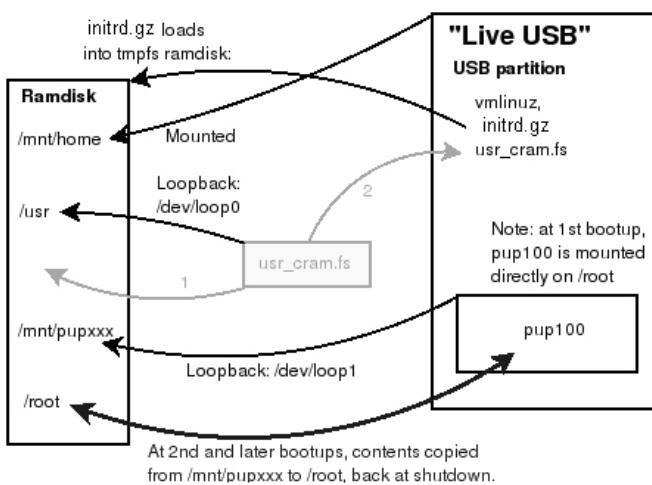


Fig. 8. Layout of Operating System at second and later boot ups from a USB device

Operating systems is especially designed to have no writes to the Flash drive during a session, enormously extending its life span. When operating system boots from USB, the steps are much the same as for the live-CD. The kernel vmlinuz is loaded into RAM, initrd.gz is uncompressed and loaded into a ramdisk, the ram disk is responsible of loading all the operating system modules. Take a look at this fig.8 that boots the operating system from the USB drive and the structure remains same for upcoming extraction.

Operating system with no writes to flash device. usr_cram.fs will find in the USB partition and will mount it on /usr. If there is enough RAM, it will copy usr_cram.fs into the ramdisk and then mount it on /usr. This will slow down bootup slightly, but will improve running speed even if the operating system does leave usr_cram.fs on the USB drive and mounts it from there onto /usr, that is not a problem as /usr is read-only [18]. There will be no writes to /usr, so the lifetime of the Flash drive is not compromised.

4.6 Packaging

Once it has the kernel with all required functionalities ready, it requires writing it into a medium like USB disk drives. Remember that the boot loader always loads the executable in one specific sector of the drive. So it is important that to place the program in correct sector for the boot loader to find it and load it into memory [19]. Tools like masm and debug allows writing the executable program to specific sectors that specify.

5. CONCLUSION

The design parameters of an operating system, with minimal components are described by considering all issues. This is of great thrust to completely design and develop the underlying principles of an operating system and boot strap loader in USB drive. And this understanding of the working platform is critical to developing better software that runs on it.

REFERENCES

- [1] Silberschatz, Galvin, and Gagne, "Operating System Concepts," Wiley, Seventh Edition Wiley, 2006
- [2] A S Tanenbaum, "Operating System Concepts," 3rd ed., Oxford:Clarendon, 1992
- [3] Dominic Giampaolo, "Practical File System Design with the Be File System," Morgan Kaufmann publishers, 1999
- [4] William Stallings "Computer Organization and Architecture: Designing for Performance", Prentice Hall, 2009

- [5] Butler W. Lampson, and Howard E. Sturgis "Reflections on an operating system design", Communications of the ACM, Volume 19, Issue 5, pp. 251-265 , 1976
- [6] A. Messer, and T. Wilkinson, "Components for operating system design", Proceedings of the 5th International Workshop on Object Orientation in Operating Systems, IEEE Press, 1996
- [7] Christine Morin, "Design and Implementation of First Advanced Version of LinuxSSI", INRIA, Campus de Beaulieu, France, 2008
- [8] William Stallings, "Operating Systems: Internals and Design Principles," Prentice Hall, Fifth Edition, 2005
- [9] Lex Stein, "Stupid File Systems Are Better", Proceedings from the Eighth Workshop of Hot Topics in Operating Systems, IEEE Press, 2005
- [10] A Bruce Carlson, Paul B Crilly, and Janet Rutledge, "Communications Systems," Mc Graw Hill, 2001
- [11] Craig Larman, Victor R. Basili, "Iterative and Incremental Development: A Brief History", IEEE Computer Society Press, Volume 36, Issue 6, pp. 47-56, 2003
- [12] Jeshope C, Shafarenko A, "Concurrency engineering", Proceedings from IEEE Computer Systems Architecture Conference," IEEE Press, 2008
- [13] Tanenbaum, A.S, Herder, J.N, Bos, H, "Can we make operating systems reliable and secure?", Computer, Volume 39, Issue 5, pp 44-51, IEEE Press, 2006
- [14] Geer, D "The OS Faces a Brave New World," Computer, Vol 42, Issue 10, pp 15-17, IEEE Press, 2009
- [15] W. Tukey, "Bias and confidence in not-quite large samples," *Annals of Mathematical Statistics*, vol. 29, p. 614, 1958.
- [16] B. Efron, "The jackknife, the bootstrap, and other resampling plans," in *Proc. Conf. Rec. SIAM*, Philadelphia, PA, 1982.
- [17] Y. D'Asseler, C. J. Groiselle, H. C. Gifford, S. Vandenberghe, R. Van De Walle, I. L. Lemahieu, and S. J. Glick, "Evaluating human observer performance for list mode PET using the bootstrap method," *IEEE Trans.Nucl. Sci.*, submitted for publication.
- [18]<http://www.puppylinux.com/development/howpuppyworks.html>
- [19] C. J. Groiselle, Y. D'Asseler, H. C. Gifford, and S. J. Glick, "Performance evaluation of the channelized Hotelling observer using bootstrap list-mode PET studies," in *Proc. IEEE Medical Imaging Conf.*, Portland, OR, 2003, pp. 2511–2515.

A SURVEY OF CONNECTIONLESS NETWORK SERVICE PROTOCOLS FOR MOBILE AD HOC NETWORKS

MANIYAR SHIRAZ AHMED¹, DR. SYED ABDUL SATTAR², FAZEELATUNNISA³

¹ Lecturer, Department of Computer Science & Information Systems,
Najran University, Najran, Saudi Arabia,

²Professor and Dean of Academics

³Lecturer, Department of Computer Science & Information Systems,
Najran University, Najran, Saudi Arabia

Abstract:

A Mobile Ad Hoc Network (MANET) is a network that changes locations and configures itself on the fly. It means MANETs are used where the infrastructure is not available such as military or police exercises, disaster relief operations and urgent business meetings. The stipulation of connectionless network service (CLNS) is much more demanding in mobile ad hoc networks. A lot of research have been done so as to provide CLNS by designing various MANET protocols. However, efficient performance evaluations and relative analysis of these protocols in a common pragmatic environment have been performed only in a limited manner. In this survey the relative features, functions and reliability of each CLNS protocols are studied and discussed.

Keywords: CLNS, MANETS, Reliability

Introduction:

Recent advancements such as Bluetooth introduced a new type of wireless systems known as mobile ad hoc networks. Mobile ad hoc networks or "short live" networks operate in the absence of fixed infrastructure. They offer quick and easy network deployment in situations where it is not possible otherwise. Ad hoc is a Latin word, which means "for this or for this only". Mobile ad hoc network is an autonomous system of mobile nodes connected by wireless links; each node operates as an end system and a router for all other nodes in the network.

Ad Hoc networks can provide communication for civilian applications, such as message exchanges among business meeting, medical and security personnel involved in rescue missions. These

applications rely only on connectionless services because of no infrastructure available.

Connectionless network service provides network layer services to the transport layer. When support is provided for CLNS, routing uses routing protocols to exchange routing information. CLNS does not perform connection setup or termination because paths are determined independently for each packet that is transmitted through a network. In addition, CLNS provides best effort delivery, which means that no guarantee exists that data will not be lost, corrupted, disordered, or duplicated. CLNS relies on transport layer protocols to perform error detection and correction.

Following this, we recap the operation, key features & functions and major protocols in selecting a connectionless network service. We

focus on journal articles and peer-reviewed conferences, thereby hopefully extracting the most useful and important rift of the candidate solutions.

(I) Issues need to be considered while providing CLNS:

Connectionless network service refers to communication between two network end points in which messages can be sent from one end point to another without prior arrangement.

CLNS are:

- Stateless having no previously defined protocol
- Easily accessible.

But the CLNS is not ensured that the recipient is available to receive the data. The Data has to be resent several times. It's hard to filter malicious packets using firewalls. No acknowledgement will be given during the data transfer. The main advantage of using CLNS is that it is mainly used in "real time" applications where data sending is more important.

CLNS is a type of network service at the layer 3 of the OSI model. This service does not have the reliability of the connection-oriented method, but it is useful for periodic burst transfers. Neither system must maintain state information for the systems that they send transmission to or receive transmission from. LANs operate as connectionless systems. A computer attached to a network can start transmitting frames as soon as it has access to the network. It does not need to set up a connection with the destination system ahead of time. However, a transport-level protocol such as TCP may set up a connection-oriented session when necessary. Contrast this with Connectionless service, which does not require establishing a session and a virtual circuit^[1]. This can be found in the network layer or transport layer, depending on

the protocol. You can think of a connectionless protocol as being akin to mailing a post card. You send it and hope that the receiver gets it.

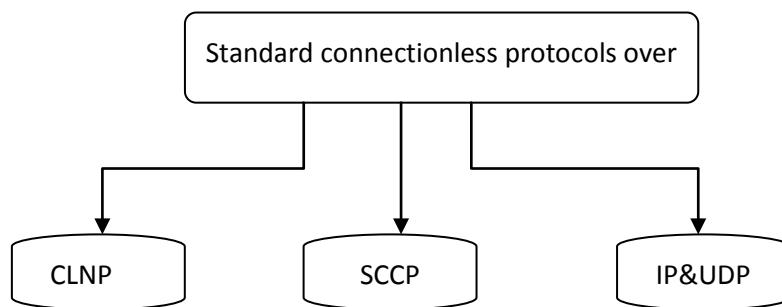
Features of a connectionless service :

- Packets do not need to arrive in a specific order
- Reassembly of any packet broken into fragments during transmission must be in proper order
- No time is used in creating a session
- No Acknowledgement is required.

The largest connectionless network in use today is the Internet.

4. Protocols providing CLNS

Protocol classification



(a) Connectionless network protocol (CLNP):

CLNP, is a Public Data Network protocol that provides the connectionless mode network service.

Aim of CLNP:

CLNP performs two services: breaking data into packets and addressing packets across networks. It is known as a "datagram" service, which refers to the process of splitting up data into chunks for transmission and adding a header to it. The addressing responsibilities of

the protocol follow the Network Service Access Point (NSAP) protocol^[6].

Functions of CLNP:

CLNP is the equivalent to the Internet Protocol definition of the TCP/IP (Transmission Control Protocol/Internet Protocol) stack.

"Connectionless" systems simply send out data to an address without checking whether the data actually arrived. Connectionless Network Protocol (CLNP)^[10] is an ISO network layer datagram protocol. CLNP provides the Connectionless-mode Network Service. CLNP is intended for use in the Sub network Independent Convergence Protocol (SNICP) role, which operates to construct the OSI Network Service over a defined set of underlying services, performing functions necessary to support the uniform appearance of the OSI Connectionless-mode Network Service over a homogeneous or heterogeneous set of interconnected sub networks^[7].

CLNP uses Network service access point (NSAP) addresses and titles to identify network devices. The Source Address and Destination Address parameters are OSI Network Service Access Point Addresses (NSAP address). A network-entity title is an identifier for a network-entity in an end-system or intermediate-system. Network-entity titles are allocated from the same name space as NSAP addresses, and the determination of whether an address is an NSAP address or a network-entity title depends on the context in which the address is interpreted. CLNP (Connectionless Network Protocol) provides the same maximum datagram size as IP, and for those circumstances where datagrams may need to traverse a network whose maximum packet size is smaller than the size of the datagram, CLNP (Connectionless Network Protocol) provides mechanisms for fragmentation (data unit identification,

fragment/total length and offset). Like IP, a checksum computed on the CLNP header provides a verification that the information used in processing the CLNP datagram has been transmitted correctly, and a lifetime control mechanism ("Time to Live") imposes a limit on the amount of time a datagram is allowed to remain in the Internet system.

CLNP has the following PDU(protocol data unit) structure:

| Header part | Address part | Segmentation part | Option part | data |
|-------------|--------------|-------------------|-------------|------|
|-------------|--------------|-------------------|-------------|------|

Header part

NLP ID - Network Layer Protocol Identifier.
The value of this field is set to binary 1000

| 8 | 16 | 24 | 24 | 35 | 40 | 56 | 72bit |
|---------------|---------------|-------------|--------------|-------|----------|--------------------|--------------|
| N LP ID | Leng th ID | Versi on | Life time | Flags | Ty pe | Seg. Leng th | Check sum |

0001 to identify this Network Layer protocol as ISO 8473, Protocol for Providing the Connectionless- mode Network Service and the value of this field is set to binary 0000 0000 to identify the Inactive Network Layer protocol subset.

- Length ID - Length Indicator is the length in octets of the header
- Version - Version/Protocol Id Extension identifies the standard Version of ISO 8473
- Lifetime - PDU Lifetime representing the remaining lifetime of the PDU, in units of 500 milliseconds.
- Flags - three flags: segmentation permitted, more segments, error report
- Type - The Type code field identifies the type of the protocol data unit, which could be data PDU or Error Report PDU

- Seg. Length - The Segment Length field specifies the entire length, in octets, of the Derived PDU, including both header and data (if present).
- Checksum - The checksum is computed on the entire PDU header.

Address Part

It contains information of destination and source addresses, which are defined in OSI 8348/AD2 with variable length.

Segmentation Part

If the Segmentation Permitted Flag in the Fixed Part of the PDU Header^[2] (Octet 4, Bit 8) is set to one, the segmentation part of the header, illustrated in Figure 6, must be present: If the Segmentation Permitted flag is set to zero, the non-segmenting protocol subset is in use.

Option Part

The options part is used to convey optional parameters.

Data Part

The Data part of the PDU is structured as an ordered multiple of octets.

(b) Signaling connection control part (Sccp):

Signaling Connection Control Part (SCCP), is a Signaling System 7 protocol that provides the connectionless mode network service as described in ITU-T Recommendation X.213. Signaling Connection Control Part (SCCP), a routing protocol in SS7 protocol suite in layer 4, provides end-to-end routing for TCAP messages to their proper database and relies on

the services of MTP for basic routing and error detection. SCCP provides connectionless and connection-oriented network services above MTP Level 3.

SCCP allows routing using a Point Code and Subsystem number or a Global Title. A Point Code is used to address a particular node on the network, whereas a Subsystem number addresses a specific application available on that node. SCCP employs a process called Global Title Translation to determine Point Codes from Global Titles so as to instruct MTP on where to route messages.

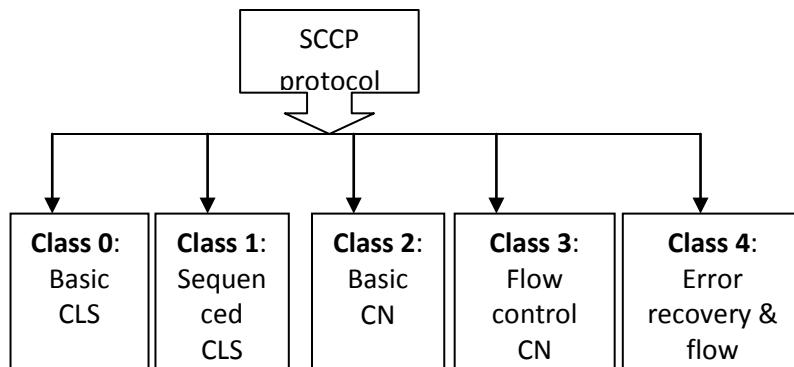
SCCP^[10] messages contains 3 parameters which describe the type of addressing used, and how the message should be routed:

SCCP message parameters:

- (i) Address Indicator
- (ii) Global title indicator
- (iii) Routing indicator
- (iv) Address Indicator Coding

SCCP Protocol classes:

SCCP provides 5 classes of protocol to its applications:



*CLS – Connectionless

*CN – Connection Oriented

Class 0: Basic connectionless

The SCCP Class 0 protocol class is the most basic of the SCCP protocol classes. Network Service Data Units passed by higher layers to the SCCP in the originating node are delivered by the SCCP to higher layers in the destination node. They are transferred independently of each other. Therefore, they may be delivered to the SCCP user out-of-sequence. Thus, this protocol class corresponds to a pure connectionless network service. As a connectionless protocol, no network connection is established between the sender and the receiver.

Class 1: Sequenced connectionless

SCCP Class 1 builds on the capabilities of Class 0, with the addition of a sequence control parameter in the NSDU which allows the SCCP User to instruct the SCCP that a given stream of messages should be delivered in sequence. Therefore, Protocol Class 1 corresponds to an enhanced connectionless protocol with assurances of in-sequence delivery.

Class 2: Basic connection-oriented

SCCP Class 2 provides the facilities of Class 1, but also allows for an entity to establish a two-way dialog with another entity using SCCP.

Class 3: Flow control connection oriented

Class 3 service builds upon Class 2, but also allows for expedited (urgent) messages to be sent and received, and for errors in sequencing (segment re-assembly) to be detected and for SCCP to restart a connection.

Class 4: Error recovery and flow control connection oriented

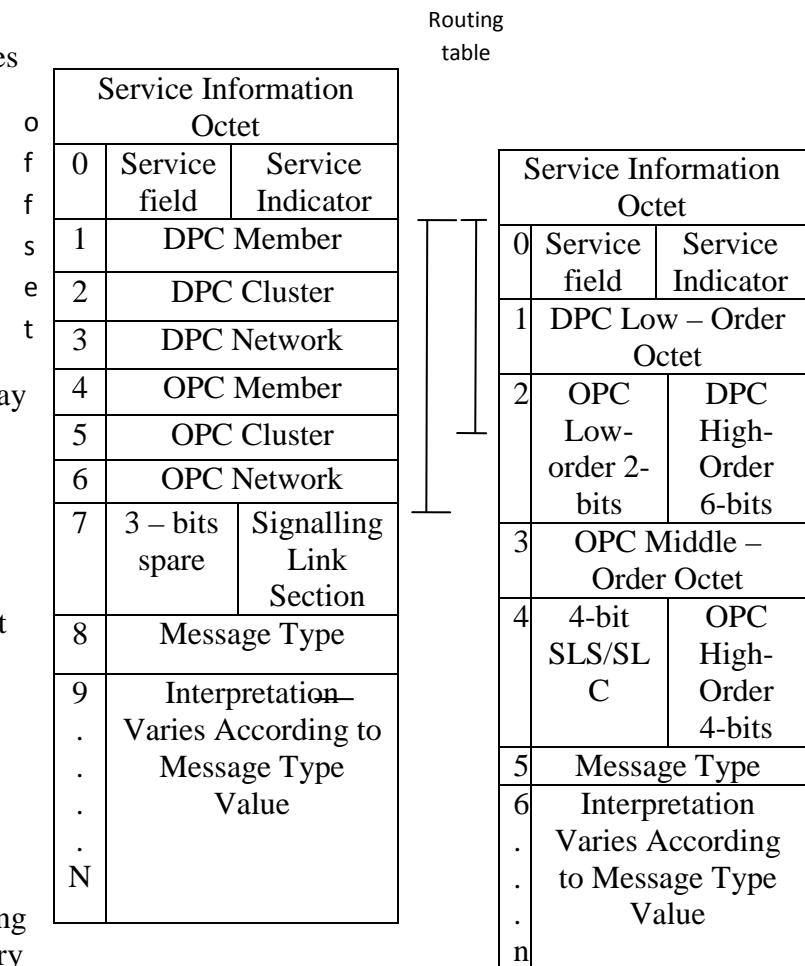
Class 4 service is never used in real time. Signaling Connection Control Part provides reliable delivery of packets between end stations in a telephone network. SCCP makes it possible to address a

message to a specific type of device, such as a conventional telephone set, cell phone set, VoIP end station, fax machine, or computer. SCCP maintains the correct sequencing of packets, even during times of high network traffic or partial network failure. SCCP is used as the transport layer for services such as 800/888/877 (free-phone) numbers, phone cards (calling cards) and roaming in cellular networks.

Protocol Structure

SCCP messages are contained within the Signaling Information Field (SIF) of an MSU. There are two formats for the SCCP messages: one is defined by ANSI^[3] and the other is defined by ITU-T.^[10]

SCCP Header Structure



The signaling information field(SIF) contains the routing label followed by the SCCP message header with the following structure:

| |
|-------------------------|
| Routing label |
| Message type |
| Mandatory fixed part |
| Mandatory variable part |
| Optional part |

- **Routing label** - A standard routing label.
- **Message type code** - A one octet code which is mandatory for all messages. The message type code uniquely defines the function and format of each SCCP message.
- **Mandatory fixed part** - The parts that are mandatory and of fixed length for a particular message type will be contained in the mandatory fixed part.
- **Mandatory variable part** - Mandatory parameters of variable length will be included in the mandatory variable part. The name of each parameter and the order in which the pointers are sent is implicit in the message type.
- **Optional part** - The optional part consists of parameters that may or may not occur in any particular message type. Both fixed length and variable length parameters may be included. Optional parameters may be transmitted in any order. Each optional parameter will include the parameter name (one octet) and the length indicator (one octet) followed by the parameter contents.

(C) IP & UDP:

Internet Protocol and User Datagram Protocol essentially provide the connectionless mode network service as described earlier^[8].

The Internet Protocol (IP) is the principal communications protocol used for relaying datagrams (packets) across an internetwork using the Internet Protocol Suite. Responsible for routing packets across network boundaries, it is the primary protocol that establishes the Internet. Historically, IP was the connectionless datagram service in the original Transmission Control Program introduced by Vint Cerf and Bob Kahn in 1974, the other being the connection-oriented Transmission Control Protocol (TCP)^[4]. The Internet Protocol Suite is therefore often referred to as TCP/IP.

The first major version of IP, now referred to as Internet Protocol Version 4 (IPv4) is the dominant protocol of the Internet, although the successor, Internet Protocol Version 6 (IPv6) is in active, growing deployment worldwide.

Services provided by IP

The Internet Protocol defines an addressing methods and structures for datagram encapsulation. Addresses identify hosts and provide a logical location service. Each packet is tagged with a header that contains the meta-data for the purpose of delivery. This process of tagging is also called encapsulation. IP is a connectionless protocol and does not need circuit setup prior to transmission.

IP Reliability

As a consequence of this design, the Internet Protocol only provides best effort delivery and its service can also be characterized as unreliable. In network architectural language it is a connectionless protocol^[11]. The lack of reliability allows any of the following fault events to occur:

- Data corruption
- Lost data packets
- Duplicate arrival
- Out-of-order packet delivery; meaning, if packet 'A' is sent before packet 'B', packet 'B' may arrive before packet 'A'. Since routing is dynamic and there is no memory in the network about the path of prior packets, it is possible that the first packet sent takes a longer path to its destination.

In addition to issues of reliability^[9], this dynamic nature and the diversity of the Internet and its components provide no guarantee that any particular path is actually capable of, or suitable for, performing the data transmission requested, even if the path is available and reliable.

UDP

The User Datagram Protocol (UDP) is the TCP/IP connectionless transport protocol. Connectionless transport protocols are used for multimedia applications. Networking protocols are grouped by function into a protocol stack^[5]. There are several transport layer protocols available.

Features & Functions of UDP

After a connection has been established, data integrity can be managed by sequencing data packets for the same session. Without establishing a connection, these data management functions are not possible. UDP merely sends out packets at one end and receives them at the other. Whether those packets are out of sequence or have damaged or lost data is not controlled.

The purpose of UDP is to offer a lightweight alternative to TCP. Where applications perform their own data integrity checks, or have alternative connection-establishing procedures, UDP is used. UDP became popular with multimedia applications like video streaming and Internet telephony, which

have separate procedures for data integrity and session management.

IV. Future challenges

MANETs are probably to expand their applications in the future communication environments. The support of CLNS will thus be an important and desirable component of MANETs. Several important research issues and open questions need to be addressed to facilitate CLNS support in MANETs. Use of location, mobility, power consumption and route availability are some of the issues that are currently being examined and need further exploration. Other challenges and open issues include robustness and security, and support for multiple levels of services in CLNS routing schemes.

V. Conclusion

In this paper, we focused on the basic concepts in CLNS routing in MANETs and the various issues that are needed to be faced during the provision of CLNS. The through overview on various CLNS routing protocols have been made. We have summarized the classifications, features and functions of these protocols. There are still many issues and challenges which have not been considered. This will be subjected to further investigations.

References:

1. Kavita Taneja, Mobile Ad hoc Networks: Challenges and Future, COIT-2007, RIMT-IET, Mandi Gobindgarh. March, 2007
2. Nishu Garg, MANET Security Issues, IJCSNS, VOL.9 No.8, August 2009 – 241

3. Santhi.G, A SURVEY OF QoS ROUTING PROTOCOLS FOR MOBILE AD HOC NETWORKS, CONFERENCE 26.2.2010

4. Moukhtar A.Ali, A Survey of Multicast Routing Protocols for Ad-Hoc Wireless Networks, Minufiya Journal of Electronic Engineering Research (MJEER), Vol. 17, No. 2, July 2007.

5. H. Yang, Security in Mobile Ad Hoc Networks: Challenges and Solutions, IEEE Wireless Communications, Vol.11, Issue 1, pp. 38-47, 2004

6. Shino Sara Varghese, A Survey on Anonymous Routing Protocols in MANET, International Conference on Networking VLSI & Signal Processing ICNVS'10

7. Pradeep Rai Shubha Singh, A Review of ‘MANET’s Security Aspects and Challenges’ IJCA Special Issue on “Mobile Ad-hoc Networks” MANETs, 2010

8. G.Vijaya Kumar, Current Research Work on Routing Protocols for MANET: A Literature Survey, (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 03, 2010, 706-713

9. Edward W. Page, Charles K. Watt, Automated Network Management for MANETs: Challenges and Opportunities, Electronic Systems Support, LLC.

10. Wikipedia, [http://en.wikipedia.org/wiki/..](http://en.wikipedia.org/wiki/)

11. A. K. Dwivedi, Performance of Routing Protocols for Mobile Adhoc and Wireless Sensor Networks: A Comparative Study, International Journal of Recent Trends in Engineering, Vol 2, No. 4, November 2009.

An authorization Framework for Grid Security using GT4

Debabrata Singh¹, Bhupendra Gupta², B.M.Acharya³, Sarbeswar Hota⁴
S 'O'A University, Bhubaneswar

Abstract

A Grid system is a Virtual Organization that is composed of several autonomous domains .It concerned with the sharing and coordinated use of diverse resources in distributed "virtual organizations." The dynamic and multi-institutional nature of these environments introduces challenging security issues that demand new technical approaches. In particular, one must deal with diverse local mechanisms, support dynamic creation of services, and enable dynamic creation of trust domains. We review the Globus Toolkit version 2 (GT2) approach; then, Globus Toolkit version 3 (GT3), then, for authorization in such virtual organization a system needs to be flexible and scalable(reviewed the blacklist/whitelist authorization) to support multiple security policies. Basing on the Web Services security specifications such as XACML, SAML, and the special security needs of the Grid computing we review Globus Toolkit version 4 (GT4), provides a simple, open and efficient method for Grid service access control.

Key Words : OGSA, XACML, SAML, Blacklist / Whitelist Authorization

1. Introduction

The term "Grid" refers to systems and applications that integrate and manage resources and services distributed across multiple control domains [1]. It is a virtual organization comprising several independent autonomous domains [2]. Authorization is an important part of the Grid security system. In a grid computing environment, every autonomous domain may have its own policy and may change its policy dynamically. Hence, the authorization mechanism of the Grid system needs to support multiple security policies and needs to have the flexibility to support dynamic changes in security policies, which suggest new challenges to the Grid computing platforms. With the merging of Grid and Web Services, many new standards and concepts in Web Services are introduced into Grid computing area. Basing on the authorization related specifications in Web Services and the special authorization requirements of Grid, we reviewed a flexible multipolicy authorization framework in Globus Toolkit release 4.

The recent definition of the Open Grid Services Infrastructure specification and other elements of the Open Grid Services Architecture (OGSA) [3] within the Global Grid Forum introduces new challenges and opportunities for Grid security. In particular, integration with Web services and hosting environment technologies introduces opportunities to leverage emerging security standards and technologies such as the Security Assertion Markup Language (SAML) [4] and Web services security. Integration of GSI with OGSA enables the use of Web services techniques to express and publish policy , allowing applications to determine automatically what security policies and mechanisms are required of them. Implementing security in the form of OGSA services allows those services to be used as needed by applications to meet these requirements. To show the flexibility & scalability of the framework we studied blacklist/whitelist based authorization mechanism .

GT3's security implementation uses Web services security mechanisms for credential exchange and other purposes, and introduces a tight least privilege model that avoids the need for any privileged network service.

GT4 Authorization implements SAML (security assertion markup language),and uses the XACML(extensible access control markup language). XACML Authorized framework architecture implementation of the Open Grid Services Architecture, an initiative that is recasting Grid concepts within a service oriented framework based on Web services. The blacklist/whitelist authorization system established under the GT4 authorization framework can provide a simple, open, scalable, and flexible and efficient method for Grid service access control.

The rest of the paper is organized as follows: section 2 discusses some related work; section 3 introduces the XACML specification ; section 4 describes the design concepts, the structure, and the components of the authorization framework; section 5 discusses the design and implementation of the blacklist/whitelist-based authorization mechanism; section 6 summarizes of all.

2. Related Work

In Globus Toolkit, the security functionality is called the Grid Security Infrastructure (GSI) [5], and authorization is developing together with GSI. From version 1 in 1998 to the 2 release in 2002 and now the 4 release, GSI has been developing rapidly. In GT1, GSI mainly provided message protection and authentication. In GT2, GSI introduced X.509 proxy certificates to support dynamic creation of computing entities and provided Community Authorization Service (CAS) to implement access control in dynamic created overlaid trust domains.

In GT3, the Grid technology worked with the emerging Web services technology. Security functionalities of GSI3 are defined as OGSA(Open Grid Services Architecture) services [6]. In GT4, additional Web Services security specifications are implemented. Web Services has provided several security standards that have great influence to the Grid computing. XACML (extensible Access Control Markup Language) and SAML (Security Assertion Markup Language) are the two important authorization related standards [7]. There are also several authorization systems that support Grid Computing, such as Akenti[8], PERMIS , Shibboleth[9], VOMS . Akenti, PERMIS and Shibboleth use user attributes to make authorization decisions; VOMS provides user attributes which can be used for authorization. These authorization systems support their own policies, and can be integrated into GT4 authorization framework as authorization services.

3. The XACML Authorization Model

GT4 implements the WSRF specification. GT4 authorization framework was constructed based on the OASIS XACML and SAML standards [10]. The architecture of the framework uses the XACML authorization model that is shown in Figure 1.

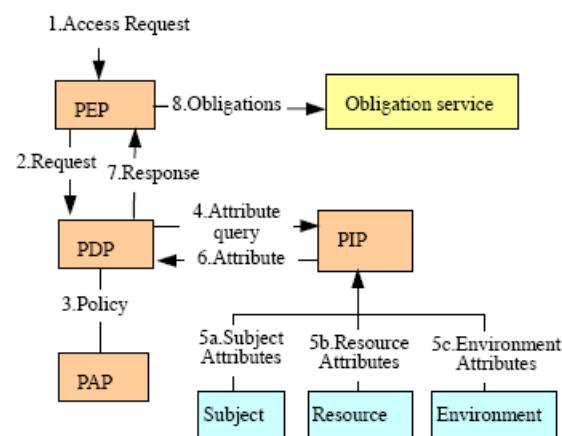


Figure 1. XACML authorization model

The XACML authorization model mainly contains PEP (Policy Enforcement Point), PDP (Policy Decision Point), PIP (Policy Information Point), and PAP (Policy Administration Point). The PEP intercepts the access requests from users and sends the requests to the PDP. The PDP makes access decisions according to the security policy or policy set written by PAP and, using attributes of the subjects, the resource, and the environment obtained by querying the PIP. The access decision given by the PDP is sent to the PEP. The PEP fulfills the obligations and either permits or denies the access request according to the decision of PDP. XACML also defines a policy language. Policies are organized hierarchically into Policy Sets, Policies and Rules, combined using combining algorithms. A rule is composed of a target, an effect and a condition. A Policy consists of a target, one or more rules, and an optional set of obligations.

3.1 XACML Policy Language Model

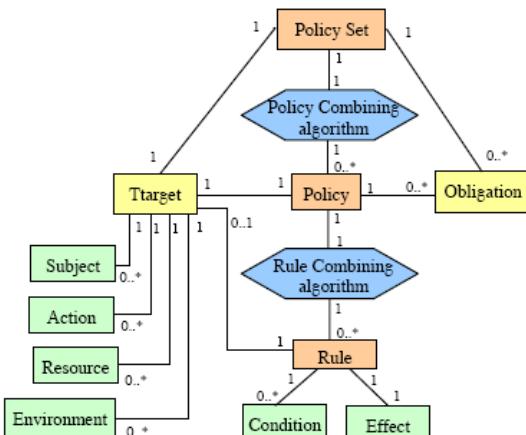


Fig 2 XACML Policy Language Model

The access control framework mainly contains PEP (Policy Enforcement Point), PDP (Policy Decision Point), PIP (Policy Information Point), and PAP (Policy Administration Point). The PEP intercepts the access requests from users and sends the requests to the PDP. The PDP makes access decisions according to the security policy or policy set written by PAP and, using attributes of the subjects, the resource, and the environment obtained by querying the PIP. The access decision given by the PDP is sent to the PEP. The PEP fulfills the obligations and either permits or denies the access request according to the decision of PDP. XACML also defines a policy language. The policy model is shown in Fig. 2. The main components of the model are the rule, the policy, and the policy set. A rule is the most elementary

unit of the policy and can be evaluated on the basis of its contents. The main components of a rule are as follows:

A target that defines the set of resources, subjects, actions and environment An effect that indicates the consequence of the satisfied rule A condition that further refines the applicability of the rule Rules are combined into a policy, which comprises four main components: a target, a rule-combining algorithm, a set of rules, and obligations. A policy set comprises four main components: a target, a policy-combining algorithm, a set of policies, and obligations. The rule-combining algorithm specifies the procedure by which the results of evaluating the component rules are combined when evaluating the policy. The policy-combining algorithm has a function similar to that of the rule-combining algorithm. Obligations are the actions that must be performed by the PEP in conjunction with the enforcement of an authorization decision; obligations are the mechanism for achieving finer-level access control.

4. The GT4 Authorization Framework

The convergence of Grid and Web services introduces both new opportunities and new challenges for Grid security. On the one hand, these specifications have provided standard and interoperable methods for Grid security. On the other hand, in order to establish an authorization mechanism suitable for Grid computing, these specifications may also need to be extended or changed to some extent, since Grid has its own special application requirements.

In a Grid system, each domain has its own security policy, such as the grid-mapfile, ACL (Access Control List), CAS, SAML authorization decision assertions, and XACML policy statements. Hence, the GT4 authorization framework needs to support multiple security policies and also needs to be flexible, so that it can be changed easily for different application environments. These special authorization requirements are not addressed in the XACML specification. Based on the XACML specification and the Grid access control requirements, we designed and implemented the GT4 authorization framework.

4.1. The Framework Architecture

The GT4 authorization framework[12][13] implements SAML and uses the XACML model, as shown in Figure 3. It is composed of a PEP, PDPs, and PIPs. For each existing authorization policy, the framework constructs a PDP for evaluating that kind of policy. The Master PDP is

responsible for coordinating the PDPs to render a final decision. The Master PDP and the PEP are collectively called the authorization engine. The framework provides different kind of PIPs. A subset of PIP, referred to as Bootstrap PIPs, collect information only about the request, such as the peer subject, the requested action, and the resource. An example of one such PIP, is the X509BootstrapPIP, which extracts the subject DN of the peer from the X509 certificate. When a request of the Grid resource comes, the PEP intercepts it and sends a decision request to the master PDP. The master PDP collects information needed by calling the Bootstrap PIPs and other PIPs and then invokes the corresponding PDPs with the request and the information collected. The PIPs and the PDPs used are all specified in the security configuration file. When the master PDP receives the decisions returned by each PDP, it combines the decisions, using a policy combination algorithm, such as deny override or permit override, to render a final decision and returns the decision to the PEP. The PEP then executes the decision, either denying or permitting the request.

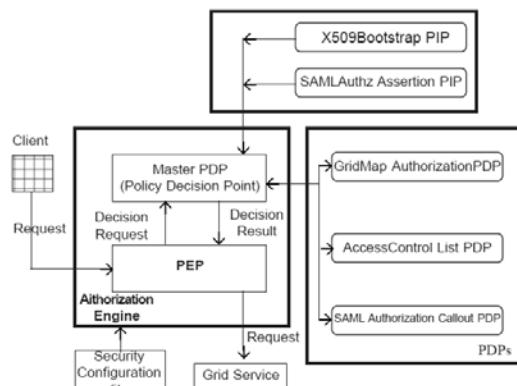


Fig 3. GT4 authorization framework

4.2. The Authorization Framework of PDP

The PDP is the core of the authorization framework. In order to make the framework support different kind of policies and be scalable, we built a multipolicy framework[11] as shown in Figure 3. Because every policy essentially needs its own custom decision evaluator that understands the intrinsic semantics of the policy expressions, it is necessary to encapsulate the policy into an independent PDP. At the same time, we abstract the common characteristic of the policies and define an abstract PDP. The PDP abstraction (the PDP class in Figure 3) defines a common interface that can be used to interact with the PEP or with other PDPs. Each specific policy is a subclass of the PDP abstraction, which implements the common interface inherited from PDP with its own policy and evaluation mechanism.

The policy framework is object-oriented. New policies can be added and modified at any time. Also, since PDP instances are queried through the same interface and the mechanism-specific details of the PDPs are all hidden behind the public interface, a change to the policy framework has no effect on the Master PDP: it can cooperate with any specific PDPs designated by the security configuration files. This multipolicy framework thus provides users with a flexible and scalable authorization mechanism. In Grid systems, there are several frequently used simple authorization policies or mechanisms, we provided PDPs that implement these existing policies, such as the Access Control List PDP and the Grid Map Authorization PDP.

Some authorization systems like Shibboleth, VOMS and PARMIS.

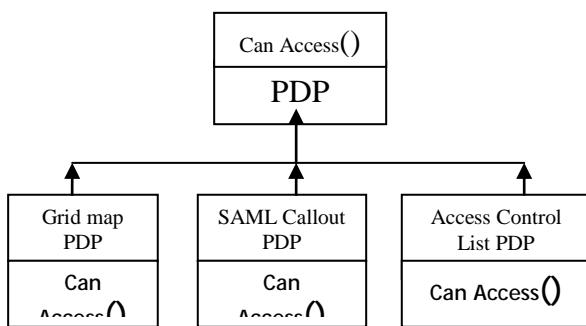


Fig 4. Authorization policy framework

5. Blacklist/Whitelist Based authorization Framework

Blacklist and whitelist mechanisms are simple and well known in the security area. The most obvious advantages of this technology are simplicity and efficiency. They can also be introduced into the Grid services access control area for establishing a simple and effective authorization mechanism. If the authorization mechanism detects the requestor on the blacklist or whitelist, it will make an access decision immediately. Based on the blacklist and whitelist concept, we designed and implemented a prototype BlackList PDP and WhiteListPDP under the GT4 authorization framework. The Blacklist/whitelist-based authorization structure is shown in Figure 5.

The BlackList PDP and the WhiteListPDP are inherited from the PDP abstraction introduced in Section 4.2. The implementation of these two PDPs has two layers: the functional layer and the implementation layer. The blacklist/whitelist access interface, which now contains a member testing method, is defined at the functional layer. The implementation layer contains two levels: the first level is JNDI, which can integrate various naming and directory services and provide a common interface; the second level is composed by

different naming and directory services. In our prototype we use an LDAP server to store and manage the blacklist and the whitelist. The URL of the LDAP server is passed to the BlackList PDP and WhiteListPDP through a configuration file. The blacklist and whitelist are composed of attributes of requestors, such as DN (Distinguished Name, which can be abstracted from the requestor's X.509 certificate), name, and email address. We chose the DN as the identity attribute. Other attributes such as username and group membership can also be used as the identity attributes. This can be achieved by establishing a blacklist/whitelist PIP, which obtains these identity attributes by querying an outside attribute authority using the requestor's DN, and then provides the identity attributes to the BlackList PDP or WhiteListPDP. This will provide more flexibility for users in different application environments. The blacklist/whitelist-based authorization can also be used together with other authorization mechanisms to make an efficient and rigorous authorization system. The Master PDP will first call the Blacklisted or the WhiteListPDP; if the requestor is not found here, other PDPs will be called to do further decision making.

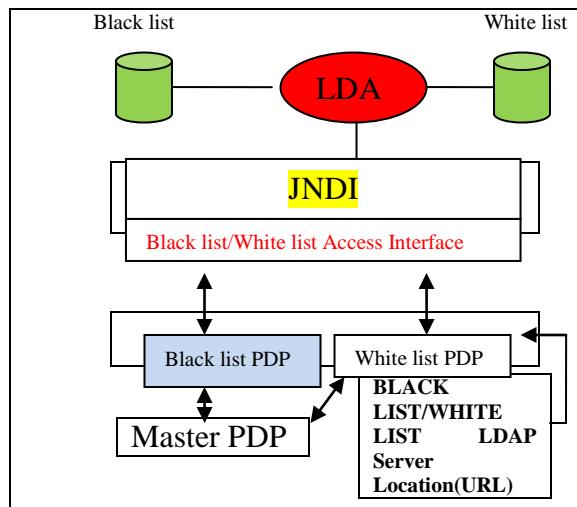


Fig 5 . Blacklist/Whitelist-based authorization structure

Based on this ,the prototype blacklist & whitelist PDP under the GT4 authorization framework ,which shown in fig 4.The blacklist & whitelist PDP are inherited from the PDP abstraction .The implementation of these two PDPs has two layers; the functional layer & implementation layer. Functional layer helps for the Blacklist/Whitelists access interface & implementation layer helps for Java layer & Directory Interface & LDAP& Handel them.

6. Conclusion

We find that for a flexible multipolicy authorization framework for GT4, The framework is based on the XACML and SAML specifications. The blacklist/whitelist authorization system established under the GT4 authorization framework can provide a simple and efficient method for Grid service access control. Also, this work illustrates that the GT4 authorization framework is open, scalable, and flexible.

References

- [1]. Foster, I. and Kesselman, C. Computational Grids. Foster, I. and Kesselman, C. eds. *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann, 1999, 2-48.
- [2] I. Foster, C. Kesselman, S. Tuecke. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. *International J. Supercomputer Applications*, 15(3), 2001.
- [3]. Foster, I., Kesselman, C., Nick, J. and Tuecke, S. The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration, Globus Project, 2002.
- [4]. Security Assertion Markup Language (SAML) 1.0 Specification, OASIS, November 2002. <http://www.oasisopen.org/committees/security/>
- [5] V. Welch, F. Siebenlist, I. Foster, J. Brresnahan, K. Czajkowski, J. Gawor, C. Kesselman, S. Meder, L. Pearlman, and S. Tuedke, Security for Grid Services. *Twelfth International Symposium on High Performance Distributed Computing (HPDC-12)*, June 2003.
- [6] I. Foster, C. Kesselman, J. Nick, and S. Tuecke, The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. *Open Grid Service Infrastructure WG*, Global Grid Forum, June 22, 2002.
- [7] M. Naedele, Standards for XML and Web Services Security, *Computer*, vol.36, No.4, PP96-98, April, 2003.
- [8] M.Thompson, A. Essiari, S. Mudumbai , Certificate-based Authorization Policy in a PKI Environment, *ACM Transactions on Information and System Security (TISSEC)*, Volume 6, Issue 4, pp: 566-588, November 2003.
- [9] V. Welch, T. Barton, K. Keahey, and F. Siebenlist. Attributes, Anonymity, and Access: Shibboleth and Globus Integration to Facilitate Grid Collaboration. In *4th Annual PKI R&D Workshop*, April 2005.
- [10]Security for Grid Services ,Von Welch1 Frank Siebenlist , Ian Foster , John Bresnahan ,Karl Czajkowski3, compute network journal (*HPDC-12*), June 2008.
- [11] Bo Lang, Ian Foster, Frank Siebenlist, Rachana Ananthakrishnan Tim Freeman :A Multipolicy Authorization Framework for Grid Security, *J Grid Computing* (2009), 7:501–518
- [12] Sarjeet Singh, Seema Bawa, “A Framework for Handling Security Problems in Grid Environment using Web Services Security Specifications”, Second International Conference on Semantics, Knowledge and Grid, SKG2006.
- [13] A Privacy, Trust and Policy based Authorization Framework for Services in Distributed Environments,

Sarjeet Singh, and Seema Bawa, *International Journal of Electrical and Computer Engineering* 2:2 2007.

Author Information

Debabrata Singh is an Assistant professor in the Department of Information Technology, holds MTech in Computer Science & Engg. (BPUT,BBSR) He has nearly four years experience in teaching, software development and research. Presently, he is working as Assistant professor in ITER,SOA University, Bhubaneswar,Orissa He has published 13 papers on multi agent technologies & Grid Computing environment in national & international journals and conferences.

Bhupendra Kumar Gupta is an Assistant professor in the Department of Computer Applications, holds MTech in Computer Science & Engg. (Utkal University,BBSR) He has nearly six years experience in teaching, and research. Presently, he is working as Assistant professor in ITER,SOA University, Bhubaneswar,Orissa He has published 6 papers on Wireless Mesh Networks, MANETs, multi agent technologies & Grid Computing environment in national & international journals and conferences.

Biswa Mohan Acharya is an Assistant professor in the Department of Computer Applications, holds MTech in Computer Science & Engg. (IIT, Guwahati) He has nearly 10 years experience in teaching, and research. Presently, he is working as Assistant professor in ITER,SOA University, Bhubaneswar,Orissa He has published 12 papers on Wireless Mesh Networks, MANETs, multi agent technologies & Grid Computing environment in national & international journals and conferences.

Sarbeswar Hota is an Assistant professor in the Department of Computer Applications, holds MTech in Computer Science & Engg. (ITER, Bhubaneswar) He has nearly 8 years experience in teaching, and research. Presently, he is working as Assistant professor in ITER,SOA University, Bhubaneswar,Orissa He has published 5 papers on Wireless Mesh Networks, MANETs, multi agent technologies & Grid Computing environment in national & international journals and conferences.

Anti-Trust Rank: Fighting Web Spam

Ms. Jyoti Pruthi¹ and Dr Ela Kumar²

¹MCA, Manav Rachna College of Engineering,
Faridabad, India-121001

²SICT, Gautam Buddha University,
Gr Noida, India

Abstract

The Web is both an excellent medium for sharing information as well as an attractive platform for delivering products and services. This platform is, to some extent, mediated by search engines in order to meet the needs of users seeking information. Search engines are the “dragons” that keep a valuable treasure: information [8]. Given the vast amount of information available on the Web, it is customary to answer queries with only a small set of results (typically 10 or 15 pages at most). Search engines must then rank Web pages, in order to create a short list of high-quality results for users. Web spam can significantly deteriorate the quality of search engine results. Thus there is a large incentive for commercial search engines to detect spam pages efficiently and accurately. Here we present the main techniques recently introduced for Web Spam detection.

Keywords: Web Graph Model, Biased Page Rank, Trust Rank, Anti Trust Rank.

1. Introduction

Web spam refers to hyperlinked pages on the WorldWideWeb that are created with the intention of misleading search engines. With the search engines' increasing importance in the people's life, there are more and more attempts to mischievously influence the page rankings. This kind of action called web spamming is always illegal, since it misleads both search engines and users seriously. Web spamming, the practice of introducing artificial text and links into web pages to affect the results of searches. It is also a serious problem for users because they are not aware of it and they tend to confuse trusting the search engine with trusting the results of a search.

Furthermore, it has a negative economic and social impact on the whole web community. It has been found that a good percentage of web pages are spam. Spammers are playing tricks on the search engines by all means, for example, term spamming, link spamming, cloaking and redirection [1].

For example, a web site may spam the web by adding thousands of keywords to its home page, often making the text invisible to humans. A search engine will then index the extra keywords, and return the web page as an answer to queries that contain some of the keywords. Another web spamming technique is the creation of a large number of bogus web pages, all pointing to a single target page. Since many search engines take into account the number of incoming links in ranking pages, the rank of the target page is likely to increase, and appear earlier in query result sets. For instance, consider a cluster of web sites that link to each other's pages repeatedly. These links may represent useful relationships between the sites, or they may have been created with the express intention of boosting the rank of each other's pages. In general, it is hard to distinguish between these two scenarios.

For all the reasons we have mentioned, Web spam detection is a challenging problem. Since spammers are constantly coming up with more and more sophisticated techniques to beat search engines.

Many anti-spamming techniques have been proposed so far [2, 3, 4, 5, 6, 7]. Trust Rank [2] improves the PageRank by using good seeds. It can effectively demote the pages that adopt link spamming tricks. Baoning Wu and Brian D. Davison propose algorithms for detecting link farms automatically by first generating a desirable seed set and then expanding it [5]. In actual fact, almost all of these biased ranking algorithms employ a seed set and this set plays an important role in identifying web spam.

2. Related Work

Recent work [1], addressed this problem by exploiting the intuition that good pages i.e. those of high quality are very unlikely to point to spam pages or pages of low quality. They propagate Trust from the seed set of good pages recursively to the outgoing links. However, sometimes spam page creators manage to put a link to a spam page on

a good page, for example by leaving their link on the comments section of a good page. Thus, the trust propagation is soft and is designed to attenuate with distance. The Trust Rank approach thus starts with a seed set of trusted pages as the teleport set [2] and then runs a biased page-rank algorithm. The pages above a certain threshold are deemed trustworthy pages. If a page has a trust value below a chosen threshold value then it is marked as spam.

The taxonomy of web spam has been well defined by Zolt'an Gy'ongyi, Hector Garcia-Molina [9]. There are many pieces of work on combating link spam. The problem of trust has also been studied in other distributed fields such as P2P systems [10]. Other approaches rely on detecting anomalies in statistics gathered through web crawls [11]. The data mining and web mining community has also worked on identifying link farms. Various farm structures and alliances that can impact ranking of a page have been studies by Zolt'an Gy'ongyi, Hector Garcia-Molina [12]. Baoning Wu, Brian D. Davison identifies link farm spam pages by looking for certain patterns in the webgraph structure.

In our work, we exploit the same intuition, in a slightly different way. Thus we start with a seed set of spam pages and propagate Anti Trust in the reverse direction with the objective of detecting the spam pages which can then be filtered by a search engine. We found that the average page-rank of spam pages reported by Anti-Trust rank was typically much higher than those by Trust Rank. This is very advantageous because filtering of spam pages with high page-rank is a much bigger concern for search engines, as these ages are much more likely to be returned in response to user queries.

3. Preliminaries

3.1 Web Graph Model

The web can be modeled as a directed graph $G = \{V, E\}$ whose nodes correspond to static pages (V) on the web, and whose edges correspond to hyperlinks (E) between these pages. The web graph (G) is massive containing billions of nodes and edges. In addition, G is dynamic or evolving, with nodes and edges appearing and disappearing over time.

In the web graph, each page has outgoing links referred to as outlinks and incoming links referred to as inlinks. The number of inlinks of a web page is called its indegree and the number of outgoing links is referred as outdegree of the page. Several studies on the analysis of the structure of web graph have shown that these links exhibit a power-law degree distribution.

One study [14] models the structure of the web as a Bowtie structure. In this model, the majority of the web pages are a strongly connected graph. Some pages do not have inlinks called unreference pages. Pages without any outlink are referred as non-referencing pages. Also, pages that do not have either inlink or outlink are called as isolated pages.

Mathematically, the graph structure can be encoded as a matrix Eq (i) where

$$G[i, j] = \begin{cases} 1 & \text{if } i \text{ connects to } j \\ 0 & \text{Otherwise} \end{cases} \quad (i)$$

In addition, transition matrix (T) Eq (ii) and inverse transition matrix (I) Eq (iii) captures the outdegree and indegree of the web graph and they can be defined as: Transition Matrix.

Transition Matrix:

$$T[i, j] = \begin{cases} 1/\text{outdegree}(j) & \text{if } j \text{ connects to } i \\ 0 & \text{if } j \text{ does not connect } i \end{cases} \quad (ii)$$

Inverse Transition Matrix:

$$I[i, j] = \begin{cases} 1/\text{indegree}(j) & \text{i connects to } j \\ 0 & \text{if } i \text{ do not connect } j \end{cases} \quad (iii)$$

3.2 Biased Page Rank

Page Rank [15] is one of the most popular link based methods to determine a page's global relevance or importance. Page rank assigns an importance score (page rank) proportional to the importance of other web pages which point to it. While page rank is a good approach to measure the relevance of a page, it is also vulnerable to adversarial IR, by way of link spamming, which can enable web pages to achieve higher than deserved scores. Page rank r is defined as the first eigenvector of the matrix A where A is defined as follow:

$$A_{ij} = \beta T_{ij} + (1 - \beta)/N \quad (iv)$$

where T is the transition matrix,

N is the total number of web pages and β is a decay factor and $0 < \beta < 1$.

While page rank assigns a score proportional to generic popularity of a page, biased page rank or topic-specific page rank [16] measures the popularity within a topic or

domain. Here the equivalent random surfer model is as follows. When the random surfer teleports, he picks a page from a set S of web pages which is called the teleport set. The set S only contains pages that are relevant to the topic. Corresponding to each teleport set S , we get a different rank vector.

In matrix Eq (v) representation:

$$A_{ij} = \begin{cases} \beta T_{ij} + (1 - \beta)/|S| & \text{if } i \in S \\ \beta T_{ij} & \text{otherwise} \end{cases} \quad (v)$$

where A is a stochastic matrix as before. Here, we have weight all pages in the teleport set S equally, but we could weight them differently if we wish.

4. Trust Rank

The Trust Rank algorithm is an approach to find differentiates trustworthy pages from spam pages [17]. The algorithm involves running a biased pagerank algorithm with the teleport set being a manually labeled set of trustworthy pages. This work exploits the intuition that good pages are unlikely to point to spam pages. Thus the approach looks to propagate Trust along forward link, attenuating with distance. Running the biased pagerank as mentioned achieves this effect. Finally, a thresholds value is chosen and all pages below the threshold are marked as spam pages.

4.1 Inverse Page Rank

Inverse page-rank is computed by reversing the in-links and out-links in the webgraph. In other words, it merely involves running pagerank on the transpose of the web graph matrix. Thus, a high inverse page-rank indicates that one can reach a huge number of pages in a few hops along outlinks starting with the given page. Thus, this metric was found to be useful in selecting a seed set of pages in the Trust Rank algorithm.

4.2 Selecting the Seed Set of Spam pages

It was pointed out that there are two important issues in selecting the seed set of pages in the Trust Rank algorithm [17].

It is important to choose pages in the seed set, which are well connected to other pages and can therefore propagate trust to many pages quickly. Since the Trust Rank approach makes trust flow along the outlinks of a pages, it was therefore important to choose pages that had a large number of outlinks.

It is generally more important to ascertain goodness of pages with higher pageranks, since these pages will typically appear high in search query results. It was observed [17] that choosing pages with high pageranks would be more useful towards this goal, since the pages pointed to by high page rank pages are likely to have high pagerank themselves.

5. Antitrust Rank

Our approach is broadly based on the same approximate isolation principle [17], i.e. it is rare for a good page to point to a bad page. This principle also implies that the pages pointing to spam pages are very likely to be spam pages themselves. The Trust Rank algorithm started with a seed set of trustworthy pages and propagated Trust along the outgoing links. Likewise, in our Anti-Trust Rank algorithm, Anti-Trust is propagated in the reverse direction along incoming links, starting from a seed set of spam pages. We could classify a page as a spam page if it has Anti-Trust Rank value more than a chosen threshold value.

Alternatively, we could choose to merely return the top n pages based on Anti-Trust Rank which would be the n pages that are most likely to be spam, as per our algorithm. Interestingly, both Trust and Anti-Trust Rank approaches need not be used for something very specific like detecting link spam alone. The approximate isolation principle can in general enable us to distinguish good pages from the not-so good pages. Thus, for the purpose of our work we consider pages in the latter category as spam as well.

5.1 Selecting the Seed Set of Spam pages

We have similar concerns to [17], with regard to choosing a seed set of spam pages. We would like a seed set of pages from which Anti-Trust can be propagated to many pages with a small number of hops. We would also prefer if a seed set can enable us to detect spam pages having relatively high pageranks. In our approach, choosing our seed set of spam pages from among those with high pagerank satisfies both these objectives. We select our seed set of spam pages from among the pages with high pagerank. This helps us nail our twin goals of fast reachability and detection of spam pages with high pagerank.

5.2 The AntiTrust Algorithm

1) Obtain a seed set of spam pages labeled by hand. Assign pages with high pageranks labeling by a human to get a seed set containing high pagerank pages.

Let $N = \{n, \text{ where } n=0, 1, 2, \dots\}$
 $n \rightarrow \text{spam page in the seed set}$

2) Compute T

Let S = matrix of binary webgraph
 Then T = transpose of S or S'

3) Run the biased pagerank algorithm on the matrix T, with the seed set as the teleport set.

4) Rank the pages in descending order of pagerank scores. This represents an ordering of pages based on estimated Spam content. Alternatively, set a threshold value and declare all pages with scores greater than the threshold as spam.

5.3 Example

Initially, the Anti-Trust Rank value is equally distributed among all the pages of seed set. The subsequent Anti-Trust Rank computation is simply the Inverse-Page rank computation with the teleport set chosen to be our seed set. In the example in figure 1, let's assume that seed set of spam pages is 1. Thus Anti-Trust would propagate to page 5, from which it would propagate to node 4 and subsequently to node 2 and then to 3. As it can be expected, the Anti- Trust rank would constantly attenuate with distance from the seed set, as a result of which the good nodes would get relatively low Anti-Trust scores, in the given example. In the given example blue nodes represent spam pages and orange nodes represent the good pages.

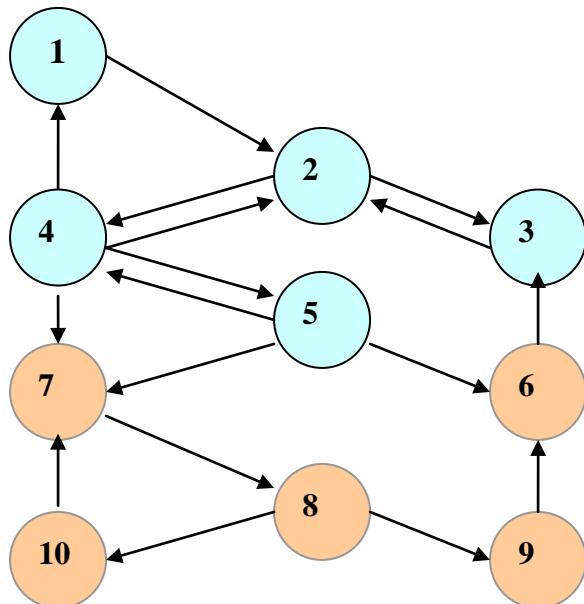


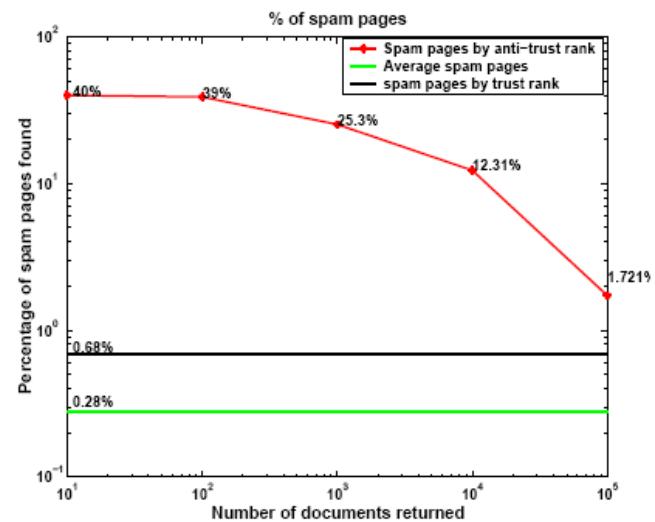
Fig.1 Web Graph with Spam and Good Pages

5.4 Results and Analysis

We ran our experiments on the WebGraph dataset, [18]. We chose data corresponding to a crawl of the “in” domain containing about 20 millions nodes and 400 million links. Clearly, the only perfect way of evaluating our results is to manually check if the pages with high Anti-Trust score are indeed spam pages and vice-versa. It was observed in [17] that this process is very time consuming and often hard to do in practice.

We however solve this problem by coming up with a heuristic which in practice selects spam pages with nearly 100% precision and also a recall which is a reasonable fraction of the set of true spam pages, on our dataset.

The Heuristic: We compiled a list of substrings whose presence in a URL almost certainly indicated that it was a spam page, on our dataset. This heuristic enables us to measure the performance of our Anti-Trust Rank algorithm and compare it against the Trust Rank algorithm with a good degree of reliability. As per this heuristic, 0.28 % was spam pages.



We can see that both Anti-Trust Rank and Trust Rank are significantly better than the naive baseline corresponding to a random ordering of the pages, for which the precision of reporting spam would merely be the percentage of spam pages in the corpus. However we also see that Anti-Trust rank typically does much better than Trust Rank at different levels of recall.

6. Conclusion

We have proposed the Anti-Trust Rank algorithm, and shown that it outperforms the Trust Rank algorithm at the task of detecting spam pages with high precision, at various levels of recall. Also, we show that our algorithm tends to detect spam pages with relatively high pageranks, which is a very desirable objective.

It would be interesting to study the effect of combining these both the Trust Rank and Anti-Trust Rank methods especially on data containing a very high percentage of spam pages.

Acknowledgment

I express my sincere gratitude and acknowledgement towards Dr. Ela Kumar, Associate Professor, who guided me. It was her constant support and inspiration without which my efforts would not have taken this shape. I sincerely thank her for this, and seek her support for all my future endeavors.

References

- [1] Z. Gyongyi and H. Garcia-Molina. Web spam taxonomy. In *First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*”, 2005.
- [2] Z. Gyongyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *vlbd'2004: Proceedings of the Thirtieth international conference on Very large data bases*, pages 576–587. VLDB Endowment, 2004.
- [3] V. Krishnan and R. Raj. Web spam detection with anti-trust rank. In *AIRWeb'06*, August 2006.
- [4] B.Wu and K. Chellapilla. Extracting link spam using biased random walks from spam seed sets. In *AIRWeb'07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pages 37–44, New York, NY, USA, 2007. ACM.
- [5] B.Wu and B. D. Davison. Identifying link farm spam pages. In *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 820– 829, New York, NY, USA, 2005. ACM.
- [6] B. Wu, V. Goel, and B. D. Davison. Propagating trust and distrust to demote web spam. In *Proceeding of Models of Trust for the Web (MTW)*, May 2006.
- [7] L. Zhang, Y. Zhang, Y. Zhang, and X. Li. Exploring both content and link quality for anti-spamming. In *CIT '06: Proceedings of the Sixth IEEE International Conference on Computer and Information Technology*, page 37, Washington, DC, USA, 2006. IEEE Computer Society.
- [8] Marco Gori and Ian Witten. The bubble of web visibility. *Commun. ACM*, 48(3):115–117, March 2005.
- [9] The EigenTrust algorithm for reputation management in P2P networks. S. Kamvar, M. Schlosser, and H. Garcia-Molina. In *Proceedings of the Twelfth International Conference on World Wide Web*, 2003.
- [10] Spam, Damn Spam, and Statistics. Dennis Fetterly, Mark Manasse and Marc Najork. *Seventh International Workshop on the Web and Databases (WebDB 2004)*, June 17-18, 2004, Paris, France.
- [11] Link Spam Alliances. Zoltan Gyongyi, Hector Garcia-Molina. . *31st International Conference on Very Large Data Bases (VLDB)*, Trondheim, Norway, 2005.
- [12] Identifying Link Farm Spam Pages. Baoning Wu, Brian D. Davison. *WWW 2005*, May 1014, 2005, Chiba, Japan.
- [13] PageRank Computation and the Structure of the Web: Experiments and Algorithms. A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener. *Proc. WWW9 conference*, 309–320, May 2000.
- [14] The PageRank citation ranking: Bringing order to the web. L. Page, S. Brin, R. Motwani and T. inograd. Technical Report, Stanford University, 1998.
- [15] Topic-sensitive Page Rank. Taher Haveliwala. In *WWW 2002*.
- [16] Combating Web Spam with Trust Rank. Z. Gyongyi, H. Garcia-Molina and J. Pedersen. In *VLDB 2004*.

Jyoti Pruthi is pursuing PhD (computer science). She received her MPhil (computer science) degree in 2008. She holds an MCA from MDU, Rohtak. She is currently working as Lecturer with Manav Rachna College of Engineering, Faridabad, MCA Department. Earlier, she worked with CSSL, Gurgaon as a Software Engineer. Her area of specialization includes Search Engine, Spamming & Anti-Spamming Techniques.

Dr Ela Kumar PhD (Artificial Intelligence) from University of Delhi. She holds an M.Tech (CSE) and B.E (ECE) from IIT, Roorkee. She is currently working as Associate Professor with Gautam Buddha University, Gr Noida. Earlier, she worked with CSSL, Gurgaon as a Software Engineer. Her current areas of specialization include Search Engine, Spamming & Anti-Spamming Techniques. She holds to her credit around 19 Years of experience of teaching and research. Besides publishing 30 research papers in journals of repute, she has written 3 books.

Triangular Pyramid Framework For Enhanced Object Relational Dynamic Data Model for GIS

Ms. Barkha Bahl¹, Dr. Navin Rajpal ² and Dr. Vandana Sharma³,

¹Research Scholar, GGS IP University,Delhi, India

²Professor, GGS IP University,Delhi, India

³Dy.Director General, National Informatics Centre,Delhi, India

ABSTRACT

GIS (Geographic Information System) data modeling is a methodology for designing spatial databases. Efficient database design is crucial to efficient GIS implementation. A model is a simplification of the complexity of reality, which makes the reality more understandable and operational.

The Objective of this paper is to present an overview of various data models and to evolve an approach to design a model which can overcome the identified gaps of various models being studied for a GIS application. The proposed data model i.e. Enhanced Object Relational Dynamic Data Model enhances Object Relational Data Model (EORDM) by introducing reusability and dynamicity for the database . Reusability concept allows storage of distinct objects having same spatial representation onto a single storage space and dynamicity allows creation of relations on user's requirement. The conceptual data model for the same has been designed named Triangular Pyramid Framework which has been explained in section IV of this paper.

Keywords: Dynamic database, Spatial database, reusability, vector database, object relational database.

1. Introduction

A geographic Information System, GIS, is a system for capturing, storing, querying, analyzing and displaying geospatial data [12]. This technology has the capability of spatial searches, overlays, association of spatial data with non spatial to generate new information. These features make this technology different from the CAD systems and conventional database applications. The CAD application becomes GIS only when an image becomes georeferenced. "Every object present on the

Earth can be geo-referenced", is the fundamental key of associating any database to GIS.

In the Geographic Information Systems (GIS) community the concept of data modeling is well developed and integrated in the design of databases and related applications. Various data models for GIS applications have been introduced Viz. Multi-Dimensional location referencing system data model[10], Multiple representations object data model [5],[9]. Spatio– temporal data model [14] etc .

Multi-Dimensional location referencing system data model allows organizations to implement improved solutions for transportation systems using advanced technologies.

Multiple representations, object data model represents same object more than once at different abstraction levels tailored towards the need of different users or analysis. Another data model i.e.Spatio-temporal model has been developed using object-oriented concepts. Spatial objects have both spatial and temporal dimensions. The geometric attributes and relationship change over time in real world objects. Hence , these aspects have been covered in temporal model .

Conventional GIS Systems represent the real world in a snapshot only ,which is inadequate for analyzing changes and patterns of change over a period of time. Every update creates a view map and no attempt is made to establish any link between the snapshots. These systems are therefore not suitable for many applications where data need to be interpreted in the context of time, such as urban application, environmental monitoring, agricultural applications, forest management, etc. [15].

Varying approaches have been used in the design of spatial data models, but no model or abstraction of reality can represent all aspects of reality, as it is impossible to design a generic data model. For example

implementation of spatial data model are good for plotting digital environment, but are inefficient for analytic purposes and further for producing graphics. Secondly , it has been observed that in Survey of India, data capture in the digital domain is done in the dgn format with the help of microstation based softwares. The database of each colour separate is generated separately, in which all the cartographic elements, i.e. points, lines, areas and text are present in the same digital file. Unfortunatey in .dgn file interrelationship between the basic cartographic elements do not exist as it does not have any topology built into it [14]. Department of Science & Technology, Survey of India, Department of Space and GSI are coordinating to bring out an exchange format which will cater to the exchange of Vector, Raster and GIS data.

Bouille approach includes all identifiable entities and their relationships in deriving the phenomenal structure for phenomenal based design which is extremely complex[2]. Whereas Mark adopted a philosophy of including only those entities and relationships that are relevant for an application for designing a data model which is known as minimalist approach and has minimum complexity[11].

The data model is more robust and flexible if it represents reality more perfectly i.e. all entities and possible relations are incorporated while designing the database. However incorporating essential entities and relations for single application a model will be considered more efficient with respect to storage of space and ease of use. Thus, the selection or design of data model must therefore be based both on the nature of the phenomenon the data represents and the specific manipulation processes which will be required to perform on the data. In this paper , we propose a Triangular framework for enhanced object relational dynamic data model that provides these and other features.

The rest of this paper is organized as follows: Section 1 covers the survey of various GIS data models. Section 2 discusses proposed data model and new features introduced in the same. Section 3 describes Triangular Pyramid Framework for the proposed data model. Section 4 concludes the paper.

Survey of GIS Models

Data models can be vector data model or raster data model. .Vector data represents discrete feature and raster data represents continuous features. This section will discuss various Vector Data Models viz. The Spaghettie Model[4], Topology Model[19], Polyvrt[13], Relational[18], Georelational [16], Object based [1]and Geodatabase [20] etc.

1.1 The Spaghettie Model

In Spaghettie a digital cartographic data file is constructed referred to as a Spaghettie file which is a collection of coordinates, strings heaped together with no inherent structure[4]. This model is inefficient for most types of spatial analysis, since any spatial relationships must be derived through computation. On the other hand the lack of stored spatial relationships, which are extraneous to the plotting process marks the spaghetti model efficient for reproducing the original graphic image. Thus they are used for simpler forms of computer assisted cartographic production.

1.2 Topological model

In topological model the information allows the spatial definitions of points, lines, and polygon type entities to be stored in a non-redundant manner[19]. The GBF/DIME (Geographic Base File/Dual Independent Map Encoding) model is built upon the topological concept. The model represents a directed graph,in which an explicit direction is being assigned by recording a from-node and to-node which automatically check for missing segments and other errors in the file. In this model the basic logical entity is a straight line , where street, river, etc is represented as a series of straight line segment which are spatially defined. The main problem with this model is that the individual line segment do not occur in any particular sequence order, so to retrieve all line segments which define the boundary of a polygon, an exhaustive search must be done as many times as there are line segments in the polygon boundary.

1.3 POLYVRT

Peucker and Chrisman (1975) developed POLYVRT[13]. This model had overcome the very major retrieval and manipulation inefficiencies seen in simpler topologic structures by explicitly and separately storing each type of data entity in a hierarchical data structure. It made distinctions between types of entities both logically and topologically meaningful, so that a chain is denoted as the basic line entity. It facilitates easy search and retrieval and there is partitioned storage. Leads to storage overhead (pointers) and integrity of pointers. It is a multipurpose database model.

1.4 Relational Data model

The power and elegance of the relational model stems from the fact that it uses a single construct, the relation [18]. Five functional closed operations are defined in

relations, namely, union, difference, selection, projection and Cartesian product.

For spatial applications, however, the resulting representation is inadequate.

For example, if layers are represented with plain relation, operations such as overlaying and reclassification cannot be derived from the fundamental relational databases.

In the relational model these operations are hidden in the physical level. As a result important information is lost and the system is tied to some specific implementation. Thus relations are inadequate as the sole modeling construct for geographical applications.

1.5 GeoRelational data model

Data representation for GIS applications includes the spatial and attribute component[16]. Spatial data describes the location of spatial features, whereas attribute data describes the characteristic of spatial features. The GeoRelational data model stores spatial and attribute data separately in a split system. Spatial data is stored in graphic files and attribute data is in a relational database. A GeoRelational data model uses the feature label or ID to link the two components as shown in figure below. The two components must be synchronized using some ID so that they can be queried, analyzed and displayed in unison.

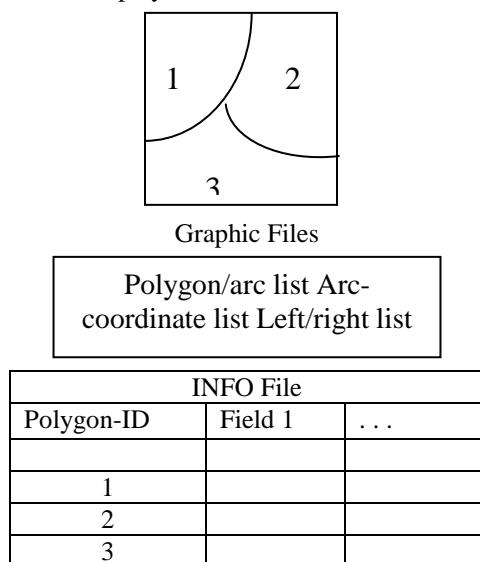


Fig 1 : Georelational database design

1.6 Object Based Data Model

A fundamental requirement for spatial database design is the ability to model spatial properties, i.e., to associate parts of space with an attribute [1].

Parts of space are usually represented by points, lines and regions and are known as geometric features. Spatial applications deal with two, orthogonal and generalizations of spatial properties. One is association of the whole of space with an attribute and the other is associations of sets of attribute and geometric feature. The former is modeled with concepts oriented toward objects [17]. Object Based Data Model has been used as a means of conceptual structuring of geographic information. In particular it models real-world objects (or entities) with a precise and ‘crisp’ spatial location and extent.

The object based data model differs from the georelational data model in two important aspects. First, the object-based data model stores both the spatial and attribute data of spatial features in a single system i.e. an object rather than a split system. Second, the object-based data model allows a spatial feature to be associated with set of properties and methods. The same is represented in Fig 2 .

The shape field stores the spatial data of land-use polygons, other field store attribute data such as landuse_id and category.

Since both spatial data and attribute data is stored in a single system the problem of data synchronization is eliminated that is found in split system (Georelational datamodel).

| Objec ted | Shape | Land use_ ID | Categ ory | Shape_ Length | Shape_ Area |
|-----------|---------|--------------|-----------|---------------|-------------|
| 1 | Polygon | 1 | 5 | 14,607.7 | 5,959,800 |
| 2 | Polygon | 2 | 8 | 16,979.3 | 5,421,216 |
| 3 | Polygon | 3 | 5 | 42,654.2 | 21,021,728 |

Fig. 2 : Object based data model .

1.7 Geodatabase Data Model

This model is built on arc-objects. It uses the geometries of point, polyline and polygon to represent vector-based spatial features [20]. The data structure of geodatabase distinguishes the feature classes and feature datasets. A feature class stores spatial data of the same geometry type and its datasets store feature classes that share the same coordinate system and area extent. In this model, feature classes can be standalone feature classes or members of a feature dataset. The geodatabase constitutes a uniform repository of both spatial and attribute data in a single database system. Objects in the geodatabase can have behavior associated with them. Integration with object-oriented concepts and COM technology allows great level of customization and reuse

of the model to create application-specifications, which may (fig. 3) provide the framework for interoperability. The main problem we encountered was the custom domains, but this was not investigated. In addition, although the ArcObjects Library is extensive, more functionality needs to be added to allow high level of customization.

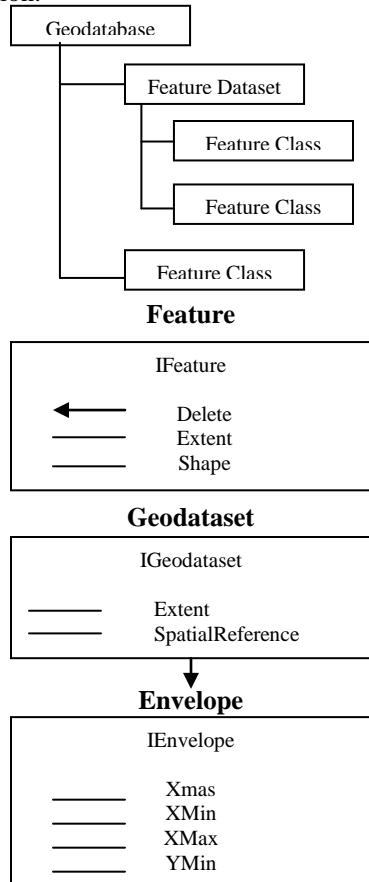


Fig . 3 Geodatabase Model

In a geodatabase , feature classes can be standalone feature classes or members of a feature dataset. A feature object implements the Ifeature interface. A geodataset object supports IGeodataset and an Envelope object supports IEnvelope.

1.8 Bi Level Data Model

There are two separate layers [3] in this model as mentioned below:

Higher level data model (Geographic object data model): This level consists of the geographic objects and a set of semantic spatial functions through which the topological relationships among objects can be defined or derived.

Lower level data model (Geometric object data model): This level consists of geometric objects which are actual spatial representations of the geographic objects. It also has a set of functions for retrieval, manipulation, computation of geometric objects. In this model, relationship between geographic and spatial objects is investigated, Spatial object is not PART-OF a geographic object , but is just a representation of the geographic object , similar to the mapping from one object to any other object(s). Where as Triangular pyramid model, the proposed model has three abstraction levels represented using three components – the object component, Geometric component and the location component. The details of the same have been discussed in the next section. In this model “uses” relationship type has been introduced, where the common reference table is used for representing the location component for various maps.

2. PROPOSED DATA MODEL – “ENHANCED OBJECT RELATIONAL DYNAMIC DATA MODEL”

An Object relational databases lie in between relational databases and object oriented databases. The approach is eventually that of relational database only.

The new construct added to the core functionality of traditional database includes user defined data types (for creating layers) and the ability to create the same dynamically through database programming interface. The reusability concept of Object oriented paradigm is incorporated using common reference location table. The basic selections, retrievals, manipulations have been included to the capabilities of the function associated with object classes.

For the rest of this section, we first introduce the features & assumptions. Features of our Enhanced Object relational dynamic data model are followed by the four modeling aspects of the proposed data model.

2.1 Features and Assumptions

Below are the features and assumptions, which we have considered for our new data model for GIS applications.

- All the Maps have the same extent. i.e. same (Xmin,Ymin) and (Xmax, Ymax)
- Each layer of the map is treated separately as virtual map sheets.
- All the maps have common reference coordinate table(x, y).
- All the spatial calculations use same scale of measurement.

- Distinct geographic objects having exactly the same spatial representation meaning occupying same physical space on the earth's surface will be represented by distinct geometric objects for their spatial representations.
- The data model is designed so that it is possible to have the dynamic creation of tables for user-defined layers, so that those data may be queried and visualized further.
- It defines the relationship between attribute and spatial data in a manner that is efficient with respect to space and time while executing complex queries.

2.2 Four Modeling Aspects

The purpose of this research paper is to develop a data model called enhanced object relational vector data model that better facilitates the access and storage of both spatial and attribute data than conventional GIS database models. The modeling aspect of this data model includes four concepts. These are ***vector data***, ***dynamicity***, ***reusability***, and ***object-relational concept*** as discussed below:-

1) Vector data

Vector data represents one major category of data managed by GIS [12]. It represents each feature as a row in a table and feature shapes are defined by x,y location in the space. Features can be discrete points, lines, and polygons.

Points: - having a pair of coordinates. Locations such as the address of a customer or the spot of crime are represented as points.

Lines: - series of coordinate pairs. Streams or roads are represented as lines.

Polygons: - are defined by borders and are represented by closed regions. They can be defined such as parcel of land, counties, watersheds etc.

- 2) **Dynamicity:** - As we know, response time and storage space are the major concerns for GIS applications. The proposed model provides a solution for improving the performance of response time and storage space by creating tables dynamically. The dynamic database can be easily created as per the user's requirement [7].

The design of a dynamic data model increases the availability of spatial data to the users by providing the latest updated information [6].

- 3) **Reusability:** In the proposed model different object classes could be defined using the "uses" relation type which introduces reusability aspect of object oriented model. The reusability has been incorporated through common reference location table. The details are discussed in section 4.

- 4) **Object relational concept:** In our model existing relational systems SQL server is enhanced in the direction of object orientation or in terms of front end interfaces by providing an object oriented flavor or outlook to a conventional relational system. Thus, both object and relational views are provided for designing the data model.

2.3 Triangular Pyramid Framework for EORDM

Unlike the data models being discussed, the proposed data model has three levels of abstraction . They are : The Object Component (Highest Level),The Geometric Component (Middle level) The Location Component (Lowest Level).

The model (fig. 4) covers the representation of whole information required by GIS application in three components: Object, Geometric and location. Diagrammatic representation have been given below where the mapping of the model with the database table have been mentioned:

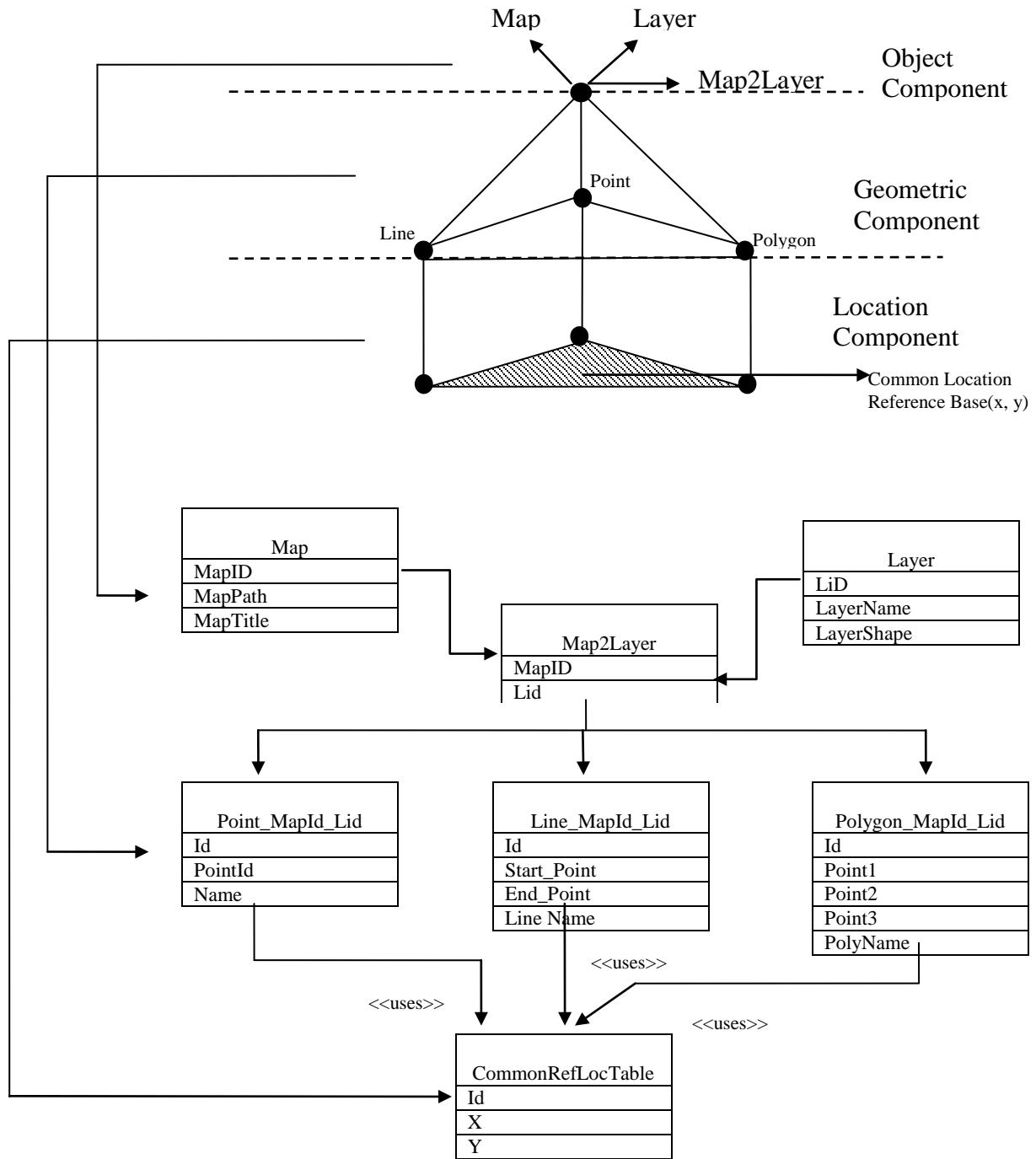


Fig.4 Triangular Pyramid Framework for EORDM

The Object Component (Highest Level)

A map is a symbolic depiction highlighting relationships between elements of that space what we here call them as layers such as rivers, lakes, mountains, districts, forests, temples, Etc. Thus map is a combination of

different types of layers. These maps and layers are called objects as they are real life entities having both attribute and behavior.

Thus object level is the highest level of abstraction, and at this level geographic information is represented by

layers representing the relative position of spatial objects.

The Geometric Component (Middle Level)

The Geometric component is the middle level as it is the interface between the object and its actual spatial existence on the earth.

Each geographic object in the higher level has its corresponding geometric object. The information at geometric level represents the shape of the geographic object, which is categorized into three: - point, line polygon.

The first is point data where each object is associated with a single location, Example a city, district, school, hospital etc.

The second is line data where the location is described by the string of points, Example: - road, river, railway line etc.

The third is polygon data, where the location of object is represented by a closed string of coordinates. They are thus associated with areas over define space, Example: - mountains, parks, wasteland area etc.

Distinct geographic objects may have same physical representation, for an instance Roads and rivers both are distinct geographic objects but are represented by same geometric object i.e. line.

Thus In geometric level, the internal information represents the geometry of the object keeping this information hidden from the object level.

The Location Component (Lowest Level)

The lowest level of the proposed data model is Location Component which represents the actual screen coordinate value of the geometric object in middle level. In the proposed model, we are considering that the maps have same extent i.e. they have same (Xmin, Ymin) and (Xmax, Ymax) screen coordinates. If we have different maps having same map extent on the screen and corresponding to them layers and their screen coordinates are stored. It may happen that many of the layers on different maps have same value of (x,y) coordinate on the screen.

For an example consider maps of four different cities viz: .India, South America, North America, Asia. With an assumption that they have same extent on the screen, layers for these maps are drawn .It may happen that for

the screen coordinates (100,200), India and South America may have a same point .Corresponding to (275,145) we have points in South America, North America, Asia. So if we store (100,200) twice and (275,145) thrice ,then it will be a redundant effort of storing the data in the database table., which leads to both wastage of space and time, hence a common reference coordinate table has been developed, which contains all the coordinates that have information corresponding to them in any of the maps.

The main concern in GIS applications is the response time and in turn the response time is dependent on how efficiently underlying database model is designed.

The proposed data model is designed keeping the same in mind. This leads to development of a database model which is dynamic and at the same time it uses the concept of reusability.

3 DDL FOR TRIANGULAR PYRAMID FRAMEWORK. The above mentioned model has been realized using following create command(DDL).

DDL for the proposed model

Table creation :

CREATE TABLE MAP

```
(  
    Mapid : int <<PK>>  
    Mapname : varchar(50)  
    Mappath : varchar(200)  
)
```

CREATE TABLE LAYER

```
(  
    Layerid : int<<PK>>  
    Layername : varchar(50)  
    LayerShape : varchar(50)  
)
```

Create Table Map2Layer

```
(  
    Id : int  
    MID : Int <<PK>><<FK>> References MAPID(Map)  
    LID : Int <<PK>><<FK>> References Layerid(Layer)  
)
```

CREATE Table Commonreference Table

```
(  
    LocID : Int <<PK>>  
    X: Int <<UNIQUE KEY>>  
    Y: Int <<UNIQUE KEY>>  
)
```

DDL for the EORDM.

Fig (4) represents a schema used to model GIS applications consisting of Roads , national highways , parks, waste land areas, schools and hospital . Each of these are real world entities/objects being represented by different layers. Layers are functionally related map features. Features types (point, line or polygon) and feature themes (What the feature represents) are the two most common considerations for organizing data layers. In our schema Roads and national highways are both lines (feature type) , but they have to be stored in separate layers , because of different attributes(feature theme).

SCHOOLS (Name: String , P: Point)
HOSPITALS (Name: String , P: Point)

ROADS (RoadNo: integer , Length: Real , L: Line)

NATIONALHIGHWAYS (Name: String , Length: Real , L: Line)

PARKS (Area: Integer , P: Polygon)

WASTELAND (Area: Integer , P: Polygon)

Fig (4) : GIS Schema

Select *Map2layer.Lid*,
Layer.layerName , *Map.Mapname*
From *Map2layer*,*layer.Map*
Where *Map2layer.Mid* = *Map.MapID* and
Map2layer.LID = *Layer.LID*
And *Mapname* = 'Delhi'

Sample queries for the proposed model

Query 1 : All the layers of Delhi Map

Query 2: Select total wasteland area of UP Map

Select Sum(*UP_Wastelands.Area*) as *TotalArea*
From *UP_Wastelands*

Query 3: Select No of Hospitals in Maharashtra

Select count(*) as *NoOfHospitals*
From *Maharashtra_Hospitals*

Query 4: Display (X,Y) coordinates of Delhi Schools

Select *CRT.X* , *CRT.Y* , *CRT.Name*
From *CommonRefreneceTable* as *CRT* ,*Delhi_Schools* as *DS*
Where *DS.PointId* = *CRT.XY_Id*

CONCLUSION

In this paper, a new data model named Triangular Pyramid framework for enhanced object relational dynamic vector data model has been introduced for representing the complete information being required for representing the data for GIS based application.

Here , we have suggested three layer components namely object , geographic and geometric. The details of the same have been explained in the previous section. Besides, we also have specified DDL for implementing the proposed model . With this proposal , we have shown that the layer/component architecture of EORDM is more suitable for modeling GIS application . In the dynamic database less storage space is required as the instances of the databases are created at the time of requirement only.Secondly , the storage space and response time is reduced since the common reference table is being introduced to handle data efficiently.

REFERENCES

- [1] Banerjee, J.,1987, "Data Model issues for object-oriented applications", ACM Transactions on office information system 5(1) pages 445-456, 1987.
- [2] Bouille, F. 1978. "Structuring Cartographic data and Spatial Processing with the hyper graph-based data structures" Proceedings of the First International Symposium on Topological Data Structures for Geographic Information Systems, Harvard University, Cambridge.
- [3] Choi Amelia , Lub.w.s, 1991 " A Bi-level object oriented data model for GIS Application ", IEEE,0730-157/90/0000/02385/0,pp 238-244
- [4] Dangermond J.1982," A Classification of software components commonly used in geographic information systems", proceedings,US, Australia workshop on the design and implementation of computer-Based Geographic Information System, Honolulu, Hawaii.,PP 70-91,
- [5] Egenhofer , J.M.,Clementini, E., Di Felice P. 1994, "Evaluating inconsistency among multiple representations.", Sixth International Symposium on Spatial Data Handling ,Edinburgh, Scotland, pp. 901-920.
- [6] Huang, B. 2003 . "An object Model with Parametric Polymorphism for Dynamic

- Segmentation.” International Journal of Geographic Information Sciences 17:343-60
- [7] Ka-Wai Kwan and Wen-Zhing Shi,2002, “A Study of Dynamic Database in Mobile GIS”, Symposium on Geospatial Theory, Processing and Applications.Pg 1 -10
- [8] Koncz,N.A. and T.M. Adams 2002.”A data Model for Multidimensional Transportation Applications.”, International Journal of Geographic Information Sciences 16:551-70
- [9] Kilpelainen , T. ,2000, Maintenance of multiple representations database for topographic data. The Cartographic Journal, Vol. 37, No.2 , pp. 101-107.
- [10] Newell . R.G and Sancha T.L,1990 “The difference between CAD and GIS ”,Computer – Aided Design , Volume 22, Issue 3, April 1990, Pages 131-135.
- [11] Mark, D.M. 1975.” Computer analysis of topography: a comparision of terrain storage methods”, Geografiska Annaler, vol . 579, pp. 179-188.
- [12] Oliver Kersting and J'urgen D'ollner, Nov2002, “ Interactive 3d visualization of vector data in gis.”, In Proceedings of the 10th ACM International Symposium on Advances in Geographic Information Systems (ACMGIS 2002), ISBN 1-58113-591-2,107-112
- [13] Peucker, T. and N. Chrisman, 1975.” Cartographic Data Structures”, The American Cartographer, Vol. 2, pp 55-69.
- [14] Perumal . A. “GIS Applications for resources management-National Initiatives”, Natural Resources Data Management System, Department of Science and
- [15] Raza, A., 2001, “Object Oriented temporal GIS for urban applications”. PhD Thesis, ITC Publication Number 79.
- [16] Robinson, A.H.,J.L.Morrison, P.C.Muehrcke, A.J. Kimerling, and S.C.Guption. 1995. “ Elements of Cartography”, 6th ed. New York: Wiley.
- [17] Shi, W. , B. Yang and Q.Li.2003 ”An object Oriented Data Model for complex objects in Three Dimensional Geographic Systems”. International Journal of Geographic Information Sciences 17:411-30
- [18] Smith, Karen and Zdonik Stanley, 1987, “ Intermedia: A case study of the differences between Relational and object-oriented Database Sytems.”, In OOPS SLA 37 Proceedings, Pages 452-465,
- [19] US Dept. of commerce, Bureau of the census, 1970. the DIME geocoding system, in report no. 4, census use study.
- [20] Zeiler, M. 1999. “Modeling our world:” The ESRI Guide to Geo database Design. Red lands, CA: ESRI Press.

Simulation and performance Analysis of a Novel Model for Short Range Underwater Acoustic communication Channel Using Ray Tracing Method in Turbulent Shallow Water Regions of the Persian Gulf

Mohammad Javad Dargahi , Abdollah Doosti Aref , Dr Ahmad Khademzade

Islamic Azad university ; Central Tehran Branch
Department of electrical engineering

Abstract

High data rate acoustic transmission is required for diverse underwater operations such as the retrieval of large amounts of data from bottom packages and real time transmission of signals from underwater sensors. The major obstacle to underwater acoustic communication is the interference of multipath signals due to surface and bottom reflections. High speed acoustic transmission over a shallow water channel characterized by small grazing angles presents formidable difficulties. The reflection losses associated with such small angles are low, causing large amplitudes in multi-path signals. In this paper, based on the results obtained from practical measurements in the Persian Gulf and available data about sound speed variations in different depths, we propose a simple but effective model for shallow water short-range multipath acoustic channel. Based on the Ray theory, mathematical modeling of multipath effects is carried out. Also in channel modeling, the attenuation due to the wave scatterings at the surface and its bottom reflections for deferent grazing angles and bottom types is considered. In addition, we consider the attenuations due to the absorption of different materials and ambient noises such as see-state noise, shipping noise, thermal noise and turbulences. We use a three-dimensional hydrodynamic model (COHERENS) in a fully prognostic mode to study the

circulation and water mass properties of the Persian Gulf – a large inverse estuary. Maximum sound speed occurs during the summer in the Persian Gulf which decreases gradually moving from the Strait of Hormuz to the north western part of the Gulf. A gradual decrease in sound speed profiles with depth was commonly observed in almost all parts of the Gulf. However, an exception occurred in the Strait of Hormuz during the winter. The results of the model are in very good agreement with our observations.

Keywords: Persian Gulf , shallow water , acoustic channel , Ray theory.

1- Introduction

High data rate acoustic transmission is required for diverse underwater operations such as the retrieval of large amounts of data from bottom packages and real time transmission of signals from underwater sensors. However, acoustic signals transmitted in shallow water are corrupted by interference from reflection and scattering at the water surface and bottom. For this reason, the shallow underwater channel is a difficult medium in which to achieve the high data rates needed for many applications. An especially difficult

problem is that the acoustic signal transmitted over a shallow water channel has associated with it inherently small grazing angles and small reflection losses. This results in significant corruption due to large amplitude multi-path signals [1], [2]. Therefore, it is very important to construct a good model of the channel and design a satisfactory communication system for this environment. This paper introduces a model that describes a shallow underwater channel using geometric and environmental parameters.

The model utilizes the impulse response of the channel with weighting according to the attenuation due to reflection, absorption etc., and windowing techniques to determine the signal-to-corruptive multi-path signal ratio (SMR) in the observation window [3], [4].

In the deep ocean multi-path signals are attenuated by spreading and reflection losses at relatively large grazing angles. The effects of this multi-path interference can be reduced by using a directional transmitter and/or receiver. However, spatial discrimination of direct path and multi-path signals by directional arrays in shallow water is virtually

impossible. On the other hand, under certain conditions the primary multi-path signals can add constructively to increase the strength of the received signal. This indicates that a high rate, coherent communication is possible for a wide range of channel geometries and parameters. We show the relationship between the achievable error-free transmission rate and the SMR. The results of computer analysis indicate that transmission rates in excess of 8 k-bits/s are possible over a distance of 13 km

and in a water depth of only 20 meters using the phase-shift-keying system (PSK).

Sea water acts as an acoustic waveguide and transmits sound signal in itself. Sound channel as a sound waveguide is a channel with random parameters. But this subject does not have the meaning of its unpredictability. The most important characteristic of sea water is its inhomogeneous nature. In the whole classifications, its inhomogeneity can be classified into two regular and random groups. Regular variations of sound speed in different layers of water leads to the formation of sound channel and this phenomenon causes the long distance sound propagation. Random homogeneity causes the scattering of sound waves and sound fields variation. Hence, in this section, as an introduction, viewpoint and basic step, investigate the variation of sound speed profile in the Persian Gulf. The Persian Gulf, referred to in some local countries as the Arabian Gulf, is an important military, economic and political region owing to its oil and gas resources and is one of the busiest waterways in the world. Countries bordering the Persian Gulf are the United Arab Emirates, Saudi Arabia, Qatar, Bahrain, Kuwait and Iraq on one side and Iran on the other side (Fig. 1). The Persian Gulf is a semi-enclosed, marginal sea that is exposed to arid, sub-tropical climate. It is located between latitudes 24°–30° N, and is surrounded by most of the Earth's deserts. The most known weather phenomenon in the Persian Gulf is the Shamal, a northwesterly wind which occurs year round [5]. In winter, the Shamal is of intermittent nature associated with the passage of synoptic weather systems, but it seldom exceeds a speed of 10 m/s. The summer Shamal is

of continuous nature from early June through to July. Seasonal variations of the Shamal are associated with the relative strengths of the Indian and Arabian thermal lows [6]. Tectonic driven subsidence deepened the seafloor of the Strait on its southern side (200–300m depths are seen in some localized seafloor depressions) and produced a 70–95m deep trough along the Iranian side of the eastern part of the Gulf. A southward widening channel leads from the Strait south across a series of sills (water depth of ~110 m) and shallow basins to the shelf edge [7]. The narrow Strait of Hormuz restricts water exchange between the Persian Gulf with the northern Indian Ocean [8].

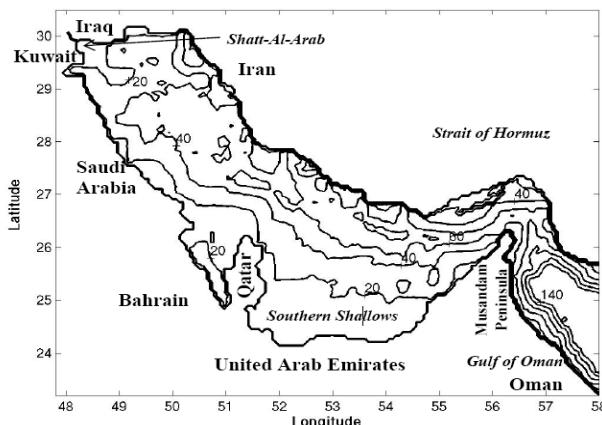


Fig.1: Bathymetry used in this study

According to the obtained measurements, as can be shown from fig. 2, sound speed is maximum during the summer (~1560 m/s) in the southern part of the Persian Gulf [9]. Also sound speed in the northern part of the Gulf varies from 1524-1528 m/s at the bottom and from 1547-1552 m/s at the surface [5]. These show that it has a 20 m/s difference from surface to bottom. In the

southern part of the Gulf except of the region surrounding Bahrain, it varies from 1548-1552 m/s at the bottom layer and from 1553-1557 m/s at the surface layer. These regions are shallow (up to 100 m) and the difference between surface and bottom temperature is negligible, therefore, sound speed is a function of both temperature and salinity [9]. Besides, Around Bahrain, it varies from 1539-1544 m/s at the bottom and from 1554-1558 m/s at the surface. In this part of the gulf, sound speed varies more than other parts because this region is shallow and salinity has more effects than temperature on sound speed [10],[11]. In the neighborhood of Iraq and Kuwait, due to the salinity of the water, the sound speed at the bottom is more than the surface. It can be obviously seen that, from the Strait of Hormuz to the western regions of the Persian Gulf, the sound speed versus the depth is reduced.

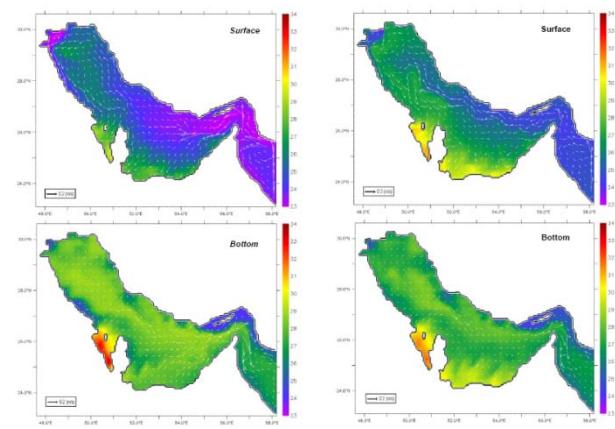


Fig.2: Sound speed (from 1500 m/s) in surface and bottom layers in deferent parts of the Persian Gulf [9]

2- Modification of Medwin formula in order to adapting with the measured data inThe Persian Gulf.

In this section, based on the performed measurements [12], to adapt the output of Medwin's equation [13] with the measured sound speed profile in the Persian Gulf, Medwin equation is modified. In Fig.4, profile of the sound speed variations versus depth in the Strait of Hormoz is shown on 56.7E° and 25.4N°, in conditions that water depth was 85m and the ADCP measurement tool, belonged to the NOAA submarine on Aug, 2, 2007, placed at the depth of 10m. By considering the subject of fix variation of sound speed versus temperature and salinity in shallow waters, which is between 1500 to 1502 m/s, the horizontal axis of this figure shows the difference of the measured sound speed with 1500 m/s. Also, the vertical axis of the profile depicts depth variations.

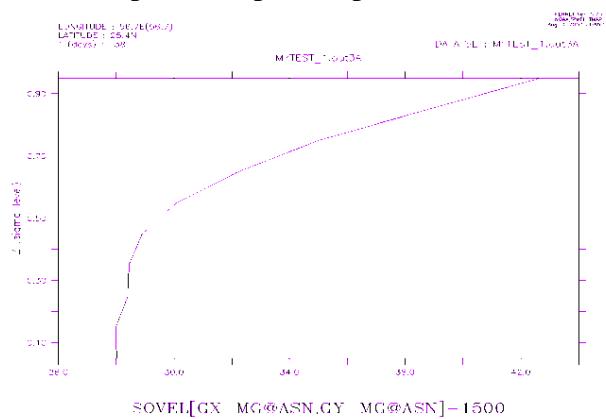


Fig.3: Sound speed profile of the Persian Gulf [12]

The relation of sound speed variations versus salinity and temperature in Medwin's formula is given in Eq. (1) [13]. According to the salinity and measured temperature from practical experiments for the profile of Fig. 3, the

obtained results (temperature=33.56°C, salinity=38.37 ppt) from Eq. (1) are in the range of 1500 to 1501 (m/s) actually.

$$C = 1449.2 + 4.6T + 0.055T^2 + 0.00029T^3 \\ +(1.34 - 0.01T)(S - 35) \\ (1)$$

Eq. (2) is the modified version of Medwin formula that we have presented. In this equation, dependence of sound speed to the salinity and temperature is adapted with the Medwin formula. However, in this case, the sound variations versus depth is approximated by a 10th order polynomial [8].

$$C = 1449.2 + 4.6T + 0.055T^2 + 0.00029T^3 \\ +(1.34 - 0.01T)(S - 35) - 8.28 \times 10^{-6} D^{10} \\ - 0.0024D^9 + 0.31D^8 - 0.24D^7 + 1.2 \times 10^{-3} \\ D^6 - 4.2 \times 10^{-4} D^5 + 10^{-6} D^4 - 1.6 \times 10^{-7} D^3 \\ + 1.1 \times 10^{-9} D + 3.1 \times 10^{-9} \\ (2)$$

Fig. 4 illustrates the output of the Medwin formula, the output of the obtained modified formula, and the experimental data. In this figure, the approximation error (3%) for each point is well acceptable.

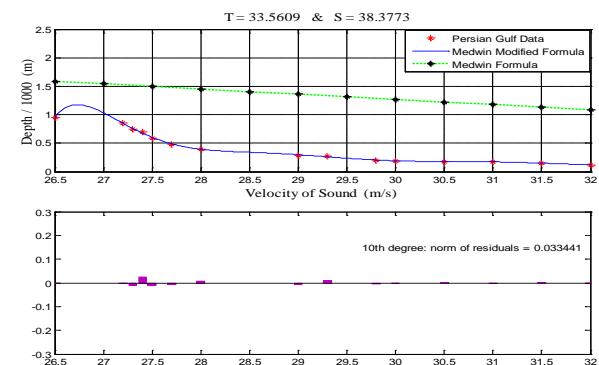


Fig.4: Obtained profile from Medwin's formula, Medwin's modified formula and measured data. The difference between real data and the presented model

3- Channel modeling

In channel modeling, the attenuations due to the frequency absorption, ambient noises and loss due to the wave scatterings at the surface and bottom for deferent grazing angles and bottom types are considered. Also, Ray theory is the basis of the mathematical model of multipath effects.

3-1- loss modeling

The acoustic energy of a sound wave propagating in the ocean is partly:

- Absorbed, i.e. the energy is transformed into heat.
- lost due to sound scattering by inhomogeneities.

On the basis of extensive laboratory and field experiments [13], the attenuation due to the absorption effects of Boric acid, $B(OH)_3$, Magnesium sulphate, $MgSO_4$ and pure water, H_2O is considered. The total loss is the sum of the each material loss. Experimental measurements and obtained profile for each material and the total loss are shown in Fig 5.

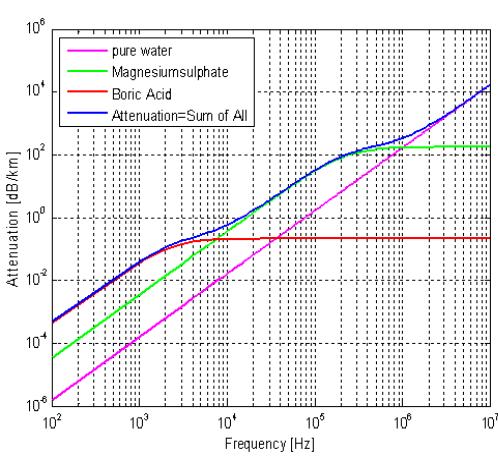


Fig.5: Amount of measured loss in different frequencies

From Fig. 5, it can be observed that for the *Boric acid region*, Attenuation is proportional to f^2 . And for the regions

Magnesium sulphate and pure water also Attenuation is proportional to f^2 . In the transition domains it is proportional to f . Attenuation increases with increasing salinity and temperature, Fig. 6. Attenuation increases with increasing frequency.

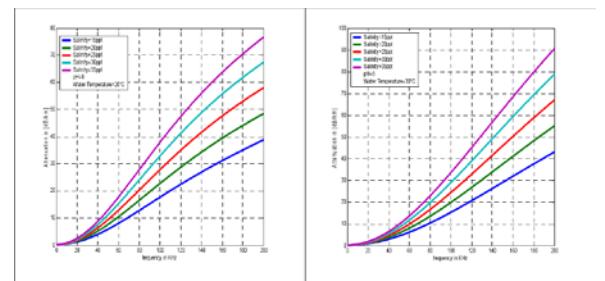


Fig. 6: Attenuation plot for various salinities & for temperature a) 20°C b) 30°C

3-2- Noise modeling

The model considered for noise is the combination of the thermal noise, shipping noise sea state noise and turbulences [14]. Eq. (3) depicts the general relation of the ambient noise:

$$NL = 10 \log_{10} \left(10^{0.1NL_{traffic}} + 10^{0.1NL_{turbulence}} + 10^{0.1NL_{see-state}} + 10^{0.1NL_{thermal}} \right) \quad (3)$$

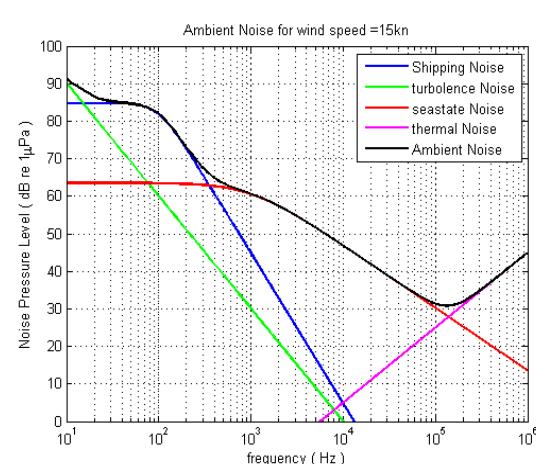


Fig. 7: Ambient noise Level for deferent frequency domains [4]

3-3- Scattering modeling in the surface and bottom reflections

To calculate loss due to the wave scattering in the surface, we use the probably density function of Gaussian Normal distribution for the surface displacement variable.

In the simulation, average of the reflection coefficient is calculated from Eq. (4) [15]:

$$R_{Gauss} = \text{Re}^{-2(kh)^2 \cos^2 \phi} \quad (4)$$

Where k denotes the wave number, h is the effective value of the surface wave height, ϕ is the angle of the collision to the normal surface, R is the pressure reflection for the normal surface. We consider R=-1 and h is obtained from the spectral density of the water surface displacement. The most famous spectrum in this case is Neumann-Pierson spectrum. For the calculation of wave bottom reflection coefficient, we use the Jackson pattern to select bottom water type which is simulated based on the Strait of Hormoz conditions and Hamilton-Bachman model [15],[16].

From Fig. 8, it can be observed that with an increase of grazing angle the scattering

loss also increases. In the same way with the increase of wind speed, there is an increase

in scattering loss. Similarly we can also observe the dependence of Bottom reflection coefficient, on grazing angle φ_m and bottom type bt . This is illustrated in Fig. 9.

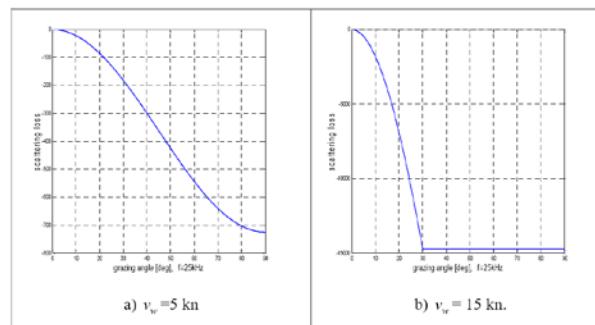


Fig. 8: Diagram illustrating dependence of surface reflection coefficient on grazing angle, frequency and two wind speeds

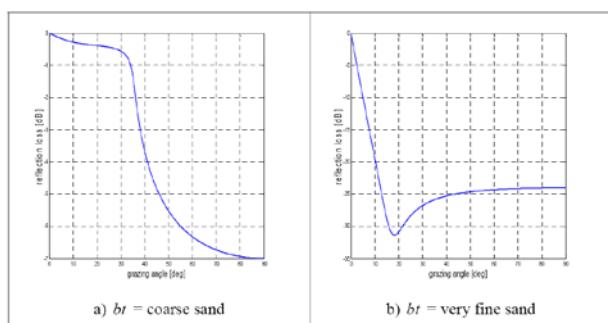


Fig. 9: Diagram illustrating dependence of bottom reflection coefficient on grazing angle and two bottom types

3-4- Modeling of Multipath

In the performed simulation, the number of the channel paths is varied and multiple of four. In the considered model, for the wave propagation from transmitter to receiver, we use four rays (Eigen rays). The transmitted wave is either one of the four Eigen rays or the multiple of them.

In the second case, after several reflections, the wave is reached to the receiver in one of the four cases shown in Fig.8. In the image method, according to Fig.9, surface and bottom are considered as the two mirrors. In the cylindrical coordinates for the channel with depth of D, surface is at Z=0 and bottom is at Z=D [12]. Assume that transmitter is at (0, Z_s) and receiver is at (0, Z). Therefore, the first image of transmitter which is due to the mirror

effect of the surface, is located in $(0, -z_s)$. Then transmitter and this image in relation to the bottom are located at $(0, 2D-z_s)$ and $(0, 2D+z_s)$ and making the second and third images, respectively. In general case, number of images or the sources of virtual transmitter equal to infinity and in each of the image repeating, four new images are generated that each of them are related to one of the eigen rays. According to this theory, the field of the pressure sound is depicted with Eq. (5) [15],[16].

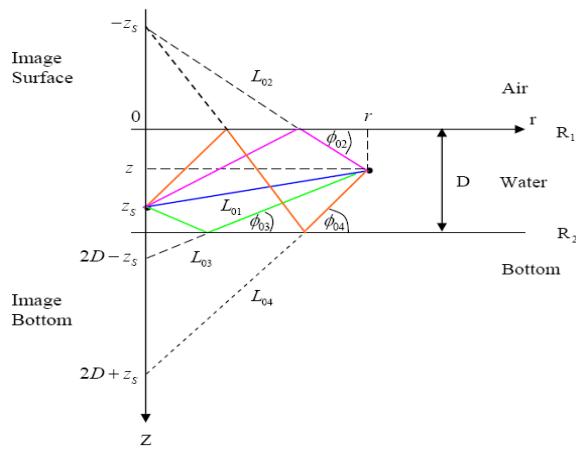


Fig.10: Schematic of transmitter and first three images in the image theory for a Homogenous channel

$$P(r, z, \omega) = A(\omega) \sum_{m=0}^{\infty} \left\{ \begin{array}{l} \hat{R}_1^m(\varphi_{m_1}, \omega) \hat{R}_2^m(\varphi_{m_1}, \omega) \frac{e^{-jkL_{m_1}}}{L_{m_1}} \\ + \hat{R}_1^{m+1}(\varphi_{m_2}, \omega) \hat{R}_2^m(\varphi_{m_2}, \omega) \frac{e^{-jkL_{m_2}}}{L_{m_2}} \\ + \hat{R}_1^m(\varphi_{m_3}, \omega) \hat{R}_2^{m+1}(\varphi_{m_3}, \omega) \frac{e^{-jkL_{m_3}}}{L_{m_3}} \\ + \hat{R}_1^{m+1}(\varphi_{m_4}, \omega) \hat{R}_2^{m+1}(\varphi_{m_4}, \omega) \frac{e^{-jkL_{m_4}}}{L_{m_4}} \end{array} \right\} \quad (5)$$

In this equation, A is the amplitude of the sound wave, R_1, R_2 are the reflection

coefficients of the surface and bottom respectively. $\phi_{m1}, \dots, \phi_{m4}$ are the reflection angles of the four eigen rays, K is the wave number and $L_{m1}, L_{m2}, L_{m3}, L_{m4}$ are the length of the displacement vectors of eigen rays RSRBR, RBR, RSR, DP in the $(m+1)^{th}$ stage of the production cycle of virtual resources, respectively. By considering the location of generated image in the m^{th} stage, the displacement vectors length of propagation paths are in accordance with Eq. (6) [15],[16].

$$\begin{aligned} L_{m_1} &= \sqrt{r^2 + (2Dm - z_s + z)^2} \\ L_{m_2} &= \sqrt{r^2 + (2Dm + z_s + z)^2} \\ L_{m_3} &= \sqrt{r^2 + (2D(m+1) - z_s - z)^2} \\ L_{m_4} &= \sqrt{r^2 + (2D(m+1) + z_s - z)^2} \end{aligned} \quad (6)$$

In the performed simulation, each of the reflection coefficients of the surface or bottom is calculated based on the introduced pattern of section. For Persian Gulf channel with considering $m=1$, i.e. eight paths, we concluded that from sixth path, due to the strong attenuation of transmitted wave, there is no signal reception. Hence, five-path channel pattern is suitable for the Persian Gulf.

4- Simulation

In this section, the obtained results for the channel, which its speed profile was shown in Fig.4, in conditions that channel depth was 5m and range of 1Km is presented. In this case, the transmitter and receiver use QPSK modulation with bandwidth of 5KHz and carrier frequency of 27KHz. Also, the transmitter and receiver are at depth of 5m and 70m from the surface, respectively. In Fig.11, the power spectral density of the transmitted signal

in the channel for each of the special paths is shown. As expected, in the RSRBR path, the largest attenuation takes place and power level in this path in comparison with direct path has 23dB loss. Also, received waves in the output of the channel in each of the 8 paths are shown in Fig.12. It can be seen that for sixth path and other paths after it, signal is strongly attenuated.

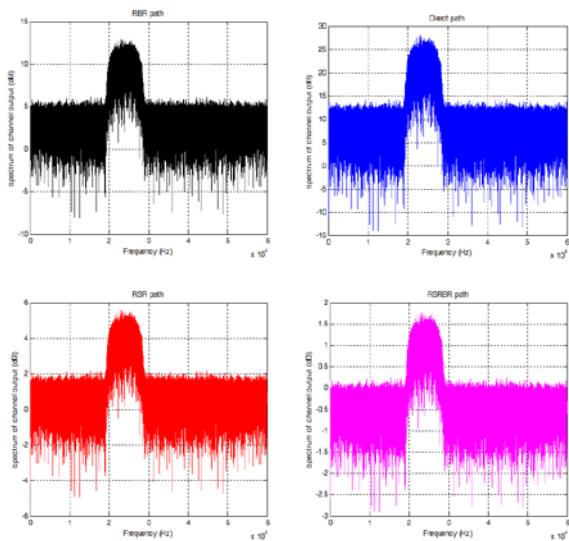


Fig.11: Power spectrum density of the four special paths used in the simulation

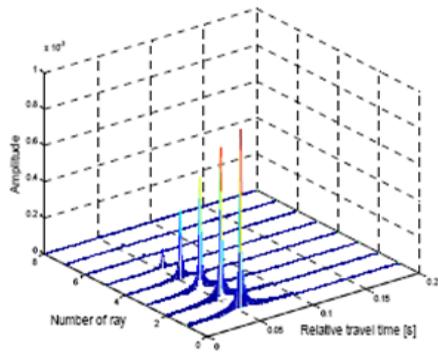


Fig.12: Received signals from different paths of channel. In this model, channel has 8

paths. First signal is for the direct path and the delays of other paths are calculated based on the traveling time of the first path.

Conclusion

Findings presented in this paper and summarized in the following provide new insight into a novel model for shallow water short-range multipath acoustic channel in the Persian Gulf. Our results, which are in good agreement with the results obtained from field measurement in the Strait of Hormuz, suggest the following.

Based on the simulations carried out in this paper, in the first place, the densest water in the Persian Gulf forms during winter in shallow waters along the coast of United Arab Emirates (Southern Shallows) and around Bahrain. This is associated with atmospheric cooling of extremely saline water masses in shallow water. Overall, the evaporative salinity increase throughout the Gulf leads to a steady component of dense water outflow through the Strait of Hormuz. In summer and autumn, the bottom outflow extends the entire length of the Gulf. Secondly we present Eq. (2) as a suitable empirical formula which describes the sound speed profile in the Strait of Hormoz. In addition, in accordance with the patterns introduced in sections 1 to 4, considering the five-path channel model is proper for the mentioned region. To further improve understanding of the circulation including seasonal variations in the Persian Gulf, more field observations are required to close data gaps that exist for autumn months for the entire Gulf and year-round for the Southern Shallows. Also required is a better knowledge of current river discharge rates of the Shatt-al-Arab.

Future theoretical studies should investigate effects of both varied river discharge and synoptic-scale wind and heat-flux forcing on the circulation in the Persian Gulf. The focus hereby should

be placed into investigation of 1) heat fluxes and dense water formation in the Southern Shallows and 2) atmospheric conditions that promote formation of a coastal jet along the Iranian coast in the northern Gulf, not adequately described in our model simulations, and how this interacts with the gulf-wide circulation.

References

- [1] R. J. Urick, Principles of Underwater Sound, 3d ed. New York: McGraw-Hill, 1983, p. 129.
- [2] R. Coates, Underwater Acoustic Systems. New York: Macmillan Education Ltd., 1990, pp. 26-28.
- [3] R. J. Urick, "Intensity summation of modes and images in shallow-water sound transmission," *J. Acoust. Soc. Amer.*, vol. 46, no. 3, pp. 780-788, Apr. 1999.
- [4] A. Zielinski, R. Coates, L. Wang, and A. Saleh, "High rate shallow water acoustic communication," *Oceans 93*, vol. 111, pp. 432437.
- [5] Emery, K. O., Sediments and water of the Persian Gulf, *AAPG Bull.*, 40, 2354–2383, 1956.
- [6] Seibold, E. and Ulrich, J.: Zur Bodengestalt des nordwestlichen Golfs von Oman. "Meteor" Forsch. Ergebnisses, Reihe C, 3, 1–14, 2005.
- [7] Reynolds, R. M.: Physical oceanography of the Gulf, Strait of Hormuz, and the Gulf of Oman – Results from the Mt Mitchell expedition, *Mar. Pollution Bull.*, 27, 35–59, 2007.
- [8] Johns, W. E., Yao, F., Olson, D. B., Josey, S. A., Grist, J. P., and Smeed, D. A.: Observations of seasonal exchange through the Straits of Hormuz and the inferred freshwater budgets of the Persian Gulf, *J. Geophys. Res.*, 108(C12), 3391, doi:10.1029/2003JC001881, 2007.
- [9] Swift, S. A. and Bower, A. S.: Formation and circulation of dense water in the Persian/Arabian Gulf, *J. Geophys. Res.*, 108(C1), 3004, doi:10.1029/2002JC001360, 2003.
- [10] Sadrinasab, M. and Kenarkooohi, J.: Three-dimensional flushing times in the Persian Gulf, *Geophys. Res. Letters*, 31, L24301, doi:10.1029/2004GL020425, 2008.
- [11] M. Sadrinasab, K. Kenarkooohi, "Numerical Modelling of sound velocity profile in different layers in the Persian Gulf", *Asian journal of applied sciences*, 232- 239, 2009, ISSN 1996-3343. Knowledgia Review, Malaysia 2009.
- [12] A. Doosti Aref: Survey underwater acoustic communication with design and simulation of an Underwater acoustic communication system in the Persian Gulf, MSc thesis, Malek-e Ashtar Industrial University, Tehran, Iran, 2010.
- [13] H. Medwin and C.S. Clay, "Fundamentals of Acoustical Oceanography", Academic Press, San Diego, pp. 220-228, 2000.
- [14] R.J. Urick, "Ambient Noise in the sea", enims publishing, P.O.Box 867, California 94023, third edition, pp. 155-205, 1984.
- [15] L.M. Brekhovskikh and Yu. P. Lysanov, "Fundamentals of Ocean Acoustics", second edition. Springer- Verlag . , pp. 130-155, 1999.
- [16] K. C. Hegewisch, N. R. Cerruti and S. Tomsovic, "Ocean acoustic wave propagation and ray method correspondence: internal wave fine structure," *The Journal of the Acoustical Society of America*, Vol. 114, Issue 4, p.2428, Oct.2007.

Classification of Electrocardiogram Signals With Extreme Learning Machine and Relevance Vector Machine

S. Karpagachelvi,

Doctoral Research Scholar, Mother Teresa Women's University,
Kodaikanal, Tamilnadu, India.

Dr. M. Arthanari,

Director, Bharathidasan School of Computer Applications,
Erode- 638 116, Tamilnadu, India.

M.Sivakumar,

Doctoral Research Scholar, Anna University – Coimbatore,
Tamilnadu, India

Abstract—The ECG is one of the most effective diagnostic tools to detect cardiac diseases. It is a method to measure and record different electrical potentials of the heart. The electrical potential generated by electrical activity in cardiac tissue is measured on the surface of the human body. Current flow, in the form of ions, signals contraction of cardiac muscle fibers leading to the heart's pumping action. This ECG can be classified as normal and abnormal signals. In this paper, a thorough experimental study was conducted to show the superiority of the generalization capability of the Relevance Vector Machine (RVM) compared with Extreme Learning Machine (ELM) approach in the automatic classification of ECG beats. The generalization performance of the ELM classifier has not achieved the nearest maximum accuracy of ECG signal classification. To achieve the maximum accuracy the RVM classifier design by searching for the best value of the parameters that tune its discriminant function, and upstream by looking for the best subset of features that feed the classifier. The experiments were conducted on the ECG data from the Massachusetts Institute of Technology–Beth Israel Hospital (MIT– BIH) arrhythmia database to classify five kinds of abnormal waveforms and normal beats. In particular, the sensitivity of the RVM classifier is tested and that is compared with ELM. Both the approaches are compared by giving raw input data and preprocessed data. The obtained results clearly confirm the superiority of the RVM approach when compared to traditional classifiers.

Index Terms—Electrocardiogram (ECG) signals classification, feature detection, feature reduction, generalization capability, model selection issue, extreme learning machine (ELM), relevance vector machine (RVM).

I. INTRODUCTION

The recognition of the ECG beats is an extremely important task in the coronary intensive unit, where the classification of the ECG beats is essential tool for the diagnosis. ECG is a technique which captures transthoracic interpretation of the electrical activity of the heart over time and externally recorded by skin electrodes. It is a non persistent recording produced by an electrocardiographic device. ECG offers cardiologists with useful information about the rhythm and functioning of the heart. Therefore, its analysis represents an

efficient way to detect and treat different kinds of cardiac diseases. A typical structure of the ECG signal is shown in Figure 1. The ECG signal is usually divided into two phases: depolarization and repolarization phases. The depolarization phase corresponds to the P-wave and QRS-wave while repolarization phase corresponds to the T-wave and U-wave. The ECG is measured by placing ten electrodes on preferred spots on the human body surface. For regular ECG recordings, the deviations in electrical potentials in 12 different directions out of the ten electrodes are measured. These 12 different electrical observations of the activity in the heart are normally referred to as leads. Trained physicians are able to recognize certain patterns in a patient's ECG signal and use them as the basis for diagnosis. Researchers have tried as the inception of computers to develop techniques and algorithms for automated processing of ECG signals for various medical applications.

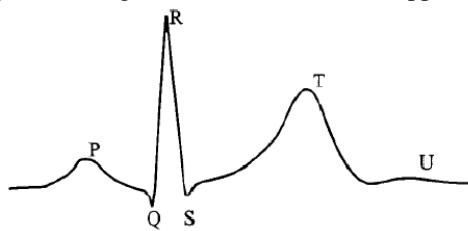


Figure 1: Structure of the ECG signal

Up to now; numerous algorithms have been introduced for the recognition and classification of ECG signal. Some of them use time and some use frequency domain for depiction. Based on that many specific attributes are defined, allowing the recognition between the beats belonging to different pathological classes. The ECG waveforms may be different for the same patient to such extent that they are unlike each other and at the same time alike for different types of beats [1]. Artificial neural network (ANN) and fuzzy-based techniques were also employed to exploit their natural ability in pattern recognition task for successful classification of ECG beats [2].

In this paper, the approach to ECG beat classification presented thorough experimental exploration of the RVM capabilities for ECG classification. Further the performances of the RVM approach in terms of classification accuracy are evaluated: 1) by automatically detecting the best discriminating features from the whole considered feature space and 2) by solving the model selection issue. Unlike traditional feature selection methods, where the user has to specify the number of desired features, the proposed system gives a method for extraction of features called as "feature detection". Feature selection and feature detection have the common characteristic of searching for the best discriminative features. The latter, however, has the advantage of determining their number automatically. In other words, feature detection does not require the desired number of most discriminative features from the user apriori. The detection process is implemented through AR Modeling framework that exploits a criterion intrinsically related to RVM classifier properties. This framework is formulated in such a way that it also solves the model selection issue, i.e., to estimate the best values of the RVM classifier parameters, which are the regularization and kernel parameters.

The rest of the paper is organized as follows. The AR method for ECG feature extraction, the basic mathematical formulation of ELMs for solving binary and multiclass classification problems and the working methodology of RVM is given in Section III. The experimental results obtained on ECG data from the Massachusetts Institute of Technology–Beth Israel Hospital (MIT–BIH) arrhythmia database [9] are reported in Sections IV. Finally, conclusions are drawn in Section V.

II. LITERATURE STUDY

In the literature survey, several methods have been proposed for the automatic classification of ECG signals. Among the most recently published works are those presented as follows

F. de Chazal et.al,[4] investigates the design of an efficient system for recognition of the premature ventricular contraction from the normal beats and other heart diseases. This system comprises three main modules: denoising module, feature extraction module and classifier module. In the denoising module of the system, it proposed the stationary wavelet transform for noise reduction of the electrocardiogram signals. In the feature extraction of the ECG module a proper combination of the morphological-based features and timing interval-based features are proposed. As the classifier, many supervised classifiers are investigated; they are: a number of multi-layer perceptron neural networks with different number of layers and training algorithms, support vector machines with different kernel types, radial basis function and probabilistic neural networks. Also, for comparison the proposed features, the author has considered the wavelet-based features. It has done complete simulations in order to achieve a high efficient system for ECG beat classification

from 12 files obtained from the MIT–BIH arrhythmia database. Simulation results show that best results are achieved about 97.14% for classification of ECG beats.

R. V. Andreao *et.al.*[5] proposed a novel embedded mobile ECG reasoning system that integrates ECG signal reasoning and RF identification together to monitor an elderly patient. As a result, the proposed method by andreao has a good accuracy in heart beat recognition, and enables continuous monitoring and identification of the elderly patient when alone. Furthermore, in order to examine and validate our proposed system, the author proposes a managerial research model to test whether it can be implemented in a medical organization. The results prove that the mobility, usability, and performance of author's proposed system have impacts on the user's attitude, and there is a significant positive relation between the user's attitude and the intent to use the proposed system.

L. Khadra *et.al.*[3] proposed a high order spectral analysis technique for quantitative analysis and classification of cardiac arrhythmias. The algorithm is based upon bispectral analysis techniques. The bispectrum is estimated with the use of an autoregressive model, and the frequency support of the bispectrum is extracted as a quantitative measure to classify atrial and ventricular tachyarrhythmias. Results illustrate a significant difference in the parameter values for different arrhythmias. Furthermore, the bicoherency spectrum shows different bicoherency values for normal and tachycardia patients. In particular, the bicoherency points out that phase coupling decreases as arrhythmia kicks in. The ease of the classification parameter and the attained specificity and sensitivity of the classification scheme reveal the importance of higher order spectral analysis in the classification of life threatening arrhythmias.

S. Mitra et.al, [6] puts forth a three stage technique for detection of premature ventricular contraction (PVC) from normal beats and other heart diseases. This method comprises a denoising module, a feature extraction module and a classification module. In the first module the author investigates the application of stationary wavelet transform (SWT) for noise reduction of the electrocardiogram (ECG) signals. The feature extraction module obtains 10 ECG morphological features and one timing interval feature. Then several multilayer perceptron (MLP) neural networks with different number of layers and nine training algorithms are designed. The performances of the networks for the speed of convergence and its accuracy classifications are evaluated for seven files from the MIT–BIH arrhythmia database. Among the various training algorithms, the resilient back-propagation (RP) algorithm illustrated the best convergence rate and the Levenberg–Marquardt (LM) algorithm achieved the best overall detection accuracy.

Chuan-Min Zhai et al, [20] presented a classification approach using ELM. In his paper, a machine learning algorithm referred to as the extreme learning machine (ELM) is used to classify plant species through plant leaf Gabor texture feature. A comparative study on system performance is

carried out among ELM and the main conventional neural network classifier - backpropagation neural networks. Results illustrate that the classification accuracy of ELM is higher than that of BP network. For given network architecture, ELM doesn't have any control parameters (i.e, stopping criteria, learning rate, learning epoches, etc.) to be manually tuned and can be implemented easily.

Sheng-Wu Xiong et.al.,[8] proposed in their paper that fuzzy support vector machines based on fuzzy c-means clustering. They employed the fuzzy c-means clustering technique to each class of the training set. At the time of clustering with a appropriate fuzziness parameter q , the more important samples, such as support vectors, become the cluster centers respectively.

Xin Zhou et al, [16] presented a novel approach on modulation classification using RVM. In his paper, a novel classification method based on relevance vector machine (RVM) is used in the MPSK signals classification. Compared with the SVM, RVM is sparse model in the Bayesian framework, not only the solution is highly sparse, but also it does not need to adjust model parameter and its kernel functions don't need to satisfy Mercer's condition. The fourth order cumulants of received signals are used as the classification vector firstly, and then multi-class classifier of RVM is designed. The authors first introduce the sparse Bayesian classification model, then transform the RVM learning to the maximization of marginal likelihood, and select the fast sequential sparse Bayesian learning algorithm. With the results of the experiment compared with SVM classifier proves the advantage of RVM.

Ke Wang et al, [15] discussed on the image classification technique RVM. In his paper, four building categories for database is prepared. Firstly the author uses the Gabor filter for image processing to extract the image features, and then divide the images to different subregions for histogram-based Gabor features. At last, for image classification, Support Vector Machine (SVM) and Relevance Vector Machine (RVM) are known to outperform classical supervised classification algorithms. SVM has excellent performance to solve binary classification problems. RVM could be more sparsity than SVM. A new method based on relevance vector machine- No-balance Binary Tree Relevance Vector Machine (NBBTRVM) is proposed to define a class in this classification task. NBBTRVM could do a good performance according to our experiment results.

III. METHODOLOGY

3.1. Feature extraction

Automatic ECG beat recognition and classification is performed in the part either by the neural network or by the other recognition systems relying in various features, time domain representation, extracted from the ECG beat [2], or the measure of energy in a band of frequencies in the spectrum (frequency domain representation) [10]. Since these features are very susceptible to variations of ECG morphology and the

temporal characteristics of ECG, it is difficult to distinguish one from the other on the basis of the time waveform or frequency representation. In this paper three different classes of feature set are used belonging to the isolated ECG beats including; third-order cumulant, auto-regressive model parameters and the variance of discrete wavelet transform detail coefficients for the different scales (1–6 scales).

3.1.1. Wavelet transformation

Physiological signals used for diagnosis are frequently characterized by a non-stationary time behavior. For such patterns, time and frequency representations are desirable. The frequency characteristics in addition to the temporal behavior can be described with respect to uncertainty principle. The wavelet transform can represent signals in different resolutions by dilating and compressing its basis functions. While the dilated functions adapt to slow wave activity, the compressed functions captures fast activity and sharp spikes. The most favorable choice of types of wavelet functions for pre-processing is problem dependent. In this paper Daubechies wavelet function (db5) which is called compactly supported orthonormal wavelets [11]. By making discretization the scaling factor and position factor the DWT is obtained. For orthonormal wavelet transform, $x(n)$ the discrete signal can be expanded in to the scaling function at j level, as follows:

$$x(n) = D_{j,k}[x(n)] + A_{j,k}[x(n)], n \in Z \quad (1)$$

where $D_{j,k}$ represents the detailed signal at j level. Note that j controls the dilation or contraction of the scale function $\Phi(t)$ and k denotes the position of the wavelet function $\Psi(t)$, and n represents the sample number of the $x(n)$. Here $n \in Z$ represents the set of integers. The frequency spectrum of the signal is classified into high frequency and low frequency for wavelet decomposition as the band increases ($j = 1, \dots, 6$). Wavelet transform is a two-dimensional timescale processing method for non-stationary signals with adequate scale values and shifting in time [12].

Multi resolution decomposition can efficiently provide simultaneous characteristics, in term of the representation of the signal at multiple resolutions corresponding to different time scales. Feature vectors are constructed by the normalized variances of detail coefficients of the DWT which belongs to the related scales.

3.1.2. Higher-order statistics and AR modeling

The main problem in automatic ECG beat recognition and classification is that related features are very susceptible to variations of ECG morphology and temporal characteristics of ECG. In the study [1] the set of original QRS complexes typical for six types of arrhythmia taken from the MIT/BIH arrhythmia database, there is a great variations of signal among the same type of beats belonging to the same type of arrhythmia. Therefore, in order to solve such problem, the author will rely on the statistical features of the ECG beats. In this paper for this aim, third-order cumulant has been taken

into account, which can be determined (for zero mean signals) as follows

$$C_{2x}(k) = E\{x(n)x(n+k)\} \quad (2)$$

$$C_{3x}(k, l) = E\{x(n)x(n+k)x(n+l)\} \quad (3)$$

$$\begin{aligned} C_{4x}(k, l, m) &= E\{x(n)x(n+k)x(n+l)x(n+m)\} \quad (4) \\ &\quad - C_{2x}(k)C_{2x}(m-l) \\ &\quad - C_{2x}(l)C_{2x}(m-k) \\ &\quad - C_{2x}(m)C_{2x}(l-k) \end{aligned}$$

Where E represents the expectation operator, and k, l, and m are the time lags. In this paper, third-order cumulant of selected ECG beats is used. Normalized ten points represents the cumulant evenly distributed with in the range of 25 lags. Each succeeding samples of a signal as a linear combination of previous samples, that is, as the output of an all-pole IIR filter is modeled by linear prediction. This process locates the coefficients of an n^{th} order auto-regressive linear process that models the time series x as

$$\begin{aligned} x(k) &= -a(2)x(k-1) - a(3)x(k-2) - \dots \quad (5) \\ &\quad - a(n+1)x(k-n-1) \end{aligned}$$

where x represents the real input time series (a vector), and n is the order of the denominator polynomial a(z). In the block processing, autocorrelation method is one of the modeling methods of all-pole modeling to find the linear prediction coefficients. This method is as well called as the maximum entropy method (MEM) of spectral analysis.

3.2. Extreme Learning Machine

A new learning algorithm called the Extreme Learning Machine for Single-hidden Layer Feed forward neural Networks (SLFNs) supervised batch learning. The output of an SLFN with $\sim N$ hidden nodes (additive or RBF nodes) can be represented by

$$f_{\tilde{N}}(X) = \sum_{i=1}^{\tilde{N}} \beta_i G(a_i, b_i, X), \quad X \in \mathbb{R}^n, \quad a_i \in \mathbb{R}^n, \quad (6)$$

where a_i and b_i are the learning parameters of hidden nodes and β_i is the weight connecting the i^{th} hidden node to the output node. $G(a_i, b_i, X)$ is the output of the i^{th} hidden node with respect to the input x. For the additive hidden node with the activation function $g(x): \mathbb{R} \rightarrow \mathbb{R}$ (e.g., sigmoid or threshold), $G(a_i, b_i, X)$ is given by

$$G(a_i, b_i, X) = g(a_i \cdot X + b_i), \quad b_i \in \mathbb{R} \quad (7)$$

Where a_i represents the weight vector connecting the input layer to the i^{th} hidden node and b_i is the bias of the i^{th} hidden node. $a_i \cdot x$ denotes the inner product of vectors a_i and x in \mathbb{R}^n . For an RBF hidden node with an activation function $g(x): \mathbb{R} \rightarrow \mathbb{R}$ (e.g., Gaussian), $G(a_i, b_i, X)$ is given by

$$G(a_i, b_i, X) = g(b_i ||x - a_i||), \quad b_i \in \mathbb{R}^+ \quad (8)$$

Where a_i and b_i are the i^{th} RBF node's center and impact factor. R^+ indicates the set of all positive real values. The RBF network is a special case of the SLFN with RBF nodes in its hidden layer. Each RBF node has its own centroid and impact factor and output of it is given by a radially symmetric function of the distance between the input and the center.

In the learning algorithms it uses a finite number of input-output samples for training. Here, N arbitrary distinct samples are considered $(x_i, t_i) \in \mathbb{R}^n \times \mathbb{R}^m$, where x_i is an $n \times 1$ input vector and t_i is an $m \times 1$ target vector. If an SLFN with \tilde{N} hidden nodes can approximate N samples with zero error, it then implies that there exist β_i , a_i , and b_i such that

$$f_{\tilde{N}}(X_j) = \sum_{i=1}^{\tilde{N}} \beta_i G(a_i, b_i, X_j) = t_j, \quad j = 1, \dots, N \quad (9)$$

Equation (9) can be written compactly as

$$H\beta = T \quad (10)$$

Where

$$H(a_1, \dots, a_{\tilde{N}}, b_1, \dots, b_{\tilde{N}}, X_1, \dots, X_{\tilde{N}}) = \quad (11)$$

$$\begin{bmatrix} G(a_1, b_1, X_1) & \dots & G(a_{\tilde{N}}, b_{\tilde{N}}, X_1) \\ \vdots & \ddots & \vdots \\ G(a_1, b_1, X_N) & \dots & G(a_{\tilde{N}}, b_{\tilde{N}}, X_N) \end{bmatrix}_{N \times \tilde{N}}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_{\tilde{N}}^T \end{bmatrix}_{\tilde{N} \times m} \quad \text{and} \quad T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{N \times m} \quad (12)$$

H is called the hidden layer output matrix of the network [15]; the i^{th} column of H is the i^{th} hidden node's output vector with respect to inputs x_1, x_2, \dots, x_N and the j^{th} row of H is the output vector of the hidden layer with respect to input x_j .

In real applications, the number of hidden nodes, \tilde{N} , will always be less than the number of training samples, N, and, hence, the training error cannot be made exactly zero but can approach a nonzero training error. The hidden node parameters a_i and b_i (input weights and biases or centers and impact factors) of SLFNs need not be tuned during training and may simply be assigned with random values according to any continuous sampling distribution. Equation (12) then becomes a linear system and the output weights are estimated as

$$\tilde{\beta} = H \dagger T \quad (13)$$

Where $H \dagger$ the Moore-Penrose is generalized inverse [15] of the hidden layer output matrix H. The ELM algorithm which consists of only three steps, can then be summarized as

ELM Algorithm: Given a training set

$\mathcal{X} = \{(X_i, t_i) | X_i \in \mathbb{R}^n, t_i \in \mathbb{R}^m, i = 1, \dots, N\}$ activation function $g(x)$, and hidden node number \tilde{N} ,

- 1) Assign random hidden nodes by randomly generating parameters (a_i, b_i) according to any continuous sampling distribution, $i=1, \dots, \tilde{N}$
- 2) Calculate the hidden layer output matrix H .
- 3) Calculate the output weight β : $\tilde{\beta} = H \dagger T$

The universal approximation capability of ELM has been analyzed by Huang et al. [7] using an incremental method and it shows that single SLFNs with randomly generated additive or RBF nodes with a wide range of activation functions can universally approximate any continuous target functions in any compact subset of the Euclidean space R^n . $g(x) = \frac{1}{1+e^{-\lambda x}}$ is the sigmoidal function used as activation function in ELM.

3.3. Relevance Vector Machine

The relevance vector machine (RVM) classifier [14], is a probabilistic extension of the linear regression model, which provides sparse solutions. It is analogous to the SVM, since it computes the decision function using only few of the training examples, which are now called relevance vectors. However training is based on different objectives.

The RVM model $y(x; w)$ is output of a linear model with parameters $w = (w_1, \dots, w_N)^T$, with application of a sigmoid function for the case of classification:

$$y_{RVM}(x) = \sigma\left(\sum_{n=1}^N \omega_n K(x, x_n)\right) \quad (14)$$

where $\sigma(x) = 1/(1+\exp(x))$. In the RVM, sparseness is achieved by assuming a suitable prior distribution on the weights, specifically a zero-mean, Gaussian distribution with distinct inverse variance α_n for each weight ω_n :

$$p(\omega|\alpha) = \prod_{n=1}^N N(\omega_n|0, \alpha_n^{-1}) \quad (15)$$

The variance hyperparameters $\alpha = (\alpha_1, \dots, \alpha_N)$ are assumed to be Gamma distributed random variables:

$$p(\alpha) = \prod_{n=1}^N \text{Gamma}(\alpha_n|a, b) \quad (16)$$

The parameters a and b are implicitly fixed and usually they are set to zero ($a = b = 0$), which provides sparse solutions.

Given a training set $\{x_n, t_n\}_{n=1}^N$ with $t_n \in \{0,1\}$ training in RVM is equivalent to compute the posterior distribution $p(\omega, \alpha|t)$. However, since this computation is intractable, a quadratic approximation $\log p(\omega, \alpha|t) \approx (\omega - \mu)^T \Sigma^{-1} (\omega - \mu)$ is assumed and computed matrix Σ and vector μ as:

$$\Sigma = (\Phi^T B \Phi + A)^{-1} \quad (17)$$

$$\mu = \Sigma \Phi^T B \hat{t} \quad (18)$$

with the $N \times N$ matrix Φ described as $[\Phi]ij = K(x_i, x_j)$, $A = \text{diag}(\alpha_1, \dots, \alpha_N)$, $B = \text{diag}(\beta_1, \dots, \beta_N)$, $\beta_n = y_{RVM}(x_n)[1-y_{RVM}(x_n)]$ and $\hat{t} = \Phi \mu + B \beta$. The parameters α are set to the values α_{MP} that maximize the logarithm of the following marginal likelihood

$$L(\alpha) = \log p(\alpha|t) = -\frac{1}{2} [N \log 2\pi + \log |\Sigma| + \hat{t}^T \Sigma^{-1} \hat{t}] \quad (19)$$

with $\Sigma = B^{-1} + \Phi A^{-1} \Phi^T$. This, gives the following update formula:

$$\alpha_n = \frac{1 - \alpha_n \Sigma_{nn}}{\mu_n^2} \quad (20)$$

The RVM learning algorithm iteratively evaluates formulas (15),(16) and (18). After training, the value of $y_{RVM}(x) = y(x; \mu)$ can be used to estimate the reliability of the classification decision for input x . Values close to 0.5 are near the decision boundary and consequently are unreliable classifications, while values near 0 and near 1 should correspond to reliable classifications. In this experiment, the reliability measure is used

$$RE_{RVM} = |2y_{RVM}(x) - 1| \quad (21)$$

which uses values near 0 for unreliable classifications and near 1 for reliable classifications.

IV. EXPERIMENTAL RESULTS

4.1. Dataset Description

The experiment conducted on the basis of ECG data from the MIT-BIH arrhythmia database [9]. In particular, the considered beats refer to the following classes: normal sinus rhythm (N), atrial premature beat (A), ventricular premature beat (V), right bundle branch block (RB), left bundle branch block (LB), and paced beat (/). The beats were selected from the recordings of 20 patients, which correspond to the following files: 100, 102, 104, 105, 106, 107, 118, 119, 200, 201, 202, 203, 205, 208, 209, 212, 213, 214, 215, and 217. In order to feed the classification process, in this paper, the two following kinds of features are adopted: 1) ECG morphology features and 2) three ECG temporal features, i.e., the QRS complex duration, the RR interval (the time span between two consecutive R points representing the distance between the QRS peaks of the present and previous beats), and the RR interval averaged over the ten last beats [4]. In order to extract these features, first the QRS detection is performed and ECG wave boundary recognition tasks by means of the well-known ecgpuwave software available on [17]. Then, after extracting the three temporal features of interest, normalized to the same periodic length the duration of the segmented ECG cycles according to the procedure reported. To this purpose, the mean

beat period was chosen as the normalized periodic length, which was represented by 300 uniformly distributed samples. Consequently, the total number of morphology and temporal features equals 303 for each beat.

TABLE1
 NUMBERS OF TRAINING AND TEST BEATS USED IN THE EXPERIMENTS

| Class | N | A | V | RB | / | LB | Total |
|-----------------------|-------|-----|------|------|------|------|-------|
| Training beats | 150 | 100 | 100 | 50 | 50 | 50 | 500 |
| Test beats | 24000 | 245 | 3789 | 3893 | 6689 | 1800 | 40416 |

In order to obtain reliable assessments of the classification accuracy of the investigated classifiers, in all the following experiments, three different trials are performed, each with a new set of randomly selected training beats, while the test set was kept unchanged. The results of these three trials obtained on the test set were thus averaged. The detailed numbers of training and test beats are reported for each class in Table 1. Classification performance was evaluated in terms of four measures, which are: 1) the overall accuracy (OA), which is the percentage of correctly classified beats among all the beats considered (independently of the classes they belong to); 2) the accuracy of each class that is the percentage of correctly classified beats among the beats of the considered class; 3) the average accuracy (AA), which is the average over the classification accuracies obtained for the different classes; 4) the McNemar's test that gives the statistical significance of differences between the accuracies achieved by the different classification approaches. This test is based on the standardized normal test statistic [16]

$$Z_{ij} = \frac{f_{ij} - f_{ji}}{\sqrt{f_{ij} - f_{ji}}} \quad (22)$$

where Z_{ij} measures the pair wise statistical significance of the difference between the accuracies of the i^{th} and j^{th} classifiers. f_{ij} stands for the number of beats classified correctly and wrongly by the i^{th} and j^{th} classifiers, respectively. Accordingly, f_{ij} and f_{ji} are the counts of classified beats on which the considered i^{th} and j^{th} classifiers disagree. At the commonly used 5% level of significance, the difference of accuracies between the i^{th} and j^{th} classifiers is said statistically significant if $|Z_{ij}| > 1.96$.

4.2. Experimental Scheme

The proposed experimental framework was performed around the following four main experiments. The first experiment aimed at assessing the effectiveness of the RVM approach in classifying ECG signals directly in the whole original hyper dimensional feature space (i.e., by means of all the 303 available features). The total number of training beats

was fixed to 500, as reported in Table 1. In the second experiment, it was desired to explore the behavior of the ELM classifier (compared to the two reference classifiers) when integrated within a standard classification scheme based on an AR feature reduction. In particular, the number of features was changed from 10 to 50 with a step of 10 so as to test this classifier in small as well as high-dimensional feature subspaces. The third experimental part had for objective to assess the capability of the proposed RVM classification system to boost further the accuracy of the ELM classifier. The fourth experiment was devoted to analyze the generalization capability of the ELM, with and without feature reduction, and of the RVM classification system by decreasing/increasing the number of available training beats. This analysis was done through two experimental scenarios, which consisted in passing from 500 to 250 and 750 training beats, respectively. Finally, the sensitivity of the RVM classification system is analyzed.

4.3. Experimental settings

In this experiment, the ELM classifier is first used to classify the signals. The desired number of features varied from 10 to 50 with a step of 10, namely, from small to high-dimensional feature subspaces. Feature reduction was achieved by the traditional AR modeling, commonly used in ECG signal classification. In particular, it can be seen that for all feature subspace dimensionalities except the lowest (i.e., 10 feature), the RVM classifier maintains a clear superiority over the ELM. Its best accuracy was established using a feature subspace made up of the first 30 components. The corresponding OA and AA accuracies were 95.67% and 95.33%, respectively. From this experiment, three observations can be made: 1) the ELM classifier shows a relatively low sensitivity to the curse of dimensionality as compared to RVM 2) the ELM classifier still preserve its superiority when integrated in a feature reduction-based classification scheme; and 3) RVM performs better results than ELM and provides very high result than ELM when the data are preprocessed with AR modeling technique.

TABLE2
 OVERALL (OA), AVERAGE (AA), AND CLASS PERCENTAGE ACCURACIES ACHIEVED ON THE TEST BEATS WITH THE DIFFERENT INVESTIGATED CLASSIFIERS WITH A TOTAL NUMBER OF 500 TRAINING BEATS

| Method | OA | AA | N | A | V | RB | / | LB |
|---|-------|-------|-------|-------|-------|-------|-------|-------|
| ELM | 88.76 | 88.48 | 88.44 | 87.39 | 83.48 | 95.98 | 84.47 | 88.76 |
| ELM classification after Preprocessing | 89.74 | 89.78 | 89.69 | 88.96 | 85.18 | 97.69 | 86.58 | 89.74 |
| RVM | 92.67 | 93.33 | 90.52 | 93.65 | 90.97 | 96.58 | 92.99 | 92.67 |
| RVM classification after Preprocessing | 95.67 | 95.33 | 92.52 | 94.65 | 93.97 | 98.28 | 94.19 | 97.16 |

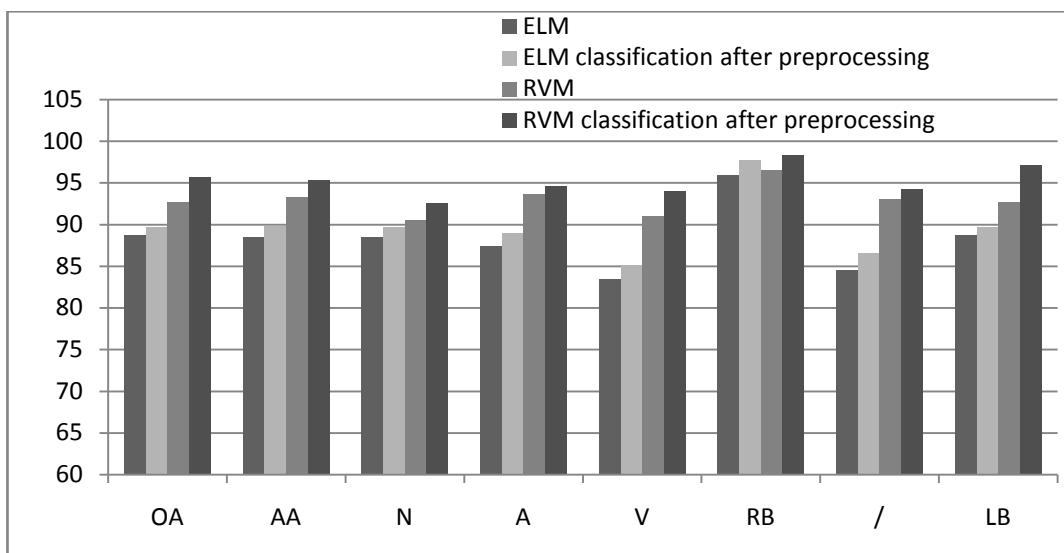


Figure2: Comparison of ELM and RVM accuracy for different datasets

The Figure 2 gives the comparison of the accuracy of classifying the ECG signals by using ELM and RVM. This shows that RVM gives much better accuracy for all datasets given as input. In which RB dataset achieves the maximum accuracy of 98.28%.

Table 3 shows the number of features detected automatically to discriminate each class from the others. The average number of features required by the RVM classifier is 47, while the minimum and maximum numbers of features

were obtained for the ventricular premature (V) and normal (N) classes with 32 and 68 features, respectively.

V. CONCLUSION

In this paper, a novel ECG beat classification system using RVM is proposed and applied to MIT/BIH data base. The wavelet transforms variance and AR model parameters have been used for the features selection. From the obtained experimental results, it can be strongly recommended that the

| Class | N | A | V | RB | / | LB | AVERAGE |
|--------------------|----|----|----|----|----|----|---------|
| #Detected Features | 68 | 49 | 32 | 50 | 47 | 41 | 47 |

use of the RVM approach for classifying ECG signals on account of their superior generalization capability as compared to traditional classification techniques. This capability generally provides them with higher classification accuracies and a lower sensitivity to the curse of dimensionality. The results confirm that the RVM classification system substantially boosts the generalization capability achievable with the ELM classifier, and its robustness against the problem of limited training beat availability, which may characterize pathologies of rare occurrence. Another advantage of the RVM approach can be found in its high sparseness, which is explained by the fact that the adopted optimization criterion is based on minimizing the number of SVs. It can also be seen that RVM accomplishes better and more balanced classification for individual categories as well in very less training time comparative to ELM. In future some advanced neural network techniques can be used to train the RVM classifier and it may enhance the classification accuracy of the ECG and reduce the training time.

REFERENCES

- [1] Osowski, S., Linh, T.H., 2001. ECG beat recognition using fuzzy hybrid neural network. *IEEE Trans. Biomed. Eng.* 48 (11), 1265–1271.
- [2] Hu, Y.H., Palreddy, S., Tompkins, W., 1997. A patient adaptable ECG beat classifier using a mixture of experts approach. *IEEE Trans. Biomed. Eng.* 44 (9), 891–900.
- [3] L. Khadra, A. S. Al-Fahoum, and S. Binajjaj, "A quantitative analysis approach for cardiac arrhythmia classification using higher order spectral techniques," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 11, pp. 1840–1845, Nov. 2005.
- [4] F. de Chazal and R. B. Reilly, "A patient adapting heart beat classifier using ECG morphology and heartbeat interval features," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 12, pp. 2535–2543, Dec. 2006.
- [5] R. V. Andreao, B. Dorizzi, and J. Boudy, "ECG signal analysis through hidden Markov models," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 8, pp. 1541–1549, Aug. 2006.
- [6] S. Mitra, M. Mitra, and B. B. Chaudhuri, "A rough set-based inference engine for ECG classification," *IEEE Trans. Instrum. Meas.*, vol. 55, no. 6, pp. 2198–2206, Dec. 2006.
- [7] C.-K. Siew and G.-B. Huang, "Extreme Learning Machine with Randomly Assigned RBF Kernels," *Int'l J. Information Technology*, vol. 11, no. 1, 2005.
- [8] Sheng-Wu Xiong, Hong-Bing Liu and Xiao-Xiao Niu, Fuzzy support vector machines based on FCM clustering. Proceedings of 2005 International Conference on Machine Learning and Cybernetics, Aug. 2005, vol. 5, pp. 2608– 2613.
- [9] R. Mark and G. Moody MIT-BIH Arrhythmia Database 1997 [Online]. Available <http://ecg.mit.edu/dbinfo.html>.
- [10] Minami, K., Nakajima, H., Toyoshima, T., 1999. Real-time discrimination of ventricular tachyarrhythmia with Fouriertransform neural network. *IEEE Trans. Biomed. Eng.* 46, 179–185.
- [11] Daubechies, I., 1998. Orthonormal bases of compactly wavelets. *Commun. Pure Appl. Math.* (41), 909–996.
- [12] Thakor, N.V., 1993. Multiresolution wavelet analysis of evoked potentials. *IEEE Trans. Biomedical. Eng.* 40 (11), 1085– 1093.
- [13] C.-W.Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.
- [14] Tipping, M.E.: Sparse Bayesian learning and the Relevance Vector Machine. *Journal of Machine Learning Research* 1 (2001) 211–244
- [15] Ke Wang and Haitao Jia, "Image Classification Using No-balance Binary Tree Relevance Vector Machine", Intelligent Interaction and Affective Computing, 2009. ASIA '09, Page(s): 79 – 82
- [16] Xin Zhou, Ying Wu and Guopeng Yang, "Modulation Classification of MPSK Signals Based on Relevance Vector Machines", *Information Engineering and Computer Science*, 2009, Page(s): 1 – 5
- [17] <http://www.physionet.org/physiotools/ecgpuwave/src/>.
- [18] J. J. Wei, C. J. Chang, N. K. Shou, and G. J. Jan, "ECG data compression using truncated singular value decomposition," *IEEE Trans. Biomed. Eng.*, vol. 5, no. 4, pp. 290–299, Dec. 2001.
- [19] A. Agresti, *Categorical Data Analysis*, 2nd ed. New York:Wiley, 2002.
- [20] Chuan-Min Zhai and Ji-Xiang Du, "Applying extreme learning machine to plant species identification", *Information and Automation*, 2008. ICIA 2008, Page(s): 879 – 884

AUTHORS PROFILE

Karpagachelvi.S: She received the BSc degree in physics from Bharathiar University in 1993 and Masters in Computer Applications from Madras University in 1996. She has 13 years of teaching experience. She is currently a PhD student with the Department of Computer Science at Mother Teresa University. She has published four papers in international journals.

Dr.M.Arthanari: He has obtained Doctorate in Mathematics in Madras University in the year 1981. He has 35 years of teaching experience and 25 years of research experience. He has a Patent in Computer Science approved by Govt. of India.

Sivakumar M : He has 10+ years of experience in the software industry including Oracle Corporation. He received his Bachelor degree in Physics and Masters in Computer Applications from the Bharathiar University, India. He holds patent for the invention in embedded technology. He is technically certified by various professional bodies like ITIL, IBM Rational Clearcase Administrator, OCP – Oracle Certified Professional 10g and ISTQB.

Negotiation in Multi-Agent System using Partial-Order Schedule

Ritu Sindhu¹, Abdul Wahid², G.N.Purohit³

¹Ph.D Scholar, Banasthali University
Rajasthan, India

²Department of CS, Gautambudh University
Ghaziabad, India

³G.N. Purohit, Dean, Banasthali University
Rajasthan, India

Abstract

In systems composed of multiple autonomous agents, negotiation is a key form of interaction that enables groups of agents to arrive at a mutual agreement regarding some belief, goal or plan, for example. In general, multi-linked negotiation (including both the directly linked and the indirectly linked relationships) describes situations where one agent needs to negotiate with multiple agents about different issues, where the negotiation over one issue influences the negotiations over other issues. The characteristics of the commitment on one issue affect the evaluation of a commitment or the construction of a proposal for another issue. In this paper we built a partial order schedule representation with the help of which we effectively manage interacting negotiation issues. Also we explored how flexibility is an important factor for ordering and managing negotiation issues in a successful negotiation and enables an agent to reason explicitly about the interactions among multiple negotiation issues in order to achieve higher performance.

Keywords: Multi agent systems, Partial order schedule, ACL, Negotiation.

1. Introduction

For understanding negotiation clearly we should know that there are three broad topics for research on negotiation, that serve to organize the issues under consideration. First, negotiation protocols are the set of rules that govern the interaction. This covers, the permissible types of participants (e.g., the negotiators and relevant third parties), the negotiation states (e.g., accepting bids, negotiation closed), the events

that cause state transitions (e.g., no more bidders, bid accepted), and the valid actions of

the participants in particular states (e.g., which can be sent by whom, to whom and at when). Second, negotiation objects are the range of issues over which agreement must be reached. The next level, however, offers flexibility to change the values of the issues in the negotiation object, through counter-proposals, changing the structure of the negotiation object (by adding guarantees, for example), and so on. Finally, the agents' reasoning models provide the decision making apparatus by which participants attempt to achieve their objectives.

For a negotiation to be completed successfully all parties (i.e. agents) must clearly understand the rules of engagement or negotiation protocol. For example, in a simple contract-net protocol, in which a manager issues a call for proposals and waits for a full set of replies (or timeouts), each bidder must be prepared to honour its bid for the duration of the bid's validity. Otherwise, acceptance of the bid will require a second negotiation, which may itself succeed or fail. For example, KQML It is clear that the design of a communications language can restrict or enable different forms of high level reasoning among the agents involved in negotiation. For example, in KQML the sender must decide whether the recipient will respond directly, broker the query, recommends an agent to send the query to, or

recruit an agent who will send response directly to the original agent. The relationships among these negotiation issues can be classified as directly-linked relationship in which first issue affects second issue directly because later issue B is a necessary resource (or a subtask) of former issue, like (such as cost, duration and quality). Secondly indirect-linked relationship in which one issue relates to another issue because they compete for use of a common resource. and at last a facilitates relationship that is the combination of both relationship as described in figure in which the relationship from "Y11" to "Y12" means that the completion of "Y11" will positively affect the execution "y12" by reducing its cost, shortening its process time and/or improving its quality.

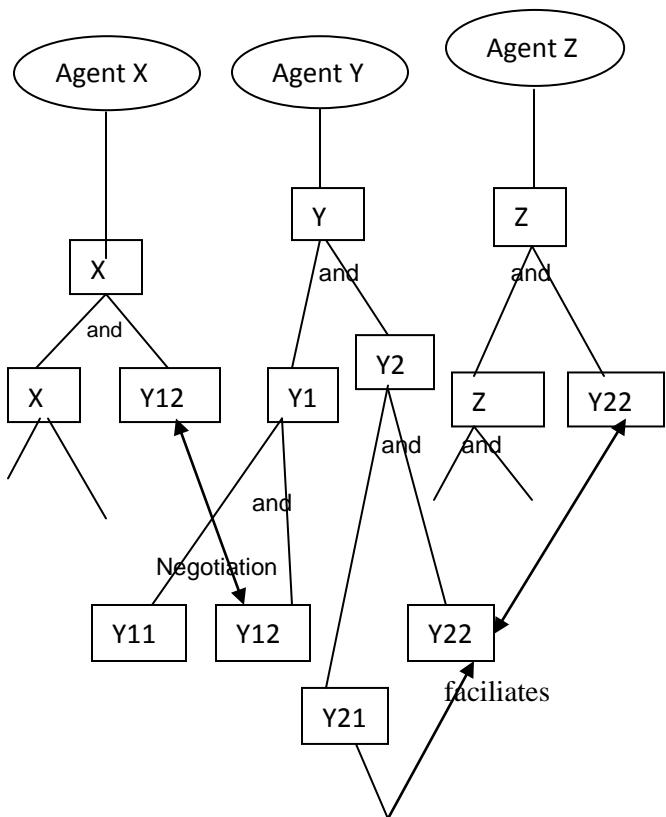


Fig .1 Negotiations between Agents with using facilitates Relationship

In general, multi-linked negotiation (describes situations where one agent needs to negotiate with multiple other agents about different subjects, where the negotiation over one issue has influence on the negotiation so over other issues.

How can the agent deal with these interrelated negotiations? One approach is to deal with these **negotiations independently ignoring their interactions**.¹ If these negotiations are performed concurrently, there could be possible conflicts among the solutions to these negotiations; hence the agent may not be able to find a combined feasible solution that satisfies all constraints without re-negotiation over some already "settled" issues. For example, in Figure 1, suppose the Car Producer Agent negotiates with the Consumer Agent and promises to finish Purchase Car by time 20, and concurrently the Car Producer Agent also negotiates with the Transporter Agent about task Deliver Car and gets a contract that task Deliver Car will be finished at time 30, then the Car Producer Agent will find it is impossible for task Par Computer be finished by time 20 given that its subtask Deliver Car will not be finished until time 30. If done effectively, it permits the agent to minimize the possibility of conflicts among the different negotiations and thus achieve better performance. The multi-linked negotiation problem also an important one because it actually happens in a number of application domains. For example, in a supply chain problem, negotiations go on among more than two agents. The consumer agent negotiates with the producer agent, and the producer agent needs to negotiate with the supplier agents. The negotiations between the producer agent and the supplier agents have a direct influence on the negotiation between the producer agent and the consumer agent.

In general, a Multi-linked negotiation problem occurs when an agent needs to negotiate with multiple other agents about different subjects, and the negotiation over one subject has influence on the negotiations over other subjects. The commitment of resources for one subject affects the evaluation of a commitment or the Construction of a proposal for another subject.

2. Negotiating agents' communication

An agent Communication Language (ACL) is a language with precisely defined syntax, semantics and pragmatics that is the basis of communication between independently designed and developed software agents [8]. Functional agents in a MAS use a common ACL to transfer information, share knowledge and negotiate with each other. Knowledge Query and Manipulation Language (KQML) and the ACL defined by Foundation for Intelligent Physical Agents/ Agent Communication Language (FIPA ACL) are the most widely used and studied ACLs. Each of them offers a minimal set of performatives to describe agent actions and allows users to extend them if the new defined ones conform to the rules of ACL syntax and semantics. In KQML there are no predefined performatives for agent negotiation actions. In FIPA ACL there are some performatives, such as proposal, CFP and so on, for general agent negotiation processes, but they are not sufficient for our purposes. For example, there are no performatives to handle third party negotiation. In this section we present a negotiation performative set designed for MAS dealing with supply chain management.

2.1 Criteria for performative definition and selection

The criteria we used to define negotiation performatives are the following:

1. Compatible with existing performatives. From a practical perspective, to extend either KQML or FIPA ACL performative sets involve a similar process. Since in FIPAACL there is a category containing four negotiation performatives we choose to construct the negotiation performative set based on this subset.
2. Defining new negotiation performatives based on a negotiation protocol. There is no clear means to judge the advantages and

disadvantages of a particular extension of a standard.

ACL, just as it is difficult to judge how to add new words and phrases to a language used by human beings.

2.2 Negotiating agents communication or Negotiating performance for pair-wise negotiation protocol

In a supply chain negotiation process, negotiating agents use an Agent Communication Language (ACL) [5] to bargain with each other. The table below presents the performatives designed for the negotiating agents based on FIPA ACL [4]. A negotiation protocol, formally described using Color Petri Net (CPN) is also given. Also Performatives for pair-wise negotiation protocol are used when two functional agents negotiate directly. The performatives definitions conform to the FIPA ACL specification. We explain their name and corresponding meaning as follows:

- *Accept Proposal*: The action of accepting a previously submitted proposal to perform an action
- *CFP*: The action of calling for proposals to perform a given action
- *Proposal*: The action of submitting a proposal to perform a certain action, given certain preconditions
- *Reject-proposal*: The action of rejecting a proposal to perform some acting during a negotiation
- *Terminate*: The action to finish the negotiation process

Initially, one agent starts negotiation by sending a CFP message to the other agent. After several rounds of conversation in which proposes and counter-propose are exchanged, the negotiation between two agents will end when one side accepts (rejects) the other side's proposal or terminates the negotiation process without any further explanation. It is not necessary that the functional agent responds to each message. A functional agent can simply ignore the incoming messages. It is the sender's responsibility to handle the lost message or in cases of lack of

responses. The conversation scenario is described in the following figure:

The pair-wise negotiation protocol simulates a conversation between two persons, in which one Side sends an “ask” and the other side send a “reply.” The difference is in the pair-wise negotiation protocol we have to limit the response message types after a functional agent receives incoming messages so that the negotiation process does not become irrelevant to the topic, and at the same time simplify the message handling process. The expected responses or performatives performed by Pairwise protocol when certain action comes:

- Accept Proposal: Terminate | NONE;
- CFP :Proposal | Terminate | NONE;
- Proposal: Accept-proposal | Reject-proposal | Terminate | NONE;
- Reject: Terminate | NONE;
- Terminate: NONE.

2.3 Negotiating performance for third party negotiation protocol Performatives

Negotiating performance for third party negotiation protocol Performatives for a third party negotiation protocol are used when functional agents negotiate through the third party (auctioneer). Some performatives defined for the pair-wise negotiation protocol, e.g. accept-proposal, reject-proposal, are still used for this protocol. One new performative, BID, is introduced. The syntax of BID is as follows:

□Bid: the action for a bidder to send a corresponding response to an auctioneer

Bid

: sender <word>
 : receiver <word>-----auctioneer
 : content <expression>-----price for a goods
 : Language <word>-----e.g. Knowledge Interchange Format (KIF)
 : ontology <word>-----system
 : in-reply-to <word>-----auction number
 : protocol <word>-----the default value is English auction.

Initially, one functional agent (seller) starts the negotiation by sending an INFORM message to

the auctioneer. This message includes the goods that it wants to sell and the highest desired price (Or the contract that it wants to be bought and the lowest desired price) and the preferred auction protocol. After receiving the message, the auctioneer will broadcast it to potential bidders (assuming the auctioneer knows that information by querying the information agent) and organize an auction according to the requirement the seller submits. After several rounds of conversation, the negotiation process will end with a deal that was reached between seller agent and bidders. It is the auctioneer’s responsibility to notify both the seller and bidders of the winner and the losers.

The scenario is described in the following figure:

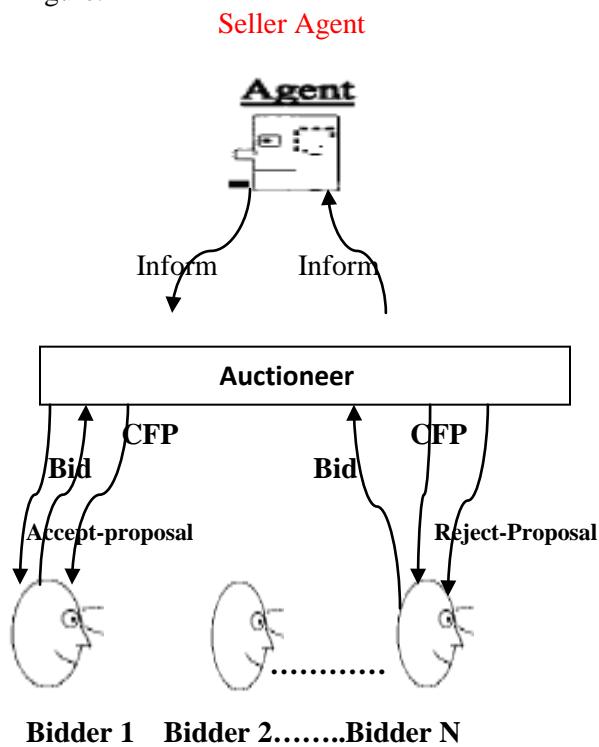


Fig2. Scenario between Agent and Third Party

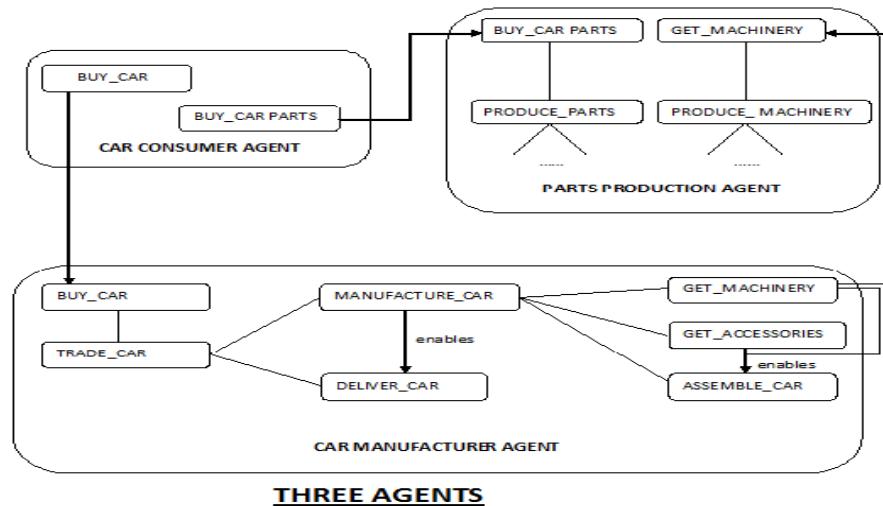
3. The Agent Scenario using SCM

Consider the following example shown in Figure 1, which is a simplified supply chain containing five agents. The consumer (dealer) agent represents the environment that generates tasks

to be completed by the other four agents. The manufacturing agent, the production agent, the purchase agent and finally the inventory agent. When a new task is generated by the consumer Agent, it indicates how much it is worth and its deadline.

When the Car Producer Agent (manufacturing agent) receives a task Purchase Car from the Consumer(Dealer) Agent, it also needs to subcontract parts of the task Get Hardware and

Deliver car to the Hardware Producer Agent (production agent+ purchase agent) and the Transporter Agent(inventory agent)respectively. The following three negotiations are interrelated: the negotiation .



between the Car Producer Agent and the Consumer Agent(dealer) on task Purchase car, the negotiation between the Car Producer Agent and the Hardware Producer Agent(production agent+ purchase agent) on task Get Hardware, and the negotiation between Car Producer Agent and the Transporter Agent(inventory agent) on task Deliver Car.

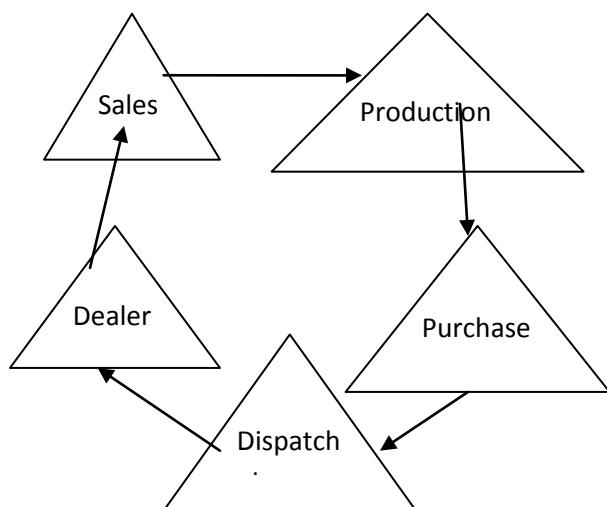


Fig3.Supply chain example

There are three agents in Figure 3:

1. **Car Consumer Agent** generates three types of new tasks: Purchase Car task for Car Production Agent, Get Machinery task for Parts

Production Agent, and Deliver car for Transporter Agent.

2. **Car Manufacturer Agent** receives the Purchase Car task from Car Consumer Agent, and needs to decide if it should accept this task and, if it does, what the promised finish time of the task should be. Figure shows the local plan for producing computers; it includes a non-local task Get Machinery that requires negotiation with Parts Production Agent. It also includes a non-local task Deliver Car that requires negotiation with Transporter Agent.

3. **Parts Production Agent** receives two types of tasks: Get Machinery from Car Manufacturer Agent and Buy Car Parts from Car Consumer Agent. It needs to decide whether to accept a new task and what is its promised finish time.

4. **Other Agents**

There are two other agents also involved in the process:

a) **Transporter Agent**

Its task is to deliver car from Car Manufacturer Agent to Car Consumer Agent.

b) **Trader Agent**

Its task is to establish a trade between Car Consumer Agent and Car Manufacturer Agent for buying a new car.

We first define two generalized terms to make the following description easier. In the following description, we will use the term contractor agent to refer to the agent who performs the task for another agent and gets rewarded for the successful completion of the task; and contractee agent to refer to the agent who has a task that needs to be performed by another agent and pays a reward to the other agent. The contractor agent and the contractee agent negotiate about a task and a contract is signed (a commitment is built and confirmed) if an agreement is reached during the negotiation. In this work, the negotiation process between agents is based on an extended contract net model.

- Contractee agent announces a task by sending out a proposal.
- Contractor agent receives this proposal, evaluates it, responds to it in one of three ways: by accepting it, by simply rejecting it, or by rejecting it but at the same time making a counter-proposal.
- Contractee agent evaluates the responses, it either chooses to confirm

an accepted proposal, or chooses to accept a counter-proposal.

- Contractee agent awards the task to the chosen contractor agent based on the commitment (the mutually accepted proposal/counter-proposal) which is confirmed by both agents; the negotiation process then ends successfully. If a mutually agreed proposal/counter-proposal cannot be found, the negotiation process fails.

This process can be extended to a multi-step process by introducing an extended series of alternative proposals and counter-proposals. However, in this paper, we only focus on the two-step (proposal, counter-proposal) negotiation process. A proposal which announces that a task (t) needs to be performed includes the following attributes:

1. **earliest start time (est):** the earliest start time of task t ; task t cannot be started before time est.
2. **Deadline (dl):** the latest finish time of the task; the task needs to be finished before the deadline dl.
3. **Minimum quality requirement (minq):** the task needs to be finished with a quality achievement no less than minq.
4. **Regular reward (r):** if the task is finished as the contract requested, the contractor agent will get reward r.
3. **Early finish reward rate (e):** if the contractor agent can finish the task by the time (ft) as it promised in the contract, it will get the extra early finish reward. $(e * r * (dl - ft), r)$.
4. **Decommitment penalty rate (p):** if the contractor agent cannot perform the

task as promised in the contract (i.e. the task could not be finished by the promised finish time), it pays a decommitment penalty ($p * r$) to the contractee agent. Similarly, if the contractee agent needs to cancel the contract after it has been confirmed, it also needs to pay a decommitment penalty ($p * r$) to the contractor agent.

Suppose Car Manufacturer Agent has received the following two tasks in the same scheduling time window.

task name : Purchase Car A

arrival time : 10

earliest start time : 15(arrival time +estimated negotiation time (7))

deadline : 50

reward : $r = 10$

decommitment penalty : $p = 0.8$

early finish reward rate: $e = 0.01$

task name: Purchase Car B

arrival time: 12

earliest start time: 17 (arrival time +estimated negotiation time (5))

deadline: 80

reward: $r = 10$

decommitment penalty rate: $p = 0.9$

early finish reward rate: $e = 0$

The agent's local scheduler reasons about these two new tasks according to the above information: their earliest start times, deadline, estimated process times and the rewards. It then generates the following agenda which includes the following tasks:

The Car Manufacturer Agent checks the local plans for these tasks as shown and finds there are five negotiations:

1. Negotiate with Car Consumer Agent about the promised finish time of Purchase Car_A.
2. Negotiate with Car Consumer Agent about the promised finish time of Purchase Car_B.

3. Negotiate with Parts Production Agent about whether it can accept the task Get_Machinery_A and if it accepts this task, what is the promised finish time.
4. Negotiate with Parts Production Agent about the task Get_Machinery_B, with the same concerns as above.
5. Negotiate with Transporter Agent about whether it can accept the task Deliver Car A, and if it accepts this task, what is the earliest start time and what the promised finish time is.

These five negotiations are all related. The potential relationships among multiple negotiation issues can be classified as two types.

4. Partial order schedule

A partial-order schedule is the basic reasoning tool that we use for multiple related negotiations. Here we present the formalization of the partial-order schedule and use an example to explain how it works for a multi-linked negotiation. Figure 5 shows the partial ordered schedule from the example in Figure 4.

A poset consists of a set together with a binary relation that indicates that, for certain pairs of elements in the set, one of the elements precedes the other. These relations are called partial orders to reflect the fact that not every pair of elements of a poset need be related: for some pairs, it may be that neither element precedes the other in the poset. Thus, partial orders generalize the more familiar total orders, in which every pair is related. A finite poset can be visualized through its Hasse diagram, which depicts the ordering relation between certain pairs of elements and allows one to reconstruct the whole partial order structure. A partial order is a binary relation " \leq " over a set P which is reflexive, antisymmetric, and transitive, i.e., for

all a, b, and c in P, we have that: In mathematics, especially order theory, a partially ordered set (or poset) formalizes the intuitive concept of an

ordering, sequencing, or arrangement of the elements of a set.

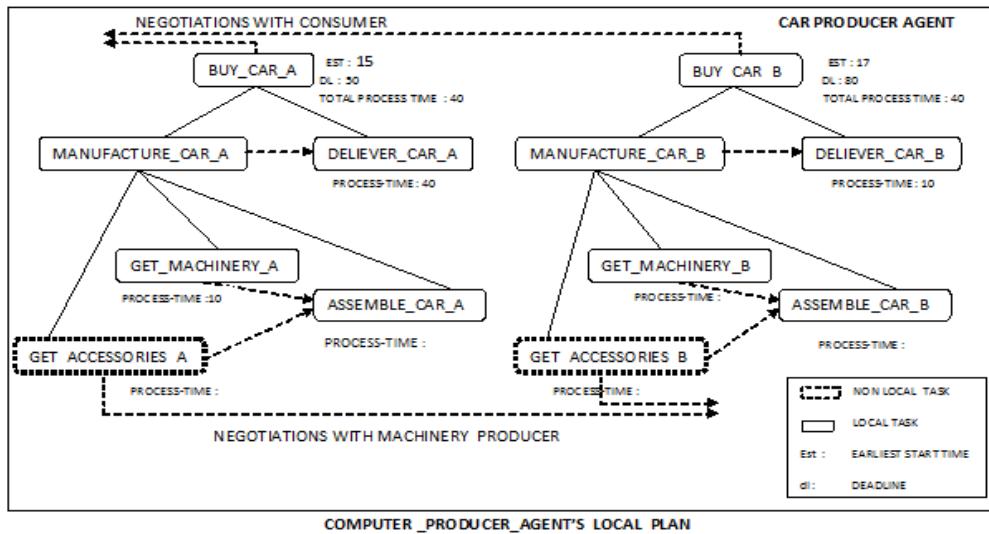


Fig: 4 Car Producer Agent's Local Plan

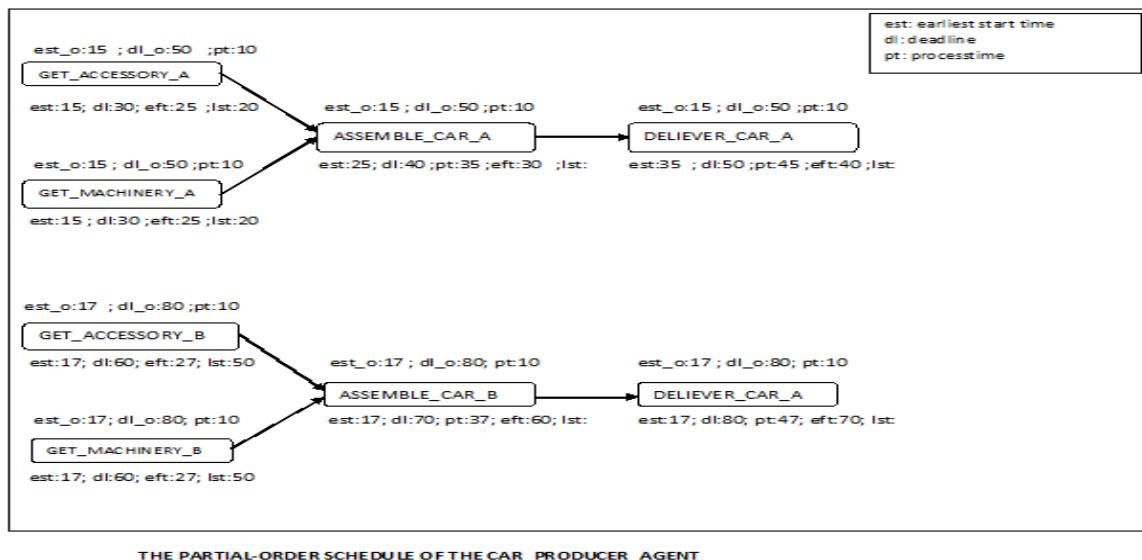


Fig: 5 The Partial-Order Schedule of the Car Manufacturer Agent

$a \leq a$ (reflexivity);
if $a \leq b$ and $b \leq a$ then $a = b$ (antisymmetric);
if $a \leq b$ and $b \leq c$ then $a \leq c$ (transitivity).

In other words, a partial order is an antisymmetric preorder. In other words we can say that a Partial-Order Schedule represents a group of tasks with specified precedence relationship among them using a directed acyclic graph: $G = (V;E)$. $V = \{u\}$, each vertex in V represents a task. $E = \{< u,v > / u,v \in V, u \neq v\}$ belongs E). Each edge (u, v) in E denotes the precedence relationship between task u and task v ($P(u;v)$), that is task u has to be finished before task v can start.

Some terms are used to explain how negotiation is done using partial order schedule.

Task (t) is represented as a node in the graph; it is the basic element of the schedule. A task (t) needs a certain amount of process time ($t.\text{process time}$). A task can be a local task or a nonlocal task: a local task is performed locally (i.e., the “Get Accessories A” task) and a nonlocal task (i.e. the “Get Machinery A” task) is performed elsewhere and hence does not consume local process time.

Pretasks of task t is a set of tasks that need to be finished before task t can start: $\text{Pre}(t) = \{s / s, \forall v \in V \mid s < t, s \rightarrow v\}$, task t can start only after all tasks in $\text{Pre}(t)$ have been finished. For example, the pre tasks of task “Install Accessories A” includes task “Get Machinery A” and task “Get Accessories A”.

The Posttasks of task t is a set of tasks that only can start after task t has been finished: $\text{Post}(t) = \{r / r \in V \mid t < r, t \rightarrow r\}$. For example, the post tasks of task “Assemble Car A” includes task “Deliver Car A”. A task t has constraints of earliest-start-time ($t.\text{est}$) and deadline($t.\text{dl}$).

The earliest-start-time of task t (t.est) is determined by the earliest-finish-time of its pre-tasks ($\text{eft}[\text{Pre}(t)]$) and its outside earliest-start-time constraint($t.\text{est_o}$)
$$t.\text{est} = \max(\text{eft}[\text{Pre}(t)], t.\text{est_o})$$

The earliest-finish-time of a task t (t.ef) is defined as : $t.\text{ef} = t.\text{est} + t.\text{process time}$.

The earliest-finish-time of a set of tasks V (eft[V]) is defined as the earliest possible time to finish every task in the set V , it depends on the earliest-start-time and the duration of each task. For example, in Figure 6, outside-earliest-start-time constraint for task “Assemble Car A” is 10 (same as its super task ‘Purchase Car A’), the earliest-finish-time for its pretasks is 20 (assume “Get Machinery A” could finish at its earliest possible time), then the earliest-start time for task “Install Accessories A” is 20.

The deadline of task t (t.dl) is determined by the latest-start time of its post tasks ($\text{lst}[\text{Post}(t)]$) and its outside-deadline-constraint($t.\text{dl_o}$)
$$t.\text{dl} = \min(\text{lst}[\text{Post}(t)], t.\text{dl})$$

The latest-start-time of a task t (lst(t)) is defined as: $t.\text{lst} = t.\text{dl} - t.\text{process time}$.

The latest-start-time of a set of tasks V (lst[V]) is defined as the latest time for the tasks in this set to start without any task missing its deadline, it depends on the deadline and the duration of each task.

The Flexibility of Task t represents the freedom to move the task around in this schedule.

$F(t) = (t.\text{dl} - t.\text{est} - t.\text{process time}) / t.\text{process time}$.
For example, $F(\text{Get Accessories A}) = (50 - 10 - 10) / 10 = 3$.

The Flexibility of a Schedule S measures the overall freedom of this schedule; it is the sum of the flexibility of each activity weighted by its process time of the process time of the schedule. The flexibility of the task with a longer process time has a bigger influence on the flexibility of the schedule.

$$F(S) = \sum_{t \in S} F(t) * t.\text{process time} / (\sum_i t_i.\text{process time})$$

5. Conclusion

In this abstract we have described an approach to modeling the supply chain management problem in the real business environment using software agents. We use the concept of negotiating agent to model the self-interested entities in the market place. The system framework we designed allows negotiating agents join, stay or leave the system freely. The

basic ideas and methods to attack the aspects of negotiating agent negotiation behaviors including the communication and problem solving parts have been given and studied. To deal with the multiple related negotiation issues, the agent needs to analyze the relationships among these negotiation issues and find what the influence of one issue on the others is. So for this the agent builds a partial-order schedule generated by the agent's local scheduler, so that the agent knows what these tasks are and how they are related to each other. The agent sorts its current negotiation issues according to their importance, their flexibilities or the difficulties of negotiation processes⁸, and finds the influence of the previous issue on the later issues. Also we explored how flexibility is an important factor for ordering and managing negotiation issues in a successful negotiation so as to achieve higher performance. Negotiation performatives for pair-wise and third party protocol have been designed.

6. References

- [1] Decker, K., Lesser, V. R. Quantitative Modeling of Complex Environments. In International Journal of Intelligent Systems in Accounting, Finance and Management. Special Issue on Mathematical and Computational Models and Characteristics of Agent Behavior., Volume 2, pp. 215-234, 1993.
- [2] Deshmukh, A. V., Talavage, J. J., and Barash, M. M. Complexity in Manufacturing Systems: Part 1 - Analysis of Static Complexity IIE Transactions, vol 30, number 7, pp.645-655, 1998.
- [3] Horling, Bryan, Lesser, Victor, Vincent, Regis. Multi-Agent System Simulation Framework. In 16th IMACS World Congress 2000 on Scientific Computation, Applied Mathematics and Simulation, EPFL, Lausanne, Switzerland, August 2000.
- [4] Pritsker, A.A.B. GERT Networks (Graphical Evaluation and Review Technique) The Production Engineer, October 1968
- [5] Sandholm, T. and Lesser, V. 1996. Advantages of a Leveled Commitment Contracting Protocol. Thirteenth National Conference on Artificial Intelligence (AAAI-96), pp.126-133, Portland, OR, .
- [6] Sandholm, T. 1996. Negotiation among Self-Interested Computationally Limited Agents. Ph.D.

Dissertation, University of Massachusetts at Amherst, Department of Computer Science.

[7] Sandip Sen and Edmund H. Durfee A Formal Study of Distributed Meeting Scheduling Group Decision and Negotiation, volume 7, pages 265-289, 1998.

[8] Wagner, Thomas and Lesser, Victor. Relating Quantified Motivations for Organizational Situated Agents. In Intelligent Agents VI: Agent Theories, Architectures, and Languages, Springer

[9] Vincent, R.; Horling, B.; Lesser, V. An Agent Infrastructure to Build and Evaluate Multi-Agent Systems: The Java Agent Framework and Multi-Agent System Simulator. In Lecture Notes in Artificial Intelligence: Infrastructure for Agents, Multi-Agent Systems, and Scalable Multi-Agent Systems. Volume 1887, Wagner & Rana (eds.), Springer, pp. 102-127, 2000

[10] Zhang, Xiaoqin and Lesser, Victor. Multi-Linked Negotiation in Multi-Agent Systems. Technical Report of Computer Science Department, UMass., TR-2002-02.

[11] Xiaoqin Zhang, Victor Lesser "Multi-Linked Negotiation in Multi-Agent Systems" AAMAS'02, July 15-19, 2002, Bologna, Italy.

Copyright 2002 ACM 1-58113-480-0/02/0007.

First Author(Ritu Sindhu): Ph.D Scholar, Banasthali University, Rajasthan. Completed her B.Tech (CSE) from U.P.T.U, Lukhnow, M.Tech (CSE) from Banasthali University, Rajasthan.

Second Author(Abdul Wahid): Presently working as a professor in Computer science department in Gautambudh University, greater noida India. Completed his MCA, M.Tech. and Ph.D. in Computer Science from Jamia Millia Islamia (Central University), Delhi..

Third Author(G.N.Purohit): Dean, Banasthali University, Rajasthan

Hybrid CHAID a key for MUSTAS Framework in Educational Data Mining

G.Paul Suthan¹ and Lt.Dr. Santhosh Baboo²

¹ Head, Department of Computer Science, Bishop Appasamy College
Race Course, Coimbatore, Tamil Nadu 641018, India

² Reader, PG and Research Department of Computer Application, DG Vishnav College
Arumbakkam, Chennai 600106, Tamil Nadu, India

Abstract

Currently there is an increased interest in Educational Data Mining due to the compelling need for quality in higher education and the need to know student behavioural pattern to cater individual needs. The performance prediction of student kind model is quite familiar and mostly it is associated with academic performance. Our proposed framework Multi Dimensional Student Assessment (MUSTAS) has unique feature to measure the student's performance through multidimensional attributes. Each dimension and its associated factors are carefully designed to predict the student's behaviour. We propose the Hybrid CHAID algorithm, a combination of CHAID and Latent Class Modeling (LCM) as the best matched technique for our MUSTAS framework in educational data mining.

Keywords: *Data Mining, Educational Data Mining, CHAID Prediction Model, Latent Class Model.*

1. Introduction

Educational system as of now, especially in India is going through a radical transformation due to the efforts taken by UGC and HRD ministry. The reason behind this is that, quality of education is not met in higher educational Institutions. Due to this many Institutes want to be centre of excellence by going through accreditation process, such as ISO etc., to enhance their quality of education.

Many affiliated institutions want to become autonomous and in due course to become unitary university; thereby enabling them to have more freedom in syllabus and course selection. This also gives them flexibility to have tie ups with foreign Universities. The private institutions are on an increase now to cater to the growing population of youth in countries like India where population growth is high. The parents on the other hand are now looking for quality education so as to enable their child to be placed in good Multi-National Companies.

The private institutions are now having no choice of selection of students in the entry level due to enormous new institutions coming up every year in higher education. This contributes to the low calibre of students in the entry level and making the faculty to take enormous efforts to cater to these students. The faculty with station seniority is also on a decline in private institutions due to change over for higher salary or migrating to Government or other lucrative jobs. On the present scenario the institution has no choice but to have quality in education for attracting students. Therefore to meet the quality needs of the institution, the staffs have to know the behaviour pattern of the students in shorter time, so as to give coaching in accordance to their specific need.

The student needs were earlier known by knowing the students personally or through some response mechanism like feedback. Later statistical methods were used to analyse these behaviour. Now with the advent of data mining techniques and tools, this process of finding patterns in the behaviour of students can be taken to a next higher level. The tools and algorithms used exclusively for educational purpose in data mining is categorised as Educational Data Mining (EDM).

2. EDM for Higher Education

Data mining is finding hidden patterns in a large collection of data. Data Mining can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students as described by Alaa el-Halees [2]. Mining in educational environment is called Educational Data Mining. Han and Kamber [15] describes data mining software that allow the users to analyze data from different dimensions, categorize it and summarize the relationships which are identified during the

mining process. New methods can be used to discover knowledge from educational databases. Student data can be used to analyze trends and behaviors toward their education [2]. Lack of deep and adequate knowledge in higher educational system may prevent management to achieve quality objectives, data mining methodology can help bridging this knowledge gaps in higher education system.

Traditional classroom environments are being widely used. Here face to face contact is established between the teacher and the student. Johnson.S, Arago Shaik and Palma-Rivas [18] says that educations are of different types as public, private, elementary, primary, adult, higher, tertiary and academic education. Most of these types uses passive learning and ignore individual differences. They also sometimes do not cater to the need of students. Here the teachers monitor the student learning process by analyzing the paper records and on observation.

3. Related Studies in EDM

Educational data mining has emerged as an independent research area in recent years, culminating in 2008 with the establishment of the annual International Conference on Educational Data Mining, and the Journal of Educational Data Mining. Romero and Ventura [30] provides a comprehensive study of EDM from 1995 to 2005. It describes the need for analyzing the student data which can be used by students, educators and administrators.

Galit [11] developed a system to warn weak students. Han and Kamber [15] discovered relationship among data. Henrik [16] found hidden relationships. Walters and Soyibo [21] discovered relationship between academic performance and nature of their schools. Z.N. Khan [36] found Girls with high socio-economic status were relatively higher achievers in science stream and boys with low socio-economic status were relatively higher achievers in general. Hijazi and Naqvi, [33] using regression found factors like mother's education and student's family income were highly correlated with the student academic performance. A.L Kristjansson, Sigfusdottir and Allegrante [1] found that Body Mass Index (BMI) affects higher academic achievement. Moriana et al. [17] used Analysis of variance (ANOVA) and it was observed that group involved in activities outside the school yielded better academic performance. Al-Radaideh, et al. [29] prescribed Decision Tree model had better prediction than other models. Cortez and Silva [25] found Decision Tree and Neural Networks in some areas give same accuracy. Gong, Rai, Beck and Heffernan [12] found Impact of self discipline on learning co-related with higher incoming knowledge and fewer mistakes but the actual impact of learning was only marginal. Perera et al. [28] got the Big 5

theory for teamwork as a driving theory to search for successful patterns of interaction within student teams. Madhyastha and Tanimoto [21] investigated the relationship between consistency and student performance with the aim to provide guidelines for scaffolding instruction. Beck and Mostow [6]; Pechenizkiy et al. [27] discovered which types of pedagogical support are most effective, either overall or for different groups of students or in different situations. McQuiggan et al. [24], found whether students are experiencing poor self-efficacy. Baker [3] identified students who are off-task. D'Mello et al. [8] studied on students who are bored or frustrated. Dekker et al. [7] Romero et al. [31]; Superby et al. [34] found factors that predict student failure or non-retention in college courses. Barnes [5] developed algorithms which automatically discover a QMatrix from data. Desmarais & Pu [9] and Pavlik et al [26] developed algorithms for finding partial order knowledge structure (POKS) models that explain the interrelationships of knowledge in a domain. Walters and Soyibo [35] conducted a study to determine Jamaican high school students and found positive significant relationship between academic performance of the student and the nature of the school. Ryan S.J.D. et al. [32] explore that prediction and discovery model are increasing while relationship mining are not used much.

4. CHAID Prediction Model

Chi-squared Automatic Interaction Detection (CHAID) analysis which was first proposed by Kass, 1980[10] is one of post hoc predictive segmentation methods. The CHAID, using of decision tree algorithms, is an exploratory method for segmenting a population into two or more exclusive and exhaustive subgroups by maximizing the significance of the chi-square, based on categories of the best predictor of the dependent variable. Segments obtained from CHAID analysis are different from cluster type models because the CHAID method, which is derived to be predictive of a criterion variable, is defined by combinations of predictor variables as described by Magidson, [22]. CHAID technique depends on interactions among the independent variables, finding those that explain the greatest differences within the dependent variable. Thus, a CHAID decision tree demonstrates how the predictors are differently formed and predicts a dependent variable that shows nominal and continuous scaling. Educators can identify the key influencers or significant drivers in certain students using CHAID analysis, which results in a tree like diagram commonly called a decision tree. Decision trees have several advantages as explained by Bakken [4]. The type of representation makes the resulting classification model easy to use. Moreover, decision trees are suited for

exploratory knowledge discovery because they are non-parametric and make no assumptions about the underlying probability distribution. Decision trees are also efficient to higher-order interactions. They are relatively quickly constructed for large datasets compared to other classification models as presented by Magidson and Vermunt, [23]

5. Latent Class Modeling- LCM

Latent class (LC) modeling was initially introduced by Lazarsfeld and Henry [19] as a way of formulating latent attitudinal variables from dichotomous survey items. In contrast to factor analysis, which posts continuous latent variables, LC models assume that the latent variable is categorical, and areas of application are more wide-ranging. The methodology was formalized and extended to nominal variables by Goodman [13,14], who also developed the maximum likelihood (ML) algorithm that serves as the basis for many of today's LC software programs. In recent years, LC models have been extended to include observable variables of mixed scale type (nominal, ordinal, continuous and counts), covariates, and to deal with sparse data, boundary solutions, and other problem areas.

6. MUSTAS Framework

The Multidimensional Students Assessment (MUSTAS) framework is a novel model, which consist of demographic factors, academic performance of the student and dimensional factors. The dimensional factors has further sub divided into three dimensions respectively self assessment, institutional assessment and external assessment. The main objective of this framework is to identify the contribution of selected dimensions over academic performance of the student, which helps to teachers, parents and management about the student's pattern. Understanding of the pattern may facilitate to redefine the education method, additional care on weakness, and promoting their abilities. The academic performance shows the present ability of the student and the demographic factors shows his personal lifestyle. Construction of this framework strongly believes each aspect considered for this framework is closely related to one another.

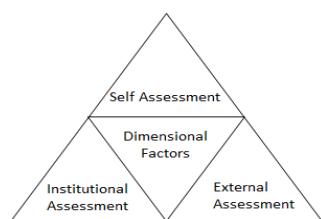


Fig.1 Dimensions of MUSTAS

The dimensional factor helps to measure the student's attitude. Self assessment is measured through five questions, which express their personal interest towards studies. Institutional assessment is specially designed for lecturers/faculty and institution's support towards studies. The third dimension is external assessment, which is designed to measure an external attribute contribution towards their studies.

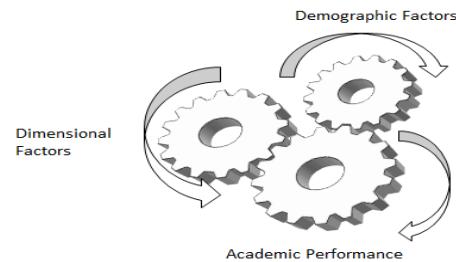


Fig.2 MUSTAS Framework

7. Proposed Model

CHAID based Performance Prediction model in EDM was analysed by Ramaswamy, [20] and the results have proved to be accurate when compared to some other models in terms of accuracy in prediction. One limitation of CHAID is that segments are defined based on a single criterion variable. Given situations where multiple criteria exist, it is not clear how one should go about obtaining a single common segmentation. Using one dependent variable as the criterion may result in one set of segments, while use of an alternative dependent variable will likely yield a different set of segments. Moreover, the categories of a predictor may merge in different ways depending upon which dependent variable is used, again leading to different segments.

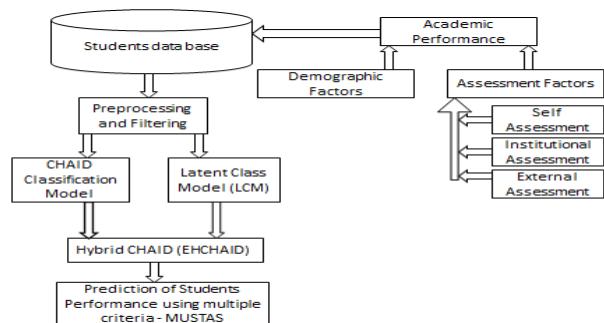


Fig.3 Hybrid CHAID in Educational Data Mining

In addition, when multiple dependent variables do exist, they may be of different scale types (nominal,

ordinal, continuous, count, etc.). Using a 3-category response variable as an example Magidson [22] showed that CHAID segments resulting from treating the dependent variable as ordinal (using profitability scores for the categories) differed substantially from segments derived from the nominal algorithm which ignored the scores. The hybrid approach resolves the need to choose between different segmentations because indicators with differing scale types can be used in extended LCMs, yielding a single LC solution. An important advantage of this hybrid approach over approaches based on specific measures for node homogeneity rather than a model is that the LC model used here can handle dependent variables of different scale types.

The evaluation of student attitude is important to predict the academic performance on 3-dimensions (Self Assessment, Institutional Assessment and External Assessment). A LCM was fit to these data, using academic performance as an active covariate and the eight demographic factors as inactive covariates. This model may be viewed as a kind of unsupervised regression with 12 dependent variables, plus the 11 attribute ratings. This LCM yielded 3 segments. The segments are Good, Average and Poor with respect to the attribute ratings and in their feedback as percentage. These percentages are displayed in the root node of the hybrid CHAID tree. The hybrid CHAID used the 3-category latent variable (segments) as the dependent variable and again utilized the 8 demographics as the predictors.

8. Conclusion

We believe the academic performances of the students are not always depending on their own effort. Our investigation shows that other factors have got significant influence over student's performance. Hence, we introduce the hybrid CHAID with MUSTAS framework in education domain as multidimensional evaluation method to classify the pattern of student through classification tree. This proposal will improve the insights over existing methods.

References

- [1] A. L. Kristjansson, I. G. Sigfusdottir, and J. P. Allegrante, "Health Behavior and Academic Achievement Among Adolescents: The Relative Contribution of Dietary Habits, Physical Activity, Body Mass Index, and Self-Esteem", *Health Education & Behavior*, (In Press).
- [2] Alaa el-Halees, 2009 Mining Students Data to Analyze e-Learning Behavior: A Case Study.
- [3] Baker, R.S.J.D., 2007. "Modeling and Understanding Students' Off-Task Behavior in Intelligent Tutoring Systems." In Proceedings of the ACM CHI 2007: Computer-Human Interaction conference, pp1059-1068.
- [4] Bakken,S.K., 2005. "Use of chi-squared automatic interaction detector in the prediction of vocational rehabilitation outcomes among veterans with substance use disorders". Doctoral dissertation, Univeristy of Wisconsin-Madison.
- [5] Barnes, T., 2005. "The q-matrix method: Mining student response data for knowledge." In Proceedings of the AAAI-2005 Workshop on Educational Data Mining.
- [6] Beck, J.E. and Mostow, J., 2008. "How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students." In Proceedings of the 9th International Conference on Intelligent Tutoring Systems, pp353-362.
- [7] Dekker, G., Pechenizkiy, M. and Vleeshouwers, J., 2009. "Predicting Students Drop Out: A Case Study." In Proceedings of the International Conference on Educational Data Mining, Cordoba, Spain, T. Barnes, M. Desmarais, C. Romero and S. Ventura Eds.,pp41-50.
- [8] D'mello, S.K., Craig, S.D., Witherspoon, A.W., McDaniel, B.T. and Graesser, A.C., 2008. "Automatic Detection of Learner's Affect from Conversational Cues." *User Modeling and User-Adapted Interaction* vol 18. pp45-80.
- [9] Desmarais, M.C. and Pu, X., 2005." A Bayesian Student Model without Hidden Nodes and Its Comparison with Item Response Theory. "International Journal of Artificial Intelligence in Education vol 15,pp291-323.
- [10] G. V. Kass, "An Exploratory Technique for Investigating Large Quantities of Categorical Data", *Applied Statistic*, Vol. 29, 1980, pp. 119-127.
- [11] Galit.et.al.,2007."Examining online learning processes based on log files analysis" : a case study. Research, Refection and Innovations in Integrating ICT in Education.
- [12] Gong, Y., Rai, D., Beck, J. and Heffernan, N. 2009. "Does Self-Discipline Impact Students' Knowledge and Learning?" In Proceedings of the 2nd International Conference on Educational Data Mining, pp61-70.
- [13] Goodman, L.A.,1974a." Exploratory latent structure analysis using both identifiable and unidentifiable models." *Biometrika*,vol61,pp215-231.
- [14] Goodman, L.A.,1974b." The analysis of systems of qualitative variables when some of the variables are unobservable". Part I: A modified latent structure approach, *American Journal of Sociology*, vol79,pp1179-1259.
- [15] Han,J. and Kamber, M., 2006. "Data Mining: Concepts and Techniques", 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor.
- [16] Henrik (2001) Clustering as a Data Mining Method in a Web-based System for Thoracic Surgery: © 2001
- [17] J. A. Moriana, F. Alos, R. Alcala, M. J. Pino, J. Herruzo, and R. Ruiz, "Extra Curricular Activities and Academic Performance in Secondary Students", *Electronic Journal of Research in Educational Psychology*,Vol. 4, No. 1, 2006, pp35-46.
- [18] Johnson, S., Arago, S., Shaik, N., & Palma-Rivas, N. (2000). "Comparative analysis of learner satisfaction and learning outcomes in online and face-to-face learning environments." *Journal of Interactive Learning Research*, 11(1), pp29-49.

- [19] Lazarsfeld, P.F., and Henry, N.W.,1968." Latent Structure Analysis. "Boston: Houghton Mill.
- [20]M. Ramaswami and R. Bhaskaran." A CHAID Based Performance Prediction Model in Educational Data Mining." Madurai Kamaraj University, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1, January 2010
- [21]Madhyastha.T.and Tanimoto, S., 2009." Student Consistency and Implications for Feedback in Online Assessment Systems."In Proceedings of the 2nd International Conference on Educational Data Mining, pp81-90.
- [22]Magidson, J.,1994." The CHAID approach to segmentation modeling: Chi-squared automatic interaction detection." In R. P. Bagozzi (Ed.), In advanced methods of marketing research. Cambridge, MA: Blackwell,pp118-159.
- [23] Magidson, J., & Vermunt, J. K.,2005. "An extension of the CHAID tree-based segmentation algorithm to multiple dependent variables." In Weihs C, Gaul W (eds), Classification: The Ubiquitous Challenge. Springer:Heidelberg.
- [24]Mcqiggan, S., Mott, B. and Lester, J. 2008." Modeling Self-Efficacy in Intelligent Tutoring Systems: An Inductive Approach. "User Modeling and User-Adapted Interaction 18, pp81-123.
- [25]P. Cortez, and A. Silva, "Using Data Mining To Predict Secondary School Student Performance", In EUROSIS, A. Brito and J. Teixeira (Eds.), 2008, pp5-12.
- [26] Pavlik, P., Cen, H. and Koedinger, K.R., 2009." Learning Factors Transfer Analysis: Using Learning Curve Analysis to Automatically Generate Domain Models." In Proceedings of the 2nd International Conference on Educational Data Mining, pp121-130.
- [27]Pechenizkiy, M., Calders, T., Vasilyeva, E. and Debra, P., 2008. "Mining the Student Assessment Data: Lessons Drawn from a Small Scale Case Study." In Proceedings of the 1st International Conference on Educational Data Mining,pp187-191.
- [28]Perera, D., Kay, J., Koprinska, I., Yacef, K. and Zaiane, O., 2009. "Clustering and sequential pattern mining to support team learning." IEEE Transactions on Knowledge and Data Engineering vol21, pp759-772.
- [29]Q. A. Al-Radaideh, E. M. Al-Shawakfa, and M. I. Al-Najjar, "Mining Student Data using Decision Trees", International Arab Conference on Information Technology(ACIT'2006), Yarmouk University, Jordan, 2006.
- [30]Romera, C. and Ventura, S., 2007." Educational Data Mining: A Survey from 1995 to 2005." Expert Systems with Applications 33, 125-146.
- [31]Romero, C., Ventura, S., Eapejo, P.G. and Hervas, C., 2008." Data Mining Algorithms to Classify Students." In Proceedings of the 1st International Conference on Educational Data Mining, pp8-17.
- [32]Ryan S.J.D. Baker and Kalina Yacef." The State of Educational Data Mining in 2009: A Review and Future Visions"
- [33]S. T. Hijazi, and R. S. M. M. Naqvi, "Factors Affecting Student's Performance: A Case of Private Colleges", Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 2006.
- [34]Superby, J.F., Vandamme, J.-P. and Meskens, N., 2006. "Determination of factors influencing the achievement of the first-year university students using data mining methods." In Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems (ITS 2006), pp37-44.
- [35]Y. B. Walters, and K. Soyibo, "An Analysis of High School Students' Performance on Five Integrated Science Process Skills", Research in Science & Technical Education, Vol. 19, No. 2, 2001, pp133-145.
- [36] Z. N. Khan, "Scholastic Achievement of Higher Secondary Students in Science Stream", Journal of Social Sciences, Vol. 1, No. 2, 2005, pp84-87.

G. Paul Suthan has done his Under-Graduation and Post-Graduation at Bishop Heber College, affiliated to Bharathidasan University and Master of Philosophy at Manonmaniam Sundaranar University. He is currently pursuing his Ph.D in Computer Science in Dravidian University, Kuppam, Andhra Pradesh. Also, he is working as the Head of the Department of MCA, Bishop Appasamy College of Arts and Science, Coimbatore, affiliated to Bharathiar University. He has organized various National and State level seminars, and Technical Symposium. He has participated in various National conferences and presented papers. He has 14 years of teaching experience. His research areas include Data Mining and Artificial Intelligence.

Lt.Dr.S.Santhosh Baboo, aged forty two, has around twenty years of postgraduate teaching experience in Computer Science, which includes Six years of administrative experience. He is a member, board of studies, in several autonomous colleges, and designs the curriculum of undergraduate and postgraduate programmes. He is a consultant for starting new courses, setting up computer labs, and recruiting lecturers for many colleges. Equipped with a Masters degree in Computer Science and a Doctorate in Computer Science, he is a visiting faculty to IT companies. It is customary to see him at several national/international conferences and training programmes, both as a participant and as a resource person. He has been keenly involved in organizing training programmes for students and faculty members. His good rapport with the IT companies has been instrumental in on/off campus interviews, and has helped the post graduate students to get real time projects. He has also guided many such live projects. Lt.Dr. Santhosh Baboo has authored a commendable number of research papers in international/national Conference/journals and also guides research scholars in Computer Science. Currently he is Reader in the Postgraduate and Research department of Computer Science at Dwaraka Doss Goverdhan Doss Vaishnav College (accredited at 'A' grade by NAAC), one of the premier institutions in Chennai.

Feature-Level based Video Fusion for Object Detection

Anjali Malviya¹, S. G. Bhirud²

¹ Dept. of IT, TSEC, Mumbai University
Mumbai, Maharashtra, India

² Dept. of Computer Engg., VJTI
Mumbai, Maharashtra, India

Abstract

Fusion of three-dimensional data from multiple sensors gained momentum, especially in applications pertaining to surveillance, when promising results were obtained in moving object detection. Several approaches to video fusion of visual and infrared data have been proposed in recent literature. They mainly comprise of pixel based methodologies. Surveillance is a major application of video fusion and night-time object detection is one of most important issues in automatic video surveillance. In this paper we analyse the suitability of a feature-level based video fusion technique that overcomes the drawback of pixel-based fusion techniques for object detection.

Keywords: *video fusion; feature-level-fusion.*

1. Introduction

Multisensor fusion attempts to combine the information from all available sensors into a unified representation. In other words, it refers to any stage in the integration process where there is an actual combination (or fusion) of different sources of sensory information into one representation. Some of the advantages to multisensory fusion are improved detection, increased accuracy, reduced ambiguity, robust operation, and extended coverage. To illustrate how these advantages come about, relationship among sensors are categorized into three types of relations, complementary, competitive, and cooperative. Moreover, fusion can take place at pixel, feature or decision level. There has been an explosion of applications in multisensor fusion and integration. Multiple-sensor based visual surveillance systems can be extremely helpful because the surveillance area is expanded. Tracking with a single sensor easily generates ambiguity due to limitations of object capturing, especially with insufficient light. This ambiguity may be eliminated from another view via other sensor.

The importance of video surveillance techniques [1-2] has increased considerably since the latest terrorist incidents. Safety and security have become critical in many public areas, and there is a specific need to enable human

operators to remotely monitor activity across large environments such as transport systems (railway transportation, airports, urban and motorway road networks, and maritime transportation), banks, shopping malls, car parks, and public buildings, industrial environments, and government establishments (military bases, prisons, strategic infrastructures, radar centers, and hospitals). Modern video-based surveillance systems [2] employ real-time image analysis techniques for efficient image transmission, color image analysis, event-based attention focusing, and model-based sequence understanding. Moreover, cheaper and faster computing hardware combined with efficient and versatile sensors create complex system architectures; this is a contributing factor to the increasingly widespread deployment of multi-camera systems. These multi-camera systems can provide surveillance coverage across a wide area, ensuring object visibility over a large range of depths. They can also be employed to disambiguate occlusions. Techniques that address handover between cameras (in configurations with shared or disjoint views) are therefore becoming increasingly more important. Events of interest (identified as moving objects and people) must be then coordinated in the multi-view system, and events deemed of special interest must be tracked throughout the scene. Wherever possible, tracked events should be classified and their dynamics (sometimes called behavior) analyzed to alert an operator or authority of a potential danger.

In the development of advanced visual-based surveillance systems, a number of key issues critical to successful operation must be addressed. The necessity of working with complex scenes characterized by high variability requires the use of specific and sophisticated algorithms for video acquisition, camera calibration, noise filtering, and motion detection that are able to learn and adapt to changing scene, lighting, and weather conditions. Working with scenes characterized by poor structure requires the use of robust pattern recognition and statistical methods. The use of clusters of fixed cameras, usually grouped in areas of interest but also scattered across the entire scene,

requires automatic methods of compensating for chromatic range differences, synchronization of acquired data (for overlapping and non overlapping views), estimation of correspondences between and among overlapping views, and registration with local Cartesian reference frames.

However, visual surveillance using multi cameras also brings problems such as camera installation, camera calibration, object matching, automated camera switching, and data fusion.

The image fusion techniques implemented earlier comprised of essentially pixel-level fusion. For video fusion, we explore feature-level fusion methodologies, along with pixel-level-fusion. Fusion at the feature level requires extraction of objects (features) from the input images. These features are then combined with the similar features present in the other input images through a pre-determined selection process to form the final fused image. Since one of the essential goals of fusion is to preserve the image features, feature level methods have the ability to yield subjectively better fused images than pixel based techniques.

2. Feature Level Fusion

Image fusion algorithms can be categorized into low, mid, and high levels [3]. In some literature, this is referred to as pixel, feature, and decision levels. Methods using pixel level either use arithmetic operations (like addition, subtraction) on corresponding pixel intensity from different input images or use the frequency domain. Using the frequency domain, the input images are first transformed in the frequency domain using various pyramid based methods like Laplacian, or Wavelet transforms. After transformation, algebraic operations are performed on the input images fusing them to one image. Then, that image is inverse transformed to the final fused image.

Feature level methods are the next stage of processing where image fusion may take place. Fusion at the feature level requires extraction of objects (features) from the input images. These features are then combined with the similar features present in the other input images through a pre-determined selection process to form the final fused image. Since, one of the essential goals of fusion is to preserve the image features, feature level methods have the ability to yield subjectively better fused images than pixel based techniques. A schematic of feature level fusion is shown in Figure 1. The typical algorithms used are feature-based template methods (like edge enhancement), Artificial Neural Networks, and knowledge based approaches.

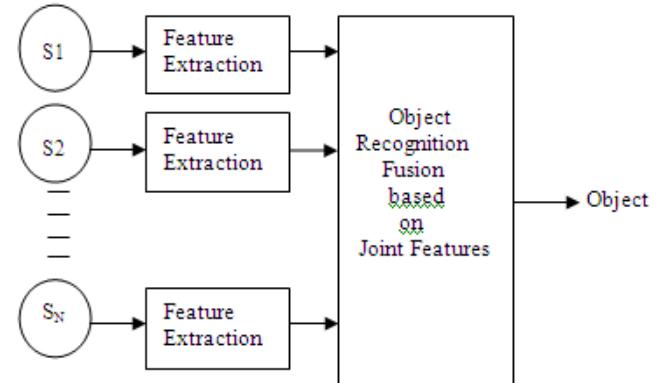


Fig. 1 Schematic of Feature Level Fusion

3. Object Detection using Feature-Level Video Fusion

Night time vision is a primary need of video fusion for surveillance application. The literature survey on video fusion drew our attention towards the problem of noise in IR video, which proves to be a huge deterrent in obtaining high quality fused videos, thereby affecting the surveillance application. We aimed at developing a procedure that could address IR video de-noising and at the same time help in pedestrian detection problem in night-time environment. Unlike most of the work on IR video de-noising, this method does not require static background assumption and Gaussian noise assumption. It involves three steps, IR video de-noising, object detection (pedestrian in this case) in IR video and visual-infrared video fusion.

4. Methodologies

Additive and multiplicative noise is an unwanted component of videos. They can occur as Gaussian noise or film grain noise and may have undesirable effects on surveillance applications. The first stage of most video processing techniques is noise removal but mere usage of spatial noise removal techniques can only give limited filtering performance [4]. Improved performance can be achieved by considering a sequence of previous and/or subsequent image frames for filtering, leading to a spatio-temporal filtering. We use a 3D window around pixel (x, y, t) for our filtering, as proposed in [5]. If noise in preceding and current frame is additive white Gaussian noise (AWGN), taking linear average of the pixels in 3-D window gives good results for de-noising.

Pedestrian motion effects should be taken into account when filtering in order to reduce temporal filtering artifacts such as blurring. Image regions that include pedestrian motion in preceding frames and /or current

frames should not be taken into account while filtering. For this purpose, we use brightness (or intensity value) threshold T_1 to determine pixels which potentially belong to the pedestrian, because we assume that the pedestrian has higher temperature than the environment, and hence the pixels corresponding to the pedestrian in infrared frames would be brighter than the background environment.

The object (pedestrian in this case) is segmented from the background, using the thermal-image features of pedestrian. As pedestrian region is brighter than the background in infrared video, the regions can be segmented according to their brightness. Shape recognition algorithms are then used to separate the pedestrian regions from other false detections. The segmentation algorithm proposed by Adams/Bischof [6] is used. The image is first searched for a “seed” pixel belonging to “pedestrian-type,” in every frame. If found one, then we segment one candidate-region of pedestrian by applying the seeded region growing algorithm proposed in [6] and continue such search until all candidate-regions are segmented.

The brightness-only information is not robust in detection of pedestrian, as there are other thermal emitters in the environment which can lead to false detection. In order to improve the robustness of detection, we fuse color information and shape information. We use area feature of bright region to detect pedestrian. If area of the bright region is greater than a threshold value and height/width ratio is in the previously-established range, then the region is regarded as a pedestrian else it is regarded as noise region.

The bright regions in infrared frames correspond to detected pedestrian. At the end of our detection method, infrared frames are fused with visual frames to provide visual context. We fuse the frames by adding an increment T_6 to pixel’s RGB values in visual frames at the corresponding geometric position, by which pedestrian region in visual video can be made more visible to human vision.

5. Implementations and Results

The dataset used is “AIC Thermal/Visible Night-time Dataset” which contains two video sequences, one in the visible spectrum and one in thermal infrared (Input 1). Both are compressed into AVI format and contain 527 frames each. It was captured from a balcony in Dublin City University campus, Ireland [7]. Figure 2 shows four frames from the IR video and the corresponding four frames from the visible video are shown in Figure 3. Figure 4 & 5 show the classification of noisy and pedestrian pixels, in the IR frame, and the denoised IR frames. Segmenting candidate region using seeded region

growing can be seen in Figure 6. Figure 7 and Figure 8 show the final fused image.

The methodology, is also tested on a day-time video (Input 2) obtained through visible camera and IR camera respectively.

6. Conclusion

Pedestrian image region is brighter than background in infrared video, thus the regions can be segmented according to their brightness [Input 1]. This is however not robust in detection of pedestrian, as there are other thermal-emitters in environment which can lead to false detection [Input 2]. Therefore to improve robustness of detection, we fuse color information (brightness) and shape information. In the fused tracking video frames [Fig. 7 and 8], detected pedestrian have been marked by a red rectangle, after enhancing its brightness. As a result this method provides a more visualized pedestrian detection result for human vision. However the algorithm fails to detect the object distinctly in the day time video where there are other bright objects in the IR video [Fig. 10 – Fig. 14].

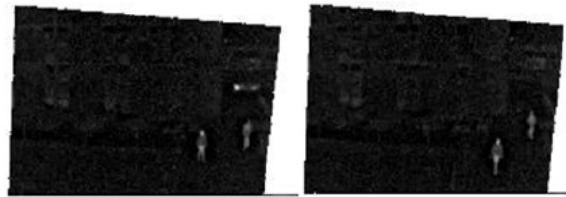
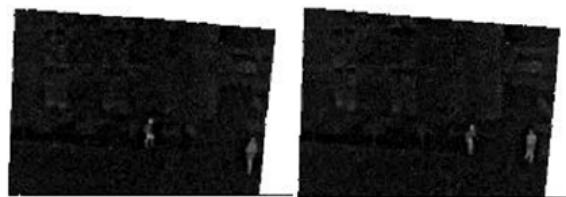


Fig.2 IR Video Frames from Input 1

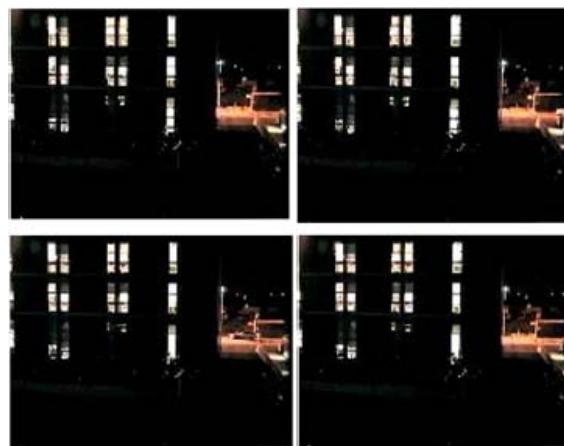


Fig.3 Visible Video Frames from Input 1



Fig. 4 Classification of IR Frames into Pedestrian and Noisy Pixels

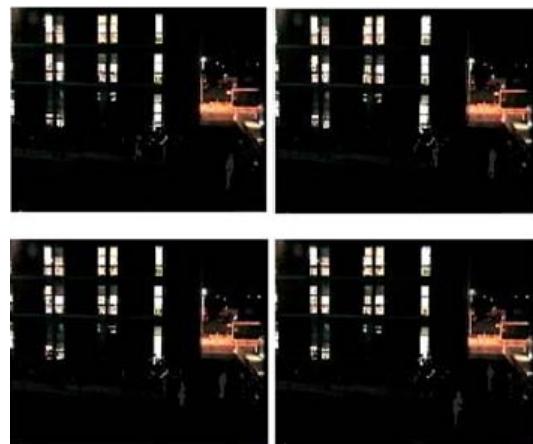


Fig. 7 Fused frames with the Pedestrian Regions



Fig.5 De-noising of IR Frames



Fig. 8 Frames from the Fused Video

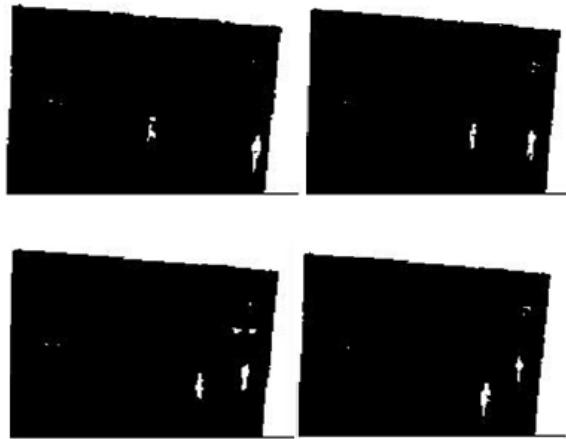


Fig.6 Seeded Region Growing



Fig.9 IR Video Frames from Input 2



Fig. 10 Visible Video Frames from Input 2

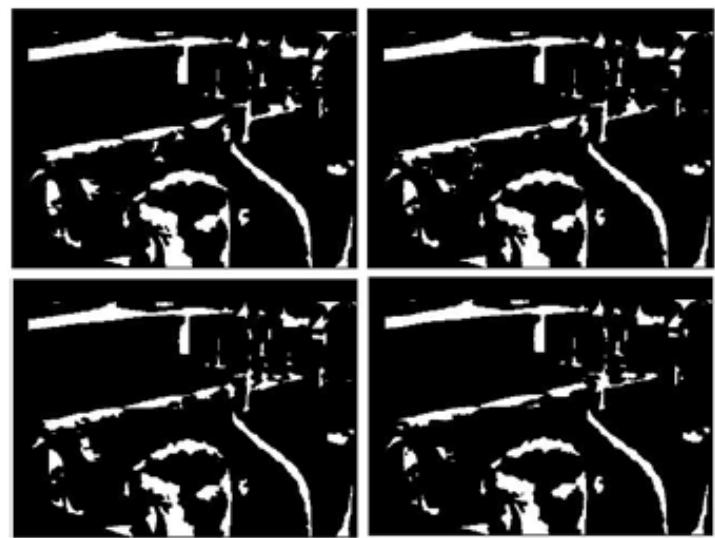


Fig.13 Seeded region growing

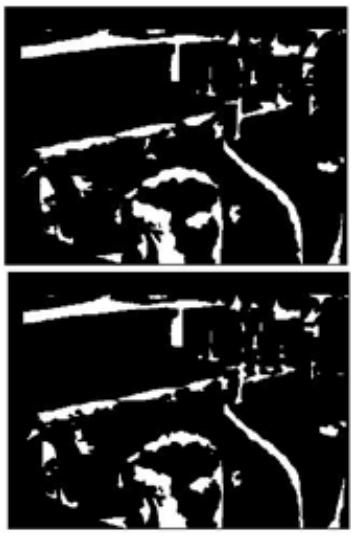


Fig.11 Classification of IR Frames into Pedestrian and Noisy Pixels

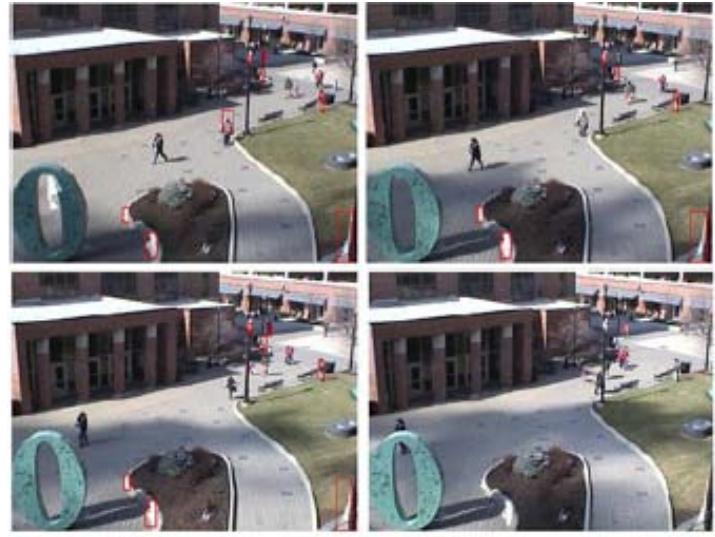


Fig.14 Fused Image

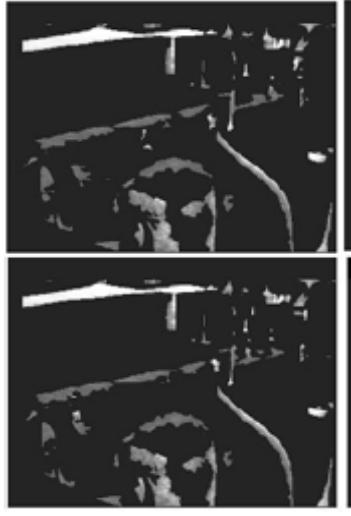


Fig.12 De-noising of IR frames

References

- [1] H.M. Chen, S. Lee, R.M. Rao, M.A. Slaman, and P.K. Varshney, "Imaging for concealed weapon detection," IEEE Signal Processing Mag., vol. 22, no. 2, pp. 52–61, Mar. 2005.
- [2] G.L. Foresti, C.S. Regazzoni, and R. Visvanathan, "Scanning the issue/technology - Special issue on video communications, processing and understanding for third generation surveillance systems," Proc. IEEE, vol. 89, no. 10, pp. 1355–1367, Oct. 2001.
- [3] Yaonan Wang, Multisensor Image Fusion: Concept, Method and Applications, Faculty of Electrical and Information Engineering, Hunan University, Changsha, 410082, China.
- [4] Brailean J., Kleihorst R., Efstratiadis S., Katsaggalegos A., and Lagendijk A. Noise Reduction Filters for Dynamic Image Sequences: A Review, Proc. Of IEEE 83 (9), (1995) 1272-1292
- [5] M. Kemal Güllü, Oğuzhan Urhan, and Sarp Ertürk, Pixel Domain Spatiotemporal Denoising for Archive Videos. ISCIS 2006, LNCS 4263, pp. 493-502, 2006.
- [6] R. Adams and L. Bischof, Seeded region growing, IEEE Trans. Pattern Anal. Machine Intell., vol. 16, no. 6, pp. 641C647, 1994
- [7] O Conaire, C., O Connor, N.E., Cooke, E., Smeaton, A.F.: Comparison of fusion methods for thermo-visual surveillance tracking. International Conference on Information Fusion (Fusion 2006)

A Fuzzy Based Stable Routing Algorithm for MANET

Arash Dana¹

Mohamad Hadi Babaei²

¹ Dept. of Elect. Eng. Islamic Azad University, Central Tehran Branch

² Scientific Association of Electrical & Electronic Engineering Islamic Azad University Central Tehran Branch

Abstract

The increasing popularity of using multimedia and real time applications in different potential commercial in MANETs, make it logical step to support Quality of Service (QoS) over wireless network. QoS support is tightly related to resource allocation and reservation to satisfy the application requirements; the requirements include bandwidth, delay, delay-jitter and packet to loss ratio. One of the notoriously difficult problems in QoS routing in Mobile Ad-hoc Networks (MANET) is to ensure that the established path for a connection does not break before the end of the data transmission. In order to reduce the number of broken routes, a novel reliable routing algorithm using fuzzy applicability is proposed to increase the reliability during the routing selection. In the proposed algorithm source chooses a stable path for nodes mobility by considering nodes position/velocity information. Also we propose novel method for rout maintenance, in this protocol before breaking packet transmitted path a new one is established. The simulation results show that the algorithm can reduce the number of broken routes efficiently and can improve route stability and network performance effectively.

Keywords: mobile ad hoc network, QoS routing, fuzzy applicability, rout maintenance

1. Introduction

In the past decade we have witnessed a phenomenal growth in the deployment of portable wireless devices and related services, including wireless multimedia. MANET, an archetypical infrastructure-less wireless packet network, enables these wireless devices to communicate with each other without the help of base stations or other pre-existing infrastructure. While a mobile node can communicate directly with the nodes lying within its transmission range, communication with the mobile nodes outside of the transmission range must necessarily be multi-hop and require the establishment of communication paths. It is

well known that most of the multimedia applications require the establishment of communication paths that satisfy a number of negotiated parameters (such as delay or bandwidth), usually referred to QoS guarantees. Due to the dynamic nature of the network topology and imprecise network state information, a lot of problems remain before more efficient solutions are found for QoS routing in MANET. One of the problems is that the established path for a connection request may break before the end of data transmission. An active path fails due to mobility when a pair of nodes forming a hop along the path move out of each other's transmits range. As general an alternative path is sought only after the current path fails. The cost of detecting a failure is high: several retries have to time-out before a path is "pronounced dead". Thus, when a path fails, packets experience large delays before the failure is detected and a new path is established. In order to confirm the stability of the whole route from source node to destination node, every link between each neighboring node should be ensured firstly. How to predict the situation of node movement in the future using current information becomes the key of predicting route stability.

Several ad-hoc routing protocols for MANETs have been proposed in recent years. Most of these routing protocols, such as Destination Sequenced Distance Vector, Optimized Link State Routing Protocol, Ad-hoc On-demand Distance Vector Routing (AODV), and Dynamic Source Routing (DSR), are belong to shortest-path routing protocols [1]. These protocols are generally based on shortest path algorithms (or minimum hop count) to determine the route paths, which may be not so robust especially for time-varying-radio-link cases. This may result in lowered throughput and increased packet loss rate. Several researchers have proposed adaptive ad hoc routing protocols; such as Associativity Based Routing (ABR) [2] and Signal Stability-based Adaptive routing (SSA) [3], to improve the Stability of discovered routes in MANETs. The goal of this paper is to select a stable path for reducing the number of broken path. To achieve this goal, source

estimate route stability coefficient (RSC) with fuzzy logic and it use RSC to select route with the highest stability for the requiring source, destination pair .Selecting the route is source responsibility for two reasons:

- 1-more flexibility in switch to a more stable path
- 2-reducing of network traffic and time for initiation route recovery

We investigate introducing preemptive route maintenance to Ad hoc routing protocols. More specifically, when two nodes are moving out of each other's transmit range, intermediate nodes of active paths warned destination a path break is likely. With this early warning, the destination can initiate route discovery and source switch to a more stable path potentially avoiding the path break altogether.

The proposed scheme utilizes GPS location information [4], and in the protocol, GPS position, velocity information is piggybacked on data packets during a live connection and is used to estimate the stability of the link between two adjacent nodes.

The rest of the paper is organized as follows. Section 2 gives a detailed description of the Route Stability Coefficient (RSC). The proposed routing algorithm is given in Section 3. The Route Maintenance algorithm and discusses some possible optimizations to it, is given in Section 4. In section 5, we do some simulations and present a comparative performance of DSR with our algorithm without Route Maintenance and with Route Maintenance. The conclusions are remarked in Section 6.

2. Route Stability Coefficient (RSC)

2.1 Description of node

In a mobile ad hoc network, the communication between two adjacent nodes needs the relative movement information of nodes. Generally speaking, the state of a node includes the position, the movement speed and the movement direction. The following are the attribute description of one node. Node: $N_i(p, v)$
 Where i denotes the No. of one node,

p denotes the position of Node i . According to GPS location information, each node has one unique position.
 v denotes the velocity of node i . It is a vector includes value and direction.



Figure 1

$$\Delta d_{i,j} = P_i - P_j$$

Where $\Delta d_{i,j}$ denotes the distance between node i and node j

$$\Delta v_{i,j} = (v_i \cos\alpha - v_j \cos\beta) - (v_i \sin\alpha - v_j \sin\beta)$$

$\Delta v_{i,j}$ means that if the velocity vectors of two nodes are similar in size and direction, the value of $\Delta v_{i,j}$ is equal zero. α denotes the angle between v_i and the line connected from node i to node j , β denotes the angle between v_j and the extended line connected from node i to node j . This part means that if the direction of v_i and v_j are face to face, the value of $\Delta v_{i,j}$ is positive , contrarily, if they are in opposite direction, the value $\Delta v_{i,j}$ is negative.

2.2 Link Stability Coefficient (LSC)

Fuzzy logic implements human experiences and preferences via membership functions and fuzzy rules. The fuzzy logic proposed to calculates the Link Stability Coefficient (LSC) of each link between source and destination. The fuzzy logic uses two input variables and one output variable. The two input variables to be fuzzified are Δd and Δv of the neighbor nodes. The inputs are fuzzified, implicated, aggregated and defuzzified to get the crisp value of LSC as the output. The linguistic variables associated with the input variables are Low (L), medium (M) and high (H) for Δd and negative (N), zero (Z) and positive (P) for Δv . For the output variable, link stability index, six linguistic variables are used. They are, very low (VL), low (L), medium (M), average (A), high (H) and, very high (VH). All membership functions are chosen to be triangular. The table 1 shows the fuzzy conditional rules for the fuzzy stability. The first rule can be interpreted as, "If (Δd is low) and (Δv is negative) then link stability is medium". Similarly the other rules have been developed.

Table 1

| Δv | N | Z | P |
|------------|----|----|---|
| Δd | | | |
| L | M | VH | A |
| M | L | H | A |
| H | VL | A | H |

2.3 Rout Stability Coefficient (RSC)

The LSC between each neighboring nodes can be computed using fuzzy logic. Here, we use $LSC_{i,j}$ denote the LSC between node i and node j . Assume one communication route between source and destination is made up of n intermitted nodes

$$RSC_{s,d} = LSC_{s,1} * LSC_{1,2} * LSC_{2,3} * \dots * LSC_{n,d}$$

$RSC_{s,d}$ denotes the Rout Stability Coefficient of the whole route.

3. Rout discovery

This process executes the path-finding algorithm to discover the stable route between source and destination. The source node initiates a route discovery process by broadcasting a Route Request (RREQ) message to all of its neighboring nodes. The RREQ packet here is similar to the RREQ in DSR protocol. Intermediate nodes receive RREQs and rebroadcast them. The destination node receives multiple RREQs within a time window, which starts from the first arrival RREQ. In this time window destination send Route Reply (RREP) per each received RREQ without delay. It creates RREP messages formatted similarly in DSR protocol for responding with the RREQs but includes a two newly field named node position and node velocity. Intermediate nodes add own position and velocity to RREP. Then the nodes forward the RREP toward the source node along the reverse route through which the selected RREQ passed. RREP packet, which contains the complete route topology information from source to destination, is sent back to the source node. The source node calculate RSC while receiving first RREP and start to transmit data packet from discovered path, by receiving next RREPs compares their RSCs with transmitted packet route RSC, in this comparing if source find route with higher RSC it will switch transmit packet path to stable path. In figure – node A is as source and G is as destination. Node A broadcast RREQ to find existing routes. Node G when received RREQ₁ sends RREP₁ without delay. In table 2 shown apparent routes between source and destination, and also time difference between receiving RREQs and RREQ₁.

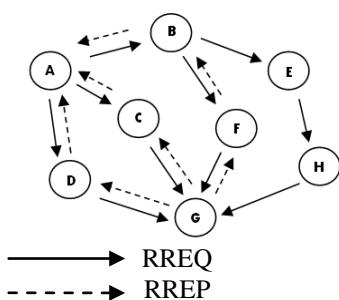


Figure 2

Table 2

| RREQ Number | Discovery path | Receive time |
|-------------|----------------|-------------------------------------|
| RREQ1 | A,D,G | t ₁ |
| RREQ2 | A,C,G | t ₂ , t ₁ < T |
| RREQ3 | A,B,F,G | t ₃ , t ₁ < T |
| RREQ4 | A,B,E,H,G | t ₄ , t ₁ > T |

It should be considered the received time of RREQ₄ (t₄) is out of time window (T), so Destination sends three RREPs to source. Source calculates RSC's after receiving RREPs. It is shown RSCs in table 3. Source starts to transmit data packet from route 1 after receiving RREP₁ and will switch to route 3 by receiving RREP₃.

Table 3

| RREP Number | RSC |
|-------------|------|
| RREP1 | 0.65 |
| RREP2 | 0.54 |
| RREP3 | 0.72 |

4. Route Maintenance

A link failure is costly because multiple retransmissions and timeouts are required to detect the failure and a new path has to be found (in on-demand routing) we propose route maintenance extension to our routing protocols. With route maintenance, recovery is initiated early by detecting that a link is likely to break and finding and using an alternative path before the cost of a link failure is incurred. Route Maintenance algorithm consists of three components: (i) detecting that a path is likely to be disconnected soon; and (ii) finding a better path (iii) switching to it. A critical component of the proposed scheme is determining when path stability is no longer acceptable. The degree of route stability is a function of neighbor nodes distance and relative velocity of neighbor nodes. If distance of two intermediate nodes (i,j) in data transmitted packet route is higher than breaking route threshold and also nodes movement be in opposite direction ($\Delta v_{i,j} < 0$) intermediate node generates breaking route warning (BRW) and sends it to destination. Destination broadcasts route recovery (RREC) when received BRW. Intermediated nodes add their position and velocity information to RREC and broadcast it. Source by receive RRECs compare RSC of discovery path (and compare with RSC of transmitted packet route) if path with higher RSC exist change transmit packet path.

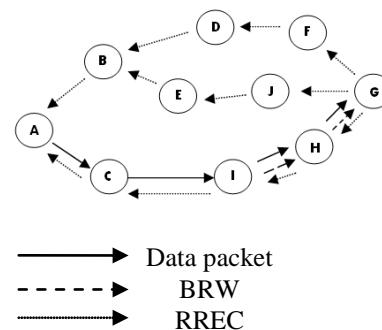


Figure 3

Generate RREC by destination and selected route by source is cause reducing initiate route recovery time. If transmitted data packet have higher RSC across exciting route between source and destination and if intermediate node generate BRW, route recovery process is start and no changing occur in transmitted data packet route, repeat this process and broadcasts RREC by destination can decrease network performance and have obverse affect. To prevent this, destination start a timer after broadcast RREC or RREP and distention drops RRECs until timer overflow.

Using preemptive route maintenance the cost of detecting a broken path (the retransmit and timeout time) is eliminated if another path is found successfully before the path breaks. In addition, the cost for discovering an alternate path is reduced (or eliminated) since the path discovery is initiated before the current path was actually broken. This can be expected to reduce the latency and jitter.

5. Performance Evaluation:

5.1 Simulation Environment and Methodology:

In order to verify the correctness of our approach and to see the performance in real application scenario, we establish a pure Ad Hoc network with 50 nodes distributed over 900mx900m area in the simulation platform during 500 seconds . A random waypoint mobility model was used: each node randomly selects a position, and moves toward that location with a speed ranging from just above 0 m/s to 10 m/s. When the node reaches that position, it becomes stationary for a programmable pause time; then it selects another position and repeats the process. We performed the simulation with various pause times 0,100,200,300,400 and 500. A mobile node is assumed to be stationary with pause time of 500 seconds. A node moves constantly with 0 second pause time. Each node has a radio propagation range of 150 meters and channel capacity is 2 Mb/s. The source-destination pairs are spread randomly over the network. Continuous Bit Rate (CBR) traffic sources were chosen, as the aim was to test the routing protocols. The sending rate used was 4 packets per second with a packet size of 512 bytes.

5.2 Performance Metrics:

The following metrics are used in computing the Performance. The metrics were derived from one suggested by the MANET working group for routing protocol evaluation.

A. Packet delivery ratio:

Packet delivery ratio is the ratio between the number of packets originated by the “application layer” CBR sources

and the number of packets received by the CBR sink at the final destination.

B. Route Stability:

Route stability is a very important performance parameter for a routing protocol. Route stability can be measured in terms of number of route failures.

C: Throughput:

It is the amount of digital data transmitted per unit time from the source to the destination. It is usually measured in bits per sec.

5.3 Simulation Results:

We present below the performance of the proposed stabled Routing Algorithm (SRA) without Route Maintenance and with Route Maintenance in comparison with DSR .

Figure 4 shows the affects of different routing algorithms on route stability. Because the dynamic of network decreases as the pause time increases, the route stability of three algorithms increases.DSR shows the worst path stability.

X-axis represents the nodes paused time and Y-axis represents the numbers of broken routes. Through changing the paused time, we can see that the broken number of SRA with Route Maintenance decreases greatly compare to DSR.

Figure 5 and 6 shows the affects of different routing algorithms on network performance. As shown in Figure 5, the packet delivery ratios of three algorithms increase as the pause time increases. SRA with Route Maintenance is the best, SRA is lower than SRA with Route Maintenance, and DSR is the worst. The reason is that the stronger route stability is, the higher packet delivery ratio is. Figure 6 shows improvement in throughput over DSR against pause time.

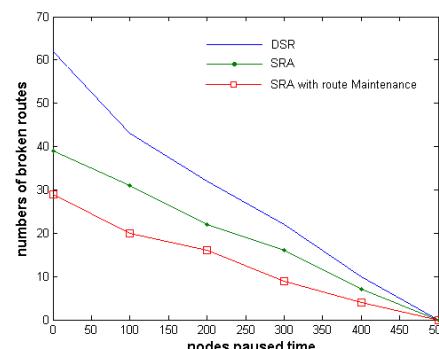


Figure 4

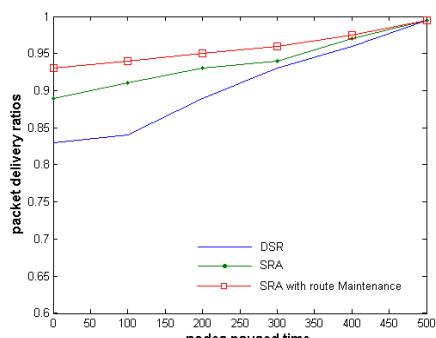


Figure 5

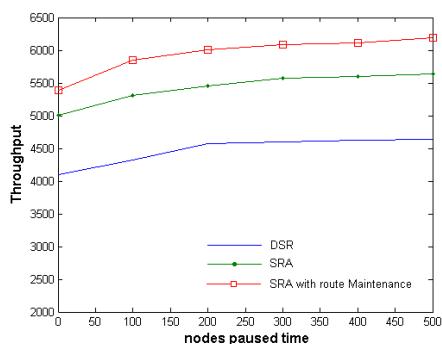


Figure 6

6. Conclusions and future work:

Mobile ad hoc networks are full of uncertainties because of dynamic topologies, dynamic traffic and different application contexts. As a well recognized decision making technique, Fuzzy logic offers a natural way of representing and reasoning the problems with uncertainty and imprecision. Fuzzy logic is a suitable way to be applied in the mobile ad hoc network routing decision. Innovation in the paper is source select rout by using important parameter. Simulation results show that the SRA with Route Maintenance algorithm improved the performance and stability of MANET networks dramatically. We believe that the proposed protocol can be further investigated based on other practical radio propagation models in order to design better adaptive mechanism for mobile ad hoc networks.

References:

- [1] C.E. Perkins, "Ad hoc networking," Addison-Wesley, 2001.
- [2] C.K. Toh, "Associativity-based routing for ad hoc mobile networks," Wireless Personal Communications, vol. 4 no. 2, pp. 103-139, 1997.
- [3] R. Dube, C.D. Rais, K.Y. Wang, and S.K. Tripathi, "Signal stability based adaptive routing (SSA) for ad hoc mobile networks," IEEE Personal Communications, vol. 4, no. 1, pp. 36-45, 1997.
- [4] E.D. Kaplan (Editor), "Understanding the GPS: Principles and Applications", Artech House, Boston, MA, 1996
- [5] Jenn-Hwan Tamg, Bing-Wen Chuang, and Fang-Jing Wu" A Radio-Link Stability-based Routing Protocol for Mobile Ad Hoc Networks"

2006 IEEE International Conference on Systems, Man, and Cybernetics
 October 8-11, 2006, Taipei, Taiwan

[6]Mehdi Zarei,Karim Faez,Javad Moosavi Nya and Morteza Abbaszadeh Meinagh"Route Stability estimation in Mobile Ad Hoc Networks using Learning Automata".16th Telecommunications forum TELFOR Serbia,Belgrade, November 25-27 ,2008

ENHANCED DIGITAL WATERMARKING ALGORITHM FOR DIRECTIONALLY SELECTIVE AND SHIFT INVARIANT ANALYSIS

B.Benita II M.E. (CSE), PSNCET

Abstract— Based on the good characteristics of dual-tree complex wavelet transform (DT-CWT), an improved digital watermarking algorithm is proposed here. The algorithm improves the embedding scheme by selecting the embedding channels and using the visual masking of Human Visual System. This also uses the spread spectrum in embedding scheme and Error correction code, in order to increase the robustness against common attacks different filtering attack etc., This algorithm increases the performance of watermark and has better robustness against common attacks.

Keywords: Spread spectrum, Error correction code

1. INTRODUCTION

WITH the fast growth development of computer network technique and multimedia technology, digital media(such as image, video, audio or text) are stored, transmitted and distributed through Internet without any loss in the quality of the content. Hence, some way of protection of copyrighted digital data is required. A digital watermarking technique has been developed to protect intellectual property from illegal duplication and manipulation. Digital watermarking means embedding information into digital media in such way that it is imperceptible to a human observer but easily detected by means of computing operations in order to make assertions about the data. The watermark is needed to be robust against intentional removal by malicious parties. Thus by means of watermarking, the data is still accessible but permanently marked [1], [2]. Watermarking schemes can be robust or fragile. Robust watermarks are designed to resist to malicious or intentional distortions, such as general

image processing and geometric distortions [3]; while a fragile watermarks are required for the purpose of authentication and verification. We can also classify watermarking schemes according to operation domain: the spatial domain and frequency domain. The simplest watermarking technique embeds a watermark directly into the spatial domain by modifying the Least Significant Bit (LSB) plane of the original image [4]. The watermarking scheme based on the frequency domains can be further classified into the Discrete Cosine Transform (DCT) [5], Discrete Fourier Transform (DFT) [6], Discrete Wavelet Transform (DWT) [7], Dual tree complex wavelet transform (DT-CWT) and others. In general, the transform domain techniques have provided more advantages and better performances than those of spatial ones in most of digital watermarking development and researches. The standard discrete wavelet transform has been exploited with great success across the scope of signal and image processing applications. For example, the DWT has the following advantages, such as good energy packing, perfect reconstruction with short support filters, no redundancy and low computation complexity. However, it lacks shift invariance (i.e., which means that small shifts in the input signal can cause major variations in the distribution of energy between DWT coefficients at different scales), and suffers from poor directional selectivity for diagonal features, because the wavelet filters are separable and real. In order to overcome these problems, complex wavelets have been proposed. Kingsbury's dual-tree complex wavelet transform (DT-CWT) is an outstanding example [8], [9]. The dual-tree complex wavelet transform is a relatively new development to the discrete wavelet transform (DWT), with important additional properties [9]:

- Approximate shift invariance;
- Good directional selectivity in 2-dimensions (2-D) with Gabor like filters (also true for higher dimensionality, m-D);
- Perfect reconstruction using short linear-phase filters;
- Limited redundancy (2:1 in 1-D and 4:1 in 2-D);

- Low computation comparing to other shift invariant transformations.

The work discussed in this paper is concerned with the design of robust and semi-blind watermarking algorithms with complex wavelet transform. We choose to use the complex wavelet transform as our watermarking domain because it is a relatively new transform and has useful properties for image processing applications. Previous work shows that DT-CWT gives good performance in image watermarking. In [10], [11], [12], [13], [14], [15], [16]. The outline of this paper is as follows: in the next section, we present the different steps for the proposed scheme. In Section III, we present the experimental results, and finally the paper is ended by a conclusion in Section IV.

II. PROPOSED WATERMARKING ALGORITHM

The new watermarking method that we propose is based on dual-tree complex wavelet transform. The overview of our watermarking scheme is illustrated in Fig. 1. In this scheme, an input gray-scale image (512x512 pixels) is split into many non-overlapping small blocks with 8x8 pixels; the sub-image (256x256 pixels) is then constructed under control of secret key "key1". On the other hand, a watermark is encrypted and decomposed into different parts which are adaptively spread spectrum and embedded in corresponding highest sub-bands of the 3-level DT-CWT transformed original sub-image. One example of oriented sub-bands of 3-level DT-CWT decomposition of an image is presented on Fig. 2. This newly proposed scheme consists of four parts, including: image preprocess, watermark preprocess, watermark embedding, and watermark detection. Details are described in the following sections.

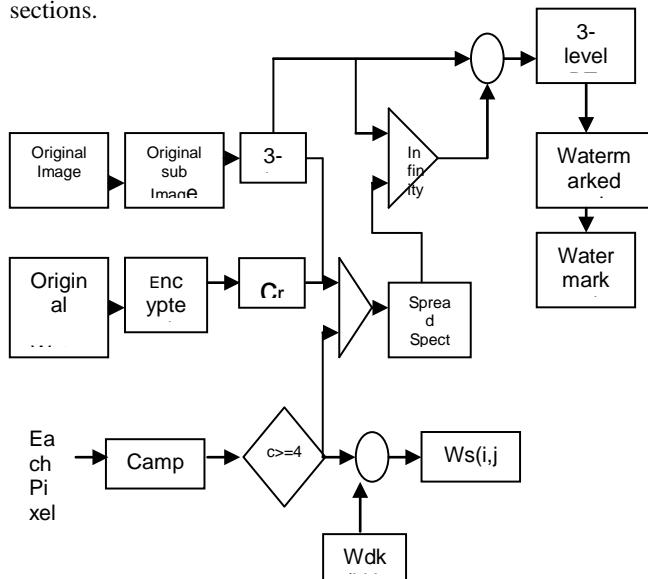


Fig. 1. Overview of the watermarking process.

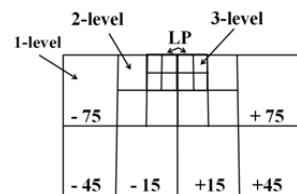


Fig. 2. 3-level DT-CWT decomposition of an image. LP corresponds to low-pass CWT coefficients.

A. Image preprocess

In our watermark scheme, we apply the dual-tree complex wavelet transform only locally, we transform the sub-image, which is extracted from the host image, in the complex waveletdomain by using 3-level DT-CWT. Modifying coefficients at levels coarser than 3 tends to be relatively ineffective and to introduce visual artifacts. To construct the sub-image, we use the following process [17]:

- 1) We first split the host image I_{orig} , into many nonoverlapping small blocks with 8x8 pixels in a scanline order. With the image has 512x512 pixels, we will get 4096 small blocks.
- 2) We label the small blocks from number 1 to number 4096, then we generate a sequence S_i , which contains 4096 elements, by using the logistic map under a special initial value "Key1". The logistic map is one of the simplest chaotic maps, described by:

$$S_{k+1} = \mu S_k(1 - S_k); (k = 0, 1, 2\dots); (1)$$

where $0 \leq \mu \leq 4$. When $3.5699456 \leq \mu \leq 4$, the map is in the chaotic state.

- 3) We multiply each element of S_i by 4096 and then round it toward infinity. Therefore, we obtain a new sequence S_n , in the integer domain [1, 4096].
- 4) We select the forefront 1024 different elements in the new sequence noted by S_1 , and we choose the small blocks accordingly. Finally, we construct the sub-image in a scanline order. Fig. 3 visualizes an example of this selection process.



Fig. 3. Example of a constructed sub-image.

B. Watermark preprocess

In recent years, the chaotic data have been used for digital watermarking to increase the security. In our approach, a fast pseudo random number traversing method is used as the chaotic mechanism to change the watermark image W , which is a binary image $\{1, 0\}$ with 93×62 pixels, into a pseudo random matrix W_d by using the Eq. 2. Then the W_d is divided into small images with size 31×31 pixels, and totally 6 independent sub-watermarks are obtained W_{dk} (Where $k=1,2,\dots,6$).

$$Key2 : W \Rightarrow W_d, W_d(Key2(i, j)) = W(i, j); i, j \in N; \quad (2)$$

"Key2" presents the second key in our watermark procedure, which is an exclusive key to recreate the watermark image.

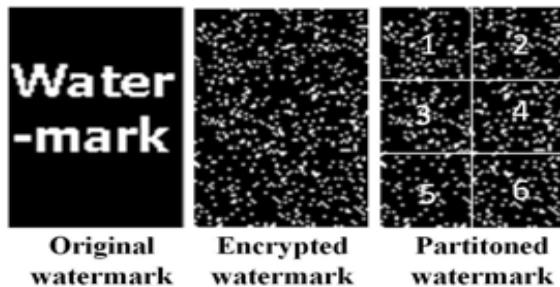


Fig. 4 shows an example of encrypted watermark image and the result of sub-watermarks.

C. Watermark embedding

Firstly, the original sub-image is decomposed by 3-level DT-CWT to obtain the 6 high-pass sub-bands. The DT-CWT coefficients are denoted by \tilde{I} . Secondly, With the Key "Key3" the position of the 6 sub-watermarks is scrambled. Based on magnitudes of the 3-level DT-CWT high-pass coefficients, the sub-watermarks W_{dk} are adaptively spread spectrum. For each pixel (i, j, k) of each highest frequency sub-band in \tilde{I} , the value is compared with those of its eight neighbors, t denotes the total

number which the value is larger than its eight neighbors, as described by the following formula:

$$W_s(i, j, k) = 1, \text{ if } (t \geq 4 \text{ and } W_{dk}(i, j, k)=1) \text{ Or } (t < 4 \text{ and } W_{dk}(i, j, k)=-1); -1, \text{ else.} \quad (3)$$

The resultant spread spectrum watermark W_s is stored (i.e., used in the extraction process) and embedded into the 6 highpass sub-band coefficients by using the following rule:

$$\hat{I}(i, j, k) = \tilde{I}(i, j, k) + \alpha * W_s(i, j, k) * \lceil I(i, j, k) \rceil, \quad (4)$$

where: $k = 1, 2, \dots, 6$.

- \hat{I} : are the watermarked real parts of the DT-CWT coefficients.
- \tilde{I} : are the original real parts of the DT-CWT coefficients.
- W_s : is the spread spectrum watermark sequence.
- α : is an intensity parameter of image watermark.

By the inverse DT-CWT, the watermarked sub-image is obtained. Finally, according to the label sequence S_1 (see sect. A), we put every small block of the watermarked sub-image into the original position of the host image. Thus, we get the watermarked image.

D. Watermark extracting

The extraction of watermark in image is an inverse process of embedding scheme. The watermark detection is accomplished without referring to the original image. Only the watermarked image, spread spectrum watermark W_s , and Keys (Key1, Key2, and Key3) need to be used. The watermark extraction algorithm can be summarized as follows:

- 1) The 3-level DT-CWT is performed on watermarked subimage, which is extracted from the watermarked image using "Key1" (see sect. A). \hat{I} denotes the DT-CWT coefficients.
- 2) Constructed the encrypted watermark image \hat{W}_{dk} : for each embed watermark pixel in \hat{I} , its value is compared with those of its eight neighbors; t denotes the total number which the value of the pixel in \hat{I} is larger than its neighbors. Encrypted watermark image can be formed as:

$$\hat{W}_{dk}(i, j, k) = 1, \text{ if } (t \geq 4 \text{ and } W_s(i, j, k)=1) \text{ Or } (t < 4 \text{ and } W_s(i, j, k)=-1); -1, \text{ else.} \quad (5)$$

- 3) Reconstructed watermark image \hat{W} : the 6 parts of the watermark \hat{W}_{dk} are collected under control of secret key "key3", then the original large watermark image \hat{W} can be reconstructed by using the inverse transform of the preprocessing with the secret key "Key2". This can be shown in Fig. 5, where the

original image, the watermarked image, the absolute difference between the original and the watermarked images, the 6 parts of spread spectrum watermark, the 6 parts of extracted encrypted watermark and the reconstructed watermarks with true and false keys. Moreover, if one secret key is changed, the final watermark can not still survive.'

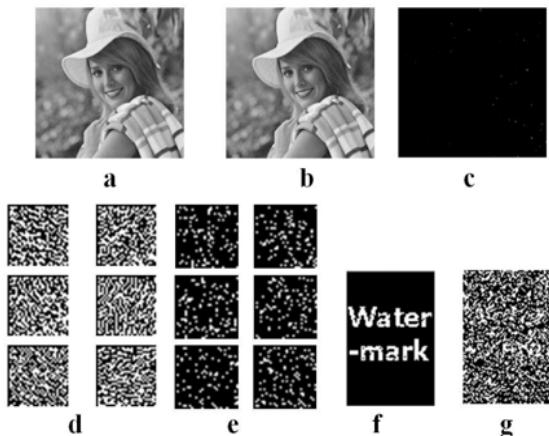


Fig. 5. An example of the mentioned watermark extracting procedure. (a) original image, (b) watermarked image, (c) absolute difference between the original image and the watermarked image, (d) the 6 parts of spread spectrum watermark, (e) the 6 parts of extracted encrypted watermark, and (f)-(g) the reconstructed watermarks with true and false keys, respectively.

IV. CONCLUSION

Proposed method describes robust and semi-blind digital image watermarking in frequency domain, which is computationally efficient. This method applies the Dual Tree Complex Wavelet Transform; the watermark image is encrypted and decomposed into different parts which are adaptively spread spectrum and added into the DT-CWT coefficients. The experimental results have confirmed that this new scheme has high fidelity and it is robust against JPEG compression, Scaling, Affine, Remove lines, PSNR attacks, and signal processing (Salt & pepper, Gaussian noise, and Median filter). The comparison of the proposed scheme [18] shows that our DTCWT approach is more effective.

REFERENCES

- [1] Juergen Seitz. "Digital Watermarking For Digital Media", Information Resources Press Arlington, VA, USA, ISBN 159140519X, 2005.
 - [2] Chun-Shien Lu. "Multimedia Security: Steganography and Digital Watermarking Techniques for Protection of Intellectual Property", Idea Group Publishing, London, ISBN 1591401925, 2004.
 - [3] J.L. Dugelay, S. Roche, C. Rey, G. Dorr, "Still image watermarking robust to local geometric distortions". IEEE transactions on image processing, 2006, 15 N9, 2831-2842.
 - [4] L. O'Gorman, H. Berghel. "Protecting Ownership Rights through Digital Watermarking". IEEE Computer 1996, 29, 101-103.
 - [5] J. R. Hernndez, M. Amado, F. Prez-Gonzlez. "DCT-domain watermarking techniques for still images: Detector performance analysis and a new structure". IEEE Trans. on Image Processing 2000, Special Issue on Image and Video Processing for Digital Libraries, 9(1), 55-68.
 - [6] J. Kusyk, A. M. Eskicioglu. "A Semi-blind Logo Watermarking Scheme for Color Images by Comparison and Modification by Comparison and Modification of DFT Coefficients". Optics East 2005, Multimedia Systems and Applications VIII Conference 2004, 23-26.
 - [7] M. Barni, F. Bartolini, A. Piva. "Improved Wavelet based Watermarking Through Pixel-Wise Masking". IEEE Transactions on Image Processing 2001, 10, 783-791.
- Benita.B(F-23)** received B.E and doing M.E in computer science and engineering from south India ,Anna University of -Tirunelveli. In 2009-2011. Her project work and research Digital Image Processing.

Antenna selection for performance enhancement of MIMO Technology in Wireless LAN

Prof. Rathnakar Achary¹, Dr. V. Vaityanathan², Dr. Pethur Raj Chellaih³, Prof. S. Nagarajan ⁴

¹ Alliance Business Academy Bangalore India

² Dept of CSE, SASTRA University Tanjavour
Tamil Nadu, India

³ Robert Bosch Engineering and Business Solutions (RBEI) Ltd,
Bangalore, 560068, India

⁴ HOD Dept. Of Computer Science Oxford College of Science
Bangalore, India

Abstract

The demand for Wireless Network hardware has experienced phenomenal growth during the past several years. The most commonly known component to enhance the performance of Wireless network is Multiple Input Multiple Output (MIMO) system. It exploits the radio channel by means of multi-path propagation; where the transmitted information bounces off walls, doors and other objects before reaching the receiving antennas multiple times via different routes and with different time delay. MIMO harnesses multi-path with a technique known as space division multiplexing. The transmitting Wireless device actually split a data stream into multiple parts called spatial streams, for transmission. This enhances the data rate. There are two features that focus on improving MIMO performance; called beamforming and diversity. By increasing the number of physical antennas at both transmitter and receiver enhances the performance of MIMO system. To achieve the desired performance the system can be designed in such a way that the number of physical antenna selection at both transmitter and receiver should maximize the performance and optimize the entire system cost.

Keywords: Wireless Network, MIMO, Spatial diversity, QoS and OFDM.

1. Introduction

There are intrinsic differences between a wired network and a wireless network with the ability to instantaneously setup and tear down a network probably being one of the most notable differences. Unlike a wired network, where resources are always available and can be dependent on, a wireless network can make no guarantees, about network resources. Most wireless technologies can support considerably less band-width than that provided by wired

technologies. Available bandwidth is a function of the wireless medium, and the condition of the environment in which the wireless device is deployed. Parameters like distance, fading, delay spread, Doppler-effect, interference by other wireless devices and noise, obstacles, blind spots, atmospheric conditions etc., can change the network behavior unpredictably. Such adverse conditions have to be overcome in order to make QoS viable in a wireless network. The challenge is to provide QoS in a wireless environment, because of the constant change in the property of the wireless media. It is understood that by increasing the number of antenna elements both at the transmitter and the receiver side, the performance and the receiver capacity of the channel can be enhanced. This system is called Multiple Input Multiple Output (MIMO) system. It comprises multiple antennas both at transmission and receiver side.

• Multi-path environment

In the transmission side, MIMO encodes a single high-rate data stream by splitting it and transmitting it across spatially separated antennas as in fig. 1. Having two streams instead of one, enables either the delivery of twice the throughput by keeping the same rate for each of the streams, or extending the reach of the original stream since each of the lower-rate streams can use lower constellations and require a lower Signal-to-Noise Ratio (SNR) to be recovered.

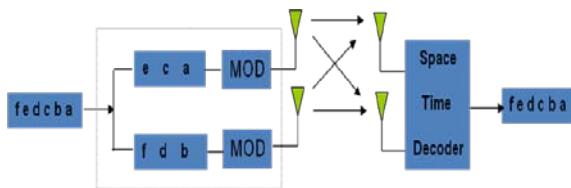


Fig. 1. Data encoding in MIMO system

The implementation cost of a MIMO system increases as the number of antennas increases. If we map the cost versus performance it is clear that, after a specific numbers of transmitting and receiving antennas the performance will remain stable and cost increases. In this paper we analyzed the performance and cost of MIMO system with different combinations of transmitting and receiving antennas.

We organize the rest of this paper as follows; section II provides a background of the need for Quality of Service (QoS) in a WLAN. In section III we explained the role of MIMO in enhancing the performance of the wireless channel. Section IV gives details about the QoS issues in wireless networks and MIMO solution, and section V illustrates the optimal number antennas required for a MIMO system followed with the conclusion.

2. Need for QoS in WLAN

A signal transmitted in a wireless network is susceptible to all kinds of fadings in an underlying channel. This will results in the variation of bandwidth, latency in data delivery in a wireless network and makes QoS guarantees very important. An intelligent adaptive QoS solution could be the answer to mitigate the unpredictable nature of the medium leading to the signal fading and better utilization of the network channel. Some of the practical applications of QoS requirements are; Video streaming, audio streaming, 1394 serial bus transmission, VoIP, wireless home networking, etc. MIMO is one of the technologies to garble these QoS issues. In which using multiple antennas at both transmitter and receiver ends we can exploit the wireless channel much more efficiently. Here multiple data streams are sent from the number of transmitters to the receivers using multiple channels, which potentially enhancing the data rate at the receiver. This MIMO advantage can be achieved without requiring extra bandwidth and power. The performance of MIMO is closely related to the multipath richness of the environment when the system is employed.

3. Multiple Input Multiple Output (MIMO)

- Higher throughput and extended reach

To meet the key requirements such as higher rate, extended range, and better spectral efficiency MIMO

utilizes spatial multiplexing [2] (multiple antennas) on top of orthogonal frequency division multiplexing (OFDM) [8]. Coding the information across both the spatial and spectral domains by using multiple transmit and receive antennas, combined with OFDM modulation on each antenna, increases the diversity and with it the robustness. This enables MIMO to withstand channel impairments such as inter-symbol interference (ISI) and other interferences.

MIMO takes the advantage of multipath propagation to increase throughput, range/coverage, and reliability. Rather than combating multipath signals, MIMO achieves this by sending and receiving more than one data signal in a same radio channel at concurrently to accomplish this.

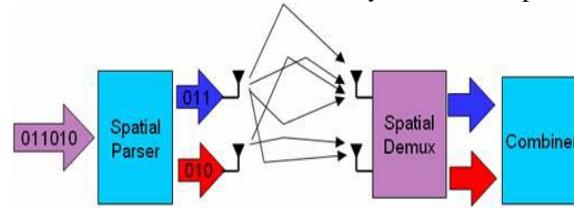


Fig.2. Multipath reflections

Communication using MIMO is the only way known to improve all three basic link performance parameters such as range, speed and reliability.

3.1 Significance of MIMO

In MIMO one coherent radio up-converter and antenna are used to transmit the multiple signals and more than one coherent radio down-converter and antenna receives the multiple signals. Using MIMO, the maximum data rate per channel grows linearly with the number of data streams transmitted in the same channel.

In addition to multiplying data rates within the same channel, properly designed MIMO systems can simultaneously improve coverage and reliability.

Wireless connection using MIMO systems enables increased spectral efficiency and link reliability for a given total transmitted power. Increased capacity is achieved by introducing additional spatial channels, which are exploited using space-time coding [4]. The spatial diversity [2] improves the link reliability by reducing the adverse effects of link fading and shadowing. The choice of coding and the resulting performance improvement are dependent upon the channel phenomenology.

MIMO systems uses an array of transmit and receive antennas for enormous gains in spectral efficiency by exploiting a rich multi-path fading environment.

4. QOS Issues in Wireless Networks and MIMO solutions

The adoption of multiple antenna techniques (MIMO) is expected to enhance the QoS of the wireless channel and the realization of re-configurable, robust and transparent operation across multiple antenna-technology wireless networks.

4.1 Shannon capacity for a wireless channels

If there is a single channel between the transmitter and receiver antenna, which is corrupted by an Additive White Gaussian Noise (AWGN) at a level of SNR denoted by ρ , the channel capacity is represented as;

$$C = \log_2(1 + \rho) \text{ Bits / Sec / Hz} \quad (1)$$

where $\rho = \frac{E_s}{\sigma^2}$ (the channel SNR). The signals in a wireless channels are time varying and subject to random fading. In such a time varying and fading channel the capacity of the channel is $C = \log_2(1 + \rho|h|^2) \text{ Bits / Sec / Hz}$. (2)

Where h is the unit power complex Gaussian amplitude of the channel at an instant observation. This expression gives the capacity of wireless systems with single transmitter and receiver channel; single input single output (SISO) case. The capacity of the SISO channel is;

$C_{SISO} = \log_2 \left\{ 1 + \frac{P}{2\sigma^2} \right\} \text{ Bits/sec/Hz}$. It is clear that capacity takes at time very small value due to fading events.

The statistics can be extracted from the random capacity related with different practical design aspects. The average capacity ' C_a ' average of all occurrences of C gives information on the average data rate offered by the link. The outage capacity C_o is defined as the data rate that can be guaranteed with a high level of capacity for reliable service; $\text{prob}\{C \geq C_o\} = 99.9\%$. Using a MIMO system with multiple antennas at both the ends we can achieve transmit and receive diversity. It results in a significant increase in both C_a and C_o .

4.2 Using multiple antennas - Transmit and Receive diversity

A receive diversity includes 'N' antennas at the receiving end and a single antenna (SIMO) at the transmitter, the channel is now included with 'N' distinct coefficients

$h = [h_1, h_2, \dots, h_N]$, Where h_i is the channel amplitude from the transmitter to the i^{th} receiver antenna, where $i = 1, 2, 3, \dots, N$. The expression for the random capacity can be generalized to,

$$C = \log_2(1 + \rho h h^*) \text{ Bits / Sec / Hz}. \quad (3)$$

Where * denote the transpose conjugate.

In transmit diversity we will have M antennas at the transmitters and one antenna at the receiver [1]. As the number of antennas are varied we can find that there is a reduction in fading and increase in SNR. This clearly indicates that by having transmit and receive diversity with the help of multiple antennas both at transmitting and receive ends will enhance the outage capacity performance, attributable to the spatial diversity effect but this effect saturates with number of antennas.

4.3 Capacity of MIMO link:

The MIMO system considered here is having M transmitters and N receivers. The channel is represented by a matrix of size $[M \times N]$ with random independent element dented by H. The capacity is derived form the expression, $C = \log_2[\det(I_M + \frac{\rho}{N} HH^*)] \quad (4)$

where ρ , is the average SNR at any receiving antenna. The advantage of the MIMO can be significant both in average and outage capacity. For a large number, of antennas where $M=N$, the average capacity increases linearly with M; $C_a \approx M \log_2(1 + \rho)$ In general the capacity will grow proportional with the smallest number of antennas $\min(M, N)$ outside and no longer inside the log function. Therefore in theory and in the case of idealized random channel, limits capacities can be realized, provided we can afford the cost and space of many antennas and RF chains.

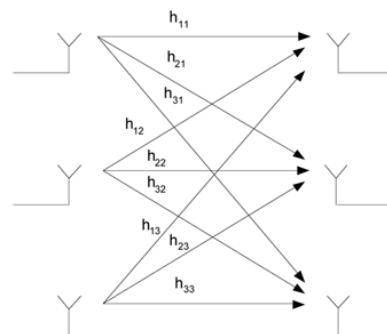


Fig. 3, *MxN channel transmission in MIMO*

Where ‘ x ’ is the channel input at the transmitter and ‘ y ’ is the output at the receiver and n is the noise corresponding to the receive antennas. The channel fading in a non-line of sight (NLOS) link is represented as h and h_{ij} is the fading of the channel corresponding to the path from transmit antenna j to receive antenna i . $y = Hx + n$. This is

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} h_{11} h_{12} \dots h_{1M} \\ h_{21} h_{22} \dots h_{2M} \\ \vdots \\ h_{N1} h_{N2} \dots h_{NM} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \\ \vdots \\ n_N \end{bmatrix} \quad (6)$$

Where

$$\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad \underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix}, \quad \underline{n} = \begin{bmatrix} n_1 \\ n_2 \\ \vdots \\ n_N \end{bmatrix},$$

$$H = \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1M} \\ h_{21} & h_{22} & \dots & h_{2M} \\ \vdots & \vdots & \dots & \vdots \\ h_{N1} & h_{N2} & \dots & h_{NM} \end{bmatrix}.$$

$$y_1 = h_{11}x_1 + h_{12}x_2 + \dots + h_{1M}x_M + n_1$$

$$y_2 = h_{21}x_1 + h_{22}x_2 + \dots + h_{2M}x_M + n_2$$

where

$$y_N = h_{N1}x_1 + h_{N2}x_2 + \dots + h_{NM}x_M + n_N$$

A MIMO system with transmit or receive beamforming [3] we have a full diversity of the order of MN , which results the antenna gain as;

$$\max(M, N) \leq \text{antennagain} \leq MN$$

The MIMO channel capacity expression with $\rho = \frac{E_s}{\sigma^2}$ is;

$$C = E_H \left\{ \log_2 \det \left[I_m + \frac{P}{T} w \right] \right\} \quad (7)$$

where $E_H \{ \cdot \}$ denotes the expectation over H

$$m = \min(M, N) \quad (8)$$

I_m is the $M \times N$ identity matrix. P is the average SNR per receiver antenna. w is given by

$$w = \begin{cases} HH^H & \text{if } N \leq M \\ H^H H & \text{if } M \leq N \end{cases}$$

where the operator H^H indicates the hermitian of matrix H . The capacity of a MIMO channel with M transmit and N receive antennas with respect to the Rayleigh distribution of fading is;

$$C_{MIMO} = \min(M, N) \log_2 \left[1 + \frac{P}{2\sigma^2} \right]. \quad (9)$$

where σ - is the complex Gaussian random variable. The multiplexing gain of a MIMO system, compared with SISO is;

$$\text{multiplexing gain} = \frac{C_{MIMO}}{C_{SISO}} \quad (10)$$

i.e, $m = \min(M, N)$

The antenna combination in MIMO are like 2×1 , 2×2 , 3×2 , 3×3 , 4×1 , 4×2 , 4×4 , what happens as the number of transmitting and receiving antennas are increased linearly?.

5. Bit Error RATE (BER) and performance relations

There are important matrices to measure the overall performance of wireless system using MIMO technology. The equation for throughput calculation is one such measure. It is given as; $\text{Throughput} = R(1 - BER)^L$ where BER is the bit error rate, L is the frame length and R is the retransmission rate.

This provides a way to calculate the throughput for MAC SAP (MAC service access point) according to the IEEE 802.11n specification.

The BER can be minimized with respect to the three variables, namely antenna size, modulation and frame aggregation constant. The value of these weights can be found out through empirical or analytical results, such that it gives us the best trade off between robustness and transmission rate depending on the channel conditions. The noisier the channel more number of antenna is required to maintain the channel to provide better performance, the same BER requirement specified by the upper layers. As the number of antenna increases the diversity order increases thus this result in diversity and spatial multiplexing gains.

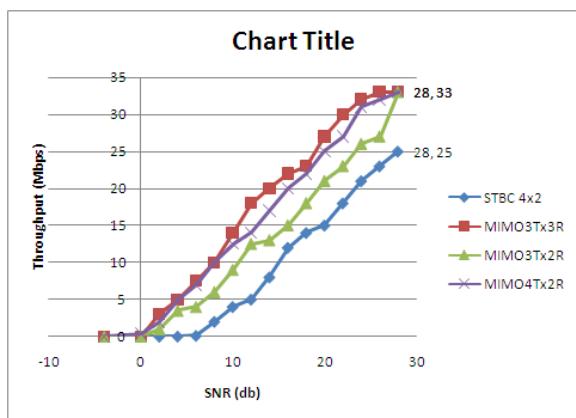


Fig. 4 Throughput variation with PHY layer

From the graph it is observed that given a specific throughput or channel utilization or BER requirement by the higher (PHY) layers. It is possible to predict the physical layer parameters that would match these requirements. Since the physical layer decides upon these parameters dynamically, depending upon the channel conditions, it presents a much smother view of the existing channel conditions it presents a much smoother view of the existing channel condition by the upper layers.

This results in an assured QoS for these layers. A dynamic physical layer also translates to better link layer performance, because of lesser retransmissions. Also in a WLAN as the mobile node (MN) moves away from the base station, the received signal strength reduces. In such noisy channels this scheme can decrease latency and decrease the channel access time. This increases the overall throughput of the WLAN system.

5.1 Optimal number of antennas required for MIMO

In a WLAN with MIMO system the throughout enhancement and latency reduction is possible by considering the physical layer parameters like the number of antenna elements. The numbers of antennas at both the

transmitter and receiver end are varied adaptively based on channel condition to minimize the BER (bit error rate). Depending upon the channel conditions the cross layer entity decides whether or not to increase or decrease the number of antennas. This in terms improves the QoS and hence there is an enhanced performance [11].

The upper layers decide the requirements of the number of antennas. For example the noisier the channel, more number of antennas are required to maintain the same BER. The increase in the number of antennas facilitates more paths between the transmitter and receiver and thus increases the diversity order. This results in the diversity and spatial multiplexing gain. As a result of this the BER reduces for a given SNR. That is BER is inversely proportional to the number of antenna elements.

If the received signal becomes correlated the system will be forced to reduce the number of streams resulting in reduced throughput. In other words MIMO system can have a larger variation in throughput even at the same distance. By increasing the number of received antennas [5] the probability than uncorrelated signal can reach the receiver increases. So, receivers with more number of antennas have higher probability of maintaining higher number of streams, with an increased throughput. Simply adding more antennas does not improve the performance linearly, but rather saturates up the number of uncorrelated signals calculated. Therefore the performance variation of the MIMO system based on the number of antennas both in sending and receiving side is,

$$\cong f_p[s \times (1 - e^{-(M \times N)})] \quad (11)$$

Where s – the number of streams (uncorrelated signals measured by the receiver chains)

M - Number of transmitter chains

N - Number of receiver chains

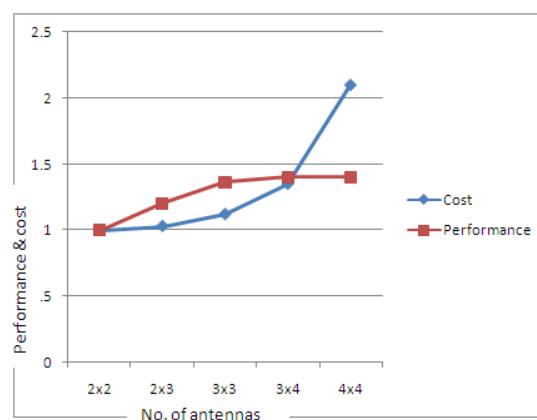


Fig. 5. Cost and performance variation based on Antenna selection

In the above expression it is clear that as we increase the number of transmitting and receiving antennas, the performance increases proportionally. By increasing the number of antenna elements the complexity of the MIMO system architecture increases which in turn increases the cost exponentially while the performance saturates. In fig. 4, a MIMO system with 3x3 antennas the performance curve is above the cost curve. However when the complexity of the MIMO system reaches 3x4 configurations the return on investment is not favorable.

6. Conclusion

MIMO technology is aimed at meeting the challenges of distributing streaming video and audio in a home environment. MIMO technology is at the core of this next-generation communication technology. The use of MIMO enables higher data transmission rates by a factor equal to the number of streams and the ability to establish a wireless connection with NLOS. Better SNR compared to legacy SISO systems can enable developing wireless video solutions to extend reach as compared to legacy approaches.

Additionally, features including new aggregation schemes to improve the MAC efficiency and support the latest QoS standards can also improve the overall performance. These features provide increased throughput, range and robustness in the face of interference, and create an enhanced, reliable user experience. As discussed in this paper, each of these features has its own merits. However, only the complete package combining all the necessary elements can provide the breakthrough abilities required to push the home wireless network to the next level, enabling reliable video delivery over WLAN with appropriate number of transmitting and receiving antennas.

References

- [1] S. Alamouti, "A Simple Transmit Diversity Technique for Wireless Communications," IEEE Journal on Select Areas in Communications, Vol. 16, No. 8, pp. 1451-1458, October 1998.
- [2] J. Choi, B. Mondal, and R. W. Heath, Jr., "Interpolation Based Unitary Precoding for Spatial Multiplexing MIMO-OFDM With Limited Feedback," IEEE Trans. Signal Processing, vol. 54, no. 12, pp. 4730-4740, Dec. 2006.
- [3] J. Choi and R. W. Heath, Jr., "Interpolation based transmit beamforming for MIMO-OFDM with limited feedback," IEEE Trans. Signal Processing, vol. 53, no. 11, pp. 4125-4135, Nov. 2005.
- [4] D. Gesbert, M. Shafi, D.-S. Shiu, P. J. Smith, and A. Naguib. From theory to practice: An overview of MIMO space time coded wireless systems. IEEE JSAC, 21(3), 2003
- [5] M. Gharavi-Alkhansari and A. B. Gershman, "Fast antenna subset selection in MIMO systems," IEEE Transactions on Signal Processing, vol. 52, no. 2, pp. 339-347, 2004.
- [6] M. Conti, G. Maselli, G. Turi, and S. Giordano, "Cross-layering in mobile ad hoc network design," IEEE Computer, vol. 37, pp. 48-51, 2004.
- [7] D. A. Gore, R. U. Nabar, and A. J. Paulraj, "Selecting an optimal set of transmit antennas for a low rank matrix channel," in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00), vol. 5, pp. 2785-2788, Istanbul, Turkey, June 2000.
- [8] H. Heiskala and J. Terry, "OFDM Wireless LANs: A Theoretical and Practical Guide," SAMS, 2002.
- [9] P. D. Karamalis, N. D. Skentos, and A. G. Kanatas, "Adaptive antenna subarray formation for MIMO systems," IEEE Transactions on Wireless Communications, vol. 5, no. 11, pp. 2977-2982, 2006
- [10] R. Monzingo and T. Miller, Introduction to Adaptive Arrays, Scitech Publishing, Raleigh, NC, 2004.
- [11] Rathnakar Acharya, Vaithianathan. V, Pethur Raj Chelliah, "Performance Enhancement of WLAN using 802.11n and MIMO Technology" www.tifac.vewlammal.org/compc/articles/34.pdf.
- [12] M. Simon, S. Hinedi, and W. Lindsey, Digital Communication Techniques Signal Design and Detection, Prentice Hall, Englewood Cliffs, NJ, 1995.
- [13] S. Toumpis and A.J. Goldsmith, "Capacity Regions For Wireless Ad hoc Networks,"
- [13] IEEE International Conference on communications (ICC), 2002

Authors

Dr. V. Vaithianathan, is a Professor in the Department of Computer Science and Engineering at SASTRA University Tanjavur India.



Prof. Rathnakar Acharya Graduate in Electrical and Electronic Engg, from Mysore University and Post Graduate in Technology and Management. Having 20 years, of experience in teaching & research. Presently pursuing research (PhD) in QoS issues in WLAN.



Dr. Pethur Raj Chelliah, PhD from Anna University, India, worked as a research associate at the Dept. of Computer Science and Automation at IISc. Bangalore and then Postdoctoral research at Japan (Nagoya Institute of Technology Kyoto University and University of Tsukuba) currently working as Lead Architect at Robert Bosch Engineering and Business Solutions (RBEI) Ltd, Bangalore, 560068, India



Mr. Nagarajan S is presently working as Associate Professor and Head, at the Oxford College of Science, Bangalore. He is also a Research Scholar at Bharathiar University at Coimbatore. He has nearly about 13 years of Industry and teaching experience. He has published two international papers in International Journals and 5 in various conferences

A Fuzzy-Ontology Based Information Retrieval System for Relevant Feedback

Comfort T. Akinribido¹, Babajide S. Afolabi², Bernard I. Akhigbe³ and Ifiok J. Udo⁴

Information Storage and Retrieval Group (ISRG),
Department of Computer Science and Engineering,
Obafemi Awolowo University, Ile-Ife, Osun State, Nigeria.

Abstract

Obtaining correct and relevant information at the right time to user's query is quite a difficult task. This becomes even complex, if the query terms have many meanings and occur in different varieties of domain. This paper presents a fuzzy-ontology based information retrieval system that determine the semantic equivalence between terms in a query and terms in a document by relating the synonyms of query terms with those of document terms. Hence, documents could be retrieved based on the meaning of query terms. The challenge has been that surface form does not sufficiently retrieve relevant document to user's query. However, the results presented showed that the Fuzzy-Ontology Information Retrieval system successfully retrieve relevant documents to user's query. This is irrespective of different meaning and varieties of domain. The System was tested on words with different meanings and some set of user's query from varied domains.

Keywords: *Information Retrieval, Synset, Probability Corpus Relevance, Term Frequency, Fuzzy techniques.*

1. Introduction

Fuzzy-Ontology Information retrieval system (FOIRS) is a system that typically measures the relevance of documents to users' query based on meaning of dominant words in each document. The weight of this dominant word in terms of both surface form (which is the matching of query terms with document terms) and domain concept should be measured according to their frequency and threshold values. Each document has a domain concept and target word that is elaborated and emphasized in the document. This word is mostly not repeated with the same spelling to prevent repetition or tautology, but it is written in different forms but with the same meaning. Hence, surface form can not only be used to determine the relevance of a document to user's query.

[12] states that context distance model compares the similarity of the contexts where a word appears, using the

local document information and the global lexical co-occurrence information derived from the entire set of documents to be retrieved. The system must be able to adapt its behaviour autonomously to the changing context, expand the query to get synonyms of the query terms and originate the search task. Then, it can filter the results from irrelevant documents, organize them, and present them as useful information to the users in their current activities. [7] reported that specifying the context of a search can significantly improve search results. Thus, assess the context of any search is an important task not to be ignored. For this, they developed a probability dominant meaning space and context vector.

According to [5], FOIRS greatly improves retrieval effectiveness by expanding the query which can be a single word, keywords or longer phrase. The query terms can be expanded through a database that contains keywords and their synonyms. In the work term frequency and corpus relevance of words, which make up user's query were determined. This they did to tackle the challenge of word polysemy in document retrieval so that words can correlate better based on their meanings rather than on their surface forms.

Also, surface form representation of query terms does not sufficiently retrieve document that is relevant to the user's query. For instance, words like bank can occur in the context of river bank and financial bank; close (of door) and close meaning near; bat (the name of a bird) and bat used in sports (e.g. in baseball); wood (for firewood) and wood for the name of a person; caterpillar can mean a heavy equipment or a worm that would later develop into a butterfly. Words in these forms normally cause the retrieval of irrelevant documents as feedback to user's query, if surface form is applied.

In literature Boolean Information retrieval system (BIRS) allows the relevance of a document to be determined as relevant or non relevant; that is as (0 and 1). This in real sense does not specify any Grade of

relevance. Also the Vector space model (VSM) for Information retrieval (IR) computes the weight of query terms for each document. However, it does not relate the meaning of words in a document with meaning of words in user's query, as well as determine the corpus relevance of the query terms. This limitation is often the case with the use of VSM for IR [15]. [12] only proposed an approach that uses the local context vector analysis. This contains occurrences of query terms based on surface form. This limitation would have been well addressed through the introduction of fuzzy techniques as suggested in this study.

Also, while [9] used the meaning of queries (that is; ontology) to disambiguate queries, and did not use the context distance to determine the grade of relevance; [2] shared how useful the introduction of the use of ontology can be in IR, but did not develop any IR application using ontology concepts. Likewise, [6] only discussed the concept of contextual retrieval (CR). The claim made was that it combines both search technologies as well as the knowledge about query and user context. With CR is able to provide the most appropriate answer for a user's information need. The provided query if expanded using ontology without any further interaction from the user [1], would provide relevant feedback. Thus, the system would be a useful IR system that can be applicable in different areas like Intelligent Distance Learning Environment.

[8] In his work only discussed issues concerning satisfaction and frustration metrics. He reported that while the satisfaction metric takes into account only relevant documents, the frustration metrics concerns non-relevant documents. The main weakness of the IR models discussed so far is in the way they represent a document. That is as a 'bag of words'. However, search engines can only find words, which have been indexed. Therefore, developers must have it at the back of their mind that the author of a document may have used other words in the same context. These words must be synonyms.

[12] Reported that an increasing number of approaches to Information Retrieval have been proposed using models that are based on concepts rather than on keywords. But in this work, the concept of Ontology was used and thus defined as objects with two fields (Keywords and Synset). For each Keyword the corresponding synset was obtained and related to words in a document. The purpose was to search for new documents that semantically correlate to user's query. Thus, with a tolerance for imprecision and a positive use of fuzzy logic, the ranking of retrieved documents in order of relevance was enhanced.

Finally, relating the meaning of terms in user's query to document concept was sufficiently taken care of using the proposed methodology for this work. Consequently, the

Fuzzy-Ontology concept used is discussed in section 2. While section 3 contains the proposed architecture and the algorithm for the system (FOIRS); both the implementation of the system and results are discussed in section 4 and 5 respectively.

2. Relating the Meaning of Terms in User's Query to Document Concept

Fuzzy-Ontology allows the easy determination of the precise meaning of a word as it relates to a document collection. [3] stated that Fuzzy-Ontology could be used in IR to locate precise information, which may be contained in a document content collection. Also, concepts represent a single sense, which is a set of synonyms called synset. Since a word is assumed to have a fixed number of senses as defined in the lexicon, such as WordNet (Thesaurus), the semantic similarity between the query terms is determined by incorporating a database that contains a dictionary of synonyms into the IR system.

2.1 Representing Ontology Properties

The concept of Ontology was used to describe the meaning of query terms by getting the synonyms of all the keywords that make up the user's query. The set of synonyms of keywords in the query is called synset. Ontology was represented by objects stored in a database with two fields (keywords and synset, which is a set of synonyms for the keywords). For each keyword the corresponding synset was obtained and then related with words in a document. The rationale for this is that most writers prefer to change words in documents without omitting the main content. Instead they use words that have the same meaning as the main content. The synset and the query terms were therefore matched with the document terms to calculate the term frequency and corpus relevance.

2.2 Term Frequency and Probability Corpus Relevance

Term frequency (TF) is the number of occurrences of the query terms in each document. It was improved by first getting the target word from query. Then the frequency of each word in a query, which appeared in the context of the target word, was divided by the frequency of the target word in each document. This was important since for instance, the document for Financial Bank would have a number of financial terms/issues than river terms (which could be assumed to be river bank). Thus, the expected document to be retrieved will not be for river bank. Current search engines do not have this technique, hence they retrieve both relevant and irrelevant documents provided they have same spelling. The emphasis therefore has been on the number of occurrences of query terms, which a

search engine matches the queries it receives against the index they create. The index consists of words in each documents, plus pointers to their locations with the documents [4], [5].

2. 2. 1 Probability of Corpus Relevance

A Corpus Relevance is how far a word is closely associated in the context of other word. For instance words like wheat, grains, cereal, and corn are mostly found in the same document. The Probability of Corpus Relevance of the target words and each word that make up the user's query was pre-computed as:

for i = 0 to p

$$R(W_q, W_{i+1}) = \text{FID}(W_q, W_{i+1}) / (\text{FID}(W_q) + \text{FID}(W_{i+1}) + \text{FID}(W_q, W_{i+1}))$$

next i.

As a result, the Probability Corpus Relevance =

$$\frac{1}{p} \left[\sum_{v=1}^p \frac{R(W_q, W_{i+1})}{R_c} \right]$$

where;

FID = frequency in document

i = position of each word in the query

R_c = Maximum corpus relevance

W = number of corpus relevance

q = number where the target word belong in the query

P = total number of word in a query

It is interesting to note that some sample word pairs well with high corpus relevance scores, while others with low corpus relevance scores. Also, important is the need to get the Corpus Relevance of words that make up user's query. This was necessary, since if the query terms have high Corpus Relevance to a document, the document will be adjudged relevant to the query and vice versa. Similarly, the Corpus Relevance was also used to obtain the weight of query terms as well as the synset in the sample document.

2. 3 Ranking Using Fuzzy Concepts

Fuzzy techniques can be used to avoid rigid definitions and to manage uncertainty in hierarchical representations of concepts and in matching processes [11]. Therefore, fuzzy techniques were applied using term frequency and Corpus Relevance result to rank relevant document in order of relevance with specified threshold value. The technique was also used to rank every sample retrieved

document in order of relevance. This was necessary, since if only relevant documents that satisfy the user's query are retrieved, users will be prevented from the burden of reading through many pages to get what they really needed. The easy applicability of the Fuzzy rule was possible, since the Probability Corpus Relevance had been achieved. Thus, if it is high then the relevance of the document to the query will be high and vice versa.

Consequently, User's preference or choice of (and access to) relevant document would be easy and precise through the use of fuzzy techniques for efficient ranking. For instance, a document that is 90% relevant will be retrieved before a document that is 80% relevant and so on. The overall implication of this that the time spent in trying to locate relevant document as mentioned earlier will be reduced.

The first step in applying the concept of Fuzzy was to determine the fuzzy set (Probability Corpus Relevance and Relevant). While the Probability Corpus Relevance was used as the fuzzy input variable, Relevant was used as the fuzzy output variable. See Table 1 and 2 below.

Table1: Membership Function of the Fuzzy-Ontology Information Retrieval System (Fuzzy Input Variable)

| Fuzzy Input Variable | Membership Function |
|----------------------|---------------------|
| Probability | High |
| Corpus Relevance | Medium |
| | Low |

Table 2: Membership Function of the Fuzzy-Ontology Information Retrieval System

| Fuzzy Output Variable | Membership Function |
|-----------------------|---------------------|
| Relevant | High |
| | Medium |
| | Low |

In order to get the Grade of relevance of retrieved document, which is the strength (advantage) of this system (FOIRS), the following was adopted:

- (i) First, if the degree of membership of one of the retrieved document is 0.7, then the document is highly relevant;
- (ii) secondly, if it is 0.5, then the document is moderately relevant; and
- (iii) thirdly, if the membership function is 0.1, then the document is not relevant.

Unlike the FOIRS, others like the BIRS will only categorize the document to be retrieved as (0.7, 0.5 and 0.1), which means (Relevant, Relevant, and not

Relevant) respectively. A second weakness with the system (BIRS), like others is that it scatters the result (retrieved documents) all over the result page. Thus making it very cumbersome for users to read through and fish out the most relevant feedback (document, which satisfy their information need).

3. The System Architecture

The diagram in figure 3.0 below is a pictorial representation of the FOIRS System Architecture.

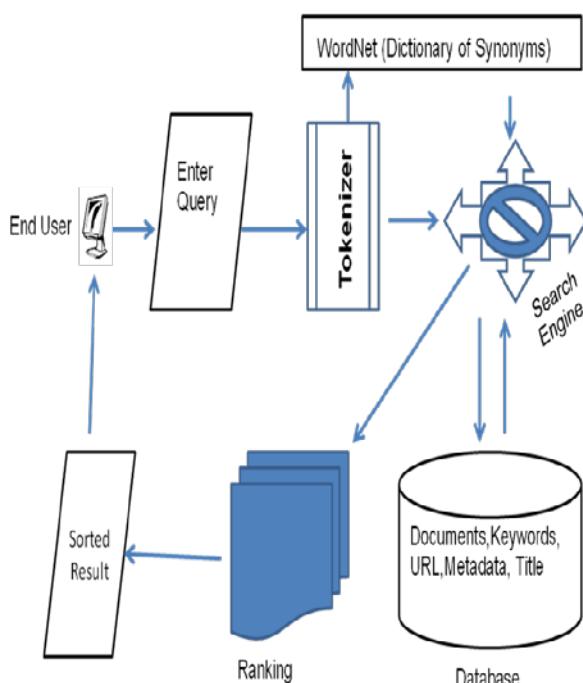


Fig. 3.0: The FOIRS System Architecture

From figure 3.0, when an end user types his/ her query through the text field (provided for user's query entry), the tokenizer divides the user's query into words. The crawler or search engine then locates the document's directories (URL), keywords, metadata, and the title of document. The query of an end user that is divided into tokens is matched with the corresponding document terms, keywords, metadata, and title in the database. The real documents whose keywords, metadata and title match the tokens of the end user's query are then retrieved to the document's directories.

A second responsibility of the search engine is to count the total number of terms in each document. Then it determines the frequency of occurrence of the tokens of the query in each document, which is the term frequency. The Synset of the tokens that make up the query was obtained using the Java WordNet library. Two parameters: The term frequency and the synset were used to calculate the Probability of corpus relevance. Also, using the fuzzy technique, the search engine pick the retrieved documents that are related to the query and ranked them based on a specified thresholds value. After the ranking the results are sorted into a list of relevant document that are arranged in a hierarchical order, and displayed in the user interface of the end-user.

3.1 Algorithm

- Step 1:** Get target word from textbox or Input box.
- Step 2:** Get the query from textbox or Input Box
- Step 3:** Break it into tokens or words
- Step 4:** Match each query term with metadata, title, keywords of document and each document term to determine their frequency.
- Step 5:** Find query terms' contextual meaning in Word Net Library by getting corresponding synset for each query terms in MySQL with database for Dictionary of Synonyms. Thus the number of synset of words in the query that appear in each document will be determined
- Step 6:** Determine Term Frequency for each query terms and their synset in the collection
- Step 7:** Find Probability Corpus Relevance.
- Step 8:** The URL of the relevant document is obtained and stored in database
- Step 9:** The probability of corpus relevance in each metadata, title, keywords and in the words that compose the document is obtained and used for ranking based on a threshold value.
- Step 10:** Display relevant document in the list box according to their level of relevance in a hierarchical order.
- Step 11:** The title of the document is linked using hyperlink to the URL of the relevant document
- Step 12:** Finally, click the title and see the documents

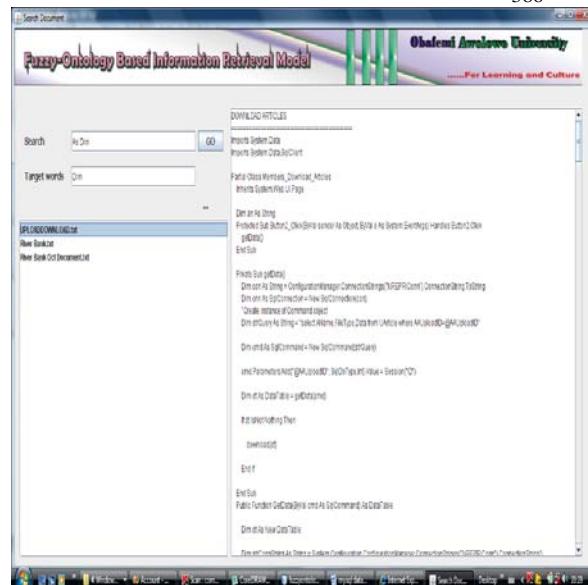


Fig. 4.1: The Graphical User Interface of FOIRS

4. Implementation

The database for the system (FOIRS) was implemented using MySQL. The database shown in figure 4.0 below contains dictionary of synonyms.

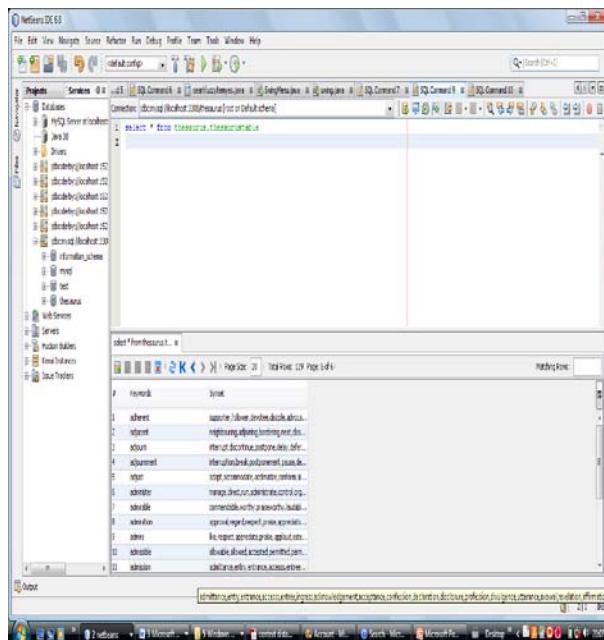


Fig. 4.0: The Database of FOIRS

Also, while Java programming language was used to develop and design the system as shown in figure 4.1 below; the Graphical User Interface was achieved using NetBeans 6.8.

4.1 System Evaluation

The system was evaluated using selected set of homonyms (words with the same spelling but different meanings). The FOIRS demonstrated the ability to strictly retrieve documents that are relevant to the specific meaning of the user's query and rank relevant document in order of relevance. Unlike FOIRS, current search engines retrieve both relevant and irrelevant document to the query of the user. This happens provided they have the same spelling with the query of the user. Two parameters: Satisfaction and frustration metrics were used for the evaluation of the system. The resultant feedback was compared with that of Google search engine. While the satisfaction metrics take into consideration only relevant documents, the frustration metrics considered non-relevant documents.

5. Result

The Table 3, presented below is used to indicate the results of some homonyms and samples of user's query tested on both FOIRS and GOOGLE.

Table 3: Results of Homonyms and User's query on both FOIRS and GOOGLE

| Homo-nyms | Sample of User's Query | Frustration (Metrics) | | Satisfaction (Metrics) | |
|-----------|------------------------|-----------------------|---|------------------------|---|
| F(FOIRS) | and G(GOOGLE) | F | G | F | G |
| Bank | Financial Bank | 1 | 5 | 11 | 7 |
| | River Bank | 0 | 4 | 7 | 3 |

| | | | | | |
|-------------------------------------|-------------------------|---------------|---------------|---------------|---|
| Bat | Bat for baseball | 0 | 6 | 10 | 4 |
| | Vampire bat | 0 | 3 | 8 | 5 |
| Cater-pillar | Caterpillar sandal | 0 | 10 | 15 | 5 |
| | Caterpillar butterfly | 0 | 9 | 17 | 8 |
| Punch | Punch a paper | 4 | 7 | 9 | 6 |
| | Punch Newspaper | 4 | 4 | 8 | 8 |
| | Weed Management | 5 | 6 | 7 | 6 |
| | Organic crop production | 1 | 4 | 7 | 4 |
| Percentage in total document | 13.2 % | 50.8 % | 86.8 % | 49.1 % | |

As shown in table 3; some sample of user's query, such as weed management and organic crop production, which contain hyponyms but not stated under the homonyms column were entered for both FOIRS and GOOGLE. The purpose was to avoid preempting the system's ability to measure up with other systems, in terms of retrieval quality and strength. Thus the irrelevant documents retrieved are as indicated under frustration metrics, while the relevant documents are stated under the satisfaction metric. The percentages of both irrelevant and relevant documents are shown in the last row of table 3. Under the frustration and satisfaction columns; F and G is used to represent FOIRS and GOOGLE respectively. Thus, the result from the table and under satisfaction metrics indicate that the percentage of relevant documents retrieved with FOIRS is 86.8%, while that of GOOGLE is 49.1%.

This result is further buttressed using the graphs in figure 5.0 and 5.1 respectively below. Therefore, the graph in figure 5.0 below shows the relationship between the irrelevant documents and the relevant ones retrieved in GOOGLE.

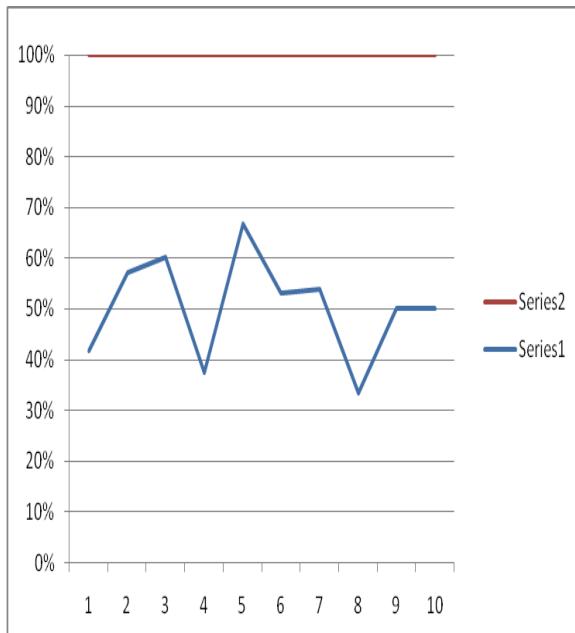


Fig. 5.0: The relationship of irrelevant and relevant documents retrieved in GOOGLE.

Also, the graph in figure 5.1 below shows the relationship between the irrelevant documents and the relevant ones retrieved in FOIRS.

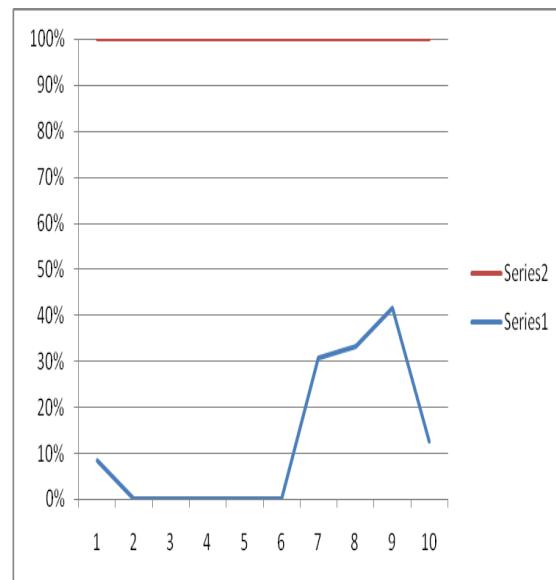


Fig. 5.1: The relationship of irrelevant and relevant documents retrieved in FOIRS.

In summary, both the relevant and irrelevant documents retrieved and as shown by the graph in figure 5.0 above is quite significant. That is, the ratio of 100% to almost 70% cannot be ignored. But from the graph in figure 5.1, the ratio of 100% and a little above 40% can be easily ignored. Thus, the ratio of irrelevant document retrieved is not significant as compared to the ratio of relevant document retrieved. This shows that there is significant improvement in the retrieval of relevant document to user's query, when FOIRS is used.

6. Conclusion

For a retrieved document to be really relevant most of the words added to the query must be related to the search context. Also, the retrieval quality of final results is likely to be high when context-based approach (CA) is applied to the design and implementation of IR systems. This could also bring some improvements in the retrieval of document with relevant feedback. CA was achieved through the introduction of fuzzy logic as proposed in this work. Thus, the use of fuzzy techniques in IR system according to the results reported, confirms that very good a result, more reasonable and satisfactory results in response to user's query will be ensured.

7. References

- [1] Ciorascu, C., Ciorascu, I., & Stoffel, K.(2003) Knowler-Ontological Support for Information Retrieval Systems Proceedings of SIGIR 2003 Conference, Work-shop on Semantic Web, Toronto, Canada.
- [2] Darius Strasunskas and Stein L. Tomasscu, (2006); Department of Computer and Information Science, Norwegian University of science and Technol-ogy,NO_7491 Trondheim,Norway
- [3] Dwi H. Widayantoro,(2001)," A Fuzzy-Ontology Based Abstract Search Engine and its User Studies, Department of Computer science Texas A &M University College Station,TX 77843-3112,USA
- [4] Liddy Elizabeth, (2005)," How a Search Engine Works", Director of Center for Natural Language Processing Professor, School of Information Studies, Syracuse University
- [5] Salton G., Wong A. and Yang C.S (1975)," A Vector Space Model for Automatic Indexing "Commun-ications of the ACM, vol.18, nr.11, pages 613-6230
- [6] Allan James (editor) et al (2002),"Challenges in information Retrieval and Language Modeling" Report of a workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst
- [7] Mohammed A. Razek, Claude Frasson, Marc Kaltenbach, (2003) " A Context-Based Information Agent for Supporting Intelligent Distance Learning Environments, Budapest, Hungary.
- [8] Korfhage, R.(1993). Information Storage and Retrieval Morgan kayfmann Publishers.
- [9] Nagypal, G (2005), "Improving Information Retrieval Effectiveness by Using Domain Knowledge Stored in Ontologies," OTM Workshops 2005,LNCS 3762, Springer-Verlag, 780-789.
- [10] Stefania Gallora(2007)" Fuzzy Ontology and Information Access on the web" Technical University of Kosice
- [11] Tzoukermann Evelyne, Hongyan Jing, (2003)."Content Distance and Morphology Approach in Information Retrieval", Columbia University.
- [12] Xu, J. and Croft W. B.(2008)" Improving the effectiveness of information retrieval with local context analysis" ACM Transactions on Information Systems (TOIS), Vol. 18, No.1.
- [13] Yi- Chun Liao, (2007) "A weight -Based approach to information retrieval and relevance feedback" Hsuan Chuang University.
- [14] Rubens N.O.,(2006) " The Application of Fuzzy Logic to the Construction of the ranking function of Information retrieval Systems; University of Massachusetts, Department of Computer Science. Computer Modeling and New Technologies, Vol 10, No.1, 20-27
- [15] Manning, C.D., Raghavan, P., Schütze, H. (2009). An Introduction to Information Retrieval. Cambridge University Press Cambridge, England. Retrieved from <http://www.nlp.stanford.edu/IR-book/pdf/00front.pdf>

7. Biography of Authors

Akinbirido C.T. studied Computer Science at Adekunle Ajasin University, Akungba-Akoko in Ondo State. She obtained Second Class Upper Division. She is currently on her M.Sc degree programme in Computer Science and Engineering in Obafemi Awolowo University, Ile-Ife. Nigeria. Her areas of interest are Information Retrieval, Artificial Intelligence, Database Organization and Operation Research.

Afolabi, B.S. (Ph. D)

He is a Senior Lecturer in the Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife. Nigeria and head of the Information Storage and Retrieval Group research team.

Akhigbe, B.I

He is a member of Information Storage and Retrieval Group in the Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife. Nigeria. He has both B.Sc and M.Sc in Computer Science.

UDO Ifiock James

Information Storage and Retrieval Group (ISRG), Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife. Nigeria.

He obtained his B.Sc in Computer Science from University of Calabar, Calabar in 2005. Udo is currently on his M.Sc degree at Obafemi Awolowo University, Ile-Ife. His area of specialization is Information system design and Data reduction in Very Large Databases (VLDB).

Distortion Analysis Of Tamil Language Characters Recognition

Gowri.N¹,

R. Bhaskaran²,

1. T.B.A.K. College for Women, Kilakarai,

2. School Of Mathematics, Madurai Kamaraj University,
Madurai.

ABSTRACT

This research work demonstrates how character recognition can be done with a back propagation network and shows how to implement this using the MATLAB Neural Network toolbox. This is a slightly enhanced version of the character recognition application based on the MATLAB Neural Network toolbox. In this research article we are focusing on the distortion analysis of Tamil Language Characters in order to recognize them effectively using the neural network we have developed. We have used the commonly used representation method for recognizing digits and uppercase English letters as suggested by Guyon et.al [1] to start with and built a method for Tamil language letters over it.

KEYWORDS: Neural Network, Natural Language Characters recognition , character shape representation, character classification, Pattern Recognition, Tamil Language Characters, Tamil Desktop Publishing, Feed forward network, Distortion

1. Introduction

The word "recognition" plays an important role in our lives and understanding actually begins with proper recognition. It is a basic property of all human beings; when a person sees an object, he or she first attempts to gather all possible information about the object and compare its properties and behaviors with the existing knowledge stored in the mind. If one finds a proper match, one recognizes it. If no match is found, it is recognized as a new object. However, when pattern recognition has to be employed for identification of fixed and known set of patterns in a commercial or research environment, it is more appropriate to have an automated way. A machine that reads checks in a bank could do the job several times faster than human beings. In effect, machines that can read symbols could provide cost effective alternative to manual methods. This kind of

application saves time and money, less error prone, and relieves human labor to work on other interesting jobs rather than doing a mechanical boring job.

A neural network is an information processing system. It consists of massive simple processing units with a high degree of interconnection between each unit. The processing units work cooperatively with each other and achieve massive parallel distributed processing. The design and function of neural networks simulate some functionality of biological brains and neural systems. The advantages of neural networks are their adaptive-learning, self-organization and fault-tolerance capabilities. For these outstanding capabilities, neural networks are used for pattern recognition applications. Some of the best known neural models are back-propagation, high-order nets, time-delay neural networks and recurrent nets.

2. Design of the Neural Network

The twenty-six by thirty five-element input vectors are defined in the MATLAB script file as a matrix of input vectors called alphabet. The target vectors are also defined in this file with variable called targets. Each target vector is a 26-element vector with a 1 in the position of the letter it represents, and 0's in other positions. For example, the letter "C" is to be represented by a 1 in the third element (as "C" is the third letter of the alphabet), and 0's in other positions. The network receives the 5×7 real values as a 35-element input vector. It is then required to identify the letter by responding with a 26-element output vector. The 26 elements of the output vector each represent a letter. To operate correctly, the network should respond with a 1 in the position of the letter being presented to the network. All other values in the output vector should be 0. In addition, the

network should be able to handle noise. In practice, the network does not receive a perfect letter as input. Specifically, the network should make as few mistakes as possible when classifying vectors with noise of mean 0 and standard deviation of 0.2 or less. We have used the commonly used representation method for recognizing digits and uppercase English letters as suggested by Guyon et.al [1]

3. Network Architecture

Feed-forward networks often have one or more hidden layers of sigmoid neurons followed by an output layer of linear neurons. Multiple layers of neurons with nonlinear transfer functions allow the network to learn nonlinear and linear relationships between input and output vectors. The linear output layer lets the network produce values outside the range -1 to +1.

The network is a two-layer network. The neural network needs 35 inputs and 26 neurons in its output layer to identify the letters. The ***log-sigmoid*** transfer function at the output layer was picked because its output range (0 to 1) is perfect for learning to output Boolean values. The hidden layer has 10 neurons. This number was picked by through several tests by changing the number of neurons and also the number of epochs and by changing the sum squared error to enhance learning.

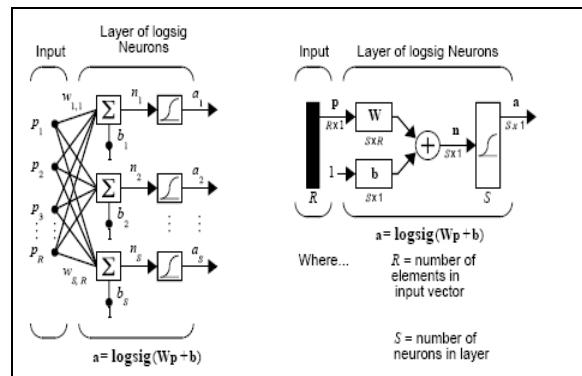


Fig1: Architecture of the network

The network is trained to output a 1 in the correct position of the output vector and to fill the rest of the output vector with 0's. However, noisy input vectors may result in the network not creating perfect 1's and 0's. After the training, the output from the network is passed through the competitive transfer function ***compet***. This makes sure that the output corresponding to the letter most like the noisy input

vector takes on a value of 1, and all others have a value of 0. The result of this post-processing is the output that is actually used.

3.1 Initialization

The two-layer network is created with newff with logsig function and by creating a set of target and input vectors.

3.2 Network Training

In order to create a network that can handle noisy input vectors it is imperative to train the network on both ideal and noisy vectors. To do this, the network is first trained on ideal vectors until it has a low sum-squared error. Then, the network is trained on 10 sets of ideal and noisy vectors each. The network is trained on two copies of the noise-free alphabet at the same time as it is trained on each noisy vector. The two copies of the noise-free alphabet are used to maintain the network's ability to classify ideal input vectors.

However after, the training described above the network may have learned to classify some difficult noisy vectors at the expense of properly classifying a noise-free vector. Therefore, the network is again trained on just ideal vectors. This ensures that the network responds perfectly when presented with an ideal letter. All training is done using back propagation with both adaptive learning rate and momentum with the function ***traingdx***.

3.3 Training without Noise

The network is initially trained without noise for a maximum of 5000 epochs or until the network sum-squared error falls beneath 0.1.

3.4 Training with Noise

To obtain a network not sensitive to noise, we trained with two ideal copies and two noisy copies of the vectors in alphabet. The target vectors consist of four copies of the vectors in target. The noisy vectors have noise of Standard Deviation 0.1 and 0.2 added to them. This forces the neuron to learn how to properly identify noisy letters, while requiring that it can still respond well to ideal vectors. To train with noise, the maximum number of epochs is reduced to 300 and the error goal is increased to 0.6, reflecting that higher error is expected because more vectors (including some with noise), are being presented.

3.5 Training Again Without Noise

Once the network is trained with noise, it makes for good judgment to train it without noise once more to ensure that the system goes back to its original state and ideal input vectors are always classified correctly. Therefore, the network is again trained with code identical to the Training without Noise section.

3.6 Estimating the System Performance

The reliability of the neural network pattern recognition system is measured by testing the network with hundreds of input vectors with varying quantities of noise. The network is tested with a noisy version of the Tamil Language Characters. The MATLAB script file tests the network at various noise levels, and then graphs the percentage of network errors versus noise. Noise with a mean of 0 and a standard deviation from 0 to 0.5 is added to input vectors. At each noise level, 100 presentations of different noisy versions of each letter are made and the network's output is calculated. The output is then passed through the competitive transfer function so that only one of the 26 outputs (representing the letters of the alphabet), has a value of 1. The number of erroneous classifications is then added and percentages are obtained. The solid line on the graph of Fig 6 and Fig 7 shows the reliability for the network trained with and without noise. The reliability of the same network when it had only been trained without noise is shown with a dashed line. Thus, training the network on noisy input vectors greatly reduces its errors when it has to classify noisy vectors. The network did not make any errors for vectors with noise of standard deviation 0.0 or 0.05. When noise of standard deviation 0.2 was added to the vectors both networks began making errors.

For example we create a noisy version (Standard Deviation 0.2) of the Tamil Language Characters. The Network is trained with distortions and a testing of networks with noise level of 0.00 to 0.50 in steps of 0.05 is done for all the input characters of the input character given. The Fig 2 and Fig 4 are the result of distorted characters of Tamil language with a maximum sum squared error output of 0.5. Fig 3 and Fig 5 is the result with a maximum distortion of 0.5 given as a test input and the character recognized by the network.

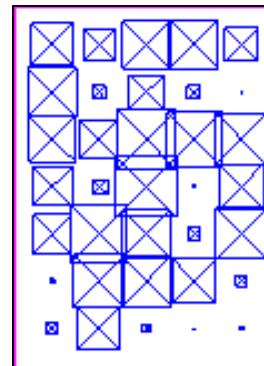


Fig 2: Tamil letter 'tha' distorted

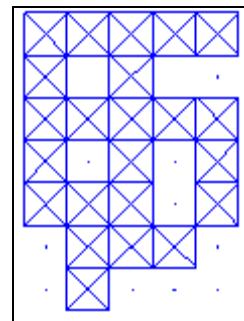


Fig 3: Tamil letter 'tha' recognized

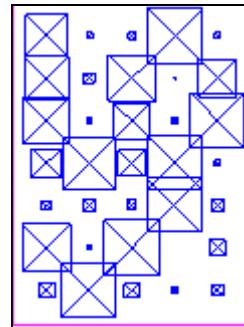


Fig 4.Tamil letter 'zha' distorted

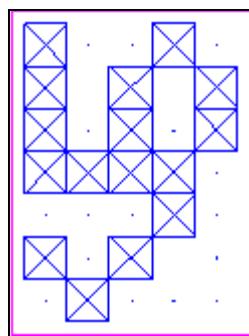


Fig 5.Tamil letter 'zha'recognized

The graphs in Fig 7 and Fig 8 show the comparison of percentage of recognition errors of training the network with and without distortion respectively and for varying values of number of epochs.

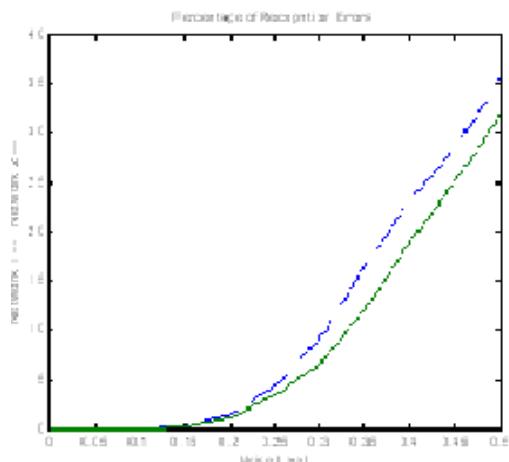


Fig 6: Percentage of recognition errors

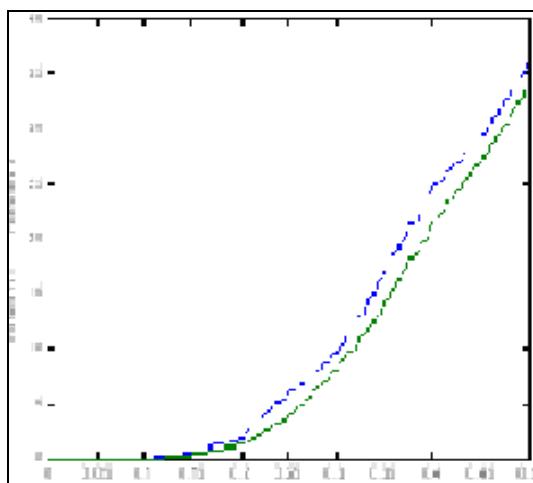


Fig 7: Percentage of recognition errors

If a higher accuracy is needed, the network can be trained for a longer time or retrained with more neurons in its hidden layer. Also, the resolution of the input vectors can be increased to a 10-by-14 grid. Finally, the network could be trained on input vectors with greater amounts of noise if greater reliability were needed for higher levels of noise.

The diagrams and the graphs demonstrate how a simple pattern recognition system could be designed. Note that the training process did not consist of a single call to a training function. Instead, the network was trained several times on various input vectors. In this case, training a network on different sets of noisy vectors forced the network to learn how to deal with noise, a common problem in the real world.

Normally, only feed-forward networks are used for pattern recognition. Feed-forward means that there is no feedback to the input. Similar to the way that human beings learn from mistakes, neural networks also could learn from their mistakes by giving feedback to the input patterns. This kind of feedback would be used to refine and reconstruct the input patterns and make them free from error; thus increasing the performance of the neural networks. Of course, it is very complex to construct such types of neural networks. These kinds of networks are called as auto associative neural networks. As the name implies, they use back-propagation algorithms.

4. Conclusion

Pattern recognition could be achieved using both normal computing techniques and neural networks. In normal computing conventional arithmetic algorithms are used to detect whether the given pattern matches a known one. It is a straightforward method. It will say either yes or no. It does not tolerate noisy patterns. On the other hand, neural networks can tolerate noise and, if trained properly, will respond correctly for noisy patterns. Neural networks may not perform miracles, but if constructed with the proper architecture and trained correctly with good data, they give amazing results in pattern recognition.

References :

- [1] I.Guyon.P.Albrecht,Y.Le.Cn, J. Denker, and W.Hubbard. *Design of neural network character recognition for a touch terminal, Pattern Recognition*, 1991.

[2] Gerald Wheatley- *Applied Numerical Analysis*

[3] A. Mitiche and J. Aggarwal. *Pattern category assignment by neural networks and the nearest neighbors rule*. International Journal on Pattern Recognition and Artificial Intelligence,10 :393–408, 1996.

[4] R. Plamondon. *A model-based segmentation framework for computer processing of handwriting*. Proceedings of the Eleventh IAPR International Conference on Pattern Recognition, pages 303–307, 1992

[5] K. Bhattacharyya and K. K. Sarma, "Innovative Segmentation of Handwritten Text in Assamese Using Neural Network", The 2009 International Conference on Genetic and Evolutionary methods (GEM'09), Las Vegas, July 2009.

[6] Z. Chen, "Handwritten Digits Recognition", The 2009 International Conference on Genetic and Evolutionary Methods (GEM'09), Las Vegas, July 2009.

[7] Tamil Standard Code for Information Interchange, <http://www.tamil.net/tscii>

[8] S. Mori, H. Nishida, H. Yamada, "Optical Character Recognition", John Wiley & Sons, 1999.

[9] R. Gonzalez, R. Woods, and S. Eddins, "Digital Image Processing Using Matlab", Prentice Hall, 2004.

Gowri N. completed M.Sc. in Physics and obtained M.Phil. in Physics from Alagappa University, Karaikudi for her thesis on "Simulation Studies in Crystal Growth". She is currently employed in TBAK College for Women in Kilakarai as Associate Professor in Computer Science Dept. and doing her Ph.D. in Pattern Recognition in School of Mathematics, Madurai Kamaraj University, Madurai.

Dr. R. Bhaskaran is Chariperson and Head of Department, School of Mathematics, Madurai Kamaraj University, Madurai.

Implementation of Clustering Through Machine Learning Tool

SREE RAM NIMMAGADDA¹, PHANEENDRA KANAKAMEDALA² and VIJAY BASHKARREDDY YARAMALA³

¹ LAKIREDDY BALIREDDY COLLEGE OF ENGINEERING
Mylavaram, Krishna Dist ,AP, India

² LAKIREDDY BALIREDDY COLLEGE OF ENGINEERING
Mylavaram, Krishna Dist ,AP, India

³ LAKIREDDY BALIREDDY COLLEGE OF ENGINEERING
Mylavaram, Krishna Dist ,AP, India

ABSTRACT

Clustering is the process of gathering or acquiring similar objects into a group known as cluster. All the objects in a cluster or group are similar to each other. The object in one cluster is dissimilar to the object in another cluster. The process of clustering is also known as un supervisory or machine learning. Weka is a popular tool for machine learning which was written in java. The Weka provides a collection of visualization tools and algorithms for data analysis and predictive modeling through a graphical user interface.

Key words: - clustering, Weka, machine learning and data analysis.

1. INTRODUCTION

Today there are mountains of stored data-containing terabytes of data. New mountains are forming daily as the transactions are performing. To deal with this terabytes of data many data mining techniques can be used. Data mining sometimes referred to as Knowledge Discovery in Databases (KDD). “Data Mining” may be defined as the process of searching, and analyzing data in order to find implicit, but potentially useful information. The advantages associated with data mining are (a) the result of analysis is objective (b) the accuracy of data is constant (c) analysis work is done routinely and (d) large quantities of data can be processed rapidly. Machine learning techniques can be used for data mining because of their ability to extract patterns relating to the concepts to be learned. Clustering is one of the Machine learning and data mining techniques which can identify the similar patterns in the data. Clustering is process of grouping similar type of objects into one group or cluster. Section 2 describes the clustering process in more detail, and section 3 provides discussion of k-means with an example problem, section 4 provides

the detailed discussion of Weka, and section 5 describes about the ARFF files and section 6 describes the process of using Weka tool to simulate clustering process.

2. CLUSTERING

Cluster analysis or clustering is the process of grouping the objects into subsets so that the objects in subset are similar in some sense. Clustering is a method of un supervisory learning and a common technique for stastical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. The following diagram represents the clustering process

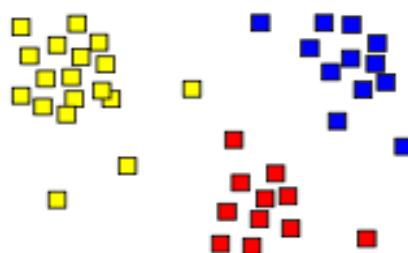


Fig 1: Result of the cluster analysis

The above diagram figure 1 represents that the result produced by the clustering process. The first step in this process is to adopt a mathematical description of similarity such as defining a proximity function. There are number of methods available to measure the similarity between the

observations. The most popular distance measure is Euclidian Distance, which is defined as

$$d(i,j) = ((x_{i1}-x_{j1})^2 + (x_{i2}-x_{j2})^2 + \dots + (x_{in}-x_{jn})^2)^{1/2}$$

Euclidian Distance has to satisfy the

- a. $d(i,j) \geq 0$: Distance is a non negative number
- b. $d(i,i)=0$: The distance of object to itself is 0
- c. $d(i,j)=d(j,i)$: Distance is symmetric function
- d. $d(i,j) \leq d(i,h)+d(h,j)$: going directly from object i to object j in space is no more than making a detour over any object h(triangular inequality).

The typical requirements of clustering are scalability, ability to deal with different types of attributes, Discovery of clusters with arbitrary shapes, Minimum requirements for domain knowledge to determine input parameters, Ability to deal with noisy data, High dimensionality and interpretability and usability. Many clustering algorithms are available. In general, the major clustering methods can be classified into following categories.

Partitioning method: given a database of n objects or data tuples a partitioning method constructs k partitions of data, where each partition represents a cluster and each group must contain at least one object and each object must belong to exactly one group.

Hierarchical methods: A hierarchical method creates a hierarchical decomposition of the given set of data objects. The following figure 2 and figure 3 represents the hierarchical clustering.

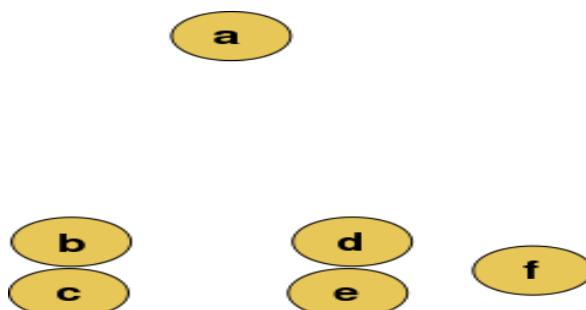


Figure 2: Raw Data

Density based methods: The general idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold; that is for each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points.

Grid based methods: Grid based method quantizes the object space into a finite number of cells that form a grid structure. All of the clustering operations are performed on grid structure.

Model based methods: Model based methods hypothesize a model for each of the clusters and find the best fit of the data to the given model. A model based algorithm may locate clusters by constructing a density function that reflects the spatial distribution of the data points.

The following diagram illustrates the hierarchical clustering.

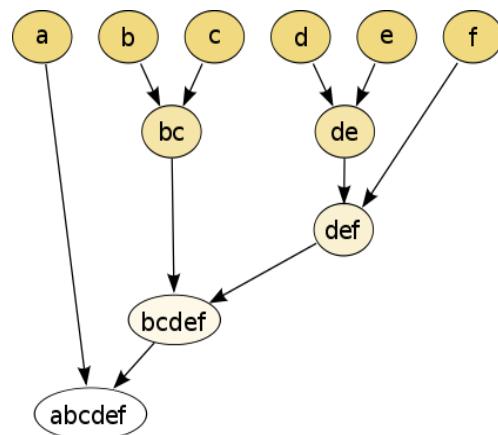


Figure 3: Hierarchical clustering

3. K-MEANS CLUSTERING

K-means is a prototype based clustering technique which performs one level partitions of the data objects. In this we first choose k initial centroids, where k represents the number of clusters desired. Each point is then assigned to the closest centroid, and each collection of points assigned to a centroid is a cluster. The centroids of each cluster is then updated based on the points assigned to the cluster. This assignment and update steps will continue until no point changes in cluster, equivalently, or centroids remain the same.

The basic steps of k-means clustering are

1. Determine the centroid coordinates.
2. Determine the distance of each object to the centroids.
3. Group the objects based on minimum distance.

4. Update the centroids.

The following flowchart represents the complete process of k-means

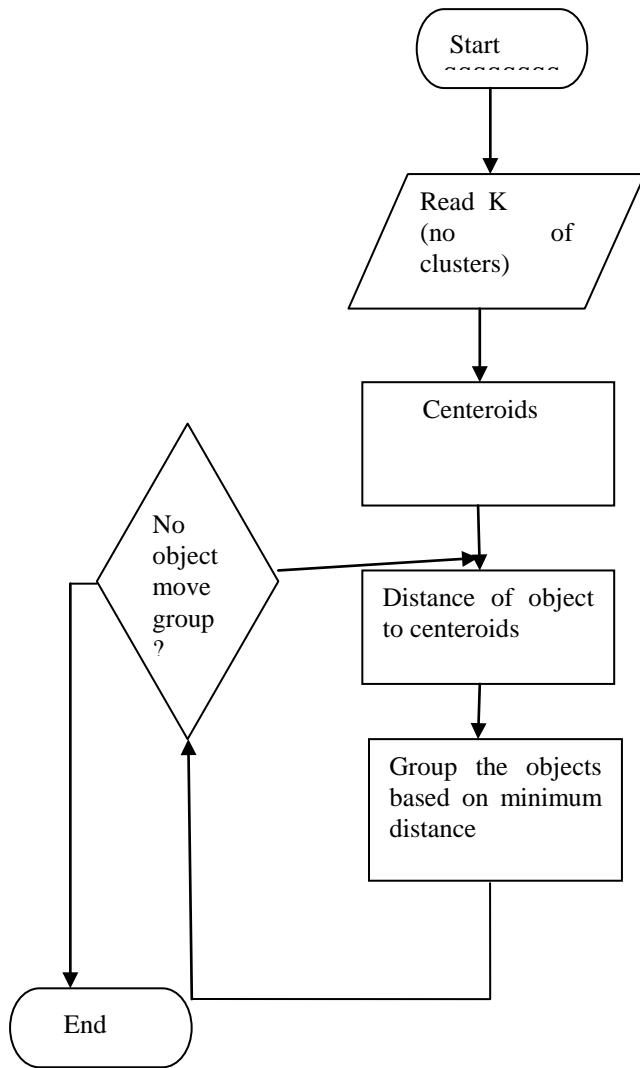


Figure 4: flowchart representing K-means

Example:

Suppose we have several objects with two attributes as shown in the following table. Our goal is group these objects in to two clusters i.e. $k=2$

| Object | Attribute 1 | Attribute 2 |
|--------|-------------|-------------|
| A | 2 | 5 |
| B | 4 | 6 |
| C | 5 | 9 |
| D | 9 | 12 |
| E | 11 | 14 |

Table 1: objects with two attributes

Each object represents one point with two attributes(x,y) that we can represent it as coordinate in an attribute space as shown in the following figure

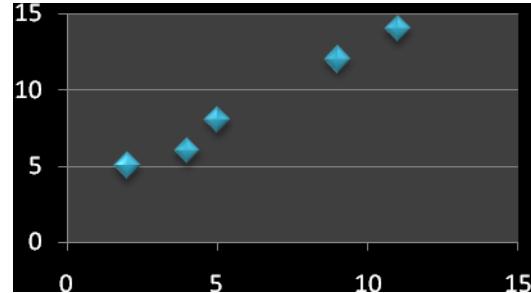


Figure 5: Attribute space of objects

Initial values of the centroids: suppose we use object A and object B as the first centroids that indicates $k=2$. Let c_1 and c_2 denote the coordinates of the centroids, then $c_1=(2,5)$ and $c_2=(4,6)$

Iteration 0:

Now calculate Euclidian Distance from c_1 to all points and c_2 to all points and construct a distance matrix D and group matrix G. Euclidian Distance between two objects can calculate as follows

$$ED = \sqrt{(x_{i1}-x_{j1})^2 + (x_{i2}-x_{j2})^2 + \dots + (x_{in}-x_{jn})^2}$$

Euclidian distance between two objects represents the geometric distance between two objects

$$ED \text{ between } (2, 5) \text{ and } (4, 6) = \sqrt{(2-4)^2 + (5-6)^2} = 2.236$$

$$D^0 =$$

| | | | | |
|-------|-------|------|------|-------|
| 0 | 2.236 | 5 | 9.9 | 12.73 |
| 2.236 | 0 | 3.16 | 7.48 | 10.63 |

Now depending on the above distance table we need to construct group tale by considering the minimum distance between centroid and object

| | | | | | |
|---------|---|---|---|---|---|
| $G^0 =$ | 1 | 0 | 0 | 0 | 0 |
| | 0 | 1 | 1 | 1 | 1 |

The above table represents that there two cluster instances and object A in one cluster and objects B, C, D, and E in another cluster.

Iteration 1:

Now group1 has only one object so that the centeroid remains same i.e. (2,5)

Group 2 has 4 objects so that we need to recomputed the centeroid as

$$((4+5+9+11)/4, (6+9+12+14)/4) = 7.25, 10.25$$

The new centeroid for group2 is (7.5, 10.25)

Now compute the distance from new centeroid and all other objects and construct distance table as

| | | | | | |
|---------|------|-------|------|------|-------|
| $D^1 =$ | 0 | 2.236 | 5 | 9.9 | 12.73 |
| | 7.42 | 5.35 | 2.57 | 2.47 | 5.3 |

Now group table is

| | | | | | |
|---------|---|---|---|---|---|
| $G^1 =$ | 1 | 1 | 0 | 0 | 0 |
| | 0 | 0 | 1 | 1 | 1 |

The above table represents that objects A, B belong to one cluster and objects C, D, E belong to another cluster.

Iteration 2:

Group 1 has two elements so that there is need to compute centeroid as $((2+4)/2, (5+6)/2)$ i.e. (3, 5.5)

Group 2 has three elements so that there is need to compute centeroid as $((5+9+11)/3, (9+12+14)/3)$ i.e. (8.33, 11.67)

Now construct distance table

| | | | | | |
|---------|------|------|------|------|-------|
| $D^2 =$ | 1.12 | 1.12 | 4.03 | 8.85 | 11.67 |
| | 9.2 | 7.2 | 4.3 | 0.75 | 3.54 |

Now group table is

| | | | | | |
|---------|---|---|---|---|---|
| $G^2 =$ | 1 | 1 | 1 | 0 | 0 |
| | 0 | 0 | 0 | 1 | 1 |

The above table represents that the objects A, B, C belong to first cluster and objects D, E belong to second cluster

Iteration 3:

The first cluster is having three objects as its members and second cluster is having two objects as its members. So that there is need to compute the centeroids of two clusters.

Centeroid of first cluster is $((2+4+5)/3, (5+6+9)/3) = (3.67, 6.67)$

Centeroid of second cluster is $((9+11)/2, (12+14)/2) = (10, 13)$

Now construct distance table

| | | | | | |
|---------|-------|------|------|-------|-------|
| $D^3 =$ | 2.36 | 0.74 | 2.68 | 7.537 | 10.36 |
| | 11.31 | 9.21 | 6.40 | 1.41 | 1.41 |

Now group table is

| | | | | | |
|---------|---|---|---|---|---|
| $G^3 =$ | 1 | 1 | 1 | 0 | 0 |
| | 0 | 0 | 0 | 1 | 1 |

The above table represents that the objects A, B, C belong to first cluster and the objects D, E belong to second cluster.

We obtain result that $G^2 = G^3$. Comparing the grouping of last iteration and this iteration reveals that the objects do not move group any more. Thus the computation of the k-means clustering has reached its stability and no more iteration is required. We get the final grouping as results as

| Object | Attribute X | Attribute Y | Cluster or group |
|--------|-------------|-------------|------------------|
| A | 2 | 5 | 1 |
| B | 4 | 6 | 1 |
| C | 5 | 9 | 1 |
| D | 9 | 12 | 2 |
| E | 11 | 14 | 2 |

This can be represented with the help of a graph

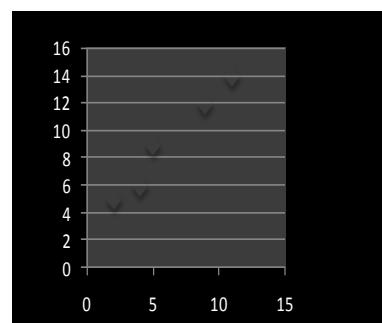


Figure 6: attribute space of clustered objects

4. WEKA

Weka is a popular open source machine learning software package implementing the many state of art machine learning algorithms. It is collection of machine learning algorithms. The algorithms can be either applied directly to data set or called from java code. Weka contains tools for data pre processing, classification, regression, clustering, association rules and visualization. Weka was written in java, developed at the University of Waikato, New Zealand. Weka is a free software available under General Public License(GNU). It provides Graphical User Interface. The fully java based version Weka 3 for which development is started in 1977, is now used in many application areas in particular for education and research.

Weka mainly consists of four interfaces

- Explorer
- Experimenter
- Knowledge Flow
- Simple CLI



Figure 7 : Weka Interface

Explorer: it is the main user interface, but the sane functionality can be accessed through the component based knowledge flow interface and from the command line.

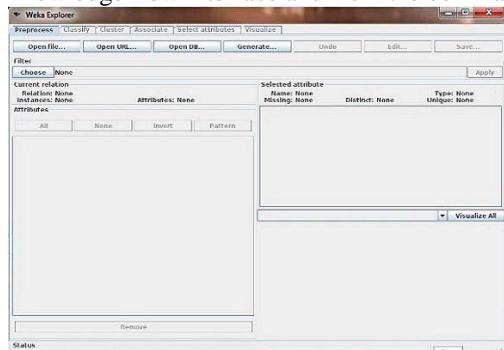


Figure 8: Explorer window

Experimenter: it allows systematic comparison of the predictive performance of the Weka's machine learning algorithms on a collection of data sets.

Knowledge Flow: In knowledge flow the user select weka components from a tool bar place them on layout canvas, and connect them into a directed graph that process and analyzes the data.

Simple CLI: It provides a command line mode to access a weka. The CLI is a text based interface to the weka environment. All weka commands are similar to the java commands. To execute a command enter it in white box below the window

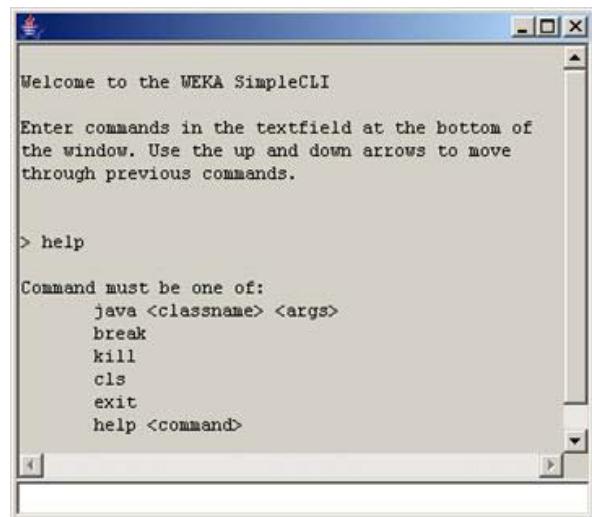


Figure 9 : Simple CLI window

The machine learning tool weka support several data mining tasks more specially data pre processing, clustering, classification, and regression. Weka can perform data mining tasks directly either on data base by using JDBC or on data sets in ARFF file

5. ARFF FILES

An ARFF (Attribute-Relation File Format) is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine learning project at the department of computer science of the University of Waikato to use with Weka machine learning software. All ARFF files two distinct sections. The first section is the header information followed by the data information. The relation and attributes are declared in the header section. The relation name is defined as the first line in the ARFF file as follows
@relation <relation-name>
Relation-name is a string.

The attribute declaration is

@attribute <attribute-name> <data type>

The data type can be any one of the four types.

- Numeric
- Nominal
- String
- Date

The comments can be specified with %

Now we define an ARFF file for our example problem in section 3

%this is header section

%it contains relation declaration and attribute declarations.

@relation xy

@attribute x numeric

@attribute y numeric

% data section started.

@data

2,5

4,6

5,9

9,12

11,14

Now save this file with .arff as its extension and name this file as sample.arff. The missing values in data section can specify with commas.

6. K-MEANS CLUSTERING WITH WEKA

This example illustrates the use of k-means with Weka. The sample data set used for this example is taken from sample.arff which contains five instances. Some implementations of K-means only allow numerical values for attributes. In that case, it may be necessary to convert the data set into the standard spreadsheet format and convert categorical attributes to binary. It may also be necessary to normalize the values of attributes that are measured on substantially different scales (e.g., "age" and "income"). While Weka provides filters to accomplish all of these preprocessing tasks, they are not necessary for clustering in Weka. This is because Weka Simple K Means algorithm automatically handles a mixture of categorical and numerical attributes. Furthermore, the algorithm automatically normalizes numerical attributes when doing distance computations. The Weka Simple K Means algorithm uses Euclidean distance measure to compute distances between instances and clusters.

First go to the Explorer interface. The explorer interface contains the various panels such as pre process, classify, cluster, associate, select attribute, and visualization. Now go to the preprocess panel. Now click on open file and browse for ARFF file which contains objects. After loading the ARFF file we can see elements in object space

by using visualization. Visualization panel represents instances in ARFF file in object space by taking their attributes. Now click on cluster and click on choose and select simple k-means and start. The result is as follows

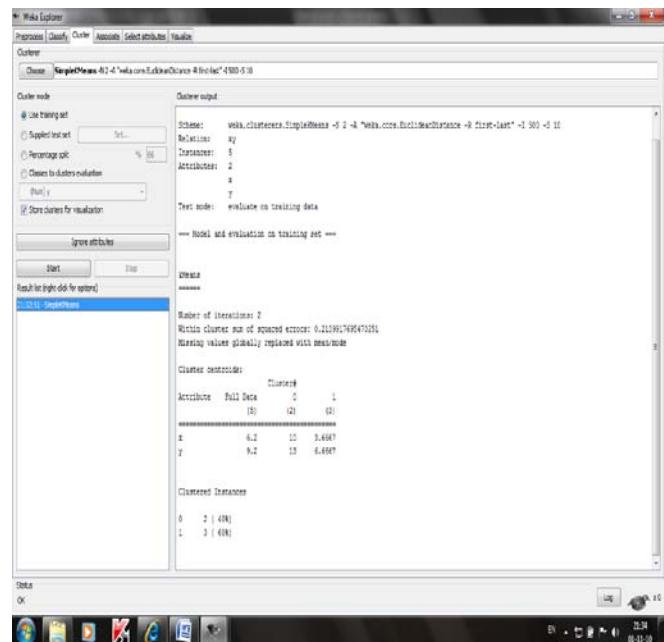


Figure 9: output screen

The output is two clusters one with two objects and another with three objects with sum of squared errors 0.21399.

7. CONCLUSION

Weka is an efficient and flexible machine learning tool. By using this we can implement all standard data mining algorithms. Here we are having one drawback for simple k-means that is we can not specify the k value i.e. number of clusters, so that for any number of objects Weka generates only two clusters. If we compare the result of weka clustering with manual procedure as we saw in section 3 we may not ensure the accuracy of the result. Thus if there is an ensure for accuracy of results and make provision to choose number of clusters it can be more and more efficient

8. REFERENCES

1. Beibei Zou, Xuesong Ma, Gen Newton, Donia precup Data Mining using relational database Management systems Supported by NSERC, CFI, NRC
2. Tapas Kanungo, senior member IEEE, David M Mount memer IEEE "An Efficient K-Means Clustering Algorithm: Analysis and implementation." IEEE Transactions on Pattern Analysis and Machine Intelligence vol 24 No 7, July 2009.

3. Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Pearson, *Introduction to Data Mining*.
4. Dmitry Lizokin, Pavel Velikhov, Maxim Grine, Denis Turdakov “Accuracy estimate and optimization techniques for SimRank Computation” Springer-Verlag 2009.
5. Remi Lehn, Viviane Lambert, Marie-Pierre Nachouki “Data Warehousing tool’s architecture: From multidimensional analysis to data mining” IEEE
6. Arun K Pujari “Data Mining Techniques” Universities pres.
7. Jiawei Han and Micheline Kamber “Data Mining Concepts and Techniques” Elsevier.

Improving Performance on WWW using Intelligent Predictive Caching for Web Proxy Servers

J. B. Patil¹ and B. V. Pawar²

¹ Department of Computer Engineering, R. C. Patel Institute of Technology
Shirpur, Maharashtra 425405, India

² Department of Computer Science, North Maharashtra University
Jalgaon, Maharashtra 425001, India

Abstract

Web proxy caching is used to improve the performance of the Web infrastructure. It aims to reduce network traffic, server load, and user perceived retrieval delays. The heart of a caching system is its page replacement policy, which needs to make good replacement decisions when its cache is full and a new document needs to be stored. The latest and most popular replacement policies like GDSF and GDSF# use the file size, access frequency, and age in the decision process. The effectiveness of any replacement policy can be evaluated using two metrics: hit ratio (HR) and byte hit ratio (BHR). There is always a trade-off between HR and BHR [1]. In this paper, using three different Web proxy server logs, we use trace driven analysis to evaluate the effects of different replacement policies on the performance of a Web proxy server. We propose a modification of GDSF# policy, IPGDSF#. Our simulation results show that our proposed replacement policy IPGDSF# performs better than several policies proposed in the literature in terms of hit rate as well as byte hit rate.

Keywords: Web caching, Replacement Policy, Hit Ratio, Byte Hit Ratio, Trace-driven Simulation.

1. Introduction

The enormous popularity of the World Wide Web has caused a tremendous increase in network traffic due to http requests. This has given rise to problems like user-perceived latency, Web server overload, and backbone link congestion. Web caching is one of the ways to alleviate these problems [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. Web caches can be deployed throughout the Internet, from browser caches, through proxy caches and backbone caches, through reverse proxy caches, to the Web server caches. In our work, we use trace-driven simulation for evaluating the performance of different caching policies for Web proxy servers.

One might argue that the ever decreasing prices of RAM and disks renders the optimization or fine tuning of cache replacement policies a “moot point”. Such a conclusion is ill guided for several reasons. First, recent studies have shown that Web cache hit ratio (HR) and byte hit ratio (BHR) grow in a *log-like* fashion as a function of cache size [5, 26, 27, 28]. Thus, a better algorithm that increases hit ratios by several percentage points would be equivalent to a several-fold increase in cache size. Second, the growth rate of Web content is much higher than the rate with which memory sizes for Web caches are likely to grow. The only way to bridge this widening gap is through efficient cache management. Finally, the benefit of even a slight improvement in cache performance may have an appreciable effect on network traffic, especially when such gains are compounded through a hierarchy of caches [6].

Cao and Irani have surveyed ten different policies and proposed a new algorithm, Greedy-Dual-Size (GDS) in [5]. The GDS algorithm uses document size, cost, and age in the replacement decision, and shows better performance compared to previous caching algorithms. In [4] and [12], frequency was incorporated in GDS, resulting in Greedy-Dual-Frequency-Size (GDSF) and Greedy-Dual-Frequency (GDF). While GDSF is attributed to having best hit ratio (HR), it is having a modest byte hit ratio (BHR). Conversely, GDF yields a best HR at the cost of worst BHR [12].

We have proposed a new algorithm called Greedy-Dual-Frequency-Size#, (GDSF#), which allows augmenting or weakening the impact of size or frequency or both on HR and BHR [13, 14, 15, 16, 17].

In this paper, we propose an extension to our algorithm GDSF#, called Intelligent Predictive Greedy-Dual-

Frequency-Size#, (IPGDSF#). We compare IPGDSF# with algorithms like LRU, GDSF, and GDSF#. Our simulation study shows that IPGDSF# outperforms all other algorithms under consideration in terms of hit rate (HR) as well as byte hit rate (BHR).

The remainder of this paper is organized as follows. Section 2 introduces IPGDSF#, a new algorithm for Web cache replacement. Section 3 describes the simulation model for the experiment. Section 4 describes the experimental design of our simulation while Section 5 presents the simulation results. We present our conclusions in Section 6.

2. IPGDSF# Algorithm

We extract *future frequency* from the Web proxy server logs. Then it is used to extend our GDSF# policy. Our idea is similar to the work of Bonchi et al. [18, 19] and Yang et al. [20]. While the Web caching algorithm in [18, 19] was designed to extend the LRU policy, Yang et al. [20] extended GDSF policy. We will be extending our policy GDSF# [13, 14, 15, 16, 17].

As pointed out early in caching research [21], the power of caching is in accurately predicting the usage of objects in the near future. In earlier works, estimates for future accesses were mostly built on measures such as access frequency, object size and cost. Such measures cannot be used to accurately predict for objects that are likely to be popular but have not yet been popular at any given instant in time. For example, as Web users traverse Web space, there are documents that will become popular soon due to Web document topology, although these documents are not yet accessed often in the current time instant [20]. Our approach is based on predictive Web caching model described by Yang et al. [20]. However, there are many noteworthy differences. Firstly, we use simple statistical techniques to find future frequency while Yang et al. use sequential association rules to predict the future Web access behavior. Secondly, for simplicity we do not try to identify user sessions. We assume that a popular document, which is used by one user, is likely to be used by many other users, which normally is the case for popular documents. We demonstrate the applicability of the method empirically through increased hit rates and byte hit rates.

Similar to the approach by Bonchi et al. [18, 19], our algorithm is an *intelligent* one as it can adapt to changes in usage patterns as reflected by future frequency. This is because the parameter *future frequency*, which is used in assigning weight (key value) to the document while storing

in the cache, can be computed periodically in order to keep track of the recent past. This characteristic of adapting to the flow of requests in the historical data makes our policy intelligent. We call this innovative caching algorithm as *Intelligent Predictive GDSF#, (IPGDSF#)*.

In GDSF#, the key value of document i is computed as follows [13, 14, 15, 16, 17]:

$$H_i = L + f_i^\lambda \times c_i / s_i^\delta.$$

where λ and δ are rational numbers, L is the inflation factor, c_i is the estimated cost of the document i , f_i is the access frequency of the document i , and s_i is the document size.

We now consider how to find future frequency, ff_i for document i from the Web logs. We mine the preprocessed Web log files. We extract the unique documents from the logs. Then we arrange these documents in the temporal order. Now for each unique document, we extract the number of future occurrences of that document. We call this parameter as *future frequency*, ff .

With this parameter, we can now extend GDSF# by calculating H_i , the key value of document i as follows:

$$H_i = L + (f_i + ff_i)^\lambda \times c_i / s_i^\delta.$$

Here we add f_i and ff_i together, which implies that the key value of a document i is determined not only by its past occurrence frequency f_i , but also by its future frequency ff_i . By considering both the past occurrence frequency and future frequency, we can enhance the priority i.e. the key value of those objects that may not have been accessed frequently enough in the past, but will be in the near future according to the future frequency. The more likely it occurs in the future, the greater the key value will be. This will promote objects that are potentially popular objects in the near future even though they are not yet popular in the past. Thus, we look ahead in time in the request stream and adjust the replacement policy.

Finally, we make the policy intelligent by periodically updating future frequency when some condition becomes false, e.g. at fixed time intervals or when there is a degradation in the cache performance.

Now we present the IPGDSF# algorithm as shown in Fig.1:

```

Initialize  $L = 0$ 
Find future frequency  $ff_i$ 
loop forever {
    do {
        Process each request document in turn:
        let current requested document be  $i$ 
        if  $i$  is already in cache
             $H_i = L + (f_i + ff_i)^\lambda \times c_i / s_i^\delta$ 
        else
            while there is not enough room in cache for  $i$  {
                let  $L = \min(H_i)$ , for all  $i$  in cache
                evict  $i$  such that  $H_i = L$ 
            }
            load  $i$  into cache
             $H_i = L + (f_i + ff_i)^\lambda \times c_i / s_i^\delta$ 
        } while (condition)
    update (future frequency)
}

```

Fig. 1 IPGDSF# algorithm.

3. Simulation Model for the Experiment

In case of proxy servers, all requests are assumed to be directed to the proxy server. When the proxy receives a request from a client, it checks its cache to see if it has a copy of the requested object. If there is a copy of the requested object in its cache, the object is returned to the client signifying a *cache hit*, otherwise the proxy records a *cache miss*. The original Web server is contacted and on getting the object, stores the copy in its cache for future use, and returns a copy to the requesting user. If the cache is already full when a document needs to be stored, then a replacement policy is invoked to decide which document (or documents) is to be removed.

Our model also assumes file-level caching. Only complete documents are cached; when a file is added to the cache, the whole file is added, and when a file is removed from the cache, the entire file is removed.

For simplicity, our simulation model completely ignores the issues of *cache consistency* (i.e., making sure that the cache has the most up-to-date version of the document, compared to the master copy version at the original Web server, which may change at any time).

Lastly, caching can only work with static files, dynamic files that have become more and more popular within the past few years, cannot be cached.

3.1 Workload Traces

For Web proxy servers, we have used: Boston University Computer Science Department client traces collected in 1995; BU272 and BU-B19 [26] and one trace collected in 1998; BU98 [30] [31].

4. Experimental Design

This section describes the design of the performance study of cache replacement policies. The discussion begins with the factors and levels used for the simulation. Next, we present the performance metrics used to evaluate the performance of each replacement policy used in the study.

4.1 Factors and Levels

There are two main factors used in the in the trace-driven simulation experiments: cache size and cache replacement policy. This section describes each of these factors and the associated levels.

Cache Size

The first factor in this study is the size of the cache. For the proxy logs, we have used ten levels from 1 MB to 1024 MB except in case of BU-B19 trace, we have a upper bound of 4096 MB. Similar cache sizes are used by many researchers [9, 22, 23, 24]. The upper bounds represent the *Total Unique Mbytes* in the trace, which is essentially equivalent to having an infinite size cache [29]. An infinite cache is one that is so large that no file in the given trace, once brought into the cache, need ever be evicted [23, 25]. It allows us to determine the maximum achievable cache hit ratio and byte hit ratio, and to determine the performance of a smaller cache size to be compared to that of an infinite cache.

Replacement Policy

We show the simulation results of LRU, GDSF, GDSF#, and IPGDSF# for the Web proxy traces for hit rate, and byte hit rate. For the last three algorithms, we consider the cost function as one. In GDSF# and IPGDSF#, we use the best combination of $\lambda = 2$ and $\delta = 0.9$ in the equation for H_i . Since we have already demonstrated that GDSF# is the champion of all the algorithms in terms of both hit rate and byte hit rate [13, 14, 15, 16, 17], we have not chosen other algorithms for the comparison. LRU is chosen as a baseline algorithm.

4.2 Performance Metrics

The performance metrics used to evaluate the various replacement policies used in this simulation are *Hit Rate* and *Byte Hit Rate*.

Hit Rate (HR) Hit rate (HR) is the ratio of the number of requests met in the cache to the total number of requests.

Byte Hit Rate (BHR) Byte hit rate (BHR) is concerned with how many bytes are saved. This is the ratio of the number of bytes satisfied from the cache to the total bytes requested.

5. Simulation Results

In this section, we present and discuss simulation results for BU272, BU-B19, and BU98 Web proxy servers.

5.1 Simulation Results for BU272

Fig. 2 gives the comparison of IPGDSF# with other algorithms.

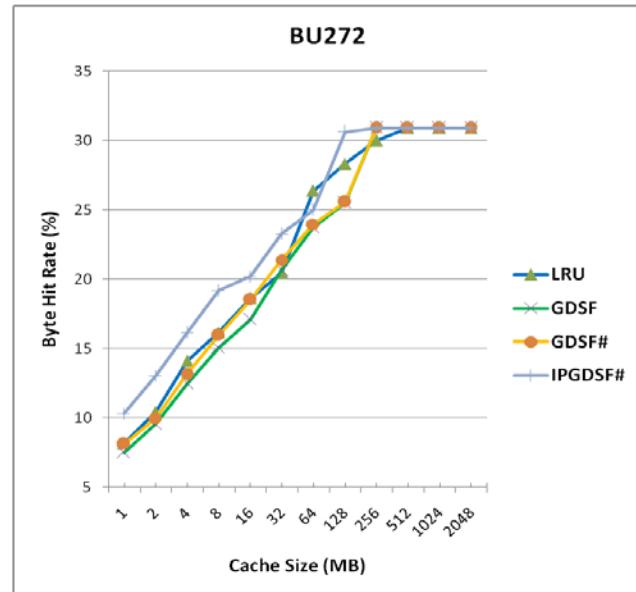
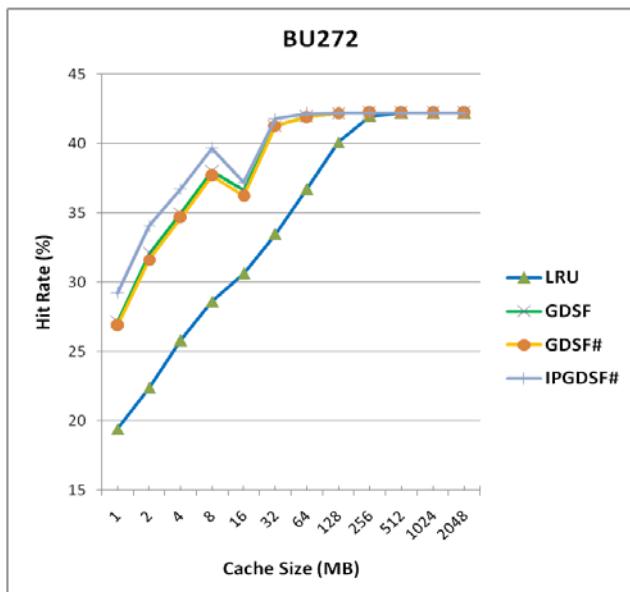


Fig. 2 Comparison of IPGDSF# with other algorithms using BU272 trace

From Figure 2, it can be seen that IPGDSF# outperforms all other algorithms in terms of hit rate as well as byte hit rate for the BU272 data. In case of hit rate, for a cache size of 16MB, there is a performance gain of 6.59% (from 30.62% to 37.21%) over LRU, 0.58% (from 36.63% to 37.21%) over GDSF and 0.99% (from 36.22% to 37.21%) over GDSF#.

In case of byte hit rate, for a cache size of 16MB, there is a performance gain of 4.62% (from 18.64% to 23.26%) over LRU, 6.16% (from 17.10% to 23.26%) over GDSF and 4.73% (from 18.53% to 23.26%) over GDSF#. The graphs, as expected, converge as the cache size grows.

5.2 Simulation Results for BU-B19

Figure 3 gives the comparison of IPGDSF# with other algorithms.

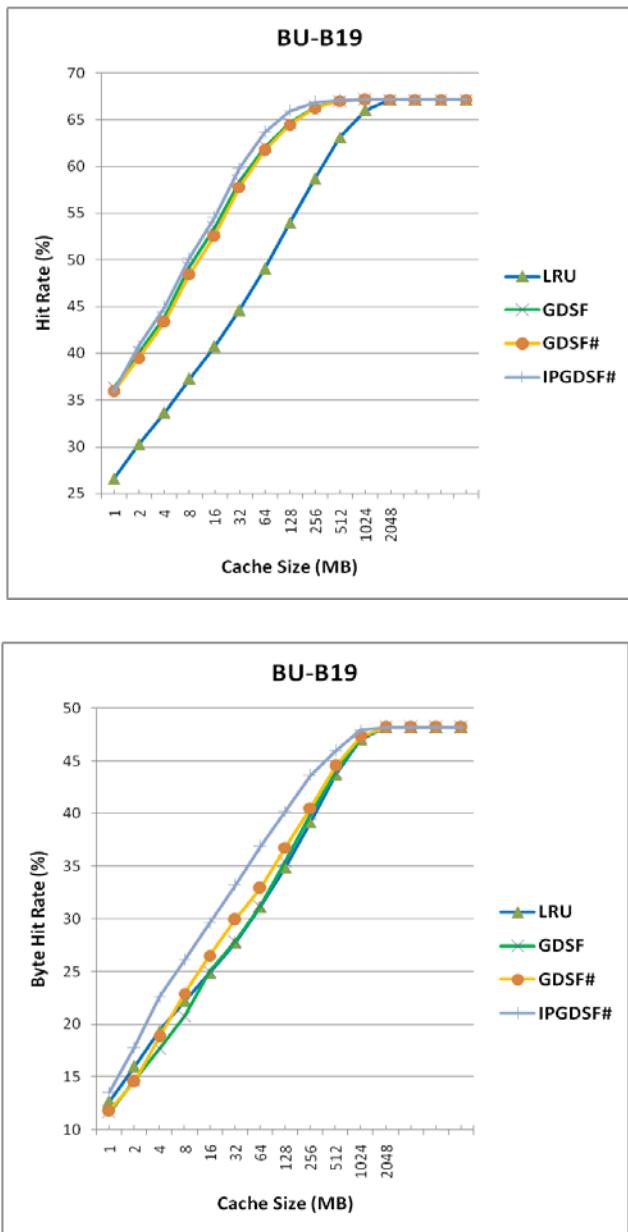


Fig. 3 Comparison of IPGDSF# with other algorithms using BU-B19 trace

Similarly, from Figure 3, it can be seen that IPGDSF# outperforms all other algorithms in terms of hit rate as well as byte hit rate for the BU-B19 data. In case of hit rate, for a cache size of 64MB, there is a performance gain of 14.6% (from 49.06% to 63.66%) over LRU, 1.6% (from 62.06% to 63.66%) over GDSF, and 1.95% (from 61.71% to 63.66%) over GDSF#.

In case of byte hit rate, for a cache size of 64MB, there is a performance gain of 5.74% (from 31.15% to 36.89%) over LRU, 5.64% (from 31.25% to 36.89%) over GDSF, and

3.97% (from 32.92% to 36.89%) over GDSF#. The graphs, as expected, converge as the cache size grows.

5.3 Simulation Results for BU98

Figure 4 gives the comparison of IPGDSF# with other algorithms.

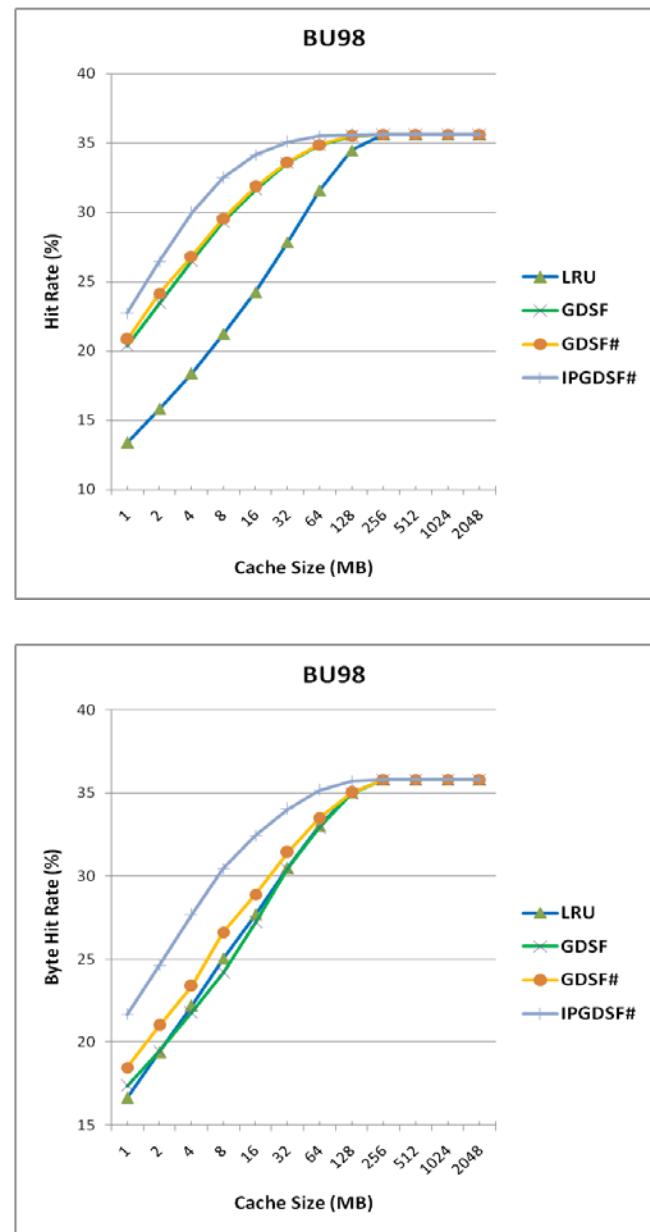


Fig. 4 Comparison of IPGDSF# with other algorithms using BU98 trace

Similarly, from Figure 4, it can be seen that IPGDSF# outperforms all other algorithms in terms of hit rate as well as byte hit rate for the BU98 data. In case of hit rate, for a cache size of 32MB, there is a performance gain of 7.86% (from 27.84% to 35.7%) over LRU, 2.13% (from 33.57%

to 35.7%) over GDSF, and 2.06% (from 33.64% to 35.7%) over GDSF#.

In case of byte hit rate, for a cache size of 32MB, there is a performance gain of 3.53% (from 30.49% to 34.02%) over LRU, 3.57% (from 30.45% to 34.02%) over GDSF, and 2.6% (from 31.42% to 34.02%) over GDSF#. The graphs, as expected, converge as the cache size grows.

6. Conclusions

In this paper, we have proposed an Intelligent Predictive Web caching algorithm, IPGDSF#, capable of adapting its behavior based on access statistics. This algorithm is based on the GDSF# algorithm, which we proposed in [13, 14, 15, 16, 17]. IPGDSF# considers future frequency in calculating the key value of the document, i.e. we look ahead in time in the request stream and adjust the replacement policy. The future frequency is mined from Web server logs using the simple statistical techniques. We make the policy intelligent by periodically updating future frequency when some condition becomes false.

We compare IPGDSF# with cache replacement policies like LRU, GDSF, and GDSF# for Web proxies, using a trace-driven simulation approach. We conduct several experiments using three Web proxy traces. We use metrics like Hit Ratio (HR) and Byte Hit Ratio (BHR) to measure and compare performance of these algorithms.

Our study shows that IPGDSF# outperforms all other algorithms in terms of hit rate as well as byte hit rate. GDSF# has improved performance in case of both HR and BHR. Now IPGDSF# has further improved both the metrics. Thus, we find that our approach gives much better performance than the other algorithms, in the quantitative measures such as hit ratios and byte hit ratios of accessed documents. We believe that use of future frequency coupled with the adaptiveness is indeed the reason that makes our approach preferable to any other caching algorithm.

References

- [1] M. Arlitt, R. Friedrich, & T. Jin, "Workload Characterization of Web Proxy Cache Replacement Policies", In ACM SIGMETRICS Performance Evaluation Review, August 1999.
- [2] M. Abrams, C. R. Standridge, G. Abdulla, S. Williams, & E. A. Fox, "Caching Proxies: Limitations and Potentials", In Proceedings of the Fourth International World Wide Web Conference, Pages 119-133, Boston, MA, December 1995.
- [3] M. Arlitt & C. Williamson, "Trace Driven Simulation of Document Caching Strategies for Internet Web Servers", Simulation Journal, Volume 68, Number 1, Pages 23-33, January 1977.
- [4] L. Cherkasova, "Improving WWW Proxies Performance with Greedy-Dual-Size-Frequency Caching Policy", In HP Technical Report HPL-98-69(R.1), November 1998.
- [5] P. Cao & S. Irani, "Cost-Aware WWW Proxy Caching Algorithms", In Proceedings of the USENIX Symposium on Internet Technology and Systems, Pages 193-206, December 1997.
- [6] S. Jin & A. Bestavros, "GreedyDual*: Web Caching Algorithms Exploiting the Two Sources of Temporal Locality in Web Request Streams", In Proceedings of the Fifth International Web Caching and Content Delivery Workshop, 2000.
- [7] S. Podlipnig & L. Boszormenyi, "A Survey of Web Cache Replacement Strategies", ACM Computing Surveys, Volume 35, Number 4, Pages 374-398, December 2003.
- [8] L. Rizzo, & L. Vicisano, "Replacement Policies for a Proxy Cache", IEEE/ACM Transactions on Networking, Volume 8, Number 2, Pages 158-170, April 2000.
- [9] A. Vakali, "LRU-based Algorithms for Web Cache Replacement", In International Conference on Electronic Commerce and Web Technologies, Lecture Notes in Computer Science, Volume 1875, Pages 409-418, Springer-Verlag, Berlin, Germany, 2000.
- [10] R. P. Wooster & M. Abrams., "Proxy Caching that Estimates Page Load Delays", In Proceedings of the Sixth International World Wide Web Conference, Pages 325-334, Santa Clara, CA, April 1997.
- [11] S. Williams, M. Abrams, C. R. Standridge, G. Abdulla, & E. A. Fox, "Removal Policies in Network Caches for World-Wide-Web Documents", In Proceedings of ACM SIGCOMM, Pages 293-305, Stanford, CA, 1996, Revised March 1997.
- [12] M. F., Arlitt, L. Cherkasova, J. Dilley, R. J. Friedrich, & T. Y Jin, "Evaluating Content Management Techniques for Web Proxy Caches", ACM SIGMETRICS Performance Evaluation Review, Volume 27, Number 4, Pages 3-11, March 2000.
- [13] J. B. Patil and B. V. Pawar, "GDSF#, A Better Algorithm that Optimizes Both Hit Rate and Byte Hit Rate in Internet Web Servers", International Journal of Computer Science and Applications, ISSN: 0972-9038, Volume 5, Number 4, Pages 1-10, 2008.
- [14] J. B. Patil and B. V. Pawar, "Trace Driven Simulation of GDSF# and Existing Caching Algorithms for Internet Web Servers", Journal of Computer Science, Volume 2, Issue 3, Page 573, March-April 2008.

- [15] J. B. Patil and B. V. Pawar, "GDSF#, A Better Algorithm that Optimizes Both Hit Rate and Byte Hit Rate in Internet Web Servers", BRI'S Journal of Advances in Science and Technology, ISSN: 0971-9563, Volume 10, No. (I&II), Pages 66-77, June, December 2007.
- [16] J. B. Patil and B. V. Pawar, "GDSF#, A Better Web Caching Algorithm", In Proceedings of International Conference on Advances in Computer Vision and Information Technology (ACVIT-2007), Co-sponsored by IEEE Bombay Section, Pages 1593-1600, Aurangabad, India, November 28-30, 2007.
- [17] J. B. Patil and B. V. Pawar, "Trace Driven Simulation of GDSF# and Existing Caching Algorithms for Web Proxy Servers", In Proceedings of The 6th WSEAS International Conference on DATA NETWORKS, COMMUNICATIONS and COMPUTERS (DNCOCO 2007), Trinidad and Tobago, November 5-7, 2007, Pages 378-384, ISBN: 978-960-6766-11-4, ISSN: 1790-5117.
- [18] F. Bonchi, F. Giannotti, C. Gozzi, G. Manco, M. Nanni, D. Pedreschi, C. Renso, and S. Ruggieri, "Web Log Data Warehousing and Mining for Intelligent Web Caching," Data and Knowledge Engineering, Volume 39, Number 2, Pages 165-189, 2001.
- [19] F. Bonchi, F. Giannotti, G. Manco, M. Nanni, D. Pedreschi, C. Renso, and S. Ruggieri, "Web Log Data Warehousing and Mining for Intelligent Web Caching," In Proceedings of International Conference on Information Technology: Coding and Computing (ITCC'01) Pages 0599-, 2001.
- [20] Q. Yang, and H.H. Zhang, "Web-Log Mining for Predictive Web Caching", IEEE Transactions on Knowledge and Data Engineering, Volume 15, Number 4, Pages 1050-1053, July/August 2003.
- [21] L.A. Belady, "A Study of Replacement Algorithms for Virtual Storage Computers," IBM Systems Journal, Volume 5, Number 2, Pages 78-101, 1966.
- [22] M. Arlitt, R. Friedrich, & T. Jin, "Performance Evaluation of Web Proxy Cache in a Cable Modem Environment", HP Technical Report, HPL-98-97(R.1), Palo Alto, 1998.
- [23] M. Busari, "Simulation Evaluation of Web Caching Hierarchies", MS Thesis, Dept of Computer Science, Uni of Saskatchewan, Canada, 2000.
- [24] R. Ayani, Y. M. Teo, & P. Chen, "Cost-based Proxy Caching", In Proceedings of International Symposium on Distributed Computing & Applications to Business, Engineering & Science, Wuxi, China, December 2002.
- [25] M. Busari & C. Williamson, "On the Sensitivity of Web Proxy Cache Performance to Workload Characteristics", In Proceedings of IEEE Infocom, Anchorage, Alaska, April 2001, 1225-1234.
- [26] C. R. Cunha, A. Bestavros, & M. E. Crovella, "Characteristics of WWW Client-based Traces", Technical Report, BU-CS-95-010, Computer Science Department, Boston University, 1995.
- [27] V. Almeida, A. Bestavros, M. Crovella, & A., de Oliveria, "Characterizing Reference Locality in the WWW", In Proceedings of PDIS'96: The IEEE Conference on Parallel and Distributed Information Systems, Miami, 1996.
- [28] L. Breslau, P. Cao, L. Fan, G. Philips, & S. Shenker, "Web Caching and Zipf-like Distributions: Evidence and Implications", In Proceedings of Conference on Computer Communications (IEEE Infocom), New York, 1999, 126-134.
- [29] H. Bahn, S. H. Noh, S. L. Min, & K Koh, "Using Full Reference History for Efficient Document Replacement in Web Caches", In Proceedings of Second USENIX Symposium on Internet Technologies and Systems, Boulder, Colorado, USA, October 1999.
- [30] A. Bradley, "BU Computer Science 1998 Proxy Trace", Technical Report, Computer Science Department, Boston University, 1999.
- [31] P. Barford, A. Bestavros, A. Bradley, & M. Crovella, "Changes in Web Client Access Patterns Characteristics and Caching Implications", World Wide Web, 2(1-2), 1999.

J. B. Patil did his M. Tech. in Computer Science and Data Processing from Indian Institute of Technology, Kharagpur in 1993 and Ph. D. in Computer Engineering from North Maharashtra University, Jalgaon in 2008. He is currently working as a Principal and Professor in Computer Engineering at R. C. Patel Institute of Technology, Shirur, India. He is a Member of Member of Institute of Engineers, India and also Life Member of Indian Society for Technical Education and Computer Society of India.

B. V. Pawar did his B. E. in Production Engineering from VJTI, Mumbai in 1986, his M. Sc. In Computer Science from University of Mumbai in 1988, and his Ph. D. in Computer Science from North Maharashtra University, Jalgaon in 2000. He is currently working as Professor and Head of Department of Computer Science, North Maharashtra University, Jalgaon. His current research interests include Natural Language Processing, Web Technologies, Information Retrieval, Web Mining, etc.

An Efficient Approach to Prune Mined Association Rules in Large Databases

D.Narmadha¹, G.NaveenSundar², S.Geetha³

¹Computer Science Department, Karunya University, Coimbatore, Tamilnadu, India

²Computer Science Department, Karunya University, Coimbatore, Tamilnadu, India

³Computer Science Department, Karunya University, Coimbatore, Tamilnadu, India

Abstract

Association rule mining finds interesting associations and/or correlation relationships among large set of data items. However, when the number of association rules become large, it becomes less interesting to the user. It is crucial to help the decision-maker with an efficient postprocessing step in order to select interesting association rules throughout huge volumes of discovered rules. This motivates the need for association analysis. Thus, this paper presents a novel approach to prune mined association rules in large databases. Further, an analysis of different association rule mining techniques for market basket analysis, highlighting strengths of different association rule mining techniques are also discussed. We want to point out potential pitfalls as well as challenging issues need to be addressed by an association rule mining technique. We believe that the results of this approach will help decision maker for making important decisions.

Keywords- CLOSET, MAFIA, FP, Ontology, User constraint Template

1. Introduction

In **data mining**, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. Association rule Mining describes analyzing and presenting strong rules discovered in databases using different measures of interestingness. An association rule is defined as the implication $X \Rightarrow Y$, described by two interestingness measures support and confidence where X and Y are the sets of items. Furthermore, valuable information is often represented by those rare-low support and discovered association rules are unexpected which are surprising to the user. As we increase the support threshold, the more efficient the algorithms are and the more the discovered rules are obvious, and hence, the less they are interesting for the user. As a result, it is necessary to bring the support

threshold low enough in order to extract valuable information. Unfortunately, the lower the support is, the larger the volume of rules becomes, making it intractable for a decision-maker to analyze the mining result. Experiments show that rules become almost impossible to use when the number of rules exceeds a limit. Thus, it is crucial to help the decision maker with an efficient technique for reducing the number of rules.

In this paper, a fairly comprehensive comparison of various association rule mining techniques is presented.

We analyzed the performance and the efficiency of different association mining approaches. Also this paper discusses the level of interestingness each technique provides. Rest of the paper is organized as follows. Section 2 contains a generalized summary of various association mining techniques and brief description of different approaches that have been taken for study. Section 3 gives a comparative analysis of various association mining techniques based on certain parameters. Section 4 discusses about the efficient approach to prune discovered association rules.

2. Methodology for Association Rule Generation

Association analysis has wide range of applications in market basket analysis, Intrusion detection, bioinformatics, web usage mining. There have been several such association mining techniques for generating association rules. Frequent item set concise representation as proposed by Burdick [2] and optimal rule discovery as proposed by Li [3], Zaki [4] introduced concise representation of frequent itemset to reduce the number of frequent itemsets and CLOSET algorithm was proposed [5] as a new efficient method for mining closed itemsets. Another solution for the reduction of the number of frequent itemsets is mining maximal frequent itemsets [6]. MAFIA algorithm is

based on depth-first traversal and several pruning methods. More recently, Bellandi [7] proposed ontology driven association rule extraction. The different approaches for the redundancy reduction of association rules are: Zaki and Hsiao used frequent closed itemsets in the CHARM algorithm [8] in order to generate all frequent closed itemsets. They used an itemset-tid set search tree and pursued with the aim of generating a small nonredundant rule set [9]. To this goal, the authors first found minimal generator for closed itemsets, and then, they generated nonredundant association rules using two closed itemsets.

Pasquier et al. [10] proposed the Close algorithm in order to extract association rules. Close algorithm is based on a new mining method: pruning of the closed set lattice (closed itemset lattice) in order to extract frequent closed itemsets. Association rules are generated starting from frequent itemsets generated from frequent closed itemsets.

From the above association rule mining techniques, few are selectively analyzed in detail in this literature.

2.1CLOSET: An Efficient Algorithm for mining Frequent closed Itemsets

This approach is an efficient algorithm for mining frequent itemsets with the development of three techniques:

- (i) Applying compressed, frequent pattern tree FP-tree structure for mining closed itemsets without candidate generation.
 - (ii) Developing a single prefix path compression technique to identify frequent closed itemsets quickly.
 - (iii) Exploring a partition based projection mechanism for scalable mining in large databases.

Optimization1: Compress transactional and conditional databases using an FP-tree structure: FP-tree compresses databases for frequent itemset mining. An FP tree is a prefix tree structure representing compressed but complete information for a database. Its construction is simple. The transactions with same prefix share the portion of a path from the root. Similarly conditional FP tree can be constructed for conditional databases.

Optimization2: Extract items appearing in every transaction of conditional database: If there exists, a set of items Y appearing in every transaction of the X-conditional database, XUY forms a frequent closed item set if it is not a proper subset of some frequent closed item set with the same support.

Fig. 1 shows how the frequent closed item sets can be extracted directly from FP tree. This reduces the size of FP-tree because the conditional databases contain less number of items after extraction and also reduces the level of recursion.

Optimization3: Directly extract frequent item sets from FP-tree:

- This allows the program to identify frequent closed item sets quickly.
 - Reduces the size of remaining FP tree to be examined.
 - Reduces the level of recursion.

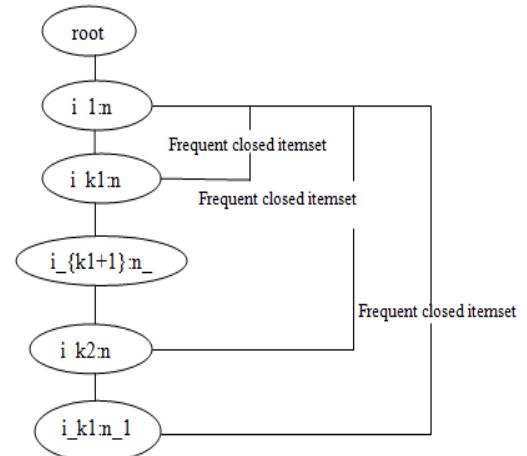


Fig. 1 Directly Extract frequent closed itemsets from FP tree

Optimization4: Prune search branches: Let X and Y are two frequent items with same support. If XCY and Y is closed itemset, there is no need to search for X conditional database because there is no hope to generate frequent item set from there. This reduces the overhead in searching for database.

2.2 Ontology-Driven Association Rule Extraction

This provides an integrated framework for extracting constraint-based Multi-level Association Rules with an ontology support. This method can improve the quality of filtered rules.

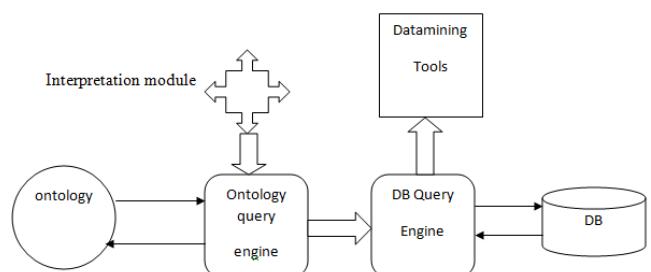


Fig. 2 System Architecture

The System Architecture as shown in Fig. 2 presents a view of set of components in the Ontology-Driven rule extraction. The ontology (OD) describes the domain of interest (D) and it is used as a means of meta-data representation. The interpretation module translates the requests of a user (user constraints) into a set of formal constraints (QD defined on OD) so that they can be

supplied to the Ontology Query Engine by means of a suitable query language. The aim of these constraints is to exclude some items from the output association rules, or to characterize interesting items according to an abstraction level. It includes both pruning constraints, used for filtering a set of non-interesting items, and abstraction constraints, which permit a generalization of an item to a concept of the ontology. By using pruning constraints, one can specify the exclusion of a set of items from the input transactions set, and, as a consequence, from the extracted rules.

There are several ways to reduce the computational complexity of Association Rule Mining and to increase the quality of the extracted rules: (i) reducing the search space; (ii) exploiting efficient data structures; (iii) adopting domain-specific constraints. The first two classes of optimizations are used for reducing the number of steps of the algorithm, for re-organizing the itemsets, for encoding the items, and for organizing the transactions in order to minimize the algorithm time complexity. The third class tries to overcome the lack of user data-exploration by handling domain-specific constraints.

This paper focuses on these optimizations by representing a specific domain by means of ontology and driving the extraction of association rules by expressing constraints. The aim of this work is to reduce the “search space” of the algorithm and to improve the significance of the association rules.

2.3 Selecting the Right Objective Measure for Association Analysis

This approach describes several key properties one should examine in order to select the right measure for a given application. An algorithm is presented for selecting a small set of patterns such that domain experts can find a measure that best fits their requirements by ranking this small set of patterns. Objective measures such as support, confidence, interest factor, correlation, and entropy are often used to evaluate the interestingness of association patterns. However, in many situations, these measures may provide conflicting information about the interestingness of a pattern. This approach describes several key properties one should examine in order to select the right measure for a given application.

The specific contributions are:

- 1) An overview of 21 objective measures is discussed in the statistics, social science, machine learning, and data mining literature. It is shown that application of different measures may lead to substantially differing orderings of patterns.
- 2) Several key properties are proposed that will help analysts to select the right measure for a given application. A comparative study of these properties is made using the twenty-one existing measures. Our

results suggest that we can identify several groups of consistent measures having similar properties.

3) This also illustrates two situations in which most of the measures become consistent with each other, namely, when support-based pruning and a technique known as table standardization are used. This method also demonstrates the utility of support-based pruning in terms of eliminating uncorrelated and poorly correlated patterns.

4) An algorithm is used for selecting a small set of tables such that domain experts can determine the most suitable measure by looking at their rankings for this small set of tables.

2.4 An Approach to Facilitate the Analysis of the Association Rules

The goal of the research presented in this paper is to enable the end users to analyze, understand and use the extracted knowledge in an intelligent system or to support in the decision-making processes. In the paper, the GART algorithm is proposed, which uses taxonomies to generalize association rules, and the RulEE-GAR computational module, that enables the analysis of the generalized rules. This method uses iterative taxonomy to generalize and then prunes redundant rules at each step.

During years, manual methods had been used to convert data in knowledge. However the use of these methods has become expensive, time consuming, subjective and non-viable when applied at large databases. The problems with the manual methods stimulated the development of processes of automatic analysis, like the process of Knowledge Discovery in Databases or Data Mining.

In the Data Mining process, the use of the association rules technique may generate large quantities of patterns which make it difficult for the analyst to analyze the resultant pattern. A way to solve the problem of the large quantities of patterns extracted by the association rules technique is the use of taxonomies in the step of post-processing Knowledge. The taxonomies may be used to prune uninteresting and/or redundant rules. In this paper the GART algorithm and the RulEE-GAR computational module is proposed. The GART algorithm (Generalization of Association Rules using Taxonomies) uses taxonomies to generalize association rules. The RulEEGAR computational module uses the GART algorithm, to generalize association rules, and provides several means to analyze the generalized rules. The RulEE-GAR computational module that provides means to generalize association rules and also to analyze the generalized rules. The screen of the interface enables the user to analyze and to explore the generalized rules sets.

3. Analysis of Association Rule mining technique

Parameters used for Comparison

an increasing amount of attention during the last few years, and quite a number of theoretical results, algorithms and implementations have been presented that explicitly aim at improving the efficiency and Scalability of multi-relational data mining approaches. Table1 shows the comparison of different association rule mining approaches.

Table 1 Comparison Table

| Parameters | 2.1 | 2.2 | 2.3 | 2.4 | Proposed Approach |
|---------------------------|-----|-----|-----|-----|-------------------|
| Scalability | Yes | Yes | Yes | Yes | Yes |
| User Interesting criteria | No | No | No | No | Yes |
| Quality | No | Yes | No | No | Yes |

Scalability: The system should be scalable with increase in amount of information.

User Interestingness Criteria: This depends on strong interaction with the user.

The comparison of different association rule mining approach is given in Table 1.

CLOSET is an efficient algorithm for mining frequent closed itemset.

Merits of CLOSET efficient algorithm:

- 1) Number of frequent items can be reduced.
- 2) Search space can be reduced.

DeMerits of CLOSET efficient algorithm:

This approach is based on statistical information and does not guarantee the rules are interesting for the user. There is no interactive approach to capture user interesting pattern.

Merits of Ontology Driven Rule extraction method:

The main advantages of the proposed framework can be summarized in terms of extensibility and flexibility.

1) The framework is extensible because data properties and concepts can be introduced in the ontology without either changing the relational database containing the transaction, or the implementation of our framework.

2) The flexibility is guaranteed from the separation of the data to analyze (the transactions) from the metadata (description of the data).

The main parameters we considered for the analysis of different association rule mining approaches are scalability, quality of filtered rules, user interesting criteria. Efficiency and Scalability have always been important concerns in the field of data mining, and are even more so in the multi-relational context, which is inherently more complex.

Demerits of Ontology Driven Rule extraction method:

- 1) The overhead in conducting pruning tests and as a result the execution time is high.
- 2) This paper uses seSQL to express user knowledge which is not as flexible as rule schema.

Merits of Taxonomy in Association Analysis Method:

- 1) Efficient approach to prune and generalize association rules.
- 2) Good approach to analyze the rules generation.

Demerits of Taxonomy in Association Analysis Method:

- 1) This method uses iterative taxonomy in order to generalize and then prune redundant rules at each step which results in more number of iterations.

Merits and Demerits of objective measure selection method:

This paper describe several key properties one should examine in order to select the right measure for a given application. Objective measure is restricted only for data evaluation not sufficient to reduce number of rules and to capture interesting one.

4. Proposed Approach

An Efficient and Interactive Post mining of Association Rules (Proposed Approach)

The proposed approach is composed of two main phases.

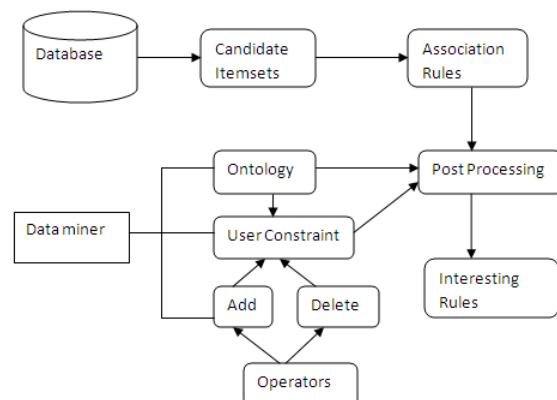


Fig. 3 Framework Description

The first phase includes the generation of support counts of item sets at each timeslot and candidate item

sets. The second phase involves mining of association rules from candidate items and post mining of association rules using ontology and user constraint template to guarantee user interesting rules as shown in Fig 3.

4.1 Mining Candidate Item sets in large databases(Steps 1-3)

The transaction database is scanned using the lattice-dominant scan method which reads a whole transaction data set from time slot t_1 to time slot t_n for calculating the support count of each item. The cost of computing the support counts of all combinations of item sets at each timeslot increases with increase in the number of items. Hence, to reduce the computational cost, the interesting properties like tight upper and lower bounds of support counts of single items are used. It is then used to estimate support count of item sets at each timeslot without examining an input data set. The upper bound and lower bound support counts at each time slot indicates the range in which the true support count of an item set can be located. The lower bounding distance is used in the pruning of candidate

4.2. Post mining of Association Rules

As the number of attributes and the number of transactions becomes large, thousands of rules are from a database. As the number of rules become huge, it is difficult for the data miner to analyze the mining results. Also it is impossible to use the results. Thus, it is crucial to help the decision-maker with an efficient technique for reducing the number of rules. The interestingness of the rule strongly depends on interactivity with the user. Existing methods do not guarantee that interesting rules can be extracted. To select the interesting rule, the user knowledge should be expressed in an accurate and understandable form. In data mining, background knowledge ontology organizes domain knowledge and plays important roles at several levels of the knowledge discovery process. Ontology provides an explicit representation of concepts in a domain, where each concept is a collection of items. Instance of a concept represents the ground level items. The subsumption relation between concepts shows is-a super class, is-a subclass relations. The concept-instance relation represents the relation between concepts and the instances. There are two types of concepts: leaf-concepts and generalized concepts from the subsumption relation¹. Leaf-concepts are connected in the easiest way to database—each concept is associated to one item in the database. Generalized concepts are described as the concepts that subsume other concepts in the ontology. A generalized concept is connected to the database through its subsumed concepts.

The Rule Schema formalism is based on the specification language for user knowledge introduced by Liu et al. The model proposed by Liu et al. is described using elements from an item taxonomy allowing an is-a organization of database attributes. Using item taxonomies has many advantages: the representation of user expectations is more general, and thus, filtered rules are more interesting for the user. However, taxonomy of items might not be enough. The user might want to use concepts that are more expressive and accurate. But, ontology includes the features of taxonomies and provides more representation power. In taxonomy, only the subsumption relationship is used to build the hierarchy. Thus, taxonomy is simply a hierarchical categorization or classification of items in a domain. In contrast, an ontology is a specification of several characteristics of a domain, defined using an open vocabulary. Dataminer develops ontology on the items in database. A user-constraint template is defined which allows the dataminer to select interesting rules according to his constraints. The user-constraint template can be represented in the following way:

UC<confectionery items=>grocery items>

We propose a matching operator for rule selection. The matching operator (M) selects the association rules that match with the user-specified constraint. When the matching operator is applied over user-constraint template M (UC), the antecedent and the consequent of the association rules should match.

1. Scan the transaction database to calculate the support count of the items at different timeslots (t_1, t_2, t_n)
2. Lower Bounding distance is found between support count of each item (at various timeslots) and the minimum_support_sequence.
3. If lower bounding distance < user -specified threshold (U), the item set is considered as the candidate item set.
4. From the candidate item set, association rules are generated.
5. In the post processing step, ontology is constructed to describe the domain in which the analysis is done.
6. User-Constraint template is created to specify the interestingness of dataminer.
7. Addition and Deletion operator is applied over the user-constraint template to select the interesting rules.

Merits of proposed method:

- 1) This approach can prune and filter the discovered rules
- 2) Guarantees the rules are interesting for the user
- 3) The use of ontology provides specification of several characteristics of a domain

Demerits of proposed method:

- 1) The task of mapping ontology concepts with the DB items is time consuming

5. Experimental Results

The study is based on the supermarket dataset. The real dataset is in .arff (attribute relation file format). The dataset contains 217 attributes and 4627 instances. T (10000) shows the total number of transactions.

Generating Association rules:

Association rule mining finds interesting associations and/or correlation relationships among large set of data items. Association rules shows attribute value conditions that occur frequently together in a given dataset.

In order to target the most interesting rules, we fix a minimum support of 2 percent, a maximum support of 30 percent, and a minimum confidence of 80 percent for the association rules mining process. Among available algorithms, we use the Apriori algorithm in order to extract association rules. The generated association rules describe the relationship between attribute.

Steps for Frequent item set and Association Rule Generation:

1. Scan the database to calculate the support of each item set.
2. Add the item set to frequent item set if support is greater than or equal to min_support.
3. At each level divide the frequent item set into left hand side and right hand side.
4. Calculate the confidence of each rule that is generated.
5. Generate strong rules satisfying min_support and min_confidence.

Fig 4. shows the user interface screen where the user can load the dataset. The dataset is in .arff file format. The input dataset contains header and data section.

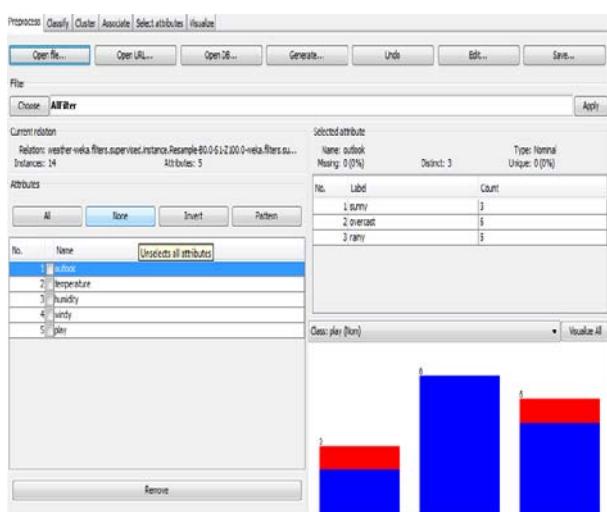


Fig. 4 Loading the Data set

$$\text{Support}(A \Rightarrow B) = P(A \cup B) \quad (1)$$

$$\text{Confidence}(A \Rightarrow B) = P(B|A) \quad (2)$$

Table 2 shows the comparison of number of interesting rules selected when matching operator is applied over user-constraint template and when not applied. The no. represents that each time different constraints are given to select different set of interesting rules. Our Experimental evaluation proves that the rules are generated and the selected rules are interesting to the user.

Table 2. Comparison of the Number of Rules with and without Applying Matching Operator

| No. | Matching Operator | Without Operator |
|-----|-------------------|------------------|
| 1. | 225 | 1000 |
| 2. | 162 | 1000 |
| 3. | 289 | 1000 |
| 4. | 77 | 1000 |
| 5. | 139 | 1000 |

5. Conclusion

This paper discusses the problem of selecting interesting association rules through huge volumes of discovered rules. This motivates the need for association analysis

This paper discusses a novel efficient approach to prune mined association rules in large databases. A fairly comparative analysis of different association rule mining techniques for market basket analysis, highlighting strengths of different approaches, potential pitfalls as well as challenging issues need to be addressed by an association rule mining technique are also discussed. We believe that the results of this evaluation will help decision maker for making important decisions. We have evaluated the algorithms based on parameters like scalability, quality of filtered rules. Our evaluation shows that an efficient approach to prune mined association rules approach should be efficient and produce user interesting rules.

Acknowledgments

First and foremost, I praise and thank ALMIGHTY GOD whose blessings have bestowed in me the will power and confidence to carry out my work. I feel it a pleasure to be indebted to my guide, **Ms. S. Geetha, M.Tech**, Assistant Professor, Department of Computer

Science and Engineering for her invaluable support, advice and encouragement.

I also thank **Mr.G.NaveenSundar, M.Tech (Ph.D)**, Assistant Professor, Department of Computer Science and Engineering for his valuable support in completing this work successfully

References

- [1] Claudia Marinica and Fabrice Guillet, "Knowledge-Based Interactive Postmining of Association Rules Using Ontologies," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, No. 6, June 2010.
- [2] D. Burdick, M. Calimlim, J. Flannick, J. Gehrke, and T. Yiu, "Mafia: A Maximal Frequent Itemset Algorithm," *IEEE Trans.Knowledge and Data Eng.*, vol. 17, no. 11, pp. 1490-1504, Nov.2005.
- [3] J. Li, "On Optimal Rule Discovery," *IEEE Trans. Knowledge and Data Eng.*, vol. 18, no. 4, pp. 460-471, Apr. 2006.
- [4] M.J. Zaki and M. Ogihara, "Theoretical Foundations Of Association Rules," Proc. Workshop Research Issues in Data Mining and Knowledge Discovery(DMKD '98), pp. 1-8, June 1998.
- [5] M.A.Domingues and S.A. Rezende, "Using Taxonomies to Facilitate The Analysis of the Association Rules," *Proc. Second Int'l Workshop Knowledge Discovery and Ontologies, held with ECML PKDD*, pp. 59-66, 2005.
- [6] J. Pei, J. Han, and R. Mao, "Closet: An Efficient Algorithm for Mining Frequent Closed Itemsets," Proc. ACM SIGMOD Workshop Research Issues in Data Mining and Knowledge Discovery, pp. 21-30, 2000.
- [7] A. Bellandi, B. Furletti, V. Grossi, and A. Romei, "Ontology- Driven Association Rule Extraction: A Case Study," *Proc. Workshop Context and Ontologies: Representation and Reasoning*, pp. 1-10, 2007.
- [8] M.J. Zaki and C.J. Hsiao, "Charm: An Efficient Algorithm for Closed Itemset Mining," Proc. Second SIAM Int'l Conf. Data Mining, pp. 34-43, 2002.
- [9] M.J. Zaki, "Generating Non-Redundant Association Rules," Proc. Int'l Conf. Knowledge Discovery and Data Mining, pp. 34-43, 2005
- [10] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Efficient Mining of Association Rules Using Closed Itemset Lattices," *Information Systems*, vol. 24, pp. 25-46, 1999.
- [11] P.-N. Tan, V. Kumar, and J. Srivastava, "Selecting the Right Objective Measure for Association Analysis," *Information Systems*, vol. 29, pp. 293-313, 2004.
- [12] A. Silberschatz and A. Tuzhilin, "What Makes Patterns Interesting in Knowledge Discovery Systems," *IEEE Trans. Knowledge and Data Eng.*, vol. 8, no. 6, pp. 970-974, Dec. 1996.
- [13] R.J. Bayardo, Jr., R. Agrawal, and D. Gunopulos, "Constraint- Based Rule Mining in Large, Dense Databases," *Proc. 15th Int'lConf. Data Eng. (ICDE '99)*, pp. 188-197, 1999.
- [14] J. S. Park, M. Chen, and P. S. Yu. An effective hash based algorithm for mining association rules. In *ACM SIGMOD Intl. Conf Management of Data*, May 1995.
- [15] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. InkeriVerkamo. Fast discovery of association rules. In U. Fayyad and et al, editors, *Advances in Knowledge Discovery and Data Mining*, pages307-328. AAAIPress, Menlo Park, CA, 1996.

Narmadha received the B.E degree in computer science and engineering from Francis Xavier Engineering College, Tirunelveli in 2006. She is currently working toward the MTech degree in Computer Science and Engineering of Karunya University. Her main research interests are Association Rule Mining and Web mining

Naveen Sundar received the B.E degree in computer Science and Engineering from C.S.I Institute of Technology, Thovalai in 2002. He received the MTech degree from Karunya University, Coimbatore in 2006. He is currently working toward the PhD degree. He is working as an assistant Professor in Computer Science department of Karunya University. His main research interests are Association Rule Mining, Databases, Web mining

S.Geetha received the M.Tech degree in Information Technology from Anna University, Coimbatore in 2009 and B.E degree in Computer Science and Engineering from SASTRA University formerly known as Shanmugha College of Engineering, Thanjavur in 2002. She is currently working as Assistant Professor in Dept. of Computer Science and Engineering, Karunya University, Coimbatore. Her areas of interests are in Data Mining and Network Security. She is a member of Computer Society of India (CSI).

Implementation of Reduced Power Open Core Protocol Compliant Memory System using VHDL

Ramesh Bhakthavatchalu¹, Deepthy G R²

¹ Dept. of ECE
Amrita Vishwa Vidyapeetham,
Amritapuri, Kollam-690525, Kerala, India

² Dept. of ECE
Amrita Vishwa Vidyapeetham,
Amritapuri, Kollam-690525, Kerala, India

Abstract

The design of a large scale System on Chip (SoC) is becoming challenging not only due to the complexity but also due to the use of a large amount of Intellectual Properties (IP). An interface standard for IP cores is becoming important for a successful SoC design. In a SoC the different IP cores are interfaced through different protocols. It increases the complexity of the design. Open Core Protocol (OCP) is an openly licensed core centric protocol intended to meet contemporary system level integration challenges. OCP promotes IP core reusability and reduces design time, design risk and manufacturing costs for SoC designs. OCP defines a highly configurable interface including data flow, control, verification and test signals required to describe an IP core's communication. This paper focuses on the design and implementation of a reconfigurable OCP compliant Master Slave interface for a memory system with burst support. The power reduction using Multivoltage design is the important feature of the paper. The proposed design was implemented in VHDL and the Synthesis is done using Synopsys ASIC synthesis tool Design Compiler.

Keywords: Memory Controller, Memory, Master, Slave, OCP compliant, Interface, Wrapper, Power Analysis, Burst Transfer, Power reduction, Multi voltage Design.

1. Introduction

Open Core Protocol is a signal exchange protocol over a family of on-chip core interfaces. OCP data transfer models range from simple request-grant handshaking through pipelined request-response to complex out-of order operations.

The OCP defines a point-to-point interface between two communicating entities, such as IP cores and bus interface modules (bus wrappers) [1]. Given the wide range of IP core functionality, performance and interface requirements, a fixed definition interface protocol cannot address the full

spectrum of system interface requirements. The need to support verification and test requirements adds an even higher level of complexity to the interface. To address this spectrum of interface definitions, the OCP defines a highly configurable interface. The OCP's structured methodology includes all of the signals required to describe an IP core's communications including data flow, control, and verification and test signals [2]. Since OCP is a core-specific, peer-to-peer protocol, OCP compliance IP cores can be verified independently with a Universal OCP monitor with OCP compliance assertions attached to OCP interface. In fact, to develop OCP compliance assertions, all possible aspects of features of the OCP protocol have to be integrated. OCP provides a master/slave connection between two cores. One core, the OCP initiator core has an OCP master interface. A master interface enables a core to generate OCP requests such as READ or WRITE and receive the READ responses. The other core, called the OCP target core, has an OCP slave interface which allows it to receive and respond to requests.

To simplify timing analysis, physical design, and general comprehension, the OCP is composed of unidirectional signals driven with respect to, and sampled by, the rising edge of the OCP clock. The OCP is fully synchronous and contains no multi-cycle timing paths with respect to the OCP clock. All signals other than the clock signal are strictly point-to-point. The OCP supports a configurable data width to allow multiple bytes to be transferred simultaneously. The OCP refers to the chosen data field width as the word size of the OCP.

2. Open Core Protocol Compliant System

The availability of a common interface platform provided by OCP has inspired system designers to use them as replacements for other interface protocols. OCP compliance is obtained by creating a wrapper around the original designs to meet the OCP specification. A wrapper is a design which satisfies all the specifications given by the Open Core Protocol. Then the designs are interfaced. For Compliance the core must include at least one OCP interface. The core and OCP interfaces must be described using an RTL configuration file. Each OCP interface on the core must comply with all aspects of the OCP interface specification. There are three types of OCP Profiles (i) High Performance (HP) (ii) Generic Profile (GP) (iii) Peripheral Profile (PP) [2]. The PP only implements the simple read/write transfer without other OCP extensions. GP extends the PP with additional data handshake phase and burst extensions for generic device with block data transfer. HP extends the GP with OCP tag extensions to support out-of-order response. OCP specifies 3 major types of interfaces. (i) Bus Bridge Interface (ii) Processor Interface (iii) Memory Interface [3]. The Bus bridge interface includes an external bus like USB or AXI and the internal bus will be OCP. In the Processor interface the interface is between processors which include only the OCP master. The memory interface is for DRAM, SRAM etc. OCP has been adopted by the industry with good results [4]. There is a large number of IPs with OCP interfaces at the top level. These OCP interfaces are different in protocol features or signals to optimize the needs of IP cores. However, all of them follow the same OCP timing and validation rules, which simplify the cost in verification and implementation [2].

3. Basic OCP Interface

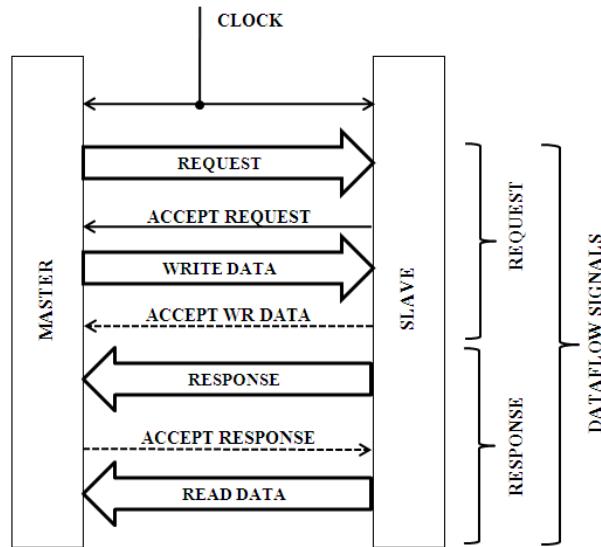


Fig. 1 Basic OCP Interface

Figure shows the basic OCP interface which consists of a Request and Response phase. The Master gives a Request to the slave, here a Write request and the Write data .the Slave accepts that request and gives a request accept signal back to the master. The slave then responds to the request and sends a response signal to the Master. Then according to the Master's request the data is read from the slave and is given to the master. The signals shown in dotted lines are optional. The basic OCP interface consists of dataflow signals only.

4. Memory Interface System

The scope of the research is that for understanding the OCP compliant system a model is needed since no such works exists in this field with experimental results. Hence a memory system with a Memory Controller as the OCP Master and Memory as OCP Slave is designed .First their performance is analyzed as the system itself is and then with the OCP wrapper.

The system includes an OCP compliant Memory Controller and a Memory where the memory controller acts as the OCP master and the Memory acts as the OCP Slave. This paper discusses the Peripheral OCP profile with Simple Write and Read transfer and Generic OCP profile with data handshaking and burst transfer [2].

The Master gives Requests and accepts responses. The slave receives and responds to the Requests provided by the master. Handshake signals are provided for both Master and Slave which indicates acknowledgements. The Memory designed can act as both program memory and

data memory and can be used as a memory system for any current day SoC design.

5. Implemented system

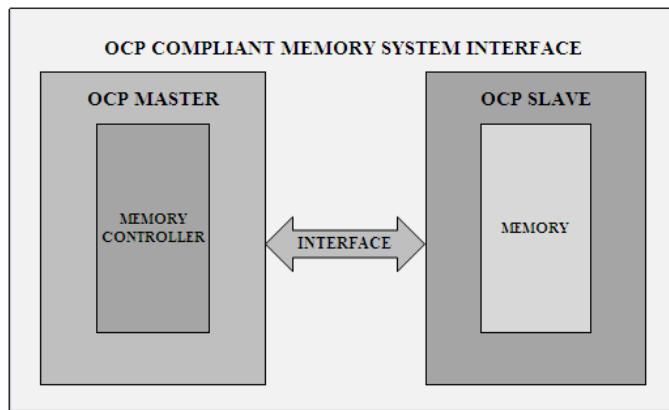


Fig. 2 Block Schematic of the Design Implemented

The proposed system is tested under 32-bit data bus and 8-bit address bus. The major control signals are Memwrite and Memread each of 1-bit length. Memory Controller gives Write and read requests as well as Address to the Memory. The Memory will respond to it by writing data into the memory address and reading data from the specified address and is given to the Memory controller. The slave responds with the response signals. During write operation the master starts a request phase by switching its command field to write and presents a valid data and address. The slave accepts the command and captures data and address and a write is performed according to the design. The master starts a Read request by switching its command field to Read. It presents a valid address and slave accepts the command. The slave captures data from the specified the address and is driven to the Master. The response is also given to master to indicate that the data is valid.

The inputs of the system are Address and Data which are in the form of Binary values. Another input is the 7-bit instruction which is given as the input of Memory Controller from which 2-bit opcode is extracted to determine the Memory operation (last 2 bits of the instruction).

There is no change in the designs of Memory Controller and Memory when OCP wrapper is introduced. The wrapper covers the existing designs to make them OCP compliant.

The proposed system is parameterizable for both address and data. However for the experiment the parameters are set as shown below.

Table 1: Design Parameters

| Design Parameter | Size(in bits) |
|------------------|---------------|
|------------------|---------------|

| | |
|--------------------|----|
| Address | 8 |
| Data | 32 |
| Instruction | 8 |
| Burst Length width | 4 |

5.1 Memory Controller without OCP wrapper

Memory Controller is the Master of the interface which controls the operations of the Memory. A simple Controller with Memwrite as the write control signal and Memread as the read control signals is designed as the Master [5], [6].

5.2 Memory without OCP wrapper

A simple SRAM is designed as the slave in the interface. It performs either write or read operation in response to the control signals from the master [5], [6].

5.3 OCP Compliant memory Controller

Memory Controller is the master who gives control signals to the memory which is the slave. The memory controller is reconfigurable. The control signals are for controlling the write and read operation of the memory. This memory controller is wrapped with an OCP wrapper which contains the basic OCP signals and will serve as an OCP master. OCP Master command is for Transfer command and this 3-bit signal indicates the type of OCP transfer the master is requesting. Each non-idle command is either a read or write-type request, depending on the direction of data flow. According to the Master command the slave will be either written into or read from [1].

5.4 OCP Compliant Memory

Memory is the slave which responds to the transfer requests provided by the master. Data can be written into the memory and read from the memory. This memory is enclosed within a wrapper which contains the basic OCP signals and will acts as an OCP slave. The response signals will be sending back to the master [1].

5.5 OCP Compliant Memory Interface

The entire system acts as a memory system interface which is a suitable alternative for a memory interface for any SoC design [3]. The design covers peripheral profile with simple read and write and Generic profile with data hand shaking and basic burst transfer [2].

5.6 OCP Compliant Memory Interface with Burst data transfer

A Memory System is not complete without Burst Transfer. Burst is a set of transfers that are linked together

into a transaction having a defined address sequence and number of transfers. There are three general categories of bursts. In Imprecise bursts, Request information is given for each transfer. Length information may change during the burst. In Precise bursts, Request information is given for each transfer, but length information is constant throughout the burst. Single request / multiple data bursts (also known as packets) is also a precise burst, but request information is given only once for the entire burst. To express bursts on the OCP interface, at least the address sequence and length of the burst must be communicated. The implemented design for burst transfer is having a word size of 32 and address width of 8. The address is incremented by 4 on each transfer and the Burst length is fixed as 4 throughout the entire transfer. In the research the Burst address sequence is selected as INCR, which is incrementing Burst. Three modes of Burst data transfer are discussed here. Burst write, Single Request Multiple Read Burst transfer and Burst Write with combined Request and data. In single request Multiple Read, the request is given only once and multiple data is read. In Burst Write with combined Request and data, the data is written and read as a burst.

6. Experimental Observations and Results

6.1 Design Setup

Table2: Design Setup

| | |
|-------------------------|--|
| Design method | VHDL based behavioral |
| Verification | Modelsim 6.3b |
| Synthesis platform | Xilinx ISE 10.1 , Synopsys Design Compiler(DC) |
| Hardware Platform | Xilinx Vertex 5 |
| Power Analysis Platform | Synopsys Design Compiler(DC) |

6.2 Simulation Results

Simulation output is shown for Memory Interface System with burst support. For the given burst transfer the simulation is done for 1500 ns.

This simple proof of concept design was used for verification of the propounded OCP compliant design.

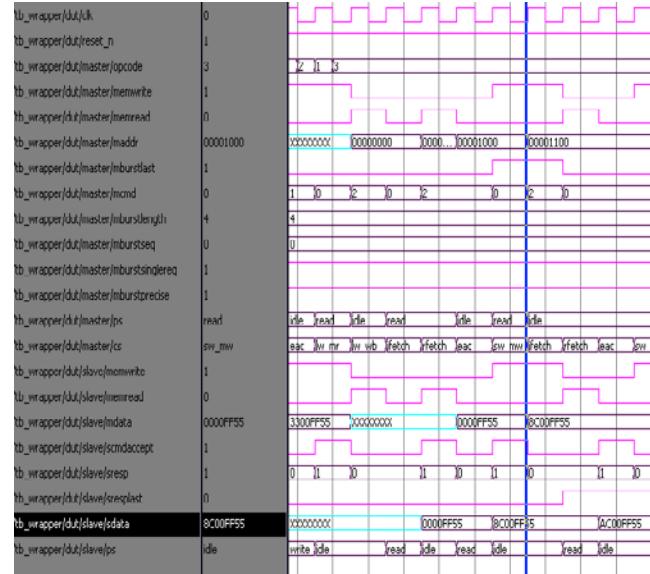


Fig. 3 Simulation Results with simple Burst transfer

Simulation results show that the signals are sampled at the rising edge of the clock signal (clk). OCP defines active low resets (reset_n). When the MCmd signal (Master Command) signal is 000 it is the idle state. When MCmd is 001 data is written into the memory and when MCmd is 010 the data is read from the memory. For precise Burst data transfer, the address is incremented by four and for 32 word size, MBurstLength is held constant as 4. When the last address comes MReqLast will be high. When the last data is read SRespLast will be high. Only the most important signals and responses are shown here.

6.3 Synthesis

Synthesis was done both in FPGA and ASIC platforms. The performance analysis is based mainly on power consumption and speed. The results obtained are as shown in the tables below.

Table3: Frequency Analysis

| Design | Frequency Of Operation |
|--------------------------|------------------------|
| With wrapper | 676.590MHz |
| With wrapper, with burst | 558.566MHz |
| Without wrapper | 633.309MHz |

Table shows that the speed of operation increases with the use of OCP wrapper. But addition of burst transfer decreases the speed of the system .It is obvious since introduction of burs will increase the minimum period of operation.

Another major analysis is on the power consumption.

Table3: Power Analysis (Xilinx Xpower Results)

| Design | Power in watts | |
|-----------------------------------|-----------------------|-------|
| Memory System Without Wrapper | Quiescent power | 0.034 |
| | Dynamic power | 0.007 |
| | Total power | 0.041 |
| Memory System With Wrapper | Quiescent power | 0.034 |
| | Dynamic power | 0.013 |
| | Total power | 0.047 |
| Memory System With Burst Transfer | Quiescent Power | 0.303 |
| | Dynamic Power | 0.042 |
| | Total power | 0.345 |

Table shows the increase in power with the OCP wrapper is not too high when compared with the original designs. Hence the proposed system can be used as an alternative to any memory systems. The burst transfer consumes high power since the frequency of operation is high as shown in the frequency analysis.

Further research is done with Synopsys tool design compiler which is based on 13 micrometer technology library. When the design is synthesized, each module is mapped to the gates and modules available in the specific technology Library.

In the analysis of power different modes of burst transfers are considered because the systems incorporating bursts are seemed to consume large power. The results are given in the table below.

Table5: Power Analysis with different burst transfers

| Design | Dynamic Power |
|---|----------------------|
| Memory Sytem Without Wrapper | 90.0266 uW |
| Memory Sytem With Wrapper | 100.7338 uW |
| Memory System With Simple Burst Transfer | 114.9890 uW |
| Memory System With Single Request Multiple Burst Transfer | 117.7919 uW |
| Memory System With Burst With Combined Request And Data | 117.8017 uW |

The results show that different Burst transfers have almost same power and hence any of them can be used efficiently according to the need.

5.4 Power Reduction:-Multi voltage Design

| Supply Voltage for Memory Controller | Design | Power In Millie Watts | % Power Saving |
|---|---------------|------------------------------|-----------------------|
| At 5V | Slave | 448.449 | 2.95% |
| | Master | 947.184 | |
| | Wrapper | 1.40e+03 | |
| At 4.75V | Slave | 442.647 | 2.95% |
| | Master | 910.250 | |
| | Wrapper | 1.35e+03 | |
| At 4.5V | Slave | 180.327 | 7.11% |
| | Master | 851.923 | |
| | Wrapper | 1.03e+03 | |

Energy efficiency has become a very important issue to be addressed in today's system-on-a-chip (SoC) designs. Multi supply voltage (MSV) is thus introduced to provide flexibility in controlling the power and performance tradeoff. One of the most effective ways is by lowering the voltage supply and has become the latest technique for power optimization. Multi voltage design provide "just enough" power to support different functional operations [11]. For dynamic power, a minor adjustment to the voltage level can result in a significant reduction in power consumption, which is proportional to the square of the voltage.

Design Compiler has a provision to set varying voltage for different designs. At the time of synthesis Synopsys uses a default voltage of 5V for 13um technology. This voltage is

| Supply Voltage for Memory Controller | Design | Power In Millie Watts | % Power Saving |
|---|---------------|------------------------------|-----------------------|
| At 5V | Slave | 448.449 | 2.95% |
| | Master | 947.184 | |
| | Wrapper | 1.40e+03 | |
| At 4.75V | Slave | 442.647 | 2.95% |
| | Master | 910.250 | |
| | Wrapper | 1.35e+03 | |
| At 4.5V | Slave | 397.279 | 7.11% |
| | Master | 816.955 | |
| | Wrapper | 1.21e+03 | |

kept the same for the Memory Block since it contains the critical path. It is evident that the decrease in voltage will increase the delay. The voltage for the Memory Controller is reduced to 4.75 and then to 4.5 V. Percentage saving in power is given as %saving = Total power - (Reduced Master Power + Original slave Power) / Total power.

The power of both slave and master is reduced .But the power reduction analysis is done by keeping the power of slave as same as that at 5V since the aim is to reduce the power of master only since it has less delay. Hence the

power delay product is maintained which is used as the parameter to find the operating voltage of the circuit at which energy dissipation is minimal.

The results of Multi voltage designs are shown in the tables below.

Table6: Power Reduction for system without Burst

Table7: Power Reduction for system with Burst

Table8: Power Reduction for system with Burst Single Request Multiple data

Table9: Power Reduction for system with Burst Combined Request and Data

| Supply Voltage for Memory Controller | Design | Power In Millie Watts | % Power Saving |
|---|---------------|------------------------------|-----------------------|
| At 5V | Slave | 139.266 | 4% |
| | Master | 923.895 | |
| | Wrapper | 1.06e+03 | |
| At 4.75V | Slave | 139.020 | 4% |
| | Master | 878.095 | |
| | Wrapper | 1.02e+03 | |
| At 4.5V | Slave | 124.771 | 12.51% |
| | Master | 788.096 | |
| | Wrapper | 912.868 | |

From the tables it is clearly shown that the power can be reduced significantly with Multi voltage design for different Memory Transfers.

6. Conclusions

A parameterizable and reconfigurable OCP compliant memory system specifically targeted to use with high speed applications is discussed here. The primary trigger to the development of such design is the lack of availability of a common interface that can be used with the different IP cores in a SoC design. This paper discusses the use of OCP for a memory system interface and concentrates on enhancing the memory system performance with different modes of Burst data transfer. Power Reduction with Multi voltage design is implemented with good results.

References

- [1] “Open Core Protocol Specification 3.0”, International Partnership, 2000- 2009 OCP-IP Association, Document Revision 1.0.
- [2] Chih-Wea Wang, Chi-Shao Lai, Chi-Feng Wu, Shih-Arn Hwang, and Ying-Hsi Lin, “On-chip Interconnection Design and SoC Integration with OCP”, Proceedings of VLSI-DAT, 2008, pp. 25 – 28, April 2008.
- [3] OCP-IP, “Open core protocol international partnership,” <http://www.ocpip.org/>, 2007.
- [4] James Aldis, “Use of OCP in OMAP 2420” <http://www.ocpip.org/>, 2005.
- [5] www.mips.com, “Computer Architecture and Engineering”, Lecture 8 ,Designing a Multicycle Processor.
- [6] David A. Patterson, John L.Hennessy, “Computer Organization and Design”, Third Edition,Morgan Kaufmann Publishers, pp.318-339.
- [7] Shihua Zhang, Asif Iqbal Ahmed and Otmane Ait Mohamed, “A Re-usable verification Framework of Open Core Protocol”, Circuits and Systems and TAISA Conference, 2009, pp. 1-4, june 28,2009.
- [8] W.-D. Weber, “Enabling reuse via an IP core-centric communications Protocol”, In Proc. IP 2000 System-on-Chip Conference, pages 217-224, Mar 2000.
- [9] Prashant D. Karandikar, “Open Core Protocol (OCP) An Introduction to Interface Specification”, 1st Workshop on SoC Architecture, Accelerators & Workloads Jan 10 2010.
- [10] Chien-Chun (Joe) Chou, Konstantinos Aisopos, David Lau, Yasuhiko Kurosawa and D. N. (Jay) Jayasimha, “Using OCP and Coherence Extensions to Support System-Level Cache Coherence”, Technical Paper, pg. nos.10, April 2009.
- [11] Qiang Ma and Evangeline F. Y. Young, Multivoltage Floor plan Design, IEEE transactions on computer-aided design of integrated circuits and systems, vol. 29, no. 4, April 2010 607

A Novel Energy Efficient Mechanism for Secured Routing of Wireless Sensor Network

Anupriya Sharma¹, Paramjeet Rawat², Suraj Malik³, Sudhanshu Gupta⁴

¹ Computer Science, GBTU, IIMT Engg. College Meerut

² Computer Science, GBTU, IIMT Engg. College Meerut

³ Computer Science, GBTU, IIMT Engg. College Meerut

⁴ Computer Science, GBTU, BIT Engg. College Meerut

Abstract

Large-scale wireless sensor networks are highly vulnerable to attacks because they consist of numerous resource-constrained devices and communicate via wireless links. As wireless sensor networks are continue to grow, so they need an effective security mechanisms. As sensor networks may interact with sensitive data and/or operate in hostile unattended environments, it is imperative that these security concerns must be addressed from the beginning of the system design. However due to inherent resource and computing constraints, security in sensor networks poses different challenges than traditional network computer security. Here we describe an energy efficient security scheme for sensor networks that is designed for long lived networks. Primary features of our scheme include autonomously computing administration keys and dynamically mapping of sensor nodes to set of keys. The scheme scales well in the size of the network and supports dynamic setup and management of arbitrary structures for secure communications in large-scale wireless sensor network. A salient feature of the security scheme is that, it supports source authentication as well as end-to-end authentication, integrity of communication, efficiently addition of the sensor nodes to the network dynamically.

Keywords: wireless, sensor, security, vulnerability, source authentication, energy efficient, integrity

1. INTRODUCTION

A wireless sensor network is a network of simple sensing devices; which are capable of sensing some changes of incidents/parameters and communication with other devices, over a specific geographic area for some specific purposes like target tracking, surveillance, environmental monitoring etc. Since sensor nodes are tightly constrained in processing ability, storage capacity and energy and secured routing over Wireless Sensor Networks (WSN) presents a unique challenge. Mature solutions [1] for key management are too complex for wireless sensor networks, as the resources required would quickly exhaust the small sensors. The suitability of these WSNs for military applications and the deployment of these networks in hostile environments have brought the challenge of securing the communication between these extremely resource constrained devices. In addition to battlefield

deployment, there are a number of future applications that will require a high level of security.

The extensive growth in using sensor networks in a wide variety of applications ranging from health care to warfare is fueling extensive research in securing these networks. The characteristics of sensor nodes and sensor networks including lack of physical protection and the resource constrained nature of sensors render most existing security solutions developed for other networks (e.g. Public-Key-based solutions) infeasible for sensor networks.

The tradeoff between managing acceptable levels of security and conserving network energy for sensor network operation is a challenging task. Recently, a number of security schemes have been developed for sensor networks [2,3,4]. We broadly classify these security schemes for sensor networks into static and dynamic keying based on whether the administrative keys (those used to establish communication keys) are distributed or updated or on the basis of initial network deployment and setup. Unlike static keying, dynamic keying schemes change all keys revealed to an attacker upon node capture.

The major advantage of dynamic keying is enhanced network survivability, so that captured keys are replaced in a timely manner. Our main contribution is purposing an power-efficient and scalable dynamic key management scheme for secured communication over sensor networks. Our scheme is a dynamic key management scheme in which a set of keys is assigned to every sensor node after its deployment and periodic refreshment of network for verifying the nodes which are alive over the lifespan of a network.

2. RELATED WORK

Key management schemes [2,4] in sensor networks can be classified broadly into dynamic or static solutions based on whether re keying (update) of administrative keys is enabled post network deployment. Schemes can also be classified into homogeneous or heterogeneous schemes with regard to the role of network nodes in the key management process. All nodes in a homogeneous

scheme perform the same functionality; on the other hand, nodes in a heterogeneous scheme are assigned different roles. Homogeneous schemes generally assume a flat network model, while heterogeneous schemes are intended for both flat and clustered networks. Other classification criteria include whether nodes are anonymous or have pre deployment identifiers and if so, when (pre- post-deployment or both) and what deployment knowledge (location, degree of hostility, etc.) is imparted to the nodes.

2.1 Static Key Management Scheme

These schemes assume that once administrative keys are[2] pre deployed in the nodes, they can not be changed. Administrative keys are generated prior to deployment, assigned to nodes either randomly or based on some deployment information and then distributed to nodes. For communication key management, most static schemes use the overlapping of administrative keys to determine the eligibility of neighboring nodes to generate a direct pair-wise communication key. In order to establish and distribute a communication key between two non-neighboring nodes and/or a group of nodes, that key is propagated one link at a time using previously established direct communication keys. All of the static schemes are homogenous and not reliant on post deployment information. Several techniques have been proposed to make use of deployment knowledge in order to improve static key management. Deployment knowledge may include node locations, neighbor locations, node cluster (or group), as well as the attack probability in certain portions of the network.

2.2 Dynamic Key Management Scheme

Dynamic key management schemes may change administrative keys periodically on demand or on detection of node capture. The major advantage of dynamic keying is enhanced network survivability, since any captured keys are replaced in a timely manner in a process known as re-keying. Another advantage of dynamic keying is providing better support for network expansion, upon adding new nodes, unlike static keying, which uses a fixed pool of keys, the probability of network capture does not necessarily increase. Both homogeneous and heterogeneous dynamic key management schemes have been proposed in the literature. The major challenge in dynamic keying is to design a secure yet efficient re keying mechanism. A proposed solution to this problem is *Jolly et al.*'s approach; key generation and assignment are the responsibility of the base station, while key distribution is performed by the cluster gateways. The proposed scheme requires very few keys to be stored at each

sensor node and shared with the base station as well as the cluster gateways.

Re keying involves reestablishment of clusters and redistribution of keys. Although the storage requirement is very affordable, the re keying procedure is inefficient due to the large number of messages exchanged for key renewals..

Another researchers Du et al. [5] proposed a novel random key predistribution scheme that exploits deployment knowledge and avoids unnecessary key assignments. It shows that the performance (including connectivity, memory usage, and network resilience against node capture) of sensor networks can be substantially improved. This scheme is based on known deployment points by choosing keys shared with nodes likely to be in close proximity.

Carman et al. [] conducted a comprehensive analysis of various group key schemes. The authors conclude that the group size is the primarily factor that should be considered when choosing a scheme for generating and distributing group keys in a WSN.

LEAP

The existing protocol LEAP [6] (Localized Encryption and Authentication Protocol) that provides the security for wireless sensor networks has the following properties:

- The design of the protocol is motivated by the observation that different types of messages exchanged between sensor nodes have different security requirements, and a single keying mechanism is not suitable for meeting these requirements. Consequently, LEAP includes support for establishing four types of keys per sensor node – individual keys are shared with the base station, pair wise keys shared with individual neighboring nodes, cluster keys shared with a set of neighbors, and a group key shared with all the nodes in the network. These keys can be used to increase the security of many non-secure protocols.
- LEAP includes an efficient protocol for inter-node traffic authentication based on the use of one-way key chains.
- A distinguishing feature of LEAP is that its key sharing approach supports in-network processing, while at the same time it restricts the security impact of a node compromise to the immediate network neighborhood of the compromised node.
- The key establishment and key updating procedures used by LEAP are efficient and the storage requirements per node are small.
- LEAP can prevent or increase the difficulty of launching many security attacks on sensor networks.

3. ISSUES WHICH NEED TO BE ADDRESSED

The following characteristics of sensor networks[1] complicate the design of secure protocols for sensor networks and make the bootstrapping problem highly challenging. We discuss the origins and implications of each factor in turn.

3.1 Impracticality of public key cryptosystems

The limited computation and power resources of sensor nodes often makes it undesirable to use public-key algorithms, such as Diffie-Hellman key agreement or RSA signatures. Currently, a sensor node may require on the order of tens of seconds up to minutes to perform these operations. This exposes a vulnerability to denial of service (DoS) attacks.

3.2 Vulnerability of nodes to physical capture

Sensor nodes may be deployed in public or hostile locations (such as public buildings or forward battle areas) in many applications. Furthermore, the large number of nodes that are deployed implies that each sensor node must be low-cost, which makes it difficult for manufacturers to make them tamper-resistant.

3.3 Lack of a-priori knowledge of post-deployment configuration

If a sensor network is deployed via random scattering[7] (e.g. from an airplane), the sensor network protocols cannot know beforehand which nodes will be within communication range of each other after deployment. Even if the nodes are deployed by hand, the large number of nodes involved makes it costly to pre-determine the location of every individual node. Hence a security protocol should not assume prior knowledge of which nodes will be neighbors in a network.

3.4 Limited memory resources

The amount of key-storage memory in a given node is highly constrained. It does not possess the resources to establish unique keys with every one of the other nodes in the network.

3.5 Over-reliance on base stations exposes vulnerabilities

In a sensor network, base stations are few and are much powerful. Hence it may be tempting to rely on them as a source of trust. However this invites attack on the base station and limits the application of the security protocol.

4. PROPOSED WORK

We describe below our assumptions regarding the sensor network scenarios in which our security protocols will be used-

Network and security assumptions:

- i. We assume that the sensor network is static, i.e. sensor nodes are not mobile.
- ii. The base station, acting as a controller (or key server), is assumed to be a laptop class device and supplied with long-lasting power. The sensor nodes are similar in their computational and communication capabilities and power resources to current generation sensor nodes. The base station is part of a trusted computing environment.
- iii. We make the assumption that the communication channel is symmetric.
- iv. The sensor nodes can be deployed via aerial scattering or by physical installation.
- v. We assume that if a node is compromised, all the information it holds will also be compromised. However we assume the base station will not be compromised.
- vi. The sensors nodes are randomly distributed and are not aware of the topology prior to the deployment.
- vii. We are not making any trust assumptions on sensor nodes or any assumptions on the capabilities of the adversary.
- viii. Sensor nodes remain stationary during the operation of the network.
- ix. In addition we assume that the base station is capable of reaching all sensor nodes within its network through broadcast.

The main goal of our protocol is to design efficient security mechanisms for supporting various communication models in sensor networks. The security requirements not only include authentication and confidentiality but also robustness and survivability. The protocol should also support sensor network optimization mechanisms such as in-network processing. Since the resources of a sensor node are very constrained, the key establishment protocols should be lightweight and minimize communication and energy consumption. It should be possible to add new sensor nodes incrementally to the sensor network.

Our goal is to efficiently source communication among sensor nodes, end-to-end authentication, and confidentiality and integrity attacks in long-lived large-scale sensor networks operating in hostile environment. Our protocol manages two types of keys, one which is shared between individual sensor node and base station and the second which a sensor node is sharing with its neighboring sensor nodes. Node capture attacks, including the capture of sensor nodes are handled through same levels of re-keying (or changing administrative keys). As previously discussed, our protocol provides multiple keying mechanisms that can

be used for providing confidentiality and authentication in sensor networks. Sensor nodes are preloaded with a unique sequence number prior to deployment and certain code (initial key) that they will share with base station. Initial communication among nodes and base station is encrypted with these keys that they will share with base station.

We first motivate and present an overview of the different keying mechanisms before describing the protocol used by our protocol for establishing these keys.

As the study reveals that no single keying mechanism is appropriate for all the secure communication that is needed in sensor networks. As such our protocol supports the establishment of two types of keys for each sensor node, an individual sensor node key shared with the base station, a pair wise key shared with another neighboring sensor node. We now discuss each of these keys in turn and describe our reasons for including it in our protocol.

A. Individual key

Every node has a unique key that it shares pair wise[2] with the base station. This key is used for secure communication between a node and the base station

B. Pair wise shared key

Every node shares a pair wise key[2] with each of its immediate neighbors. In our protocol, pair wise keys are used for securing communications that require privacy or source authentication and which provide an end to end authentication.

Key Establishment

We describe the schemes provided by our protocol for sensor nodes to establishment of individual keys and pair wise shared keys for each sensor node. There is a list of notations which are used in our protocol-

1. N is the number of sensor nodes in the network
2. $A \longrightarrow B$ message is transmitting from sensor node A to B.
3. Key_k is a key at node k used for communication by the node k .
4. $Encrpt_key$ is the encryption key which is shared between any sensor node and base station.
5. NT_k is the neighbor table maintained locally by sensor node k in the network.
6. NT_{BS} is the neighbor table maintained at the base station.
7. $KeyTable_{BS}$ is the key table maintained at the base station, which maps the set of keys which are assigned to any sensor node in the network.
8. $AliveTable$ is the table maintained at base station to keep the list of alive sensor nodes in the network

4.1 Establishing Individual Node Keys

Every sensor node has a 128 bit secret key that is only shared with the base station. This key is generated and preloaded into each node prior to its deployment. When base station needs to communicate with an individual node k , it used that key which is shared with the sensor node. Due to the computational efficiency of base station, the computational overhead is negligible.

4.2 Establishing Pairwise shared keys

A pair wise shared key belonging to a sensor node refers to a key shared only between the node and one of its direct neighbors (i.e. one-hop neighbors). Here we are interested in establishing pair wise keys for sensor nodes unaware of their neighbors until their deployment (e.g. via aerial scattering). Our approach exploits the special property of sensor networks consisting of stationary nodes that the set of neighbors of a node is relatively static and that a sensor node that is being added to the network will discover most of its neighbors at the time of its initial deployment.

Pre-deployment initialization

The pre-deployment phase securely implants the initial key in all nodes. One major advantage of our protocol is that all sensor nodes are preloaded with a unique sequence number. In addition to the unique sequence number, each node is also preloaded with a 128 bit secret key which it shared with the base station. A pre-deployment initialization includes loading the entire set of sensor nodes with the sensor node id's. It is not

required that the base station knows the location of the all sensor nodes before or after deployment.

Post-deployment initialization

After deployment the network starts a top-down bootstrapping process beginning at the base station and proceeding downwards to the sensor nodes. The communication message format is following type:

$\langle S_Addr, (key_k(D_Addr), TYPE), Encrpt_key[data] \rangle$

Where:

1. S_Addr will contain address of sending node.
2. D_Addr contains the address of the destination.
3. $TYPE$ is the type of message that is being transmitted.
4. $Encrpt_key[data]$ is the encrypted data sent from one node to the other

The different types of communication message used in the key distribution algorithm are:

1. $HELLO_BS$ - is the broadcast message from the base station to all sensor nodes in the network.
2. $HELLO_SN$ - is the broadcast message from any sensor node to all sensor nodes in the network.
3. $HELLO_SNREPLY$ - is the reply of the broadcast message from any sensor node to other sensor nodes in the network.
4. $NLIST$ - this message is generated by any sensor node in the network and it contains the neighbor list of any sensor node in the network.
5. $KEYS$ - this message is generated by the base station and contains the set of keys assigned to any sensor node in the network.

After deployment base station sends a broadcast to all sensor nodes in the network to send their neighbors information. All sensor nodes in the network then broadcast a neighbor discovering hello message in the network. All the sensor nodes which hear message will reply to their immediate neighbor nodes by sending their unique ID to it. Then each sensor node use the secret key to register with the base station by sending their ID and neighbor table information encrypted with shared key with the base station. Upon receiving such messages, the base station registers all sensor nodes and determines the number of valid sensor nodes and accordingly computes suitable set of key values for each sensor node.

ALGORITHM: Key Distribution Algorithm

N : all sensor nodes in the network

1. Set $NT_k = \emptyset$
[where k belongs to N . Neighbor table at the each sensor node is initially empty and size of NT_k is at most $|N|$.]
2. Set $NT_BS = \emptyset$

[Neighbor table at the base station is initially empty and the size of NT_BS is at most $|N| * |N|$.]

3. Base station(BS) broadcast message:
 $\langle BS, HELLO, NULL \rangle$
to all nodes to collect the neighbor information in the network .
4. Each node of sensor network broadcast a Hello message: $\langle S_Addr[k], HELLO_SN, NULL \rangle$ to collect the neighbor information.
5. If a sensor node replies with the message:
 $\langle S_Addr[j], HELLO_SNREPLY, Node_ID \rangle$ to other sensor node then it is added to the neighbour list of the previous sensor node.
6. $NT_k = NT_k + j$
[where j is the node that is replying to the node k . So its ID is added in the Neighbor table of k .]
7. Then the sensor nodes which requires keys, sends its neighbour information to the base station
 $\langle S_Addr[k], BS, Encrpt_key[NT_k] \rangle$
8. Then Base station updates its neighbour table by adding the ID of node k in its Neighbor Table.
 $NT_BS = NT_BS + k$
9. Base station sends the set of keys to the sensor nodes.
 $\langle BS, S_Addr[k], Encrpt_key[NT_k] \rangle$
10. Base stations updates its Key Table:
 $KeyTable = KeyTable + k$
and updates the Alive table:
 $AliveTable = AliveTable + k$

Then base station sends the set of keys to each sensor node encrypted with the shared key. The base station maintains a key table with it, where it will keep track of the set of keys allocated to each sensor node in the network. Base station also maintains an *AliveTable* by which it will keep track of all alive sensor nodes in the network.

5. PERFORMANCE EVALUATION

In the case of our sensor network the security requirements are comprised of authentication, integrity, privacy (or confidentiality). The recipient of a message needs to be unequivocally assured that the message came from its stated source. Similarly the recipient needs to be assured that the message was not altered in transit and that it is not an earlier message being replayed in order to veil the current environment. Finally all communications need to be kept private so that eavesdroppers cannot intercept, study and analyze and devise counter measures in order to circumvent the purposes of the sensor network. The simulation implements application using the Network Simulator-2 (NS-2) tool and the MannaSim, which is a framework made of a set of base classes that extends NS-2 to simulate sensor networks. The Mannasim Framework is a module for wireless sensor network simulation based

on the Network Simulator (NS-2). Mannasim extends NS-2 introducing new modules for design, development and analysis of different wireless sensor network applications. We have taken the observations between 500 to 5000 nodes by incrementing 500 node at each step.

In simulated environment when we scattered 500 node and further randomly add 10 nodes and delete 10 node from the network we measured some standard results of energy radiation

After gathering the data from different observations, following graphs are obtained that compares our receiving energy with receiving energy used in LEAP [7]. This graph shows that energy consumed in our protocol is less than the previous results.

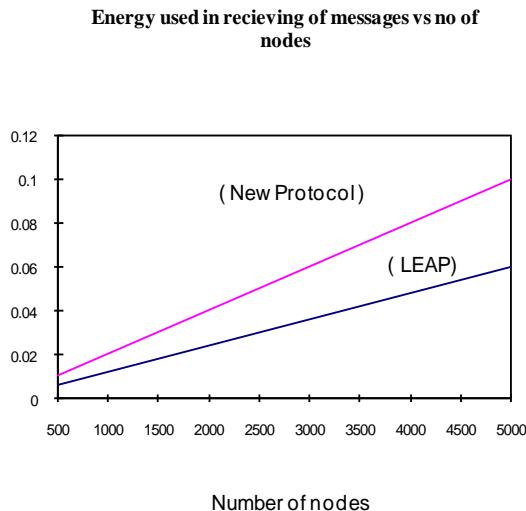


FIGURE 1

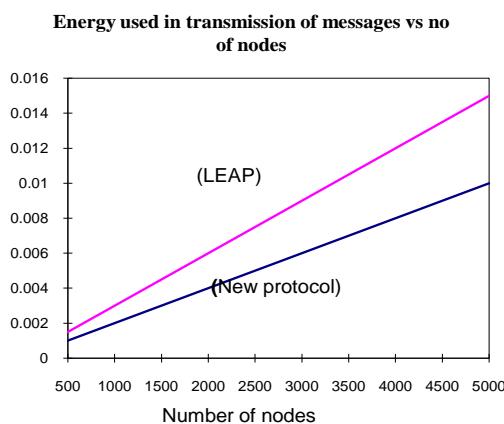


FIGURE 2

RESULT ANALYSIS

Analysis of the communication cost and storage requirement of this new key establishment scheme is described below:

5.1 COMMUNICATION COST

The analysis of communication cost for distributing keys to the sensor nodes depends on the energy used in number of message transmitted and number of message received in the key establishment phase. For establishing the sensor node administrative keys, the average number of message transmitted is $1+d$ where d is density of the network and the average number of message received are equal to $1+d+X$, where X is the number of nodes in the transmission range of sensor node in a network of size N . The average communication cost increases with the connection degree of a sensor network.

5.2 STORAGE REQUIREMENTS

In this scheme, a sensor node needs to keep two types of keys. If a node has d neighbors, it needs to store one individual key, d pair wise keys. In a sensor network the packet transmission rate is usually very small. Thus a node could store a reasonable length of key chain. Let d be the number of keys a node stores for its neighbor information. Thus the total number of keys a node stores is $d+1$. Therefore the total amount of memory needed is $128(1+d)+e$ bits, where e is a constant term used to signify the memory used by various things like encryption algorithm, neighbor table etc. Although memory space is a very scarce resource for the current generation of sensor nodes, for a reasonable degree d , storage is not an issue in this scheme. For example, when $d = 20$, a node stores 21 keys (totally 336 bytes when the key size is 128 bits). Overall, we conclude that the new scheme is scalable and efficient in computation, communication and storage.

5.3 SECURITY ANALYSIS

In analyzing the security of the keying mechanisms, firstly discuss the survivability of the network when undetected compromises occur and then study the robustness of the scheme in defending against various attacks on routing protocols. When a sensor node u is compromised, the adversary can launch attacks by utilizing node u 's keying materials. If the compromise event is detected somehow, our scheme can revoke node u from the group efficiently. Basically the base station and every neighbor of node u delete its pair wise key shared with u and update their key set. After the revocation the adversary cannot launch further attacks.

Our Pair wise keys provide source authentication as well as end-to-end authentication. The basic scheme for authentication is, every node authenticates a packet it transmits using its own key, which it is sharing with its neighboring nodes. A receiving node first verifies the packet using the same key that it shared with the sending node in the pair wise key establishment phase then authenticates the packet to

its own neighbors with its own that it shared with its neighbor's key. Thus a message gets authenticated repeatedly in a hop-by-hop fashion if it traverses multiple hops. The approach provides immediate authentication (node to node) as well as end-to-end authentication.

6. CONCLUSION

Design requirements for security scheme include energy awareness, survivability and localization of attack impact given a highly vulnerable network that mainly operates unattended and scalability to a large dynamic network. One major challenge to dynamic keying schemes is the need for the participation (to varying degrees) of a key management authority (usually the base station) post network deployment. In this paper, we presented an energy efficient security scheme for wireless sensor network which provides an end-to-end and inter node authentication for all communication in an efficient manner. The design of the security scheme is motivated by the observation that different types of messages exchanged between sensor nodes have different security requirements and that a single keying mechanism is not suitable for meeting these different security requirements. Consequently our scheme includes support for establishing two types of keys per sensor node individual keys which are shared with the base station, pair wise keys shared with individual neighboring nodes in the network. A distinguishing feature of our scheme is that it restricts the security impact of a node compromise to the immediate network neighborhood of the compromised node. The key establishment and key updating procedures for a compromised is used by our scheme.

REFERENCES

1. Akylidz, W. Su, Y. Sankarasubramaniam and E. Cayirci. Wireless Sensor Networks: A Survey. Computer Networks, March 2002.
2. Du W, Deng J, Han YS, Varshney PK. A pairwise key predistribution scheme for Wireless Sensor Networks. In Proceedings of the 10th ACM Conference on Computer and Communications Security (CCS-03), Washington D.C., October 2003.
3. Chorzempa M, Park JM, Eltoweissy M. SECK: survivable and efficient keying in Wireless Sensor Networks. IEEE Workshop on Information Assurance in Wireless Sensor Networks, WSNIA-2005, April 2005.
4. Eltoweissy M, Wadaa A, Olariu S, Wilson L. Group key management scheme for large-scale Wireless Sensor Network. Ad-Hoc Networks 2005.
5. Du W, Deng J, Han YS, Varshney PK. A pairwise key predistribution scheme for Wireless Sensor Networks. In Proceedings of the 10th ACM Conference on Computer and Communications Security (CCS-03), Washington D.C., October 2003.
6. Zhu S, Setia S, Jajodia S. LEAP: Efficient security mechanisms for large-scale distributed sensor networks. In Proceedings of the 10th ACM Conference on Computer and Communications Security (CCS-03), Washington D.C., October 2003.
7. Jolly G, Kuscu M, Kokate P, Younis M. A low-energy key management protocol for Wireless Sensor Networks. In Proceedings of the IEEE Symposium on Computers and Communications (ISCC-03). Kemer-Antalya, Turkey, June 2003.
8. D. Liu, P. Ning, and W. Du. Group-Based Key Pre-Distribution in Wireless Sensor Networks, Proc. 2005 ACM Wksp. Wireless Security (WiSe 2005), Sept. 2005.
9. M. Younis, K. Ghumman, and M. Eltoweissy. Location aware Combinatorial Key Management Scheme for Clustered Sensor Networks, to appear, IEEE Trans. Parallel and Distrib. Sys., 2006.
10. Liu and P. Ning. Improving Key Pre-Distribution with Deployment Knowledge in Static Sensor Networks, ACM Trans. Sensor Networks, 2005.
11. L. Eschenauer and V. Gligor. A Key Management Scheme for Distributed Sensor Networks, Proc. 9th ACM Conf. Comp. and Commun. Sec., Nov. 2002.
12. J. Agre, L. Clare, An integrated architecture for cooperative sensing networks, IEEE Computer Magazine (May 2000).
13. N. Bulusu, D. Estrin, L. Girod, J. Heidemann, Scalable coordination for wireless sensor networks: self-configuring localization systems, International Symposium on Communication Theory and Applications, July 2001.
14. A. Cerpa, D. Estrin, ASCENT: adaptive self-configuring sensor networks topologies, UCLA Computer Science Department Technical Report, May 2001.
15. A. Chandrakasan, R. Amirtharajah, S. Cho, J. Goodman, G. Konduri, J. Kulik, W. Rabiner, A. Wang, Design considerations for distributed micro-sensor systems, Proceedings of the IEEE 1999 Custom Integrated Circuits.
16. Chien, I. Elgorriaga, C.McConaghy, Low-power directsequence spread-spectrum modem architecture for distributed wireless sensor networks, ISLPED'01, Huntington Beach, California, August 2001.
17. S. Cho, A. Chandrakasan, Energy-efficient protocols for low duty cycle wireless microsensor, Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, January 2000.
18. R. Colwell, Testimony of Dr. Rita Colwell, Director, National Science Foundation, Before the Basic Research Subcommitte, House Science Committe, Hearing on Remote Sensing as a Research and Management Tool, September 1998.

Anupriya Sharma, M.Tech(CS), B.Tech(CS) currently working as Asst. Prof at IIMT Engg. College Meerut presently working on WSN and published few papers on WSN in reputed International Journal.

Paramjeet Rawat, Ph.D(P), M.Tech(CS), MCA, currently working as Asst. Prof at IIMT Engg College Meerut having 12 years of Teaching Experience, member of ACM, published several papers in reputed International Journals.

M-AODV: AODV variant to Improve Quality of Service in MANETs

Maamar Sedrati¹, Azeddine Bilami² and Mohamed Benmohamed³

**¹ UHL Batna , Department of computer science
Batna, Algeria**

**² UHL Batna , Department of computer science
Batna, Algeria**

**³ UMC Constantine, Department of computer science
Constantine, Algeria**

Abstract

Nowadays, multimedia and real-time applications consume much network resources and so, need high flow rates and very small transfer delay. The current ad hoc networks (MANETs), in their original state, are not able to satisfy the requirements of quality of service (QoS). Researches for improving QoS in these networks are main topics and a subject of intensive researches. In Adhoc networks, the routing phase plays an important role for improving QoS. Numerous routing protocols (proactive, reactive and hybrid) were proposed. AODV (Adhoc On demand Distance Vector) is probably the more treated in literature.

In this article, we propose a new variant based on the AODV which gives better results than the original AODV protocol with respect of a set of QoS parameters and under different constraints, taking into account the limited resources of mobile environments (bandwidth, energy, etc...).

The proposed variant (M-AODV) suggests that the discovering operation for paths reconstruction should be done from the source. It also defines a new mechanism for determining multiple disjoint (separated) routes.

To validate our solution, simulations were made under Network Simulator (NS2). We measure traffic control and packet loss rate under diverse constraints (mobility, energy and scale).

Keywords: Adhoc Networks (MANETs), protocol AODV, QoS, routing, multiple paths, Network Simulator (NS2).

1. Introduction

The new multimedia applications (videoconferencing, video telephony, web games, etc.) and real-time require respectively high throughput and reduced delays that are among fundamental quality of service (QoS) parameters.

Providing quality of service for networks, particularly for Adhoc networks (MANETs), have and still being the subject of intensive research in order to propose better solutions for these new requirements, not only in particular

level but at different levels of network architecture i.e. by various network layers (physical, network, etc.) [1] [2].

The routing function represents a main function for a network in general and for Adhoc networks (i.e. wireless networks without infrastructure) more particularly. Routing protocols in these networks have been a subject of numerous researches; several approaches have been discussed, and many protocols [3] [4] have been proposed. The most cited (quoted) in literature is probably the AODV protocol [5]. Ensuring routing QoS consists in determining one or several (paths) that satisfy best QoS constraints such as packet loss, throughput, jitter, etc.

In this paper a new variant M-AODV (M for Modified) that discovers in a first step, all possible paths between sources and destinations and maintain them during all data transfer phase. In case of a failure of the actual route, the data transfer will use one of the previously established routes (secondary routes). The failure state is declared only if all paths, found in discovery phase, cannot be used.

In this study, we focus on QoS metrics such as load control (overhead), reliability (packet loss), the packets delay transit etc... under various constraints like mobility, energy and scaling from which suffers the majority of routing algorithms in MANETs.

The remainder of this paper is organized as follows: in section 2, we give a brief review of QoS. In section 3, we discuss the most important characteristics of the AODV protocol, in section 4, we present our new protocol variant and the proposed changes. We evaluate later in section 5, the performance of this new AODV variant by simulation using NS-2 (Network Simulator), considering several contexts. We finish with conclusion and future recommendations of our researches.

2. Quality of Service (QoS) in the MANET

2.1. Qos model

Quality of Service (QoS) refers to a set of mechanisms able to share fairly various resources offered by the network to each application as needed, to provide, if possible, to every application the desired quality (the network's ability to provide a service) [6].

The QoS is characterized by a certain number of parameters (throughput, latency, jitter and loss, etc.) and it can be defined as the degree of user satisfaction.

QoS model defines architecture that will provide the possible best service. This model must take into consideration all challenges imposed by Ad-hoc networks, like network topology change due to the mobility of its nodes, constraints of reliability and energy consumption, so it describes a set of services that allow users to select a number of safeguards (guarantees) that govern such properties as time, reliability, etc.. [7][8].

Classical models like Intserv / RSVP [9] and DiffServ [10] proposed in first wired network types are not suitable (adapted) for MANETs. Various solutions or models [11] [12] namely: 2LqoS (Two-Layered Quality), CEDAR, noise, FQMM (Flexible QoS Model for MANET), SWAN (Service Differentiation in Wireless Ad-hoc Networks) and INSIGNIA have been proposed for the Ad-hoc networks. Each of these models attempts (tries) to improve one or several QoS parameters, as they may be part of one or more network layers architecture.

2.2. QoS Routing

New requirements (needs) for multimedia and real-time applications require few delay and very high data rates which require (oblige) the use of new routing protocols supporting QoS [13] [14].

The QoS support must take in consideration a number of Ad-hoc networks constraints (mobility, energy, scale, etc.). QoS can be introduced into different layers network if there is need (channel access functions at MAC layer, routing protocols at network layer, etc.).[15].

Routing operation consists to find routes between communicating entities (transmitter / receiver) able to convey data packets continuously using less bandwidth and fewer packets control. Routing in MANETs must also manage constraints of nodes energy problems, topology frequent changes due to nodes mobility and communication channel nature (air). QoS routing can be defined as the research for routes satisfying the wanted (desired) QoS. To be as eligible routes, they must satisfy a number of constraints (such that delay, bandwidth, reliability, etc.) [16]. Indeed, any path that satisfies a

number of quantitative or qualitative criteria can be described as path providing (ensuring) certain QoS.

3. Original AODV

3.1 Introduction

The AODV protocol (Ad-hoc on demand Distance Vector) [7] is a reactive routing protocol based on the distance vector Principle, combining unicast and multicast routing. In AODV, the path between two nodes is calculated when needed (if necessary), i.e. when a source node wants to send data packets to a destination, it finds a path (Discovery Phase), uses it during the transfer phase, and it must maintain this path during its utilisation (Maintenance Phase).

The finding and maintaining process of a path is based on the exchange of a set of control packets: RREQ (Route REQuest), RREP (Route Reply), RERR (Route Error), RRepAck (Route Reply Acknowledgment) and Hello messages (Hello). RREQ is initiated by the source node to find a path in multicast mode. RREP is used by an intermediate or destination node to respond to a request of path finding in unicast mode. Hello messages are used to maintain the consistency of a previously established path. Routing table is associated for each node in AODV protocol with containing: the destination address, the list of active neighbors, the number of hops (hop) to reach the destination, time of expiration after which the entry is invalidated, and so on.

To avoid the formation of infinite loop, AODV uses the principle of sequence numbers, limiting the unnecessary transmission of control packets (problem of the overhead); these numbers allow the use of fresh routes following the mobility of nodes, as they ensure the coherence and consistency of routing information [5].

It should be noted when the path breaks due to the absence of one node either by removal or a problem of energy, a local repair procedure (local repair) is called, it takes over the reconstruction of the path from this point. If this procedure cannot solve the problem, the source node try to find a new path and the number of attempts (RREQ_RETRIES) is decremented by 1, until the success or failure of the communication link.

This procedure generates a considerable amount of control packets. It should be noted that the original AODV maintains only one path to destination. To address this problem, it is preferable to have an alternative path already prepared. Two solutions are possible [8]: AODV with relief paths or multi-paths and there are two variants:

Several paths from source noted (M-AODV)

Several paths by intermediate node noted (M-AODV-I)

Paths from source or intermediate node are either completely disjoint (totally separated) or with common links.

Completely disjoined paths: a break at a route does not affect the rest of routes and therefore the use of another route is always possible to transmit data.

Paths with common links: sometimes, connection between two unspecified nodes belongs to several routes and when it break, all routes passing through this section are not in use and we have very small number of routes available adding to this we are obliged to generate very important additional traffic to notify source of this disconnection.

4. Proposed variant

4.1. Motivations

Among the key points having motivated our proposition (Modification of AODV protocol) we can cite: improving mechanisms that generate data packets loss (broken link or queues overflow associated at each nodes), the rational use of bandwidth (flow) and reducing packets latency.

Two cases are causing packet loss. The first one is due to frequent topology change by migration or remoteness (mobility) node formerly is part of link and its downstream and upstream neighbors respectively continues to send acknowledgments and data packets for a certain period before realize that link is failing (broken). The second is when a node starts the local repair procedure after detecting a broken link, the source was not aware of this situation, therefore continues to send its data packets normally, causing an overflow queue associated with the node without these data will be transmitted to their destination.

Loss can therefore be improved by changing the discovery and maintenance mechanisms of routes providing an almost continuous availability of links between communicating pairs (multi paths).

Rationalize the bandwidth use back to allow more useful data transfer and less control packets such as path discover (RREQ, RREP, etc.) path maintenance (Hello) which significantly reduces overload problem afflicting almost all wireless networks.

4.2. M-AODV protocol "M stands for Modified"

In this solution, we adds information to control packets for all routes and after exploring all possible paths, one with the shortest path hop count is first selected that respect QoS criterion required by user. In this solution control packet RREQ and RREP are routed in broadcast way. When the source wishes to transmit, it checks its routing table for any valid route to desired destination. If this is not

the case, it starts Discovery Phase (discovery route process) by broadcasting control packet RREQ (Fig. 1).



Fig. 1: Path discovery in M-AODV

As soon as an unspecified node has any path to destination, it responds to source with RREP packet and if isn't the case, destination respond to source by broadcasting a control packet RREP (Fig. 2) to trace (recall) all possible routes.

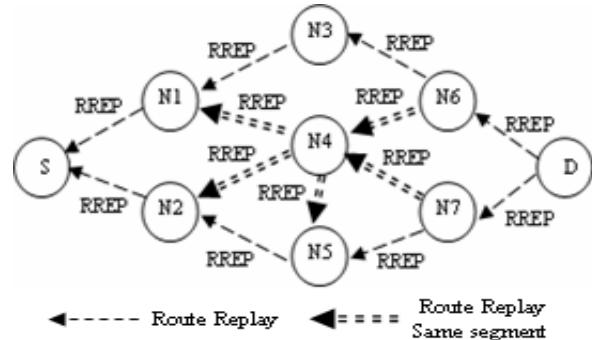


Fig. 2: Reverse Path

Different possible paths are: (S, N1, N3, N6, D) (S, N1, N4, N6, D) (S, N1, N4, N5, N7, D) (S, N1, N4, N7, D) (S, N2, N5, N7, D) (S, N2, N4, N6, D) (S, N2, N5, N4, N6, D) and (S, N2, N5, N4, N7, D). Once the node N4 is carried out, the different routes which pass (there) will not be valid.

Different Completely disjoined routes are selected from paths passed by low degree nodes: (S, N1, N3, N6, D) and (S, N2, N5, N4, N7, D), one will be taken as Primary path and other as secondary (minor) routes (Fig. 3).

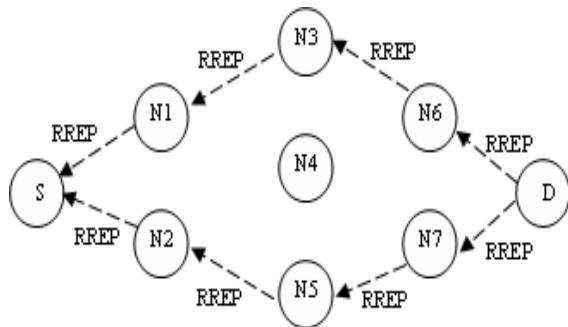


Fig. 3: Disjoint paths in M-AODV

Management process of routes is source responsibility but intermediate nodes are responsible only on their routing tables.

In link failure case, source stops transmission and repeats (reiterates) operation after selecting a new route from the spare paths available to it (minor routes).

In adopted (constraint) solution, the number of completely disjoined paths is fixed first to a number less than or equal to "n0" with a threshold of "s0". Once the threshold is reached in parallel with current data transfer, discovery phase for new route is initiated to determine other routes until reaching (to achieve) "n0". Although theoretically this solution generates a sizeable overhead but has the advantage of route availability at any time.

In original AODV when path is broken, local repair phase is initiated and if failure is declared a new discovery phase is initiated by source node.

In this modification, we propose to eliminate local repair phase to minimize modified protocol (M-AODV) task and the discovery phase is delegated in all scenarios to source node for a number of attempts RREQ_RETRIES (without M-AODV local repair).

Maintenance of all routes follows the same principle as original AODV by using "Hellos" packets (see Fig. 4).

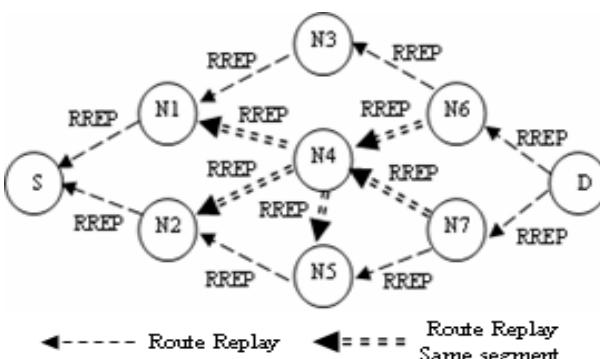


Fig. 4: Data transfer and maintenance phase

The path (S, N2, N5, N7, D) is considered as main street and (S, N1, N3, N6) as minor route or rescue.

5. Simulation

Measures taken in this simulation studies must be made on a set of parameters (metrics) like packet loss, load control, etc. and this must be observed under certain conditions (constraints) such as mobility, density, scale etc...

5.1. Constraints

5.1.1. Mobility

Once route is built (established), it will be used and maintained for fixed period or until path is broken. In an Ad-hoc network, nodes are mobile, they will not be in neighbor's scope and therefore the route or routes which they are member become invalid. In this case, we try to revive (start again) discovery phase which generates an additional volume control packets [17].

The objective here is to test original AODV routing protocol behavior and modified variant M-AODV under constraint of mobility to know if amended (proposed) version minimizes control packets volume generated when establishing routes, transit delay of data packets and at the same time data packets loss or control packets transmitted.

5.1.2. Energy

In general energy consumption is proportional to the number of packets processed and the type of treatment performed (carried) (transmission / reception), it is noted that packet emission requires more energy than reception [18]. Here the objective is to determine (know) which of the two protocols (AODV and M-AODV) manages in best manner nodes energy.

5.1.3 Density

It means the number of nodes used or involved in an Adhoc networks. This constraint is examined to determine the impact of the nodes number on the network overload, how it is easy to determine routes if nodes number is so important, etc.

5.1.4 Scale Transition (network topology)

It means space or scale and not the sense in usual networks topology (star, bus, ring ... etc....). Currently most routing protocols in MANETs suffer from the problem of scalability (from few meters to tens meters). The objective here is whether this constraint is well respected and which of the two protocols tested keep its performances face to scale transition.

5.2. Metrics

5.2.1. Packet loss

Indeed, packet loss is crucial factor [16] to evaluate routing protocol performances. Ensure zero loss transfer is desired (coveted), but is that possible in Ad-hoc networks? A protocol is efficient (powerful), if it can minimize to maximum the packet loss in all conditions to which it is confronted (large or small nodes number (density), a small or large scale, high or low mobility ... etc.).

5.2.2. The load control

As for packet loss parameter, load control is a crucial element [11] for protocol performance evaluation. Over control packet volume is large, performance degrades and more bandwidth is used by control packets than data. The measure of such parameter can justify the choice of using such or such protocol.

5.2.3. Rate (bandwidth)

Is the maximum information quantity per unit time? It's actually the maximum transfer ratio can be maintained between transmitter and receiver, this factor depends on physical links and also on flows sharing them. A better bandwidth [16] [19] management allows to pass high rate is for such reasons it is essential to measure our proposed protocol performance on this parameter.

To study and analyze our proposal behavior, we used the Network Simulator (Ns2) [20] version 2.31 installed on Debian GNU / Linux. Simulation context consists of 20 nodes in a region 800x600m². Transmission range is 250m for a perfect space (unobstructed) with Random Way Point (RWP) mobility model. Nodes moves at a maximum speed of 5m / s (average speed). Traffic was automatically generated randomly using Ns2 script cbrgen.tcl for Constant Bit Rate (CBR) of 512 bytes according to UDP protocol. Different mobility scenarios are also produced with Ns2 Setdest program. Time Simulation is set to 120 s for all test and each node has 10 joules as initial value.

5.3. Curves & discussions

The desired parameters to be evaluated by simulation under different contexts (mobility, density, etc.) are: throughput in kbps that indicates data transfer rate. Having a network system with high flow is coveted. Average end-to-end delay (e2e) reflects time taken between data packet transmission and reception. More time is short over the network is requested. Packet Delivery Ratio (PDR) describes the ratio between successfully delivered packets and the total number of transmitted packets. More the value of the PDR is small more the network is effective. Normalized Routing Load or Normalized Overhead Load

(NRL or NOL) is just the ratio of transmitted packets control on received packets number. This value expresses overhead network. Packets lost ratio is the rapport among successfully received and sent packets. This ration proves network reliability. The last parameter is the consumed energy by each node in whole network or only in routing phase to see which algorithms manage better this resource. In the following stage, we will study the mobility impact (respectively: nodes number) on the above parameters such as a high mobility (pause time equal to zero) to no mobility state (for pause time equal to 200s and more). (Respectively: by varying network size from small network with 10 nodes to denser network of 100 nodes).

Figures (Fig. 5 & 6) show that the presence of backup paths (multiple paths) in M-AODV version improves a better throughput especially for a high and medium mobility and network size less than 80 nodes; while delay in basic version (AODV) takes over for high mobility (pause time less or equal to 80 seconds) but beyond this value M-AODV takes the hand (Fig. 7). The same remark is done for network size ranging from 40 to 80 nodes (Fig. 8). Packet loss ratio is smaller in M-AODV for low and medium mobility due to available routes number and network size up to 75 nodes (Fig.9 & Fig.10). The PDR for our proposal method shows acceptable values for medium network size above 70 nodes (Fig. 12) and for low and medium mobility (Fig. 10). NRL ratio is more or less balanced i.e. sometimes is high due to generation of more packets control in both variants (more paths in M-AODV and more discovery operations in AODV) (Fig.13 & Fig.14). For consumed energy over time by all nodes is almost identical in either AODV or M-AODV (Fig. 15) but M-AODV minimizes this resource better than in routing phase (Fig .16)

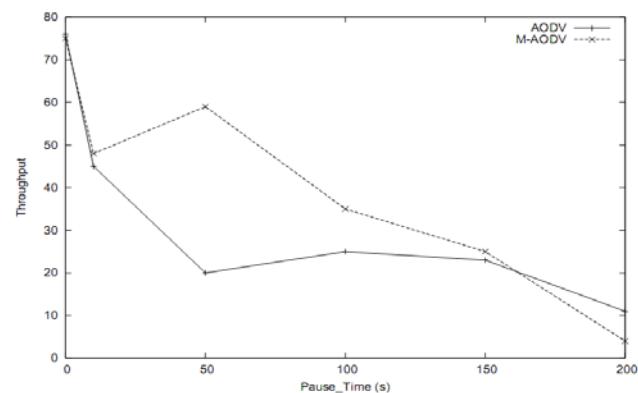


Fig. 5 Throughput Vs Pause_Time

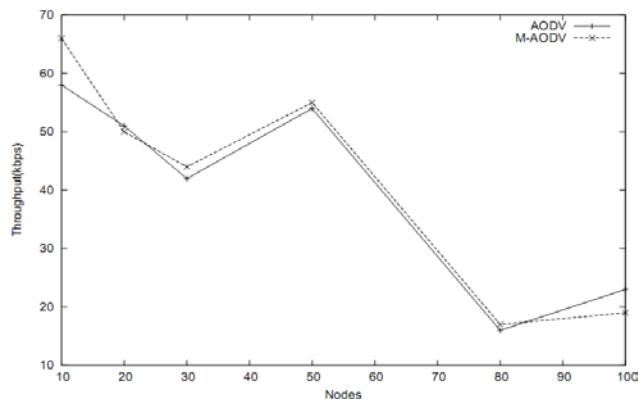


Fig. 6 Throughput Vs Nodes

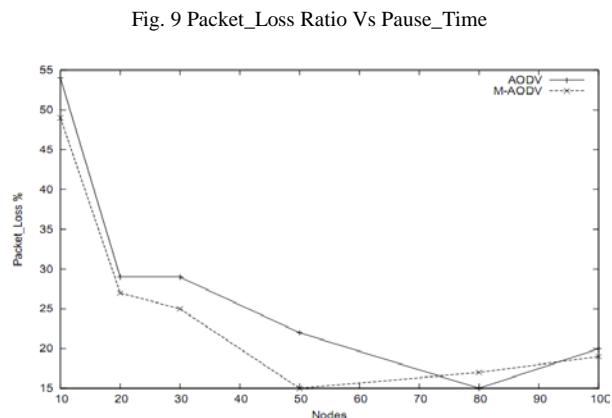


Fig. 9 Packet_Loss_Ratio Vs Pause_Time

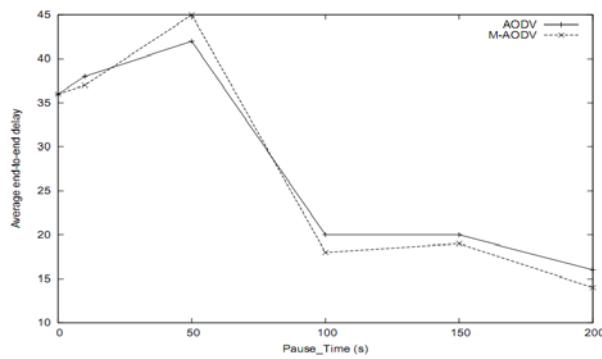


Fig. 7 Average_end_to_end_delay Vs Pause_Time

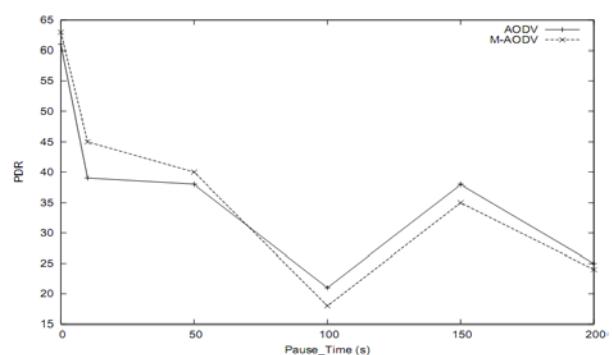


Fig. 10 Packet_Loss_Ratio Vs Nodes

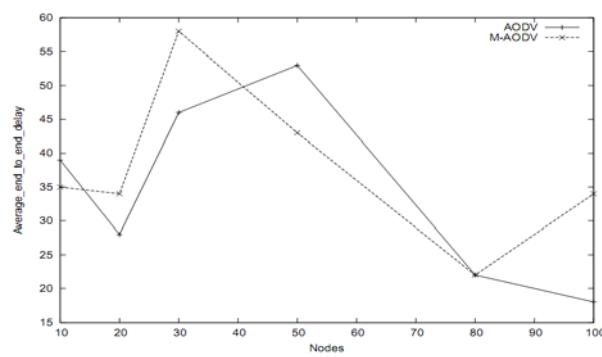


Fig. 8 Average_end_to_end_delay Vs Nodes

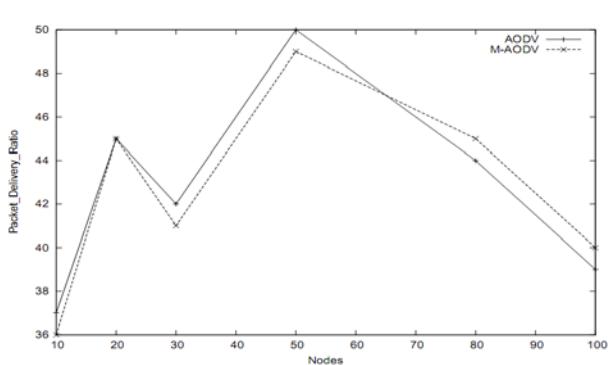


Fig. 11 Packet Delivery Ratio Vs Pause_Time

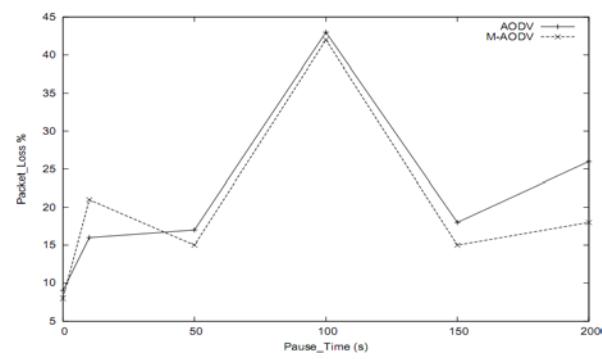


Fig. 12 Packet Delivery Ratio Vs Nodes

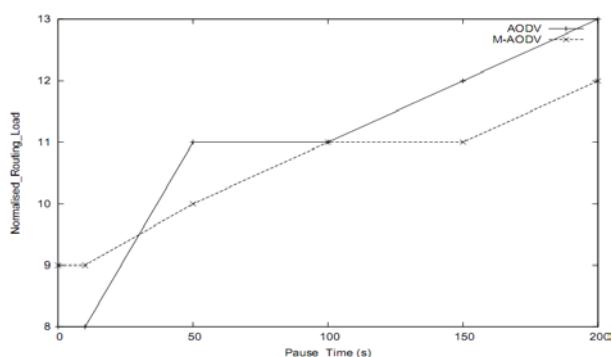


Fig. 13 Normalised_Routing_Load Vs Pause_Time

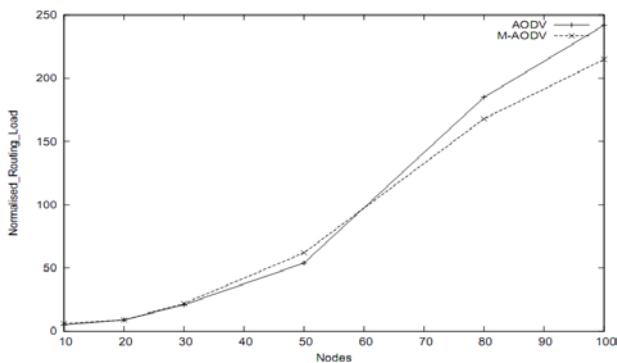


Fig. 14 Normalised_Routing_Load Vs Nodes

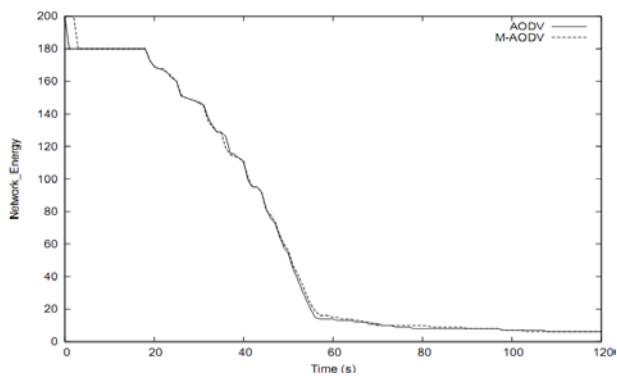


Fig. 15 Network_Energy Vs Time

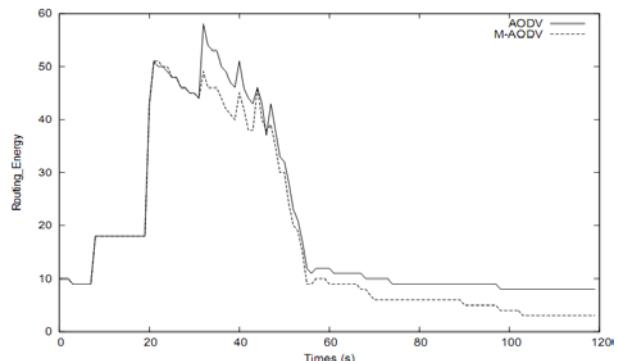


Fig. 16 Routing_Energy Vs Time

6. Conclusion & Perspectives

The proposed M-AODV generates in a first time, a number of possible paths (multiple paths) between sources and destinations with the aim to use them if needed (to minimize the repair phase overhead) i.e. when the link in use is broken; by running discovery phase in parallel with data packets transfer phase ensure to have a number of

adjacent paths in advance. The performed simulation and comparison between the original AODV protocol and M-AODV show that this last can improve the QoS in Mobile ad hoc networks under different conditions.

Future extensions of AODV protocol should be based on the prediction of a future link disconnection considering the signal quality and mobile node speed. We project also to extend the functionality of the protocol with the aim of adapting it to large scale networks.

References

- [1] S. Chakrabarti, A. Mishra. "Quality of service challenges for wireless mobile Adhoc networks". Wireless Communi and Mobile Comput., vol. 4, no. 2, pp. 129-153, March. 2004.
- [2] J.Novatnack, L.Greenwald, H.Arora. "Evaluating Adhoc Routing Protocols with Respect to Quality of Service". Technical Report DU-CS-04-05, Department of Computer Science, Drexel Univ, Philadelphia, PA 19104, Oct. 2004.
- [3] M.Lai. "Wireless Networking Final Report- Mobile Adhoc Networks Routing Protocols Simulation and Comparisón EECS Department, Univ.of California, Irvine, March. 2005.
- [4] C-K. Toh, E. Royer. "A review of current routing protocols for ad-hoc mobile wireless networks IEEE Personal Communications Magazine, pages 46–55, April. 1999
- [5] C. Perkins, E. Royer, S. Das. "Adhoc on-demand distance vector (AODV) routing". RFC 3561, IETF, July. 2003
- [6] J. Kay., J. Frolik, "Quality of Service Analysis and Control for Wireless Sensor Networks". Submitted to the 1st IEEE International Conference on Mobile Adhoc and Sensor Systems (MASS2004), Ft. Lauderdale, FL, Oct.25-27, 2004.
- [7] C. Boulkamh, A. Bilami, A. Saib, M. Sedrati, "AODV_MC: Un Protocole de Routage AODV avec Minimum de Contrôle Routage Jeesi09, Alger 19Mai. 2009
- [8] R. Ramanathan, R. Hain. "An Adhoc wireless testbed for scalable, adaptive QoS support In Proc eedings of IEEE WCNC'2000, Chicago, IL, USA, 2000.
- [9] R. Braden, L. Zhang, et al. "Integrated Services in the Internet Architecture: an Overview". Juin.1994. RFC 1633.
- [10] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss. "An Architecture for Differentiated Services". Déc.1998. RFC 2475.
- [11] C. Chaudet. "Qualité de service et réseaux Adhoc : un état de l'art Rapport de recherche INRIA, RR 4325, Nouv.2001.
- [12] K.WU, J.HARMS. "QoS Support in Mobile Adhoc Networks". Computing Science Department, University of Alberta, Crossing Boundaries—an interdisciplinary journal Vol 1, N°1, 2001.
- [13] M. Curado and E. Monteiro, "An Overview of Quality of Service Routing Issues in Proceedings of the 5 th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2001), June.2001.
- [14] A. Orda, A. Sprintson, "Precomputation Schemes for QoS Routing" IEEE/ACM Transaction on Networking, vol. 11, pp. 578 à 591, 2003.
- [15] C. Chaudet, I. Guérin Lassous. "Routage qos et réseaux Adhoc : de l'état de lien à l'état de nœud Technical Report 4700, INRIA, jan.2003.

- [16] H. Badis, A. Munaretto, K. Al Agha, G. Pujolle. "QoS for Adhoc Networking Based on Multiple Metrics: Bandwidth and Delay". In the proceedings of IEEE MWCN2003, Singapore, October 2003.
- [17] M. Cheung, J. Mark, "Effect of mobility on QoS provisioning in wireless communications networks". Wireless Communications and Networking Conference, Vol.1, pp.306–310, 1999.
- [18] H. Ozgur Sanli, H. Cam, X. Cheng, "EQoS: An Energy Efficient QoS Protocol for Wireless Sensor Networks". Proceedings of 2004 Western Simulation MultiConference, San Diego, CA, Jan.2004.
- [19] I. Kassabalis, A.K. Das, M.A. El-Sharkawi, R.J. Marks II, P. Arabshahi, A. Gray. "Intelligent routing and bandwidth allocation in wireless networks". Proc. NASA Earth Science Technology Conf. College Park, MD, August 28-30, 2001.
- [20] K. Fall, K. Varadhan. "The NS Manuál Vint Project, UC Berkeley, LBL, DARPA, USC/ISI, and Xerox PARC. April 14, 2002.

Optimization of LSE and LMMSE Channel Estimation Algorithms based on CIR Samples and Channel Taps

Saqib Saleem
CSE Department,
Institute of Space Technology, Islamabad,
Punjab, Pakistan

Abstract-- For spectrally efficient transmission over time-varying channels, the use of Adaptive Coding and Modulation (AMC) in wireless OFDM systems requires the estimation of radio channel at the receiver. This paper focuses on the use of time domain channel statistics, mainly concentrating on two schemes: Linear Minimum Mean Square Estimation (LMMSE) and Least Square Estimation (LSE) and their variants. LMMSE performs better than LSE but at the cost of computational complexity. The performance of LSE can be improved by increasing CIR samples and channel taps. To avoid the matrix inversion lemma, the channel matrix can be downsampled or regularized. Theoretical analysis and computer simulations are used for performance and complexity comparisons.

I. INTRODUCTION

Orthogonal Frequency Division Multiplexing (OFDM), a multicarrier modulation method, is considered an essential technique for a variety of high data rate communication systems like 4G WiMAX and LTE-Advanced due to its efficient management of ISI in frequency selective fading channels. OFDM can also be used as a modulation technique because of the simple equalizer design and spectrum efficiency. The combination of OFDM with Multi-Input Multi-Output (MIMO) provides the increased data rate and improved quality of service. That is why MIMO-OFDM is adopted in B3G (Beyond 3rd Generation) mobile communication systems.

Coherent OFDM, which has 3-4 dB performance gain more than non-coherent OFDM, requires channel state information (CSI) at the receiver and/or transmitter. CSI only at the transmitter is usually preferred to make the receiver design simple. Data throughput of channel depends on the quality of the channel estimator. For channel estimation there are mainly two methods proposed as, first is decision directed channel estimation and other one is pilot-assisted channel estimation. In decision directed method, the modulation is removed from subcarriers using the previously demodulated symbol, thus all subcarriers can be used for channel estimation. This method requires a large amount of data and its convergence rate is also very slow, that is why it is not well suited for real time systems. In pilot assisted method there are two modes, if all subcarriers have known pilots then it is called block pilot mode while in comb pilot mode only a few subcarriers carry known pilots.

Channel can be estimated in time domain or frequency domain. In frequency domain two algorithms are proposed Least Square Estimation (LSE) and Linear Minimum Mean Square Estimation (LMMSE). LSE algorithm is relatively easy to implement due to its less complexity and it also does not require any channel apriority probability. To achieve better performance LMMSE is proposed. LMMSE is optimum in minimizing Mean Square Error (MSE) as it uses addition information of operating SNR and the channel statistics. But its complexity is higher due to the channel correlation and the matrix inversion lemma. There can be a

compromise of complexity and performance by taking the effect of the channel taps and channel impulse response (CIR) samples. By assuming the impulse response of finite length, these two algorithms can be modified having less complexity. In mobile wireless links the channel statistics are not known, in these cases it is robust to consider the uniform Power delay profile (PDP), which also reduces complexity than LMMSE. The complexity of LSE can be reduced by regularizing the Eigen values of the matrix being inverted or by down-sampling the channel vector.

The rest of the paper is organized as: Section 2 describes OFDM signal and channel model, in Section 3, LMMSE, LSE and their different variants are discussed, followed by the simulation results in Section 4 and in the last section conclusions are drawn.

II. OFDM SIGNAL AND CHANNEL MODEL

In OFDM, the transmitted bit stream is divided into many different sub-streams and send them over many orthogonal sub-channels. Suppose the transmitted data at k -th subcarrier is $d(k)$. Then the multicarrier modulated signal will be

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} d(k) e^{\frac{j2\pi k}{N}}, \quad n = 0, 1, 2, \dots, N - 1$$

Where N is total number of sub-carriers. Before transmitting $x(n)$, guard interval (GI) is inserted to avoid Inter-symbol interference (ISI) and inter-carrier interference (ICI). This signal is then passed through a time-varying multipath channel whose impulse response is characterized by

$$g(t, \tau) = \sum_{i=0}^{L-1} \alpha_i \delta(t - \tau_i)$$

where L is total number of multi-paths and $\{\alpha_i\}$ is a complex Gaussian random variable of zero mean having a power delay profile: $Ce^{\tau_i/\tau_{rms}}$. $\{\tau_i\}$ represents time delay between different multi-paths, whose maximum value is not supposed to exceed the guard interval length.

After passing this fading channel and removing GI, the received OFDM signal in frequency domain will be

$$Y = HX + W$$

W is the complex-valued additive Gaussian noise having zero mean and σ^2 variance. H is the channel frequency response, that is DFT of the channel impulse response $g(t, \tau)$.

III. CHANNEL ESTIMATION ALGORITHMS

A. LMMSE Channel Estimation

In presence of channel noise, LMMSE estimation of the uncorrelated Gaussian channel vector g is given by [1]

$$\hat{g} = \Gamma_{gy} \Gamma_{yy}^{-1} y$$

Where

$$\Gamma_{gy} = \Gamma_{gg} F^H X^H$$

$$\Gamma_{yy} = X F \Gamma_{gg} F^H X^H + \sigma_n^2 I_N$$

Γ_{yy} is the auto-covariance matrix of y and Γ_{gy} is the cross co-variance matrix between g and y . σ_n^2 is variance of noise. For unique minimum MSE, these co-variance matrices should be positive definite,

In frequency domain the channel estimate \hat{h}_{mmse} is given by

$$\hat{h}_{mmse} = F \hat{g} = F Q F^H X^H y$$

Where F is orthonormal DFT-matrix and Q is given by [1]

$$Q = \Gamma_{gg} [(F^H X^H X F)^{-1} \sigma_n^2 + \Gamma_{gg}]^{-1} (F^H X^H X F)^{-1}$$

B. Modified LMMSE Channel Estimation

For large N the calculation of Q matrix implies high complexity. To reduce the size of Q , we can take only first L taps having significant energy. Using this approximation Γ_{gg} is reduced to $L \times L$ matrix. So modified LMMSE estimation becomes [1]

$$\hat{h}_{mmse} = T Q' T^H X^H y$$

Where T have only first L columns of DFT matrix and Q' is

$$Q' = \Gamma'_{gg} [(T^H X^H X T)^{-1} \sigma_n^2 + \Gamma'_{gg}]^{-1} (T^H X^H X T)^{-1}$$

Γ'_{gg} denotes the upper left $L \times L$ matrix of Γ_{gg} .

C. Low Complex LMMSE Channel Estimation

In LMMSE channel estimation, a matrix inversion is needed as the input data X is changed which results in high complexity. This complexity can be reduced by averaging the transmitted data x i.e. $E(XX^H)^{-1}$. If we assume same signal constellation for all frequencies, then

$$E(XX^H)^{-1} = E \left| \frac{1}{x_k} \right|^2.$$

The simplified LMMSE estimation will be [2]

$$\hat{h}_{mmse} = \Gamma_{gg} (\Gamma_{gg} + \frac{\beta}{SNR} I)^{-1} X^{-1} y$$

Where β depends upon the signal constellation.

D. Robust LMMSE Channel Estimation

In mobile wireless links, the channel changes with time depending on the particular environment. It is not possible to know the channel PDP at the design time [3]. Identical MSE performance can be obtained for all PDPs with same maximum delay. So it is robust to design the channel co-variance matrix with a uniform PDP [4].

E. LSE Channel Estimation

A prior knowledge of second order channel statistics is required for LMMSE estimator, which is not possible in many practical situations. We can design an estimator filter which is a function of available data only [5]. In LSE estimation, we use only signal model, no probabilistic assumptions are required.

LSE estimation of channel is given by

$$\hat{h}_{ls} = F Q_{ls} F^H X^H y$$

where

$$Q_{ls} = (F^H X^H X F)^{-1}$$

\hat{h}_{ls} can also be written as [1]

$$\hat{h}_{ls} = X^{-1} y$$

F. Modified LSE Channel Estimation

Though no modification are needed because of less complexity of LSE estimator but performance can be improved by considering only first L high energy channel taps. The modified LSE estimator becomes

$$\hat{h}_{ls} = T Q'_{ls} T^H X^H y$$

where

$$Q'_{ls} = (T^H X^H X T)^{-1}$$

G. Regularized LSE Channel Estimation

The problem of inversion of $N \times N$ matrix can be solved by regularizing the Eigen values of the matrix by adding a constant term to the diagonal elements. In this case, the matrix Q_{ls} will be [6]

$$Q_{reg,ls} = (\alpha I + F^H X^H X F)^{-1}$$

Where off-line constant α is chosen such that the matrix $Q_{reg,ls}$ is least perturbed.

H. Down-Sampled Impulse Response LSE Channel Estimation

The inversion of $N \times N$ matrix can be simplified by decreasing the sampling frequency, but ensuring the absence of aliasing. Only 2 out of 3 channel taps are used and the discarded taps are set to zero.

The down-sampled version of channel vector g can be [6]

$$\bar{g} = (g_0 \ g_1 \ 0 \ g_3 \ g_4 \ 0 \ \dots \ g_{L-1})^T$$

The channel transfer function can be written as

$$H^{DS} = F \bar{g}$$

Which is equivalent to

$$H^{DS}$$

$$= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & w^1 & w^3 & \dots & w^{(L-1)} \\ 1 & w^2 & w^6 & \dots & w^{2(L-1)} \\ 1 & w^3 & w^9 & \dots & w^{3(L-1)} \\ 1 & w^4 & w^{12} & \dots & w^{4(L-1)} \\ 1 & w^5 & w^{15} & \dots & w^{5(L-1)} \\ 1 & \dots & \dots & \dots & \dots \\ 1 & w^{N-1} & w^{3(N-1)} & \dots & w^{(N-1)(L-1)} \end{bmatrix} \begin{bmatrix} g_0 \\ g_1 \\ g_3 \\ g_4 \\ \vdots \\ g_{L-1} \end{bmatrix}$$

The estimated channel in this case will be

$$\hat{h}_{DS} = (F^{DS,H} X^H X F^{DS})^{-1} F^{DS,H} X^H y$$

IV. SIMULATION RESULTS

To demonstrate the effectiveness of the discussed algorithms, Matlab Simulations are provided in this section. All simulations have been performed for OFDM signal in Rayleigh Fading Channel with BPSK modulation scheme and FFT size is kept 64. To illustrate the performance of the estimators, the widely used Mean Square Error (MSE) has been used as a function of SNR, Channel Taps and Channel Impulse Response (CIR) samples. The complexity of the estimators is compared in terms of computational time.

a. Comparison of LMMSE Channel Estimators

The performance of LMMSE with its variants i.e. Modified LMMSE with 10 taps, 40 taps, Robust LMMSE and Low Complex LMMSE is shown in Fig.1. The difference between LMMSE and Modified LMMSE estimators is due to the fact that some parts of the channel statistics are not taken into account in the former estimators. For low SNR values, the performance of LMMSE is better than R.LMMSE but for higher SNRs R.LMMSE outperforms LMMSE. The performance of both LMMSE and Low Complex LMMSE is same and the difference lies in the complexity as the computational time of Low Complex LMMSE is less than that of LMMSE. The comparison of computational time of LMMSE estimators is given in Table 1. Table 1 indicates that there is a wide gap of time between LMMSE while using covariance matrix and correlation matrix.

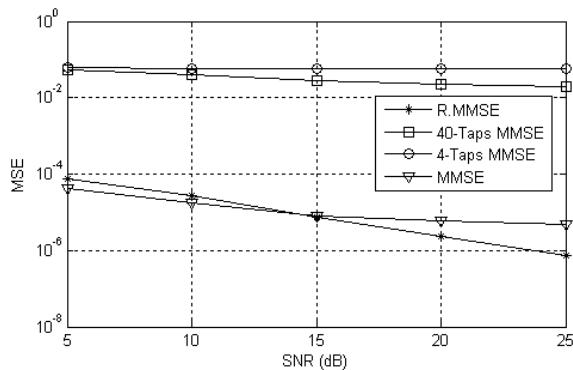


Figure 1. MSE v/s SNR for LMMSE Estimators

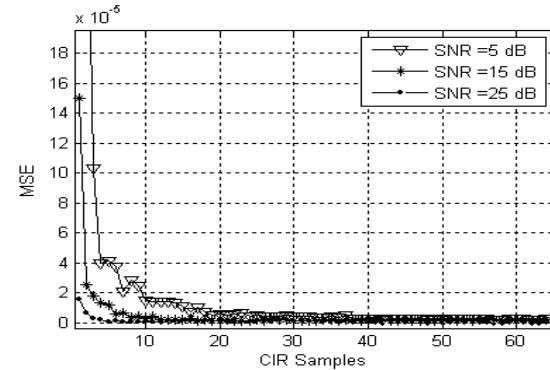


Figure 3. MSE v/s CIR Samples for LMMSE Estimator

TABLE 1 COMPUTATIONAL TIME FOR LMMSE ESTIMATORS

| Estimator | 5000 Simulation (sec) | 1 OFDM (mSec) | 1 Bit (mSec) |
|-------------------|-----------------------|---------------|--------------|
| LMMSE Modified-10 | 208.278 | 41.656 | 0.651 |
| Low Complex LMMSE | 320.713 | 64.143 | 1.003 |
| LMMSE (Corr Mtx) | 346.8 | 69.36 | 1.084 |
| LMMSE Modified-40 | 440.945 | 88.189 | 1.378 |
| R.LMMSE | 528.133 | 105.627 | 1.651 |
| LMMSE (Cov Mtx) | 529.319 | 105.864 | 1.65 |

TABLE 2 TIME V/S CIR SAMPLES FOR LMMSE ESTIMATOR

| CIR Samples | Time (mSec) |
|-------------|-------------|
| 30 | 1 |
| 40 | 1.25 |
| 50 | 1.5 |
| 60 | 1.75 |

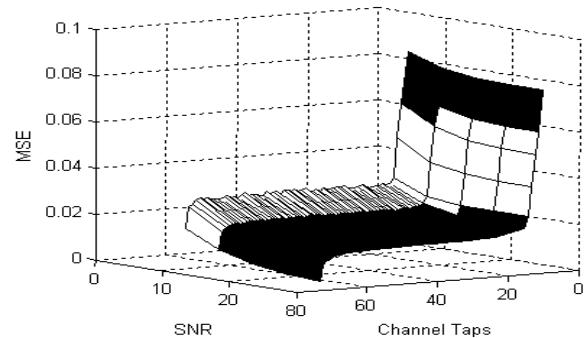


Figure 4. MSE v/s SNR v/s Channel Taps for Modified LMMSE Estimator

TABLE 3 TIME V/S CHANNEL TAPS FOR MODIFIED LMMSE ESTIMATOR

| Channel Taps | Time (mSec) |
|--------------|-------------|
| 30 | 5 |
| 40 | 6 |
| 50 | 10 |
| 60 | 12 |

The performance of LMMSE estimator in terms of CIR samples for different values of SNR is shown in Fig.3. As we notice that after a certain number of CIR samples we have the same MSE for all values of SNR. The effect of increasing CIR samples on time is shown in Table 2.

The effect of channel taps and SNR on MSE is shown in Fig.4. By increasing channel taps up to 10, there is a significant improvement in MSE but from 10 to 60, the MSE behavior remains same and after 60 we get further improvement. Since there is no improvement in MSE by increasing channel taps from 10 up to 60 as the disadvantage only comes in form of more time of computation as shown in Table 3.

b. Comparison of LSE Channel Estimators

Fig.5 shows the MSE verus SNR for LSE, Modified LS, Regularized LS and Downsampled LS estimators. Contrary to the modification of LMMSE estimator, the modification of LS estimator reduces MSE for a range of SNRs. However the same approximation effect, as in the modified LMMSE estimators, shows up at high SNRs. For every SNR, there exists an estimator which gives the smallest MSE. The effect of regularized LS is same to LSE but at higher SNR the performance of regularized LS degrades. Downsampled LS is exactly same to that of LSE, advantage of former is only less complexity. The effect of CIR samples on MSE of LS estimator is shown in Fig.6. For CIR samples 0 to 10, there is a rapid improvement in performance specially at low SNRs, but by increasing samples further there is no further improvement in terms of MSE but the cost comes in more computational complexity that is shown in Table 4. It is clear from Table 4 that by increasing number of samples, there is a gradual increment in computational time, that is a drawback of increasing samples without improving performance. The effect of CIR samples and SNR on MSE is shown in Fig.7. The combined effect of

channel taps and SNR on MSE is shown in Fig.8. For specific channel taps, the effect of CIR samples on MSE is demonstrated in Fig.9. By increasing samples from 1 to 2, there is a dominant improvement in MSE but beyond this value of samples the performance saturates. The effect of channel taps for certain values of CIR samples on MSE is shown in Fig.10.

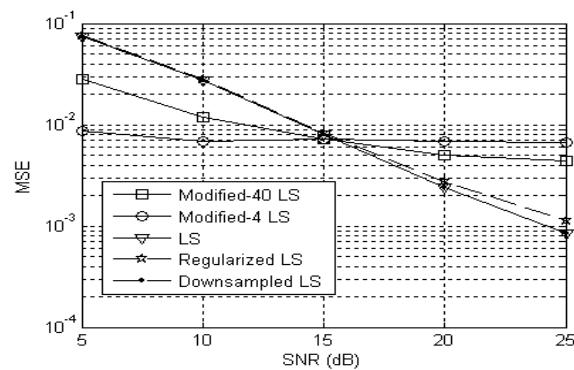


Figure 5. MSE v/s SNR for LS Estimators

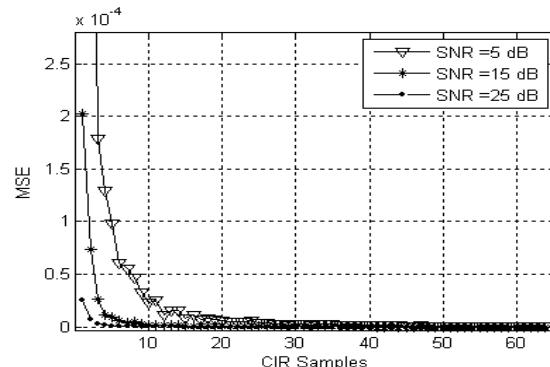


Figure 6. MSE v/s CIR Samples for LS Estimator

TABLE 4 TIME V/S CIR SAMPLES FOR LS ESTIMATOR

| CIR Samples | Time (mSec) |
|-------------|-------------|
| 30 | 0.5 |
| 40 | 1 |
| 50 | 1.25 |
| 60 | 1.5 |

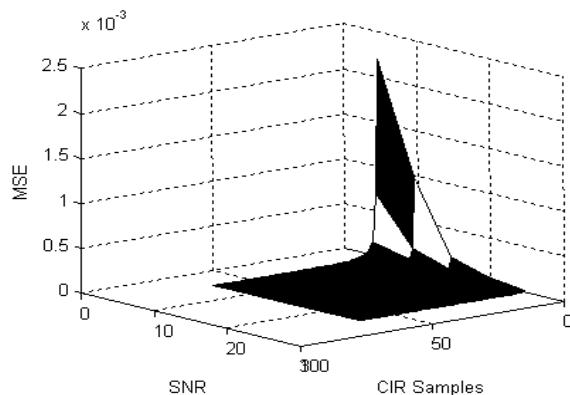


Figure 7. MSE v/s SNR v/s CIR Samples for LS Estimator

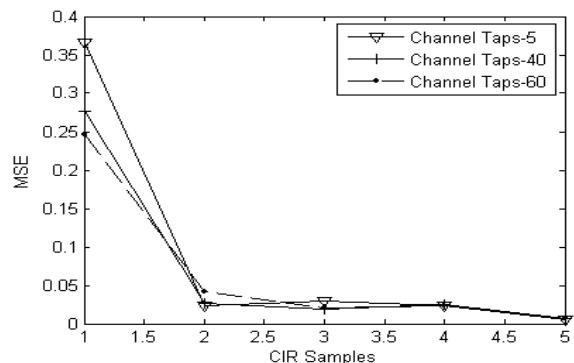


Figure 9. MSE v/s CIR Samples for Modified LS Estimator

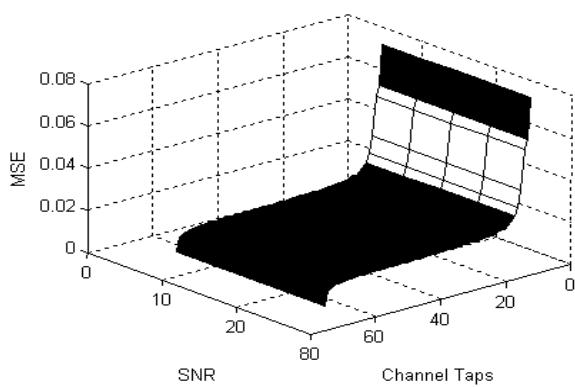


Figure 8. MSE v/s SNR v/s Channel Taps for Modified LS Estimator

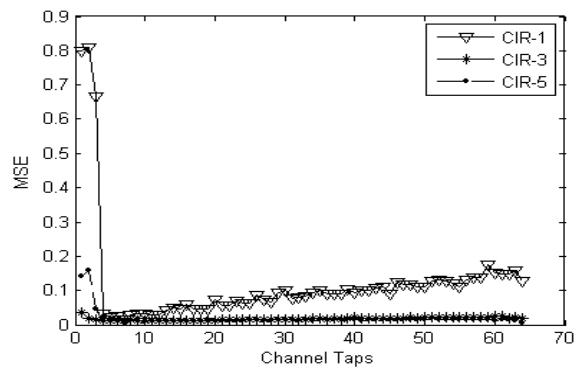


Figure 10. MSE v/s Channel Taps for Modified LS Estimator

The different downsampling rate versus corresponding MSE is shown in Fig.11. By increasing the downsampling rate, the performance is degraded while there is no significant effect on complexity.

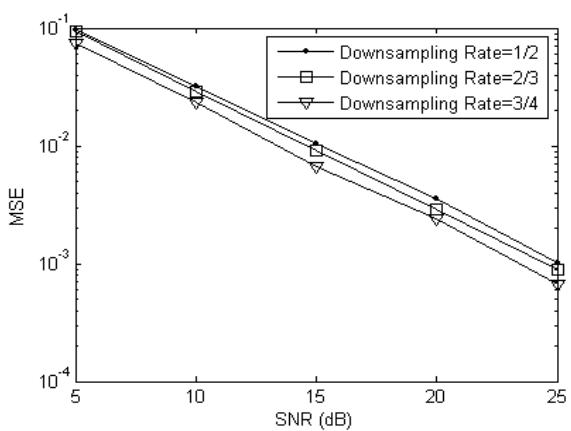


Figure 11. MSE v/s SNR for Downsampled LS Estimators

c- Comparison of LSE and LMMSE Channel Estimators

The performance comparison between LSE and LMMSE estimator is shown in Fig.12. When the channel has less number of CIR samples, then LMMSE is better to use than LSE due to less MSE, not in terms of time. But as CIR samples increases, for lower SNR values LMMSE is better in terms of MSE than LSE but for higher SNR values later one is better to use. But if we increase CIR samples further, then after certain number of CIR samples, LSE outperforms LMMSE for whole range of SNR values. The computation of both LSE and LMMSE with the increasing number of CIR samples is shown in Table 5. It is evident from Table 5 that LSE takes always less time than LMMSE, as it does not account for the channel statistics.

TABLE 5 TIME V/S CIR SAMPLES FOR LMMSE AND LS ESTIMATOR

| CIR Samples | Time (mSec) | |
|-------------|-------------|-------|
| | LS | LMMSE |
| 30 | 0.5 | 1 |
| 40 | 1 | 1.25 |
| 50 | 1.25 | 1.5 |
| 60 | 1.5 | 1.75 |

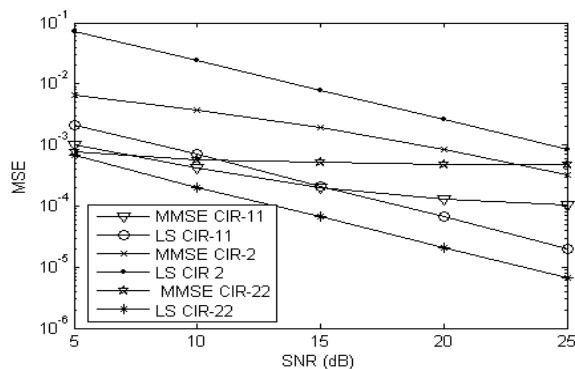


Figure 12. MSE v/s SNR for LMMSE and LS Estimators with different CIR Samples

V. CONCLUSIONS

In this paper we present LMMSE and LSE channel estimators based on CIR samples and channel taps and evaluated their comparison in terms of performance and complexity. The performance of LMMSE is better than LSE as it assumes the channel statistics which results in high complexity. The performance can be improved by increasing either CIR samples or channel taps but after a certain limit there is no prominent impact on performance while the complexity goes on increasing. As we go on increasing CIR samples, after a certain value LSE degrades LMMSE both in performance and complexity. We also noticed that the channel taps have no effect on the performance of LSE estimator for different SNR values. So if we use a channel filter of more length then we can improve the channel estimator performance even without having a prior channel information.

REFERENCES

- [1] J.J. van der Beek, O. Edfors, M. Sandell, S.K. Wilson, and P. O.Borgesson, "On channel estimation in OFDM systems," *Proc. VTC'95*, pp. 815-819.
- [2] Edfors, M. Sandell, J. J. van der Beek, S. K. Wilson, and P. O.Borgesson, "OFDM channel estimation by singular value decomposition," *IEEE Trans. Comm.*, vol. 46, no.7, pp. 931-939, July 1998.
- [3] Srivastava, C. K. Ho, P. H. W. Fung, and S. Sun, "Robust mmse channel estimation in ofdm systems with practical timing synchronization," in *Wireless Communications and Networking Conference, 2004. WCNC.2004 IEEE*, vol. 2, pp.711–716 Vol.2, 2004.
- [4] Y. Li, L. J. Cimini, Jr., and N. R. Sollenberger, "Robust channel estimation for OFDM systems with rapid dispersive fading channels," *IEEE Trans. Comm.*, vol. 46, no. 7, pp. 902-915, July 1998.
- [5] Dimitris G. Manolakis, Vinay K. Ingle. *Statistical and Adaptive Signal Processing,Spectral Estimation, Signal Modeling, Adaptive Filtering and Array Processing*, Artech House, Boston London
- [6] Ancora. A, Bona. C, Slock, D.T.M," Down sampled impulse response LS channel estimation for LTE OFDMA", *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007.ICASSP 2007.Vol.3*, pp.293-296, 2007

Classification rules for Indian Rice diseases

A.Nithya¹ and Dr.V.Sundaram²

¹Asst Professor in Computer Applications, Nehru Arts and Science College,
Coimbatore, Tamil Nadu, India.

²Dr.V.Sundaram, Karpagam College of College,
Coimbatore,Tamil Nadu,India.

Abstract

Many techniques have been developed for learning rules and relationships automatically from diverse data sets, to simplify the often tedious and error-prone process of acquiring knowledge from empirical data. Decision tree is one of learning algorithm which posses certain advantages that make it suitable for discovering the classification rule for data mining applications. Normally Decision trees widely used learning method and do not require any prior knowledge of data distribution, works well on noisy data .It has been applied to classify Rice disease based

1.Introduction

Decision trees have become one of the most powerful and popular approaches in knowledge discovery and data mining, the science and technology of exploring large and complex bodies of data in order to discover useful patterns. The area is of great importance because it enables modeling and knowledge extraction from the abundance of data available. The construction of decision tree classifiers does not require any domain Knowledge or parameter setting, and therefore is appropriate for exploratory Knowledge discovery. The Decision tree can handle high dimensional agricultural data. Their representation of acquired knowledge. The learning and classification steps of decision trees induction are simple and fast. The transfer of experts from consultants and scientists to agriculturists, extends workers and farmers represent a bottleneck

on the symptoms. This paper intended to discover classification rules for the Indian rice diseases using the c4.5 decision trees algorithm. Expert systems have been used in agriculture since the early 1980s. Several systems have been developed in different countries including the USA, Europe, and Egypt for plant-disorder diagnosis, management and other production aspects. This paper explores what Classification rule can do in the agricultural domain.

Key words: *Decision Trees, Pruning, Datamining, Classification, Expert System*

for the development of agriculture on the national. The term *Knowledge Discovery in Databases* or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization. The unifying goal of the KDD process is to extract knowledge from data in the context of large databases. Many machine learning schemes can work with either symbolic or numeric data, or a combination of both, and attempt to discover relationships in the data that have not yet been hypothesized. Once a relationship has been discovered, further statistical analysis can be performed to confirm its significance. Sometimes, both

fields work independently towards the same goal, as in the case of ID3 (Quinlan, 1986), a machine learning scheme, and CART (Breiman et al, 1984), standing for “classification and regression trees,” a statistical scheme. These methods both induce decision trees using essentially the same technique. Machine learning researchers also incorporate statistics into learning schemes directly, as in the case of the Bayesian classification system AUTO CLASS (Cheeseman et al, 1988). **C4.5** performs top down induction of Decision trees from a set of examples which have already been given a classification (Quinlan, 1992). Typically, a training set will be specified by the user. The root of the tree specifies an attribute to be selected and tested first, and the subordinate nodes dictate tests on further attributes. The leaves are marked to show the classification of the object they represent. An information-theoretic heuristic is used to determine which attribute should be tested at each node, and the attribute that minimizes the entropy of the decision is chosen. C4.5 is a well-developed piece of software that derives from the earlier ID3 scheme (Quinlan, 1986), which itself evolved through several versions

2. The ID3 algorithm

According to [9], the ID3 algorithm is a decision tree building algorithm which determines classification of objects by testing values of their properties. It builds tree in top down fashion, starting from set of objects and specification of properties. At each node of tree, the properties tested and the result is used to partition data object set. The information theoretic heuristic is used to produce shallower trees by deciding an order in which to select attributes. The first stage in applying the information theoretic heuristic is to calculate the proportions of positive and negative training cases that are currently available at a node. In the case of the root node this is all the cases in the training set. A value

known as the information needed for the node is calculated using the following formula where p is the proportion of positive cases and q is the proportion of negative cases at the node:

$$-p \log_2 p - q \log_2 q$$

The basic algorithm of ID3

Examples S, each of which is described by number of attributes along with the class attribute C, the basic pseudo code for the ID3 algorithm is:

If (all examples in S belong to class C) then make leaf labeled C

Else select the “most informative” attribute A

Partition S according to A’s values (v₁... v_n)

Recursively construct sub-trees T₁, T₂... T_n for each subset of S.

ID3 uses a statistical property, called information gain measure, to select among the candidates attributes at each step while growing the tree. To define the concept of information gain measure, it uses a measure commonly used in information theory, called entropy. The entropy is calculated by

$$\text{Entropy}(S) = \sum_{i=1}^c -p_i \log_2 p_i$$

Where S is a set, consisting of s data samples, P_i is the portion of S belonging to the class i. Notice that the entropy is 0 when all members of S belong to the same class and the entropy is 1 when the collection contains an equal number of positive and negative examples. If the collection contains unequal numbers of positive and negative examples, the entropy is between 0 and 1. In all calculations involving entropy, the outcome of all calculations involving entropy, the outcome of (0 log₂ 0) is defined to be 0. With the Information gain measure, given entropy as a measure of the impurity in a collection of training examples, a measure of effectiveness of an attribute in classifying the training data can be defined. This measure is called information gain and is the expected reduction in entropy caused by partitioning the examples according to this attribute. More precisely, the information gain is *Gain*(S, A) of an attribute A, relative to a collection of examples S.

“blast,” “helminthosporiose,” “stem rot” and “foot rot”.

Splitting Criterion

i) Information gain:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} (|S_v| / |S|) \text{Entropy}(S_v)$$

ii) Gain Ratio:

$$\text{Gain Ratio}(S, A) \equiv \text{Gain}(S, A) / \text{Split Information}(S, A)$$

$$\text{Split Information}(S, A) \equiv - \sum_{i=1}^c (|S_i| / |S|) \log_2 (|S_i| / |S|)$$

iii) Gini value:

$$Gini(D) = 1 - \sum_{j=1}^n p_j^2$$

Where p_j is relative frequency of class j in D

3. Data Domain

Rice crop is one of the crops in India, due to its importance as the main food and for exporting. The rice cultivation area in India is approximately. Rice is the main grain crop of India. India ranks second in the world in production of rice. About 34% of the total cultivated area if the nation is under rice cultivation. Out of the total production of food grains, production of rice is 42%. Rice is cultivated in areas having annual average rainfall of 125 cm and average temperature of 23 degree Celsius. Major Rice cultivating areas are north east India, eastern and western coastal regions and river basin of Ganga. West Bengal, Punjab and Uttar Pradesh are the major rice producing states. Besides, Tamil Nadu, Karnataka, Orissa, Haryana, Bihar, Chhattisgarh, Assam and Maharashtra also produce rice. Many affecting diseases infect the Indian rice crop: some diseases are considered more important than others. In this case we focus into the most important diseases for example

Table1

| Attribute | Possible Values |
|-------------|---|
| Variety | Taichung-65, Jaya (IET-723), rohini(PTB-36), Aswathi(PTB-37) |
| Age | Possible Value |
| Part | Leaves, leaves spot,nodes,panicles,grains,plant,flag leaves,leaf sheath,stem |
| Appearance | Spots,oval,fungal |
| Color | Gray,olive,brown,brownish, whitish,yellow |
| Temperature | Real Values |
| Diseases | “blast,” “helminthosporiose,” “stem rot” and “foot rot”, kernelsmut brown spot. |

if appearance=spot and color =discolor then disease =Kernel smut

if appearance=spot and color =brown<=age55

Then disease=brown-spot

4. Decision Tree Comparisons and Results

The decision tree classifier applied on the dataset uses three different splitting criteria namely

- (i) Information Gain
- (ii) Gain Ratio
- (iii) Gini Index

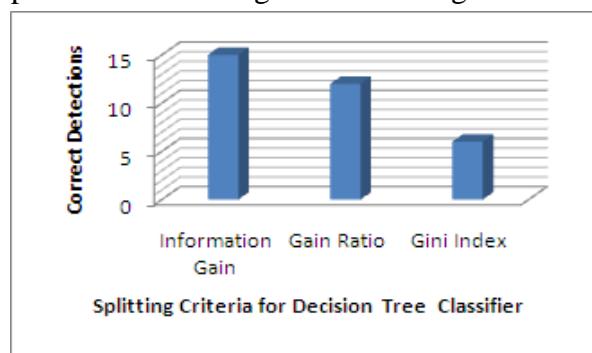
Each option resulted in a different decision tree. The resultant accuracy of each tree when applied to the testing samples also

varied. The complete results are provided below:

Table 2: Decision Tree's splitting criteria comparison data.

| Splitting Criteria | Number of Recognized Samples | Percentage% |
|--------------------|------------------------------|-------------|
| Information Gain | 15 | 65 |
| Gain Ratio | 12 | 52 |
| Gini Value | 6 | 26 |

As the above results depict the fact that change in selection criteria of best attribute while constructing learning tree may change the performance of decision tree classifier. According to above shown results information gain has the highest performance over gain ratio and gini Index



splitting criteria. Below is the performance chart of three different splitting criterions drawn against the number of correctly recognized test samples.

Figure 1: Decision Trees splitting criteria comparison bar chart

5. Conclusion

The decision tree algorithm provides many benefits of trees over many other classifiers such as neural network. The most important benefits are interpretability.

Moreover the c4.5 can effectively create comprehensive tree with greater predictive power and able to get a prediction error about 1.5% on data of test set. The enhancement in classification results over fitting error using pruning techniques and Handling the huge numbers of attribute values.

References

- [1].Gilbert Saporta. Data Mining and Official Statistics, Paper, Chaire de Statistique Appliquée, Conservatoire National des Arts et Métiers. 292 rue Saint Martin, Paris, 15 novembre 2000.
- [2].a Sikandar, Haris Vohra, Syed Samad Ahmed Bukhari, *Faiz-ul-Haque Zeya'*, Decision Tree and Neural Network Classifier Performance Comparison using Canny Cancer Detector a Diagnosis Tool
- [3] Ying Lu, Jiawei Han, "Cancer Classification Using Gene Expression Data", Information Systems vol. 28 issue 4, Elsevier Science Ltd., Oxford,UK, 2003, pp. 243 – 268.
- [4]Patricia L.Dolan, Yang Wu, Linnea K. Ista, Robert L. Metzenberg, Mary Anne Nelson, Gabriel P. Lopez, "Robust and efficient synthetic method for forming DNA microarrays", PubMed Central, Oxford University Press, USA, 2001.

Author Biographies

1. **A.Nithya** received her BSc degree in Computer Science from Bharathiar University and Msc, M.Phil degree in Computer Science from Bharathiar University. Currently doing PhD in Karpagam University. She is currently working in Nehru Arts and Science College, Coimbatore, Tamilnadu, India. Her research Interests include Data Mining, Data Warehousing.

2. Dr.V.Sundaram has more than 35 years of teaching experience in Government & various Private Engineering Colleges. He has published above 30 papers in International journals and conferences. He is currently guiding 25 research scholars in the area of Data mining and Computer Networks. He has been the HOD in-charge in Karpagam College of Engineering Coimbatore. He has served as member of the research board, Anna University, Coimbatore. Currently, he is the HOD of Karpagam College of Engineering and Technology, Coimbatore.

Predict Success or Failure of Remote Infrastructure Management

Er. Satinder Pal Ahuja¹, Dr. Sanjay P. Sood²

¹ Research Scholar, Deptt of Computer Science, Manav Bharti University, Solan, HP, India

² Department of Information Technology, CDAC, Mohali, Punjab, India

Abstract

Matching IT infrastructure with the needs of a business is a CIO's biggest challenge. The increasing complexity of IT infrastructure and the constant pressure to reduce its costs, forces CIOs to maximize the use of existing resources and to enhance productivity of key technical people. Downtime, however brief, can result in revenue losses, unhappy customers and a loss of productivity. Instead of using productive time to making strategic decisions, key personnel are forced to spend it in routine operations management of IT infrastructure. Remote Infrastructure Management (RIM) involves a combination of near-shore and offshore delivery models. RIM can reduce the costs of operations, thus enabling IT managers to consider investing in new technology. The objective of this paper is to focussed whether a RIM will ultimately lead to improvement in Software Process using ANN. The benefit of this work will be that it will save cost and time in actual implementation of RIM

Keywords—Artificial Neural Networks (ANN), Remote InfrastructureManagement (RIM) etc

1. Introduction

Remote Infrastructure Management Services provide for monitoring and management of all infrastructures pertaining to networks, data centers, servers, storage security, Applications and End user computing, outside the company's offices.

Advantages of RIM include:

- Cost Reduction - Reduced manpower, performance efficiency and capacity planning resulting in 40-60% reduction in costs
- Quality and Process - Process-based approach to solve issues, efficient handling of escalations, and

quality certifications to ensure adherence to standards and security controls

- Best practices – Experience with multiple enterprises resulting in standardization through ideas, learning and best practices
- Expertise - Domain experts who aid others in skill development
- Risk Mitigation - Improved risk-mitigation and assured business continuity
- Visibility - Timely reporting providing CIOs with greater visibility, real-time control and analysis of historical trends. Increased automation with integrated tools, providing common framework for operations
- Scalability - Absorb the peaks and troughs of manpower needs
- Service levels - Precise service-level agreements with penalty clauses for downtime
- Pre-emptive problem resolution - Through proactive monitoring and event correlation

2. Operational Issues of RIM

Following are the issues observed in RIM process.

- People related issues
- Poorly defined roles and responsibilities
- Every team does administration, development & support
- No clear responsibility matrix / SLAs. Outsourced support -
Technology related issues
- When there is no evidence of architecture documents for network, systems and security
- When there are no tools for monitoring & managing the infrastructure
- When there is poor metrics measurement
- When there are single points of failure in Internet connectivity, Load Balancers, Firewalls
- When there is weak security

- When redundancy for applications at the system/server/ database level is unavailable
- Since actual Implementation of RIM involves large costs and time investments there is a need to develop a system for predicting the success or failure of RIM

3. Objective

The objective of this work is to develop an ANN based system to verify whether a RIM will ultimately lead to improvement in Software Process. The benefit of this work will be that it will save cost and time in actual implementation of RIM

4. Literature Survey

While IT management services represent a mature subject in the IT business arena, the emerging cloud generation of management services require critical enhancements to the current processes and technologies in order to deliver IT management remotely with rapid onboarding and minimal labor involvement from experts, to be affordable and scale up to the promise of the cloud. Traditional Remote Infrastructure Management (RIM) service providers use their own Network Operations Centers (NOC) to remotely monitor and manage customers' IT infrastructure. The primary business value for RIM services is that it helps global enterprises to small and medium businesses (SMB) to outsource the burden of managing their IT infrastructure. Although the IT management service itself delivered this way is more affordable, the RIM customer on-boarding process particularly is not, taking between one to two months of expensive labor.

Management services represent a mature subject in the IT business arena. According to Gartner Dataquest, Remote Infrastructure Management (RIM) is a rapidly growing market growing at a Compound Annual Growth Rate of 36%, and projected to grow from USD \$14.3B to \$30B by 2010 [[8]]. Typical RIM service providers use their own Network Operations Centers to remotely monitor and manage customers' IT infrastructure elements such as networks, systems' hardware and operating systems, and applications. The primary business value for RIM services is that it helps global enterprises and SMBs to outsource the burden of managing their IT infrastructure, thus, cutting down costs for infrastructure management and gaining access to expert skills. The customers can focus then on their core business, shifting the responsibility for IT management to RIM, while maintaining ownership of their assets.

A RIM solution generally involves monitoring services comprising of NOC support, reporting, incident notification and escalation, while management services cover problem management and root cause analysis, configuration management, change and release management, maintenance and updates installation. Prior to providing any of these RIM services, the customer has first to select what services to subscribe to during a procedure that is called "on-boarding". Although the IT management itself is rendered more affordable when provided remotely as a service, the RIM customer on-boarding process particularly is not, taking between one to two months of expensive labor.

The current process for RIM customer onboarding consists of multiple interviews and interactions with customers to 'discover' their IT environment, identify the resources to be managed and guide the enablement of the environment for remote management. This labor intensive approach (measured in weeks) proves to be unscalable when RIM is to be delivered as Management-as-a-Service from an IT infrastructure management cloud. Cloud computing is an emerging paradigm whereby services and computing resources are delivered to customers over the internet (or intranet) from a service provider who owns and operates the cloud. Cloud-based services characteristically can scale up promptly to meet growing demand. The benefit of this will remain unrealized if RIM onboarding takes weeks as is the standard today. Since the duration to traditionally provision resources for new RIM customers is comparable to the current onboarding duration, there is little incentive to motivate change to the current on-boarding approach.

However, RIM's goal for delivery from the cloud is 'on-boarding in minutes', which means radical revision of the current approach. To this end, we have identified the following on-boarding problems: (1)lack of a standardized approach or automation for the on-boarding operation flow, (2) inaccuracies in manually assessing the environment from the customer's descriptions or semi-updated inventory files, (3) missing configuration data (e.g., credentials, directory paths, key performance indicators -- KPIs) necessary to setup the monitoring systems, (4) overhead for the SMB customer who is expected to perform complex configurations in their environment (e.g., VPN setup, monitoring data agent/collector installations), (5) evaluated price is not commensurate with the cost of the service expected to be provided.

There are many managed services providers in the marketplace. Some are local providers, others regional, and still others global. It is largely the regional and global providers that utilize RIM techniques. They are recruiting IT professionals and making them available to client projects through the use of Global Delivery centers. For the client on-boarding process, these IT professionals have

to identify the client's IT environment either manually during multiple interviews with the customer or by providing a template for exchanging inventory information (e.g., a spreadsheet). Sometimes the customer may provide one by filling out his own questionnaire. However, a better option is programmatic discovery using dedicated discovery software. The inventory information and additional configuration details are then used to configure the monitoring and management toolset. Manual information gathering methods are notoriously error-prone – some of these errors are caught during the tool configuration step which engenders more interactions with the same customer. Other causes that drive the inaccuracy of the manual environment assessment are existing inventory out of date, incomplete or invalid data, and untracked configuration changes (e.g., for the credentials, directory paths, KPIs), that may jeopardize the quality of the RIM service. We will use the manual data gathering performances in the comparison of our experimental results. When the IT environment discovery is done programmatically, the typical approaches are via stand-alone products, e.g., TADDM [1] or via services that make use of remote product download, e.g., Paglo [2]. Although more accurate in terms of discovery quality compared to the manual approach, the stand-alone products are not suitable for small and medium business or SMBs, which have tens to few hundreds of servers. These customers cannot afford nor need sophisticated tools oriented towards large IT enterprises with thousands of IT elements. SMBs prefer to use a streamlined asset discovery service to get the inventory of their IT environment, without the hassle of installing, configuring and managing a discovery product. However, the current remote discovery service providers still require software to be installed and configured by the SMBs in their environment.

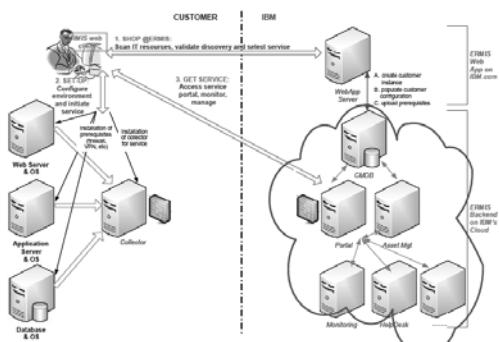


Figure 1. Enhanced on-boarding process for RIM delivered from an IT management cloud.

In [2] the provider offers the discovery tool for free in the context of their monitoring service and the customer has to take care of the discovery tool installation. This is due to the fact that the discovery must take place from a node in

the network to be probed, to circumvent firewalls, Network Address Translation (NAT), and other impediments. Many users, especially in the SMB space, prefer not to face the burden of this administrative overhead and may lack the necessary skills required. After the discovery is completed, the IT professionals have typically to manually collect additional configuration details since the discovery provides partial detection of the environment. We call this approach manual onboarding with automated discovery and will compare in Section 5 its characteristics to our approach presented in this paper as well as to the manual approach. Other related solutions that involve remote discovery include [3] which uses a browser to control a discovery process, however, it is unclear whether the provider intends for the "NDM Agent" to be running on the web server or some other machine (or whether it should or could be located in the browser). They also provide "passive discovery" which involves packet sniffing to discover applications running on client machines. JLocator [4] describes a Java applet based network discovery tool. They discover the network topology, but do not examine applications neither services on the identified elements. XAssets [5] is a service comparable to SNAPPiMON [6], where the management tool is also browser based. In [5], the discovery process uses a wide variety of discovery techniques to collect hardware items details, while unrecognized software items lists from customers are sent on a regular basis to the provider staff and these items are manually investigated and added to the discovery database. In [6] the discovery is also a combination of manual discovery, for network and server level items and credentials, and automatic look-up for OS and application configuration. Once the inventory has been discovered and validated, and the additional information on the resources to be managed gathered, the IT professionals proceed to or guide the customer through the enablement of the environment for remote management. This step consists of the installation of data collector or agents into the customer's premises, firewall configuration for site-to-site VPN set-up and NATing of endpoints.

Finally, upon performing all necessary data collection and setup for monitoring and managing the selected items in the customer's environment, the RIM provider prices the offering and starts delivering the IT management service.

Artificial Neural Networks have emerged as a major paradigm for Data Mining applications. Neural nets have gone through two major development periods -the early 60's and the mid 80's. They were a key development in the field of machine learning. Artificial Neural Networks were inspired by biological findings relating to the behavior of the brain as a network of units called neurons. The human brain is estimated to have around 10 billion neurons each connected on average to 10,000 other neurons. Each neuron receives signals through synapses that control the

effects of the signal on the neuron. These synaptic connections are believed to play a key role in the behavior of the brain.

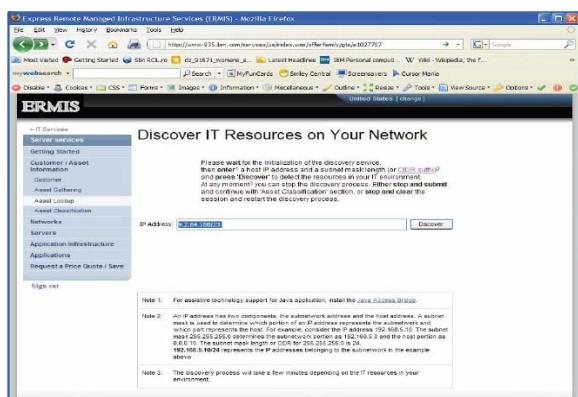


Figure 2. Applet prompts customer for network address range to scan.

Neural networks take a different approach to problem solving than that of conventional computers. Conventional computers use an algorithmic approach i.e. the computer follows a set of instructions in order to solve a problem. Unless the specific steps that the computer needs to follow are known the computer cannot solve the problem. That restricts the problem solving capability of conventional computers to problems that we already understand and know how to solve. But computers would be so much more useful if they could do things that we don't exactly know how to do.

Neural networks process information in a similar way the human brain does. The network is composed of a large number of highly interconnected processing elements(neurones) working in parallel to solve a specific problem. Neural networks learn by example. They cannot be programmed to perform a specific task. The examples must be selected carefully otherwise useful time is wasted or even worse the network might be functioning incorrectly. The disadvantage is that because the network finds out how to solve the problem by itself, its operation can be unpredictable.

On the other hand, conventional computers use a cognitive approach to problem solving; the way the problem is to solved must be known and stated in small unambiguous instructions. These instructions are then converted to a high level language program and then into machine code that the computer can understand. These machines are totally predictable; if anything goes wrong is due to a software or hardware fault.

Neural networks and conventional algorithmic computers are not in competition but complement each other. There are tasks are more suited to an algorithmic approach like arithmetic operations and tasks that are more suited to

neural networks. Even more, a large number of tasks, require systems that use a combination of the two approaches (normally a conventional computer is used to supervise the neural network) in order to perform at maximum efficiency.

5. Methodology

Matlab will be used as the simulation tool. Attempt will be made to build a classifier that can predict the success or failure of implementation of RIM. Six parameters of the RIM will be considered. Neural networks have proved themselves as proficient classifiers and are particularly well suited for addressing non-linear problems. Given the non-linear nature of real world phenomena, like predicting success of RIM, neural networks is certainly a good candidate for solving the problem. The six characteristics will act as inputs to a neural network and the prediction of success will be the target. Given an input, which constitutes the six measured values for the parameters of the matrix, the neural network is expected to identify if the RIM process will produce success or not. This is achieved by presenting previously recorded RIM parameters to a neural network and then tuning it to produce the desired target outputs. This process is called neural network training. The samples will be divided into training, validation and test sets. The training set is used to teach the network. Training continues as long as the network continues improving on the validation set. The test set provides a completely independent measure of network accuracy. The trained neural network will be tested with the testing samples. The network response will be compared against the desired target response to build the classification matrix which will provide a comprehensive picture of a system performance.

The training data

The training data set includes a number of cases, each containing values for a range of input and output variables. The first decisions you will need to make are: which variables to use, and how many (and which) cases to gather.

The choice of variables (at least initially) is guided by intuition. Expertise in the problem domain will give you some idea of which input variables are likely to be influential. As a first pass, you should include any variables that you think could have an influence - part of the design process will be to whittle this set down.

Neural networks process numeric data in a fairly limited range. This presents a problem if data is in an unusual range, if there is missing data, or if data is non-numeric. Fortunately, there are methods to deal with each of these problems. Numeric data is scaled into an appropriate range for the network, and missing values can be substituted for

using the mean value (or other statistic) of that variable across the other available training cases.

Handling non-numeric data is more difficult. The most common form of non-numeric data consists of nominal-value variables such as Outcome= {Success, Failure}. Nominal-valued variables can be represented numerically. However, neural networks do not tend to perform well with nominal variables that have a large number of possible values.

For example, consider a neural network being trained to estimate the value of houses. The price of houses depends critically on the area of a city in which they are located. A particular city might be subdivided into dozens of named locations, and so it might seem natural to use a nominal-valued variable representing these locations. Unfortunately, it would be very difficult to train a neural network under these circumstances, and a more credible approach would be to assign ratings (based on expert knowledge) to each area; for example, you might assign ratings for the quality of local schools, convenient access to leisure facilities, etc. Other kinds of non-numeric data must either be converted to numeric form, or discarded. Dates and times, if important, can be converted to an offset value from a starting date/time. Currency values can easily be converted. Unconstrained text fields (such as names) cannot be handled and should be discarded.

The number of cases required for neural network training frequently presents difficulties. There are some heuristic guidelines, which relate the number of cases needed to the size of the network (the simplest of these says that there should be ten times as many cases as connections in the network). Actually, the number needed is also related to the (unknown) complexity of the underlying function which the network is trying to model, and to the variance of the additive noise. As the number of variables increases, the number of cases required increases nonlinearly, so that with even a fairly small number of variables (perhaps fifty or less) a huge number of cases are required. This problem is known as "the curse of dimensionality," and is discussed further later.

For most practical problem domains, the number of cases required will be hundreds or thousands. For very complex problems more may be required, but it would be a rare (even trivial) problem which required less than a hundred cases. If your data is sparser than this, you really don't have enough information to train a network, and the best you can do is probably to fit a linear model. If you have a larger, but still restricted, data set, you can compensate to some extent by forming an ensemble of networks, each trained using a different resampling of the available data,

and then average across the predictions of the networks in the ensemble.

Many practical problems suffer from data that is unreliable: some variables may be corrupted by noise, or values may be missing altogether. Neural networks are also noise tolerant. However, there is a limit to this tolerance; if there are occasional outliers far outside the range of normal values for a variable, they may bias the training. The best approach to such outliers is to identify and remove them (either discarding the case, or converting the outlier into a missing value). If outliers are difficult to detect, a city block error function may be used, but this outlier-tolerant training is generally less effective than the standard approach.

Pre- and Post-processing

All neural networks take numeric input and produce numeric output. The transfer function of a unit is typically chosen so that it can accept input in any range, and produces output in a strictly limited range (it has a squashing effect). Although the input can be in any range, there is a saturation effect so that the unit is only sensitive to inputs within a fairly limited range. The illustration below shows one of the most common transfer functions, the logistic function (also sometimes referred to as the sigmoid function, although strictly speaking it is only one example of a sigmoid - S-shaped - function). In this case, the output is in the range (0,1), and the input is sensitive in a range not much larger than (-1,+1). The function is also smooth and easily differentiable, facts that are critical in allowing the network training algorithms to operate.

The limited numeric response range, together with the fact that information has to be in numeric form, implies that neural solutions require preprocessing and post-processing stages to be used in real applications. Two issues will need to be addressed:

Scaling. Numeric values have to be scaled into a range that is appropriate for the network. Typically, raw variable values are scaled linearly. In some circumstances, non-linear scaling may be appropriate (for example, if you know that a variable is exponentially distributed, you might take the logarithm). Non-linear scaling is not supported in ST Neural Networks. Instead, you should scale the variable using STATISTICA's data transformation facilities before transferring the data to ST Neural Networks.

Nominal variables. Nominal variables may be two-state (e.g., Outcome ={Success ,Failure }) or many-state (i.e., more than two states). A two-state nominal variable is easily represented by transformation into a numeric value (e.g., Success =0, Failure =1). Many-state nominal

variables are more difficult to handle. They can be represented using an ordinal encoding but this implies a (probably) false ordering on the nominal. A better approach, known as one-of-N encoding, is to use a number of numeric variables to represent the single nominal variable. The number of numeric variables equals the number of possible values; one of the N variables is set, and the others cleared. ST Neural Networks has facilities to convert both two-state and many-state nominal variables for use in the neural network. Unfortunately, a nominal variable with a large number of states would require a prohibitive number of numeric variables for one-of-N encoding, driving up the network size and making training difficult. In such a case it is possible (although unsatisfactory) to model the nominal variable using a single numeric ordinal; a better approach is to look for a different way to represent the information. In classification, the objective is to determine to which of a number of discrete classes a given input case belongs. The most common classification tasks are two-state, although many-state tasks are also not unknown. Neural networks can actually perform a number of classification tasks at once, although commonly each network performs only one. In this case the network will have a single output variable.

Multilayer Perceptrons is the type of network in which the units each perform a biased weighted sum of their inputs and pass this activation level through a transfer function to produce their output, and the units are arranged in a layered feedforward topology. The network thus has a simple interpretation as a form of input-output model, with the weights and thresholds (biases) the free parameters of the model. Such networks can model functions of almost arbitrary complexity, with the number of layers, and the number of units in each layer, determining the function complexity. Important issues in Multilayer Perceptrons (MLP) design include specification of the number of hidden layers and the number of units in these layers.

The number of input and output units is defined by the problem (there may be some uncertainty about precisely which inputs to use, a point to which we will return later. However, for the moment we will assume that the input variables are intuitively selected and are all meaningful). The number of hidden units to use is far from clear. As good a starting point as any is to use one hidden layer, with the number of units equal to half the sum of the number of input and output units.

We are going to use unary encoding in this simulation to perform symbol translation. The first six columns of data will represent the emails characteristics. The 7th column represents whether the RIM is successful or not. This data will be randomly generated. The next step will be to preprocess the data into a form that can be used with a

neural network. The next step is to create a neural network that will learn to identify if the RIMprocess will cause improvement or not.

The assumed samples will be automatically divided into training, validation and test sets. The training set will be used to teach the network. Training will continue long as the network continues improving on the validation set. The test set will provide a completely independent measure of neural network accuracy to detect success. The trained neural network will be tested with the testing samples. This will give a sense of how well the network will do when applied to data from the real world.

6. Conclusion

Due to vast changes in IT Infrastructure & technologies, ANN is useful to predict success or failure of RIM.

Neural networks have proved themselves as proficient classifiers and are particularly well suited for addressing non-linear problems. Given the non-linear nature of real world phenomena, like predicting success of RIM, neural networks is certainly a good candidate for solving the problem.

7. References

1. Manager(TADDM),
<http://www.ibm.com/software/tivoli/products/taddm>
2. Paglo,<http://www.paglo.com/opensource/paglocrawler>
3. Design of Hybrid Network Discovery Module for Detecting Client Applications an ActiveX Controls,
<http://www.springerlink.com/content/y51p5g76k25578g1/>
4. JLocator,
<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.3.9.5030>
5. XAssets, <http://www.xassets.com/discovery.aspx>
6. IBM SNAPPiMON, <http://www.snappimon.com/>
7. Head, M.R.; Sailer, A.; Shaikh, H.; Viswanathan, M.; , "Taking IT Management Services to a Cloud," Cloud Computing, 2009. CLOUD '09. IEEE International Conference on , vol., no., pp.175-182, 21-25 Sept. 2009
8. Aljunaid, A.B.; AbuElMaaly, I.; Sagahyroon, A.; , "Using ANN To Predict The Best HUB Location," Circuits and Systems, 2006. APCCAS 2006. IEEE Asia Pacific Conference on , vol., no., pp.317-320, 4-7 Dec. 2006
9. Nmap, <http://nmap.org/>
10. Gartner Dataquest, August 2006,
<http://www.gartner.com/it/products/research/dataquest.jsp>

MRI Mammogram Image Segmentation using NCut method and Genetic Algorithm with partial filters

Pitchumani Angayarkanni⁽¹⁾ M.C.A,M.Phil,(Ph.d)

Assistant Professor ,Department of Computer Science,Lady Doak College, Madurai

ABSTRACT:

Cancer is one of the most common leading deadly diseases which affect men and women around the world. Among the cancer diseases, breast cancer is especially a concern in women. It has become a major health problem in developed and developing countries over the past 50 years and the incidence has increased in recent years. Recent trends in digital image processing are CAD systems, which are computerized tools designed to assist radiologists. Most of these systems are used for automatic detection of abnormalities. However, recent studies have shown that their sensitivity is significantly decreased as the density of breast increases. In this paper , the proposed algorithm uses partial filters to enhance the images and the Ncut method is applied to

segment the malignant and benign regions , further genetic algorithm is applied to identify the nipple position followed by bilateral subtraction of the left and the right breast image to cluster the cancerous and non cancerous regions. The system is trained using Back Propagation Neural Network algorithm. Computational efficiency and accuracy of the proposed system are evaluated based on the Frequency Receiver Operating Characteristic curve(FROC). The algorithm are tested on 161 pairs of digitized mammograms from MIAS database. The Receiver Operating Characteristic curve leads to 99.987% accuracy in detection of cancerous masses.

Keywords: Filters, Normalized Cut, Segmentation, BPN, Genetic Algorithm and FROC.

varies from 0.1 mm to 1mm. “Cluster: of MCs is defines as a group of three to five MCs within regions. Generally microcalcification clusters are important indication of possible cancer. This algorithm effectively and automatically detect MCs

2. Algorithm Design:

There are four steps involved in the algorithm for the detection MCCs which is shown in the figure.

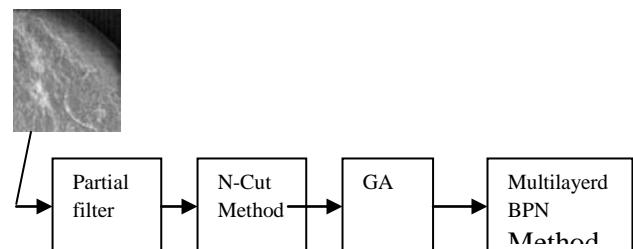


Fig 1: Flow Chart of Algorithm

2.1 Partial Filter for Image enhancement: A filter is a mathematical transformation (called a *convolution product*) which allows the value of a pixel to be

In this paper we have introduced the detection of microcalcifications. As one of the early signs of breast cancer , mirocalcifications are tiny granule like deposits of calcium, which appear as small bright spots of mmaograms. Their size

modified according to the values of neighbouring pixels, with coefficients, for each pixel of the region to which it is applied. The filter is represented by a table (matrix), which is characterized by its dimensions and its coefficients, whose centre corresponds to the pixel concerned. The table coefficients determine the properties of the filter[1]. The following is an example of a 3 X 3 filter:

| | | |
|---|---|---|
| 1 | 1 | 1 |
| 1 | 4 | 1 |
| 1 | 1 | 1 |

One of the most important problems in image processing is denoising. Usually the procedure used for denoising, is dependent on the features of the image, aim of processing and also post-processing algorithms [5].

Denoising by low-pass filtering not only reduces

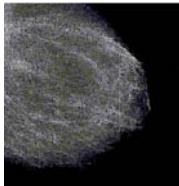
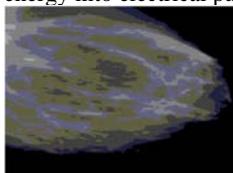
| High Pass Filter | Low Pass filter |
|---|---|
| MRI Mammogram Image Segmentation using NCut method and Genetic Algorithm with partial filters  | Low pass filtering, otherwise known as "smoothing", is employed to remove high noise from a digital image. Noise is often introduced during the analog-to-digital conversion process as a side-effect of the physical conversion of patterns of light energy into electrical patterns.  |

Figure 2:
Mammogram Image enhanced using high pass filter

Figure 3: Mamamogram Image Enhancement using Low Pass filter

the noise but also blurs the edges.

Spatial and frequency domain filters are widely used as tools for image enhancement. Low pass filters smooth the image by blocking detail information. Mass detection aims to extract the edge of the tumor from surrounding normal tissues and background, high pass filters (sharpening filters) could be used to enhance the details of images

3 Image Segmentation:

The goal of image segmentation is to cluster pixels into salient image regions, i.e., regions corresponding to individual surfaces, objects, or natural parts of objects. In this we apply Normalized Cut method of segmentation to cluster microcalcification regions[2].

Finally we outline the normalized cut approach of Shi and Malik [13].Here we seek a partition F and G = V- F of the affinity weighted,undirected graph (without source and sink nodes). In order to avoid partitions where one of F or G is a tiny region, Shi and Malik propose the normalized cut criterion, namely that F and G should minimize.

$$N(F,G) \equiv L(F,G)(1/L(F,V)+1/L(G,V))$$

$$L(F,G) = \sum a(\bar{x}_i, \bar{x}_j).$$

$$\nabla x_i \in F, \nabla x_j \in G$$

Unfortunately , the resulting graph partitioning problem.

$$F = \arg \min N(F, V-F)$$

FCV

Note any segmentation technique can be used for generating proposals for suitable regions F, for which N(F, V - F) could be evaluated. Indeed, the SMC approach above can be viewed as using S and T to provide lower bounds on the terms L(F, V) and L(G, V) (namely L(S, V) and L(T, V), respectively), and then using the S-T min cut to globally minimize L(F,G) subject to S C F and T C G. Using this method the microcalcifications are clustered

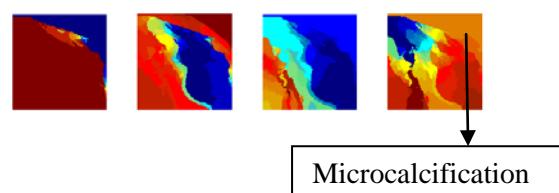


Figure 4: After Normalized Cut Segmentation

The computational efficiency 12.563 seconds on the 160x160 image.

4 Genetic Algorithm:

A partial filtering based normalized cut method is used to generate a image to separate the breast and the non breast region . The GA enhances the breast border . Border detector detects the edges in the binary images , where each pixel takes on either the intensity value of zero for a non border pixel or one for border pixel[3]. Each pixel in the binary map corresponds to an underlying pixel in the original image . In this proposed system , kernel

is extracted from border points as a neighborhood array of pixels of the size 3*3 window of binary image. The binary kernels are considered population strings for GA. The corresponding kernels are extracted from gray level mammogram image using spatial coordinate points and the sum of the intensity values are considered as the fitness value . After identifying initial population and the fitness value , the genetic operator can be applied to generate a new population. Reproduction operator produces new string for crossover. Reproduction is implemented as linear search through roulette wheel with slots weighted in proportion to kernel fitness values. In this function, a random number multiplies the sum of population fitness called as stopping point.

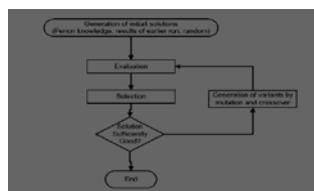


Figure 5: GA

| DSS Size | Training hit Rate | Evaluation Hit Rate |
|----------|-------------------|---------------------|
| 10 | 80.6 | 69.7 |
| 22 | 81.2 | 74.3 |
| 58 | 86.2 | 85.7 |
| 100 | 86.6 | 82.6 |

Table 1: DSS Parameter and Performance

| Parameter | Value |
|-------------------------------|-----------|
| Population Size | 400 |
| Maximum number of tournaments | 40000 |
| Mutation Frequency | 8 |
| Crossover Frequency | 94 |
| Maximum Program size | 516 |
| Instruction Set | {+,-,*,/} |

Table 2:Parameter Setting for Genetic Programming

5 Generating the Asymmetric Image:

After the images were aligned, bilateral subtraction was performed [47,48] by subtraction was performed by subtracting the digital matrix of

the left breast image from the digital matrix of the right breast image. Microcalcification in the right breast image have positive pixel values in the image obtained after subtraction, while microcalcification in the left breast image have negative pixel values in the subtracted image. As a result, two new images were generated: one with positive values and the other with negative values. The most common gray level was zero, which indicated no difference between the left and right images. Simple linear stretching of the two generated images to cover the entire available range of 1024 gray levels was then calculated. The difference between corresponding pixels contains important information that can be used to discriminate between normal and abnormal tissue. The asymmetry image can be thresholded to extract suspicious regions. To generate FROC curve, the asymmetry image is thresholded using ten different intensity values ranges from 50-150. Figure 6 shows a asymmetry image and connected regions extracted based on thresholding to obtain a progressively larger number of high difference pixels.



Figure 6 Asymmetric images

Two different techniques are used in the interpretation of mammogram. The first technique consists of systematic search of each mammogram for visual pattern symptomatic tumors. Such as, a bright, approximately circular blob with hazy boundary might indicate the presence of a circumscribed mass. The second technique, the asymmetric approach , consists of systematic comparison of corresponding regions in the left and the right breast.

6 BPN training:

In addition, a backpropagation artificial neural network (BP-ANN) was also developed and evaluated on the same data. The parameters for ANN training were published before. Figure 5 compare the ROC curves for the LGP and the BP-ANN algorithms respectively. The BP-ANN yielded an ROC area index of $Az=0.88\pm0.01$. Our GP approach achieved a statistically significantly better performance with $Az=0.91\pm0.01$.

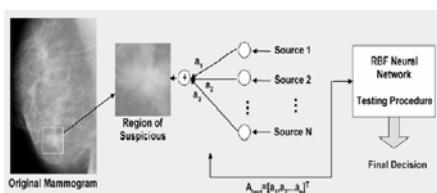
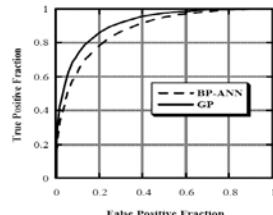


Figure 7a) Steps involved in automated Classification using Ant Colony Optimization

ROC performance evaluation of the GP and BP-ANN classifiers for the discrimination of true masses from normal breast parenchyma in screening mammograms.



7. ROC curve:

Finally the technique was evaluated on the mammograms randomly selected from the non-suspicious section of the data base. The method outlined small regions in 5 out of the 15 non suspicious mammograms. The areas identified were generally very small compared to those in abnormal mammograms

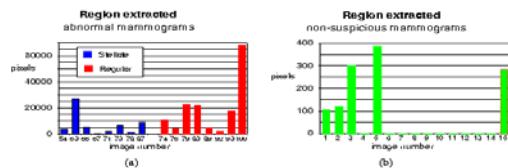


Figure 8 Lesion Areas detected for Abnormal and Non-Suspicious cases (large image extracts). [Figures (a) and (b) are presented at different ordinate scales]

Fig 8(a) shows the extracted areas for the abnormal lesions. (Image sequence 54 - 87 are stellate lesions and 74 to 100 are regular masses). We first establish whether these represent two different populations, by applying a Mann-Whitney (Wilcoxon rank sum) non-parametric test, since it is unrealistic to presume any specific underlying distribution. Median values are 450 and 1450 pixels respectively which produce a confidence level of 85% that the two data sequences emanate from distinct populations. Since this is not significant at normally acceptable levels we can compare the abnormalities as a single distribution against the non-suspicious set, Fig 8(b). Using the same test, median values of 5500 and 10 pixels for the two distributions are established, giving a confidence level of greater than 97.5% that the two distributions are different, suggesting that our protocols are an effective method of area detection.

8. Results & Discussion:

In our proposed algorithm the mammogram is segmented Partial filter based enhancement and Ncut method based segmentation with clustering using Genetic Algorithmic system that optimizes the Maximizing a Posterior Probability(MAP)[4]. The Neuron Genetic Algorithm based image segmentation method is a process seeking the optimal labeling of the image pixels. Labeling process consists of assigning same label to the kernels having similar patterns. Kernel is a 3*3 window of neighborhood pixels. The Optimum label is the one which minimizes the MAP estimate.. The system is trained using Multilayered feed forward Network was found to be 99.99% accurate.

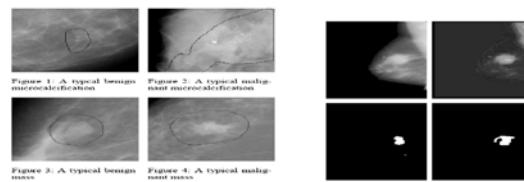


Fig 9 Original image, “voteboard”, “coarse” segmentation, “fine” Segmentation

| MIAS Category | No. of Image Pairs | No. of Abnormalities |
|--------------------------|--------------------|----------------------|
| Normal | 53 | - |
| Circumscribed | 21 | 26 |
| Masses:Speculated Masses | 19 | 18 |
| III-Defined Masses | 14 | 15 |
| Architectural | 18 | 19 |
| Distortion: Asymmetry | 14 | 14 |
| Calcification | 22 | 25 |
| Total | 161 | 117 |

Table 3: Tested Pairs Vs. Abnormalities

| Author & Reference | Methods | Computational Time | Computational Efficiency | Detection Rate |
|---------------------|--|--------------------|--------------------------|----------------|
| Ferrari & Rangayyan | Directional Filtering with Gobar Wavelet | 12 Sec | $O(n)$ | 74.4 % |
| Lau and Bischor | Asymmetry Measure | 15 Sec | $O(\log n)$ | 80.0 % |
| Sallam and Bowyer | Unwrapping Technique | 10 Sec | $O(n \log n)$ | 81.6 % |
| Proposed Approach | NCut Segmentation with GA based Neural network | 2 S | $O(n \log n)$ | 99.9 % |

Table 4: Detection Rate and Efficiency proposed

9. Conclusion:

The proposed algorithms are tested on 161 pairs of digitized mammograms from Mammographic Image analysis Society(MIAS) database. A free response receiver operating characteristic (FROC) curve is generated for the mean value of the detection rate for all the 161 pairs of mammograms in the MIAS database, to evaluate the performance of the proposed method. There is no doubt that for the immediate future mammography will continue to play a major role in the detection of breast cancer. The ultimate objective of this paper was to identify tumor or masses in breast tissue[5]. Since hamartomas consists of normal breast tissue with abnormal proportions and the first step was try to identify the different tissue type in mammography with normal breast tissue. The important features have been extracted from the Normalized cut method of the each sub image using various statistical techniques. The Genetic algorithm has been implemented and the breast border was identified from the clustered image. The tests that were carried out using a set of 117 tissues samples, 67 benign and 50 malignant. The result analysis has given a sensitivity of 99.8%, a specificity of 99.9% and an accuracy above 99.9%, which means encouraging results. The preliminary results of this approach are very promising in characterizing breast tissue.

References:

- [1] Bosch. A.; Munoz, X.; Oliver.A.; Marti. J., **Modeling and Classifying Breast Tissue Density in Mammograms**, Computer Vision and Pattern Recognition, 2006 IEEE Computer

Society Conference on Volume 2, Issue , 2006
 Page(s): 1552 – 15582.

[2] Dar-Ren Chena, Ruey-Feng Changb, Chii-Jen Chenb, Ming-Feng Hob, Shou-Jen Kuo, Shou-Tung Chena, Shin-Jer Hungc, Woo Kyung Moond,*Classification of breast ultrasound images using fractal feature*, *ClinicalImage*, Volume 29, Issue4, Pages 234-245.

[3] Suri, J.S., Rangayyan, R.M.: *Recent Advances in Breast Imaging, Mammography, and Computer-Aided Diagnosis of Breast Cancer*. 1st edn. SPIE (2006)

[4] Hoos, A., Cordon-Cardo, C.: *Tissue microarray profiling of cancer specimens and cell lines: Opportunities and limitations*. Mod. Pathol. 81(10), 1331–1338 (2001)

[5] Lekadir, K., Elson, D.S., Requejo-Isidro, J., Dunsby, C., McGinty, J., Galletly, N., Stamp, G., French, P.M., Yang, G.Z.: *Tissue characterization using dimensionality reduction and fluorescence imaging*. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4191, pp. 586–593. Springer, Heidelberg (2006).

About the Author:

Pitchumani Angayarkanni Sekaran is Assistant Professor at Lady Doak College, Madurai,Tamil Nadu,India. My research interests include “Image Processing and data mining”.

A Schematic Technique Using Data type Preserving Encryption to Boost Data Warehouse Security

M.Sreedhar Reddy¹ Prof.M.Rajitha Reddy² Prof R.Viswanath³ Prof.G.V.Chalam⁴ Prof.Rajya Laxmi⁵ Prof.Md.Arif Rizwan

¹Dept of CSE SSIETW,Hyderabad,AP.India

²Dept of CSE ,SSIETW,Hyderabad, AP.India

^{3&6}Dept of CSE ,REC,Hyderabad, AP.India

⁴Dept of commerce&BA, A N University, Guntur , AP.India

⁵Dept of IT
GITAM, VIZAG. AP.India

Abstract

An ingenious data warehouse habitually contains information which must be painstaking enormously sensitive and proprietary. Protection of this information, as important as it is, is too often thorny by the presence of assorted computing environments, managerial issues, difficulties in controlling data distribution, and slipshod attitudes towards information security. We present a method of in progression fortification based on an encryption scheme which preserves the data type of the plaintext resource. We suppose that this method is particularly companionable for multifaceted data warehouse environments

Key words: Query, prevention, detection, encryption, decryption, datawarehouse, datamining

1. Introduction

Nowadays numerous IT security professionals it sounds as if believe that their restraint has three goals: confidentiality, integrity, and availability. This paper explores a different loom to the problem of IT security goals. Guided by the desire to anchor IT security in the most familiar features of ordinary human life, the approach is two pronged.

In terms of “security”; continues by analyzing them; and, based upon this analysis, proposes a basic definition of “Information Technology security.”

Data warehouse technology has, in modern years, provides its corporate executives and business planners with extraordinarily powerful decision support tools. Data warehouses can tell us **which** products to manufacture, **where** to locate factories and **how** to gain market share. They can be used to answer questions which weren't even contemplated when the data warehouse was built. Most remarkable of all—these feats can be accomplished using data extracted from existing operational systems.

It has become gradually more perceptible however, that the data warehouse is an alluring

goal for snoopers. To envisage that you are a annoyed employee, or industrial surveillance representative who manages to gain access to an organization's computer system. Now, would we likely to find out?

- Data grouped into subject data areas so that you can quickly find the items of interest.
- Accurate and complete information that has been painstakingly reviewed for correctness.
- Time-indexed data so trends can be easily identified.
- Summarized information with the ability to drill down for details.

Conceptually, the data warehouse process consists of three simple steps:

- 1) Extract data from the operational system,
- 2) Load the extracted data into data warehouse tables, and
- 3) Query the data warehouse to obtain decision support information.

Data is at risk during each of these phases. Several factors render data warehouses particularly susceptible to attack:

- 1) Extracted data is frequently transmitted over insecure communication lines.

- 2) Extracted data is stored on a variety of computer systems and removable media which may have only minimal security.
- 3) The extraction process produces intermediate files and load files which contain sensitive information, but may not be well-protected.
- 4) Maintaining proper security attributes for the data warehouse tables is extremely time consuming in the face of constant organizational change.
- 5) Users often retrieve data from the data warehouse and create a “data mart,” leading to widely distributed copies of sensitive data.
- 6) Sound security practice is often undermined because the data warehouse development effort is a “high visibility” project with a tight schedule.

Furthermore that research into the unique aspects of data warehouse security is still in the early stages, additional vulnerabilities will certainly be identified in the future

2. Problem Statement

When to develop a cryptographic approach to data warehouse security which could be practical in the complex, heterogeneous environments encountered in the business world, we recognized certain decisive objectives:

- 1) The approach should work with any combination of commonly used relational databases. (This rules out requiring binary data storage or other database-dependent features.)
- 2) It must function on multiple hardware platforms and operating systems.
- 3) It must appropriately encrypt and decrypt data on machines with different character sets (e.g., ASCII and EBCDIC).
- 4) The strength of the encryption algorithm should be comparable to widely used, state of the art technology.
- 5) To add encryption to an existing database should require no changes to the structure of the database. (Neither should any application changes be required to access non-encrypted fields.)
- 6) Encryption should occur as early as possible in the extraction process and decryption should occur at the last possible moment.
- 7) It could not be dependent on a particular programming language.
- 8) It should be “fail safe.” (Any likely failure mode should be such that that access to the data is *denied*.) These requirements, based on our business requirements, constituted a formidable challenge. During the course of our research, however, it became apparent that the tractability of the problem could be improved significantly if

we could find a way to preserve the original data type across the encryption and decryption transformations.

3. Proposed Solution

Cipher text bears roughly the same resemblance to plaintext as a hamburger does to a T-bone steak. A social security number, encrypted using the DES encryption algorithm, not only does not resemble a social security number but will likely not contain any numbers at all. A database field which was defined to hold a nine-character social security number would not be able to store the DES-encrypted version of the data.

A Visual Basic program would not read it. A graphical interface would not display it. There would be nothing that you could do with the encrypted social security number unless you had made extensive provisions for changes in data format throughout your application and physical database design.

3.1 Basic Data type Preservation

Our method reduces the need for changes to database structures and applications by preserving the data type of the encrypted field.

Data type preservation simply means that each cipher text field is as valid as the plaintext field it replaces. The key to our approach is defining an appropriate alphabet of valid characters and performing all operations within the constraints of the defined alphabet. Each different data type requires a judicious choice of alphabet. An alphabet consisting of numeric digits (“0123456789”) could be used to encrypt most number data, such as social security numbers (e.g. 234-415-6978). (The dashes, not included in the chosen alphabet, are copied unchanged to the corresponding positions in the cipher text output.)

Other alphabets, such as all printable ASCII characters, all characters shared by ASCII and EBCDIC, or all hexadecimal digits can be used to encode a variety of common data types.

3.2 The Approach

The first processing step involves replacing each plaintext character in the string by an integer that represents its position, or index, within the chosen alphabet. This number is between zero and one less than the total number of characters in the alphabet. If a plaintext character is not in the valid alphabet, it is copied to the output and removed from the string to be encrypted.

Example:

Plain text=“crazy”

Alphabet=“abcdefghijklmnopqrstuvwxyz”

- Step1: Assign Index values
Index values=7,4,11,14,11
Step2: Add position sensitive offsets
Offsets=10,4,18,25,8
New index values=12,9,6,9,17
Step3: Shuffle the index value string
Shuffled values=3,18,17,10,9
Step4: Convert Back to desired data type
Cipher text-“crazy”

Figure 1 includes a worked example of the basic algorithm.:.

After the alphabet index values have been assigned, we add a varying integer “offset” to each. We use modular addition to ensure that we generate only valid characters (i.e. characters which are contained in the alphabet). Remember that “modular” addition means adding two numbers and then determining the remainder after division by a constant “modulus” value. In the example above the alphabet size is 26 so, for example, $18 + 11 \pmod{26} = 3$. The actual offset values are generated based on a portion of the key being used to encrypt the data. This step ensures that long series of identical characters (such as 20 blanks at the end of a character field) will not encrypt identically. After adding the offsets, the entire string is shuffled. The shuffling method varies according to a permutation-invariant property of the index values, such as a sum or exclusive-or, of all values.2

The shuffling step helps to ensure that plaintexts with common prefixes or suffixes do not produce cipher text with common prefixes or suffixes. Once the encoding process is complete, each index value is mapped to the appropriate character in the alphabet. To recover the plaintext from the cipher text, one replaces the cipher text characters by their alphabet index values, “unshuffles” the string, regenerates the offset values, subtracts modularly on an integer-by-integer basis and substitutes the appropriate alphabet character.

Two enhancements to the above algorithms may be used to deal with certain data specific situations:

- First, in order to ensure that the encoded values of two single character strings with adjacent characters are not sequential (for example, we would not want “b” to encrypt as “y” whenever “a” encrypts as “x”), the alphabet itself can be shuffled based on a portion of the encryption key.
- Second, in order to inhibit guesses based on encrypted character

permutations, we can “ripple” the data from left to right and from right to left. This is done by hashing the key into a “starter-digit” and adding adjacent values pair wise. For example, the string of index values “1, 2, 3” might be rippled into “23, 5, 40” as follows (assuming a 55 character alphabet):

starter value = 72 (obtained by hashing the encryption key)

Adding left to right:

$$72 + 1 \pmod{55} = 18$$

$$16 + 4 \pmod{55} = 20$$

$$20 + 3 \pmod{55} = 23$$

Adding right to left:

$$23 + 72 \pmod{55} = 40$$

$$20 + 40 \pmod{55} = 5$$

$$18 + 5 \pmod{55} = 23$$

Applying this same method to the permutation “3, 2, 1,” on the other hand, ripples it to “27, 7, 40” and the fact that the two strings contain the same characters is disguised.

2 A variety of techniques could be used to generate the offsets and shuffling pattern, including the use of pseudo-random number generators.

3.3 Enhanced Encryption

While the encoding scheme presented above is sufficient to deter casual attacks, more substantial protection is required to protect sensitive data in the data warehouse. The approach described above can be combined with well-known encryption algorithms, such as DES or IDEA, to significantly increase the attacker’s burden. The basic idea is to use an established algorithm of known strength to produce the “offset” values.

The DES algorithm takes as input a 64-bit input block and a 64-bit key (56 key bits and 8 parity bits) and uses these two values to produce a 64-bit output. The cipher text output can be decrypted using the same key. For all practical purposes, the only way to break the scheme is by an exhaustive search of the key space.

DES, like any block cipher, can be operated as a stream cipher in “cipher-feedback” mode. We use this mode to encrypt one index value at a time. At the end of each encryption pass, we also shift the plaintext data into the DES input register. This process is illustrated as follows:

Let the alphabet index values of the n character, plaintext input string be represented

By $i_1 i_2 i_3 i_4 i_5 \dots i_n$ Let the 64-bit DES initial value required by cipher-feedback mode be constructed based on a portion of the encryption key

$$H(K) = a_1 a_2 a_3 a_4 a_5 a_6 a_7 a_8 = A$$

Where each subscripted “a” value represents an 8-bit number (“0” to “255”). Let the output of the DES algorithm, using a key of “k” and an input of “A,” be represented by

$$E_k(a_1 \ a_2 \ a_3 \ a_4 \ a_5 \ a_6 \ a_7 \ a_8) = b_1 \ b_2 \ b_3 \ b_4 \ b_5 \ b_6 \ b_7 \ b_8.$$

The first transformed index value is the modular sum $z_1 = b_8 + i_1 \pmod{l}$. Where “l” represents the alphabet length.

At this point, a new DES input value, A, is constructed as $A_2 = b_2 \ b_3 \ b_4 \ b_5 \ b_6 \ b_7 \ b_8 \ i_1$ and a new DES output is obtained

$$E_k(b_2 \ b_3 \ b_4 \ b_5 \ b_6 \ b_7 \ b_8 \ i_1) = c_1 \ c_2 \ c_3 \ c_4 \ c_5 \ c_6 \ c_7 \ c_8.$$

The second transformed index value is the modular sum $z_2 = c_8 + i_2 \pmod{l}$.

Note that the use of an addition operator is required, instead of the usual exclusive-or operator, is required to ensure that the data type is preserved.

After n such steps, during each of which a single input index value is transformed, we have an encrypted index-value string

$$Z = z_1 \ z_2 \ z_3 \ z_4 \ z_5 \dots z_n$$

We claim that recovering the string, $i_1 \ i_2 \ i_3 \ i_4 \ i_5 \dots i_n$, from the transformed string, $z_1 \ z_2 \ z_3 \ z_4 \ z_5 \dots z_n$, without knowledge of the key, K, is as difficult as breaking the DES algorithm itself. When using cipher-feedback mode, DES decryption, per se, is never invoked. Reversing the transformation is done by subtracting the low order DES output from the transformed index value. Below is a summary of the algorithm.

Setup responsibilities:

- 1) Select an encryption key with enough bits for the encryption algorithm key, encryption algorithm initial value and any basic processing stages.
- 2) Select a suitable alphabet to support the data type of the data to be encrypted.
- 3) Shuffle the alphabet according to a scheme based on the key. For each encrypted field:
- 4) Scan the input buffer for characters which are not included in the chosen alphabet.
- Move all invalid characters unchanged to their corresponding positions in the cipher text output buffer.
- 5) Move the index values of all valid characters to adjacent positions in a work buffer.
- 6) Add position-sensitive offsets according to a key-dependent scheme.
- 7) Scuffle the work buffer positions according to a data-dependent scheme.

8) “Ripple” the work buffer by calculating a key-based starter number and modularly adding pair wise from left to right then from right to left.

9) Set the cipher-feedback initial value using the chosen key.

10) Calculate the modular sum of the first work buffer position and the lowest order

DES output byte. Store this value in a second work buffer.

11) Obtain a new DES initial value by moving the DES output to the input, shifted one byte to the left, and shifting the work buffer value into the lowest order position.

12) Repeat steps 9 through 11 using successive work buffer index values until all of the data is transformed.

13) Replace the transformed index values by their corresponding character equivalents and store them in the open cipher text positions.

Decryption is accomplished by performing the inverse of each transformation in the reverse order.

4. Implementation Issues and Usage Constraints

Perhaps the most important *caveat* for anyone who wishes to implement our proposed encryption scheme is to guard against possible misinterpretation of encrypted data.

Scrambled text fields such as names and addresses are not likely to be mistaken for real information, but numeric fields may contain quite plausible values. A legitimate user who, through some administrative oversight, is erroneously presented with encrypted data may not recognize it as such and make bad decisions as a result. One approach to this shortcoming may be to include code in the query tool to fill in encrypted fields with a default value whenever the user has not been authenticated. Another approach may be simply to restrict the application of the technique to text fields. A revenue field may be quite useless without the corresponding product data or sales region information.

Another restriction on the use of this technique is that decryption must be performed before aggregate functions, such as minimum, maximum, sum, and average, are applied. This is not a serious inconvenience in the data warehousing environment because precompiled summary tables are usually available.

One must also bear in mind that this encryption scheme is *consistent* in that the same plaintext always results in the same cipher text. This has both positive and negative implications. On the positive side, the consistency of the encrypted

data allows for relational joins and *blind keys* (described later). On the other hand, consistent encryption exposes the data to the possibility of a statistical attack. If an attacker knows the relative frequency of specific data items, such as medical tests, he can deduce the corresponding encrypted values. This kind of attack can be stymied by using a value from another field (the table's primary key, for example) to modify the encryption key. This would, of course, preclude the use of this data in relational join predicates.

5. Co-existence with Other Security Controls

We do not propose data type-preserving encryption as the ultimate solution to all data warehouse security concerns. It is presented, rather, as one of several mechanisms to be employed in a more comprehensive security strategy. Specifically, we see our technique as a *containment* device which limits potential damage in the event of a successful bypass of other security controls. In general, there are at least five categories of security controls:

Prevention- Preventative measures include anything which can be done to prevent an attack or to keep it from succeeding. This includes strengthening vulnerabilities and providing disincentives to the attacker.

Detection- Detective measures include anything which alerts the support staff to the fact that an attack is in progress or has been recently attempted.

Containment- Containment measures include anything which can serve to limit the damage of a successful attack.

Recovery- Recovery measures include anything which is done to restore normal operation and user access after an unscheduled interruption.

Investigation- Investigative measures include anything which is done to identify a malefactor and collect evidence which will be used in a disciplinary process or criminal prosecution. A good data warehouse security plan will include multiple countermeasures for each identified threat ideally, at least one from each of these categories.

6. Applications to Other Areas

In addition to data warehouse security, there may be several other areas in which this technique may prove useful, such as providing an additional check on data integrity.

By adding a check character to the beginning of each plaintext field any alteration would be immediately obvious during the decryption process. The decryption routine could be

modified to perform this check and return an error code if tampering is suspected.

Another possible application is in the use of *blind keys*. In certain situations, one needs to know that two quantities are equal without actually knowing the quantities themselves.

One may wish to match bank account numbers from multiple sources, for example, in credit check applications but not use the actual numbers themselves because of the potential for fraudulent activity.

It may also be possible to control access to commercially available data through the use of this technique. A master database could be distributed to subscribers with individually licensed components encrypted using different keys. Access to the individual components could be made available by distributing keys following payment of the proper license fees.

Acknowledgement: I really thankful our Chairman Madam Smt.R.Usharani, Director Sri R.RajiReddy and my kids Mr.M.Crazy and Kum M.Lucky who encouraged lot directly or indirectly to present this research paper

References

- [1] H. Akaike. On entropy maximization principle. *Applications of Statistics*, pages 27{41, 1977.
- [2] M. R. Anderberg. *Cluster Analysis for Application*. Academic Press, 1973.
- [3] P. S. Bradley, U. Fayyad, and C. Reina. Scaling clustering algorithms to large databases. In *Proc. 4th International Conf. on Knowledge Discovery and Data Mining (KDD-98)*. AAAI Press, August 1998.
- [4] I. P. Felligi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Society*, 64:1183{1210, 1969.
- [5] J. H. Friedman, J. L. Bentley, and R. A. Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Software*, 3(3):209{226, 1977.
- [6] C. L. Giles, K. D. Bollacker, and S. Lawrence. CiteSeer: An automatic citation indexing system. In *Digital Libraries 98 / Third ACM Conference on Digital Libraries*, 1998.
- [7] M. Hernandez and S. Stolfo. The merge/purge problem for large databases. In *Proceedings of the 1995 ACM SIGMOD*, May 1995.
- [8] H. Hirsh. Integrating mulitple sources of information in text classification using whril. In *Snowbird Learning Conference*, April 2000.
- [9] J. Hylton. Identifying and merging related bibliographic records. MIT LCS Masters Thesis, 1996.
- [10] B. Kilss and W. Alvey, editors. *Record Linkage Techniques/1985*, 1985. Statistics of Income Division, Internal Revenue Service Publication 1299-2-96.
Available from <http://www.fccm.gov/>.
- [11] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. Automating the construction of internet

- portals with machine learning. *Information Retrieval*, 2000. To appear.
- [12] A. K. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/_mccallum/bow, 1996.
- [13] A. Monge and C. Elkan. The _eld-matching problem: algorithm and applications. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, August 1996.
- [14] A. Monge and C. Elkan. An e_cient domain-independent algorithm for detecting approximately duplicate database records. In *The proceedings of the SIGMOD 1997 workshop on data mining and knowledge discovery*, May 1997.
- [15] A. Moore. Very fast EM-based mixture model clustering using multiresolution kd-trees. In *Advances in Neural Information Processing Systems 11*, 1999.
- [16] H. B. Newcombe, J. M. Kennedy, S. J. Axford, and A. P. James. Automatic linkage of vital records. *Science*, 130:954{959, 1959.
- [17] S. Omohundro. Five balltree construction algorithms. Technical report 89-063, International Computer Science Institute, Berkeley, California, 1989.
- [18] K. Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE*, 86(11):2210(2239, 1998.
- [19] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513{523, 1988.
- [20] M. Sankaran, S. Suresh, M. Wong, and D. Nesamoney. Method for incremental aggregation of dynamically increasing database data sets. *U.S. Patent 5,794,246*, 1998.
- [21] D. Sankar and J. B. Kruskal. *Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley, 1983.
- [22] J. W. Tukey and J. O. Pedersen. Method and apparatus for information access employing overlapping clusters. *U.S. Patent 5,787,422*, 1998.
- [23] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH:An e_cient data clustering method for very large databases. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, pages 103{114, 1996.

QEMPAR: QoS and Energy Aware Multi-Path Routing Algorithm for Real-Time Applications in Wireless Sensor Networks

Saeed Rasouli Heikalabad¹, Hossein Rasouli², Farhad Nematy² and Naeim Rahmani²

¹ Technical and Engineering Department, Islamic Azad University - Tabriz Branch
Tabriz, East Azerbaijan, Iran

² Technical and Engineering Department, Islamic Azad University - Tabriz Branch
Tabriz, East Azerbaijan, Iran

Abstract

Enabling real time applications in wireless sensor networks requires certain delay and bandwidth which pose more challenges in the design of routing protocols. The algorithm that is used for packet routing in such applications should be able to establish a tradeoff between end to end delay parameter and energy consumption. In this paper, we propose a new multi path routing algorithm for real time applications in wireless sensor networks namely QEMPAR which is QoS aware and can increase the network lifetime. Simulation results show that the proposed algorithm is more efficient than previous algorithms in providing quality of service requirements of real-time applications.

Keywords: *Wireless Sensor Network, Real-Time Application, Multi-Path Routing, Quality of Service, Energy Efficiency.*

1. Introduction

In the recent years, the rapid advances in micro-electro-mechanical systems, low power and highly integrated digital electronics, small scale energy supplies, tiny microprocessors, and low power radio technologies have created low power, low cost and multifunctional wireless sensor devices, which can observe and react to changes in physical phenomena of their environments. These sensor devices are equipped with a small battery, a tiny microprocessor, a radio transceiver, and a set of transducers that used to gathering information that report the changes in the environment of the sensor node. The emergence of these low cost and small size wireless sensor devices has motivated intensive research in the last decade addressing the potential of collaboration among sensors in data gathering and processing, which led to the creation of Wireless Sensor Networks (WSNs).

A typical WSN consists of a number of sensor devices that collaborate with each other to accomplish a common task (e.g. environment monitoring, target tracking, etc) and

report the collected data through wireless interface to a base station or sink node. The areas of applications of WSNs vary from civil, healthcare and environmental to military. Examples of applications include target tracking in battlefields [1], habitat monitoring [2], civil structure monitoring [3], forest fire detection [4], and factory maintenance [5].

However, with the specific consideration of the unique properties of sensor networks such limited power, stringent bandwidth, dynamic topology (due to nodes failures or even physical mobility), high network density and large scale deployments have caused many challenges in the design and management of sensor networks. These challenges have demanded energy awareness and robust protocol designs at all layers of the networking protocol stack [6].

Efficient utilization of sensor's energy resources and maximizing the network lifetime were and still are the main design considerations for the most proposed protocols and algorithms for sensor networks and have dominated most of the research in this area. The concepts of latency, throughput and packet loss have not yet gained a great focus from the research community. However, depending on the type of application, the generated sensory data normally have different attributes, where it may contain delay sensitive and reliability demanding data. For example, the data generated by a sensor network that monitors the temperature in a normal weather monitoring station are not required to be received by the sink node within certain time limits. On the other hand, for a sensor network that used for fire detection in a forest, any sensed data that carries an indication of a fire should be reported to the processing center within certain time limits. Furthermore, the introduction of multimedia sensor networks along with the increasing interest in real time applications have made strict constraints on both throughput and delay in order to report the time-critical

data to the sink within certain time limits and bandwidth requirements without any loss. These performance metrics (i.e. delay, energy consumption and bandwidth) are usually referred to as Quality of Service (QoS) requirements [7]. Therefore, enabling many applications in sensor networks requires energy and QoS awareness in different layers of the protocol stack in order to have efficient utilization of the network resources and effective access to sensors readings. Thus QoS routing is an important topic in sensor networks research, and it has been under the focus of the research community of WSNs. Authors of [7] and [8] have surveyed the QoS based routing protocol in WSNs.

Many routing mechanisms specifically designed for WSNs have been proposed [9][10]. In these works, the unique properties of the WSNs have been taken into account. These routing techniques can be classified according to the protocol operation into negotiation based, query based, QoS based, and multi-path based. The negotiation based protocols have the objective to eliminate the redundant data by include high level data descriptors in the message exchange. In query based protocols, the sink node initiates the communication by broadcasting a query for data over the network. The QoS based protocols allow sensor nodes to make a tradeoff between the energy consumption and some QoS metrics before delivering the data to the sink node [11]. Finally, multi-path routing protocols use multiple paths rather than a single path in order to improve the network performance in terms of reliability and robustness. Multi-path routing establishes multiple paths between the source-destination pair. Multi-path routing protocols have been discussed in the literature for several years now [12]. Mutli-path routing has focused on the use of multiple paths primarily for load balancing, fault tolerance, bandwidth aggregation, and reduced delay. We focus to guarantee the required quality of service through multi-path routing.

The rest of the paper organized as follows: in section 2, we explain the related works. Section 3 describes the proposed algorithm with detailed. Section 4 explore the simulation parameters and result analysis. Final section is containing of conclusion and future works.

2. Related Works

QoS-based routing in sensor networks is a challenging problem because of the scarce resources of a sensor node. Thus, this problem has received a significant attention from the research community, where many works are being made. Some QoS oriented routing works are surveyed in [7] and [8]. In this section we do not give a comprehensive summary of the related work, instead we

present and discuss some works related to proposed protocol.

One of the early proposed routing protocols that provide some QoS is the Sequential Assignment Routing (SAR) protocol [13]. SAR protocol is a multi-path routing protocol that makes routing decisions based on three factors: energy resources, QoS on each path, and packet's priority level. Multiple paths are created by building a tree rooted at the source to the destination. During construction of paths those nodes which have low QoS and low residual energy are avoided. Upon the construction of the tree, most of the nodes will belong to multiple paths. To transmit data to sink, SAR computes a weighted QoS metric as a product of the additive QoS metric and a weighted coefficient associated with the priority level of the packet to select a path. Employing multiple paths increases fault tolerance, but SAR protocol suffers from the overhead of maintaining routing tables and QoS metrics at each sensor node.

K. Akkaya and M. Younis in [14] proposed a cluster based QoS aware routing protocol that employs a queuing model to handle both real-time and non real time traffic. The protocol only considers the end-to-end delay. The protocol associates a cost function with each link and uses the K-least-cost path algorithm to find a set of the best candidate routes. Each of the routes is checked against the end-to-end constraints and the route that satisfies the constraints is chosen to send the data to the sink. All nodes initially are assigned the same bandwidth ratio which makes constraints on other nodes which require higher bandwidth ratio. Furthermore, the transmission delay is not considered in the estimation of the end-to-end delay, which sometimes results in selecting routes that do not meet the required end-to-end delay. However, the problem of bandwidth assignment is solved in [15] by assigning a different bandwidth ratio for each type of traffic for each node.

SPEED [16] is another QoS based routing protocol that provides soft real-time end-to-end guarantees. Each sensor node maintains information about its neighbors and exploits geographic forwarding to find the paths. To ensure packet delivery within the required time limits, SPEED enables the application to compute the end-to-end delay by dividing the distance to the sink by the speed of packet delivery before making any admission decision. Furthermore, SPEED can provide congestion avoidance when the network is congested.

However, while SPEED has been compared with other protocols and it has showed less energy consumption than other protocols, this does not mean that SPEED is energy efficient, because the protocols used in the comparison are

not energy aware. SPEED does not consider any energy metric in its routing protocol, which makes a question about its energy efficiency. Therefore to better study the energy efficiency of the SPEED protocol; it should be compared with energy aware routing protocols.

Felemban et al. [17] propose Multi-path and Multi-Speed Routing Protocol (MMSPEED) for probabilistic QoS guarantee in WSNs. Multiple QoS levels are provided in the timeliness domain by using different delivery speeds, while various requirements are supported by probabilistic multipath forwarding in the reliability domain.

Recently, X. Huang and Y. Fang have proposed multi constrained QoS multi-path routing (MCMP) protocol [18] that uses braided routes to deliver packets to the sink node according to certain QoS requirements expressed in terms of reliability and delay. The problem of the end-to-end delay is formulated as an optimization problem, and then an algorithm based on linear integer programming is applied to solve the problem. The protocol objective is to utilize the multiple paths to augment network performance with moderate energy cost. However, the protocol always routes the information over the path that includes minimum number of hops to satisfy the required QoS, which leads in some cases to more energy consumption. Authors in [19], have proposed the Energy constrained multi-path routing (ECMP) that extends the MCMP protocol by formulating the QoS routing problem as an energy optimization problem constrained by reliability, playback delay, and geo-spatial path selection constraints. The ECMP protocol trades between minimum number of hops and minimum energy by selecting the path that satisfies the QoS requirements and minimizes energy consumption.

Meeting QoS requirements in WSNs introduces certain overhead into routing protocols in terms of energy consumption, intensive computations, and significantly large storage. This overhead is unavoidable for those applications that need certain delay and bandwidth requirements. In our work, we combine different ideas from the previous protocols in order to optimally tackle the problem of QoS in sensor networks. In our proposal we try to satisfy the QoS requirements for real time applications with the minimum energy. Our QEMPAR routing protocol performs paths discovery using multiple criteria such as energy remaining, probability of packet sending, average probability of packet receiving and interference.

3. Proposed Protocol

In this section, we explain the assumptions and energy consumption model used in QEMPAR and describe the various constituent parts of the proposed protocol.

3.1 Assumptions

We assume that all nodes are randomly distributed in desired environment and each of them is assigned a unique ID. At start, the initial energy of nodes is considered equal. All nodes in the network are aware of their location (by GPS) and also are able to control their energy consumption. Because of this assumption has been that the nodes can communicate with other nodes outside their radio range in the absence of node in their radio transmission range.

Let us assume that nodes are aware of their remaining energy and also remaining energy of other nodes in their transmission radio range. We consider that each node can calculate its probabilities of packet sending and packet receiving with regard to link quality. Predications and decisions about path stability may be made by examining recent link quality information.

3.2 Energy Consumption Model

In QEMPAR, energy model is obtained from [20] that use both of the open space (energy dissipation d^2) and multi path (energy dissipation d^4) channels by taking amount the distance between the transmitter and receiver. So energy consumption for transmitting a packet of l bits in distance d is given by Eq. (1).

$$E_{Tx}(l,d) = \begin{cases} lE_{elec} + l\varepsilon_{fs} d^2 & , d \leq d_0 \\ lE_{elec} + l\varepsilon_{mp} d^4 & , d > d_0 \end{cases} \quad (1)$$

In here d_0 is the distance threshold value which is obtained by Eq. (2), E_{elec} is required energy for activating the electronic circuits. ε_{fs} and ε_{mp} are required energy for amplification of transmitted signals to transmit a one bit in open space and multi path models, respectively.

$$d_0 = \sqrt{\frac{\varepsilon_{fs}}{\varepsilon_{mp}}} \quad (2)$$

Energy consumption to receive a packet of l bits is calculated according to Eq. (3).

$$E_{Rx}(l) = lE_{elec} \cdot \quad (3)$$

3.3 Link Suitability

The link suitability is used by the node to select the node at the next hop as a forwarder during the path discovery phase. Let N_A be a set of neighbors of node A. Then our suitability function includes the PPS (Probability of Packet Sending), APPR (Average Probability of Packet Receiving) and I_B (Interference of link A and B) and obtained by Eq. (4).

$$N_H = \max_{B \in N_A} \{PPS_B + APPR_{N_B} + 1/I_B + \frac{E_r - B}{E_i}\}. \quad (4)$$

In here, N_H is the selected node at the next hop and B is the node at the next hop. PPS_B is the probability of packet sending of node B . Each node calculates the value of this parameter by Eq. (5). $APPR_{N_B}$ is the average probability of packet receiving of all neighbors of node B that obtained by Eq. (6). I_B is interference of link between A and B. In this paper, I_B is same signal to noise ratio (SNR) for the link between A and B.

$$PPS = \frac{\text{Number-of-Successful-Sending-Packets}}{\text{Total-Number-of-Sending-Packets}}. \quad (5)$$

$$APPR_{N_B} = \sum_{j=1}^{N_B} PPR_j. \quad (6)$$

The total merit (TM) for a path p consists of a set of K nodes is the sum of the individual link merit $l_{(AB)}$ along the path. Then the total merit is calculated by Eq. (7).

$$TM_p = \sum_{i=1}^{K-1} l_{(AB)_i}. \quad (7)$$

3.4 Paths Discovery Mechanism in QEMPAR

In multi-path routing, node-disjoint paths (i.e. have no common nodes except the source and the destination) are usually preferred because they utilize the most available network resources, and hence are the most fault-tolerant. If an intermediate node in a set of node-disjoint paths fails, only the path containing it node is affected, so there is a minimum impact to the diversity of the routes.

In first phase of path discovery procedure, each node collects the needed information about its neighbors by beacon exchange between them and then updates its neighboring table.

After this phase, each sensor node has enough information to compute the link suitability for its neighboring nodes.

3.5 Paths Assortment

After the execution of paths discovery phase and the paths have been constructed, we need to break a provided real time packet to few smaller packets, with sequence numbers assigned to each of them, in order to packet fast sending and consequently end to end delay decreasing. For this purpose, source node assortments the all paths according to hop counts of them in several classes. Then source node sends each tiny packet through separate paths. The tiny packet which its sequence number is 1 is sent through the path that has the least number of hops. Then other tiny packets with subsequent number according to the tiny packet number from packet number 2 to end through the paths with minimum hop count to maximum hop count. Because the sink to receive tiny packets consecutively. Fig.1 shows these operations.

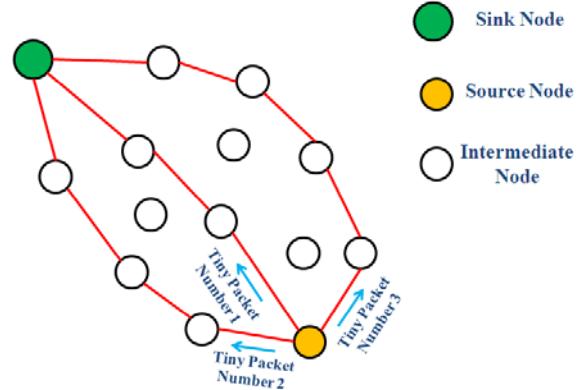


Fig. 1 Tiny packets sending through different paths

4. Simulation and Performance Evaluation

In this section, we present and discuss the simulation results for the performance study of QEMPAR protocol. We used GCC to implement and simulate QEMPAR and compare it with the MCMP protocol [18]. Simulation parameters are presented in Table 1 and obtained results are shown below.

The radio model used in the simulation was a duplex transceiver. The network stack of each node consists of IEEE 802.11 MAC layer with 40 meter transmission range.

We assume that the source node is located at (300, 300) meters.

Table 1: Simulation parameters

| Parameters | Value |
|-----------------------|------------------------------|
| Network area | 400 meters × 400 meters |
| Base station location | (0, 0)m |
| Number of sensors | 100 |
| Initial energy | 2J |
| E_{elec} | 50 nJ/bit |
| ϵ_{fs} | 10 pJ/bit/m ² |
| ϵ_{mp} | 0.0013 pJ/bit/m ⁴ |
| d_0 | 87 m |
| E_{DA} | 5 nJ/bit/signal |
| Data packet size | 512 bytes |

We investigate the performance of the QEMPAR protocol in a multi-hop network topology. We study the impact of changing the packet arrival rate on end-to-end delay, packet delivery ratio, and energy consumption. We change the real-time packet arrival rate at the source node from 5 to 50 packets/sec.

4.1 Average End-to-End Delay

The average end-to-end delay is the time required to transfer data successfully from source node to the destination node.

Fig. 2 shows the average end to end delay for QEMPAR and MCMP. In this evaluation, we change the packet arrival rate at the source node, and measure the delay.

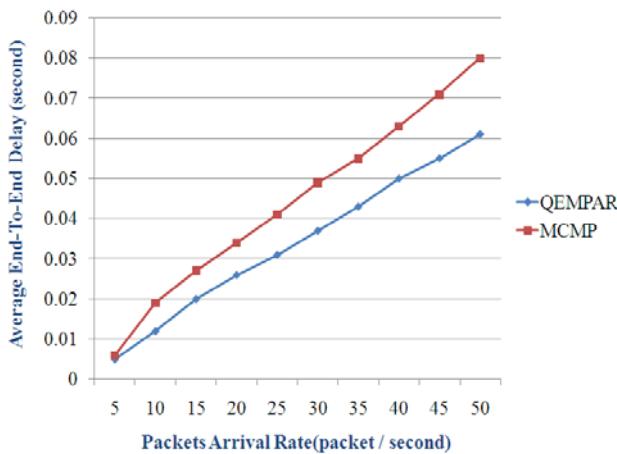


Fig. 2 Average end to end delay

As it can be seen, proposed protocol has performance better than MCMP in average end to end delay.

4.2 Average Energy Consumption

The average energy consumption is the average of the energy consumed by the nodes participating in message transfer from source node to the destination node.

Fig. 3 shows the results for energy consumption in two protocols. As it can be seen, in our protocol, energy consumption for packet sending is some deal optimize in comparison to the MCMP.

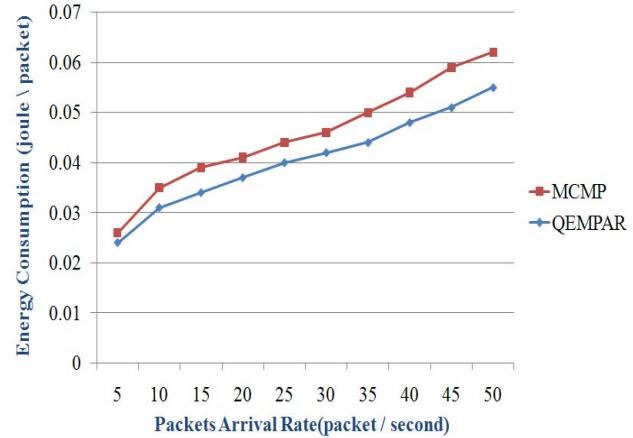


Fig. 3 Average energy consumption

5. Conclusion

In this paper, we propose the new multi path routing algorithm for real time applications in wireless sensor network namely QEMPAR which is QoS aware and can increase the network lifetime. Our protocol uses four main metrics of QoS with special relation in path discovery mechanism. Simulation Result shows that the performance of QEMPAR in end to end delay is optimized compared to the MCMP protocol.

References

- [1] T. Bokareva, W. Hu, S. Kanhere, B. Ristic, N. Gordon, T. Bessell, M. Rutten and S. Jha, "Wireless Sensor Networks for Battlefield Surveillance", In roceedings of The Land Warfare Conference (LWC)- October 24 – 27, 2006, Brisbane, Australia.
- [2] A. Mainwaring, J. Polastre, R. Szewczyk, D. Culler, and J. Anderson, "Wireless Sensor Networks for Habitat Monitoring," in the Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications (ACM-WSNA), Pages: 88-97, September 28 - 28, 2002, Atlanta, Georgia, USA.
- [3] N. Xu, S. Rangwala, K. Chintalapudi, D. Ganesan, A. Broad, R. Govindan, and D. Estrin, "A Wireless Sensor Network for structural Monitoring," in Proc. ACM SenSys Conf., Nov.2004.
- [4] M. Hefeeda, M. Bagheri, "Wireless Sensor Networks for Early Detection of Forest Fires", in the proceedings of IEEE Internatonal Conference on Mobile Adhoc and Sensor Systems, 2007. MASS 2007. Volume , Issue , 8-11 Oct. 2007 Page(s):1 – 6, Pisa, Italy.
- [5] K. Srinivasan, M. Ndoh, H. Nie, H. Xia, K. Kaluri, and D. Ingraham, "Wireless Technologies for Condition-Based

- Maintenance (CBM) in Petroleum Plants,” Proc. of DCOSS’05 (Poster Session), 2005.
- [6] Bashir Yahya, Jalel Ben-Othman, “Towards a classification of energy aware MAC protocols for wireless sensor networks”, Journal of Wireless Communications and Mobile Computing, Wiley.
 - [7] Kemal Akkaya, Mohamed Younis, “A Survey on Routing for Wireless Sensor Networks”, Journal of Ad Hoc Networks, Volume, 3, Pages: 325- 349, 2005.
 - [8] D. Chen, P.K. Varshney, “QoS Support in Wireless Sensor Networks: a Survey”, In the Proceedings of the International Conference on Wireless Networks (ICWN), 2004, pp 227-233.
 - [9] Jamal N. Al-Karaki, Ahmed E. Kamal, “Routing Techniques in Wireless Sensor Networks: A Survey”, IEEE Journal of Wireless Communications, Volume 11, Issue 6, Dec. 2004 Page(s): 6 – 28.
 - [10] Anahit Martirosyan, Azzedine Boukerche, Richard Werner Nelem Pazzi: A Taxonomy of Cluster-Based Routing Protocols for Wireless Sensor Networks. ISSPAN 2008: 247-253.
 - [11] Anahit Martirosyan, Azzedine Boukerche, Richard Werner Nelem Pazzi: Energy-aware and quality of service-based routing in wireless sensor networks and vehicular ad hoc networks. Annales des Telecommunications 63(11-12): 669-681 (2008).
 - [12] Jack Tsai, Tim Moors, “A Review of Multipath Routing Protocols: From Wireless Ad Hoc to Mesh Networks”, Proc. ACoRN Early Career Researcher Workshop on Wireless Multihop Networking, Jul. 17-18, 2006.
 - [13] K. Sohrabi, J. Pottie, “Protocols for self-organization of a wireless sensor network”, IEEE Personal Communications, Volume 7, Issue 5, pp 16-27, 2000.
 - [14] K. Akkaya, M. Younis, “An energy aware QoS routing protocol for wireless sensor networks”, In the Proceedings of the MWN, Providence, May 2003. pp 710-715.
 - [15] M. Younis, M. Youssef, K. Arisha, “Energy aware routing in cluster based sensor networks”, In the proceedings of the 10th IEEE international symposium on modeling, analysis, and simulation of computer and telecommunication systems (MASCOTS-2002), Fort Worth, 11-16 October 2002.
 - [16] T. He et al., “SPEED: A stateless protocol for real-time communication in sensor networks,” in the Proceedings of the Internation Conference on Distributed Computing Systems, Providence, RI, May 2003.
 - [17] E. Felemban, C.-G. Lee, and E. Ekici, “MMSPEED: multipath multispeed protocol for QoS guarantee of reliability and timeliness in wireless sensor networks,” IEEE Trans. on Mobile Computing, vol. 5, no. 6, pp. 738–754, Jun 2006.
 - [18] X. Huang, Y. Fang, “Multiconstrained QoS Multipath Routing in Wireless Sensor Networks,” Wireless Networks (2008) 14:465-478.
 - [19] A. B. Bagula, K. G. Mazandu, “Energy Constrained Multipath Routing in Wireless Sensor Networks”, UIC 2008, LNCS 5061, pp 453-467, 2008.
 - [20] W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, “Energy-Efficient Communication Protocol for Wireless Microsensor Networks,” Proceedings of the Hawaii International Conference on System Sciences, 2000.

DPCC: Dynamic Predictive Congestion Control in Wireless Sensor Networks

Saeed Rasouli Heikalabad¹, Ali Ghaffari², Mir Abolgasem Hadian² and Hossein Rasouli²

¹ Technical and Engineering Department, Islamic Azad University - Tabriz Branch
Tabriz, East Azerbaijan, Iran

² Technical and Engineering Department, Islamic Azad University - Tabriz Branch
Tabriz, East Azerbaijan, Iran

Abstract

Congestion occurs in wireless sensor networks (WSNs) when nodes are densely distributed, and/or the application produces high flow rate near the sink due to the convergent nature of upstream traffic. Congestion can lead to packet losses, delay, and energy waste due to a large number of packet drops and retransmissions. Therefore it is necessary to carry out congestion control which detects congestion precisely and regulates it fairly. To achieve this objective, a dynamic predictive congestion control (DPCC) algorithm is proposed in this paper. The DPCC can predict congestion in a node and will broadcast traffic on the entire network fairly and dynamically. Simulation results show that the proposed protocol is more efficient than previous ones.

Keywords: Wireless Sensor Network, Congestion Control, Predictive, Fairness.

1. Introduction

In the recent years, the rapid advances in micro-electro-mechanical systems, low power and highly integrated digital electronics, small scale energy supplies, tiny microprocessors, and low power radio technologies have created low power, low cost and multifunctional wireless sensor devices, which can observe and react to changes in physical phenomena of their environments. These sensor devices are equipped with a small battery, a tiny microprocessor, a radio transceiver, and a set of transducers that used to gathering information that report the changes in the environment of the sensor node. The emergence of these low cost and small size wireless sensor devices has motivated intensive research in the last decade addressing the potential of collaboration among sensors in data gathering and processing, which led to the creation of Wireless Sensor Networks (WSNs).

A typical WSN consists of a number of sensor devices that collaborate with each other to accomplish a common task (e.g. environment monitoring, target tracking, etc) and report the collected data through wireless interface to a

base station or sink node. The areas of applications of WSNs vary from civil, healthcare and environmental to military. Examples of applications include target tracking in battlefields [1], habitat monitoring [2], civil structure monitoring [3], forest fire detection [4], and factory maintenance [5].

However, with the specific consideration of the unique properties of sensor networks such limited power, stringent bandwidth, dynamic topology (due to nodes failures or even physical mobility), high network density and large scale deployments have caused many challenges in the design and management of sensor networks. These challenges have demanded energy awareness and robust protocol designs at all layers of the networking protocol stack [6].

The upstream traffic from sensor nodes to the sink is many-to-one multi-hop convergent. Fig. 1 shows many-to-one traffic pattern. The upstream traffic can be classified into four delivery models: event-based, continuous, query-based, and hybrid delivery. Due to the convergent nature of upstream traffic, congestion more probably appears in the upstream direction. Congestion that can leads to packet losses and increased transmission latency has direct impact on energy-efficiency and application QoS, and therefore must be efficiently controlled. Congestion control generally follows three steps: congestion detection, congestion notification, and rate-adjusting.

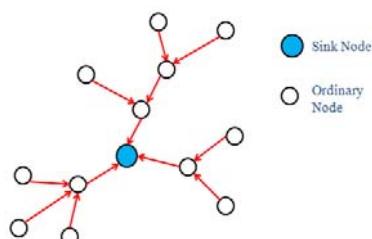


Fig. 1 Many-to-one traffic pattern in wireless sensor networks

In response to congestion, a rate adjustment mechanism must be designed and implemented properly in order to eliminate or avoid congestion. A number of different schemes were reported in the literature in last few years. A stop-and-start and hop-by-hop strategy is employed in [7]. In [8] and [11], an end-to-end and AIMD-like (Additive Increase Multiplicative Decrease) rate adjustment approach is employed. All of these mechanisms, however, aim at guaranteeing the simple fairness instead of the weighted fairness. A priority-based congestion control protocol (PCCP) is presented to achieve the weighted-fairness transmission for single-path routing WSNs in [14]. In this paper we introduce a priority-based rate adjustment algorithm called joint priority algorithm (JPA), which guarantees weighted fairness in multipath routing WSNs. In this scheme, intermediate nodes keep a record of the information on joint priorities (JP) of their neighbors. When congestion is detected, the sending rates of the upstream neighbors of the congested node are limited based on their joint priorities. In other words, upstream neighbors with important traffic will share more bandwidth than others when congestion occurs. Each data source, however, will send packets with its current equal rate when there is no congestion.

Two types of congestion could occur in WSNs [9] (As can be seen in Fig. 2). The first type is node-level congestion that is common in conventional networks. It is caused by buffer overflow in the node and can result in packet loss, and increased queuing delay. Packet loss in turn can lead to retransmission and therefore consumes additional energy. For WSNs where wireless channels are shared by several nodes using CSMA like (Carrier Sense Multiple Access) protocols, collisions could occur when multiple active sensor nodes try to seize the channel at the same time. This can be referred to as link level congestion. Link-level congestion increases packet service time, and decreases both link utilization and overall throughput, and wastes energy at the sensor nodes. Both node-level and link-level congestions have direct impact on energy-efficiency and QoS.

Congestion control protocol efficiency depends on how much it can achieve the following performance objectives: (i) First, energy-efficiency requires to be improved in order to extend system lifetime. Therefore congestion control protocols need to avoid or reduce packet loss due to buffer overflow, and remain lower control overhead that will consume additional energy more or less. (ii) Second, fairness needs to be observed so that each node can achieve fair throughput. Fairness can be achieved through rate-adjustment and packet scheduling (otherwise referred to as queue management) at each sensor node. (iii) Furthermore, support of traditional quality of service (QoS)

metrics such as packet loss ratio and packet delay along with throughput may also be necessary.

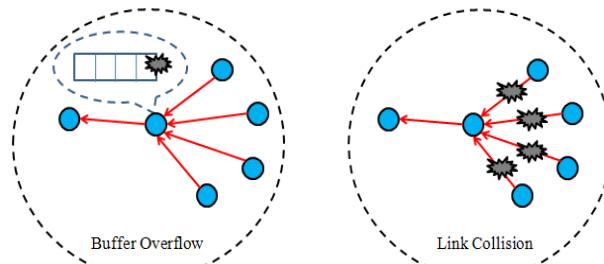


Fig. 2 Congestion in wireless sensor networks

The rest of the paper organized as follows: in section 2, we explain the related works. Section 3 describes the system models. Section 4 explores the DPCC protocol with details. Section 5 describes simulation parameters and result analysis. Final section is containing of conclusion and future works.

2. Related Works

There are several congestion control protocols [9]-[12] for sensor networks. They differ in the way that they detect congestion, broadcast congestion related information, and the way that they adjust traffic rate when congestion occurs. In this section, we review some of them and discuss their limitations.

Congestion detection and avoidance (CODA) [11] proposes an open-loop, hop-by-hop backpressure mechanism and a closed-loop, multi-source regulation mechanism in event-driven WSNs. Sensor nodes detect congestion by monitoring the channel utilization and buffer-occupancy level. In response to congestion, the congested sensor nodes send backpressure messages to their neighbors which may drop packets, reduce their sending rate and further propagate backpressure messages. If the sending rate of a source node is greater than the preset threshold, the source node must receive a continuous stream of ACKs from the base station in order to maintain that rate. By this means, the base station may limit the sending rate of a source node based on deciding how many ACKs to broadcast. CODA employs the AIMD (Additive Increase Multiplicative Decrease) coarse rate adjustment. It only guarantees simple fairness of the congestion control.

Event-to-sink reliable transport protocol (ESRT) [13] monitors the local buffer level in intermediate sensor nodes and sets a congestion notification bit in the packet when the buffer overflows. If a base station receives a packet whose congestion notification bit is set, it

broadcasts a control signal to inform all source nodes to reduce the sending rate according to certain proportion. ESRT limits sending rate of all source nodes when congestion occurs regardless of where the hot spot happens in WSNs. The best way is to regulate those source nodes that are responsible for this congestion.

Priority based congestion control protocol (PCCP) [14] defines a new variable, congestion degree as ratio of average packet service time over average packet inter-arrival time at each sensor node. Congestion degree is intended to reflect the current congestion level of each sensor node. Based on congestion degree, PCCP employs a hop-by-hop rate adjustment technique called priority-based rate adjustment (PRA) to adjust the scheduling rate and the source rate of each sensor node in a single-path routing WSN. In the tree-based network topology of single-path routing WSNs, a sensor node will only have one downstream neighbor, but it may have multiple upstream neighbors. The whole data flow generated by a source node will pass through the nodes and links along with the single routing path. Sensor nodes learn the number of upstream data sources in the sub tree roots and measure the maximum downstream forwarding rate. Finally, they calculate the per-source rate based on priority index of each source node.

In Fusion [15], congestion is detected in each sensor node based on measurement of queue length. The node that detects congestion sets a congestion notification (CN) bit in the header of each outgoing packet. Once the CN bit is set, neighboring nodes can overhear it and stop forwarding packets to the congested node so that it can drain the backlogged packets. This non-smooth rate adjustment could impair link utilization as well as fairness, although Fusion has a mechanism to limit the source traffic rate and a prioritized MAC algorithm to improve fairness.

Adaptive Rate Control (ARC), [12], is an LIMD-like (linear increase and multiplicative decrease) algorithm. In ARC, if an intermediate node overhears that the packets it sent previously are successfully forwarded again by its parent node, it will increase its rate by a constant α . Otherwise it will multiply its rate by a factor β where $0 < \beta < 1$. ARC does not use explicit congestion detection or explicit congestion notification and therefore avoids use of control messages. However the coarse rate adjustment could result in tardy control and introduce packet loss.

CCF (Congestion Control and Fairness) [9] uses packet service time to deduce the available service rate and therefore detects congestion in each intermediate sensor node. Congestion information, that is packet service time in CCF, is implicitly reported. CCF controls congestion in a hop-by-hop manner and each node uses exact rate

adjustment based on its available service rate and child node number. CCF guarantees simple fairness. That means each node receives the same throughput. However the rate adjustment in CCF relies only on packet service time which could lead to low utilization when some sensor nodes do not have enough traffic or there is a significant packet error rate (PER).

Those existing congestion control protocols for WSNs have limitations. For example, they only guarantee simple fairness, which means that the sink receives the same throughput from all nodes. However, sensor nodes may have different priority or importance due to either their functions or the location at which they are deployed.

3. System Models

This section describes network and node models, as shown in Figs. 3 and 4, respectively.

3.1 Network Model

This paper addresses upstream congestion control for a WSN that supports single-path routing. The network model to be investigated in this work is depicted in Fig. 3, where sensor nodes are supposed to generate continuous data and form many-to-one convergent traffic in the upstream direction. CSMA/CA MAC protocol is implemented in MAC layer. Each sensor node could have two types of traffic: source and transit. The source traffic is locally generated at each sensor node, while the transit traffic is from other nodes. As shown in Fig. 3, node 1 is a source node and only has source traffic, while nodes 2, 3, 4, 5, 6 and 7 are source nodes as well as intermediate nodes because they have source traffic as well as transit traffic. Each node could have two types of neighbor nodes: backward and forward. For example, the backward node of node 3 is node 1, because its data can be sent by node 3 and forward nodes of node 3 are nodes 5, 6 and 7. In this paper, $f(i)$ is the set of forward nodes of i and $b(i)$ is set of backward nodes of i . For example, in Fig.3, $b(3)$ is equal {1} and $f(3)$ is equal {5, 6, 7}.

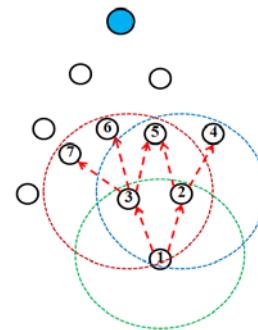


Fig. 3 General network model

3.2 Node Model

Node model of the investigated wireless sensor network is presented in Fig. 4. The source traffic of node i is generated with source traffic rate (r_s^i) by itself locally. The transit traffic of node i is received with transit traffic rate (r_{tr}^i) from its child nodes through MAC layer of node i . Both r_s^i and r_{tr}^i are converged through network layer to MAC layer as total input rate of node i (r_{in}^i). Traffic packets could be queued if r_{in}^i ($r_{in}^i = r_s^i + r_{tr}^i$) exceeds packet forwarding rate (r_f^i) at MAC layer. Congestion could take place in node i if r_{in}^i is larger than r_f^i continuously, when the buffer of node i could be filled up quickly and finally overflow. This congestion can be controlled by reducing r_{tr}^i in the DPCC protocol.

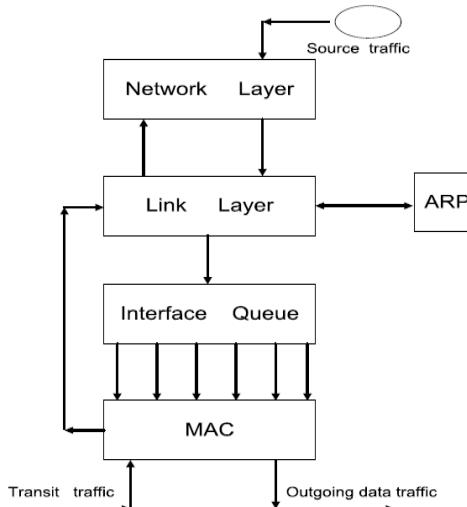


Fig. 4 General node model

4. DPCC Protocol

The DPCC protocol tries to increase throughput and reduce packet loss while guaranteeing distributed priority-based fairness with lower control overhead. The congestion control scheme of sensor node i is shown in Fig. 5. DPCC protocol consists of three components: backward and forward nodes selection (BFS), predictive congestion detection (PCD) and dynamic priority-based rate adjustment (DPRA), which are introduced with responsibility for precise congestion discovery and weighted fair congestion control.

4.1 Backward and Forward Nodes Selection

The node i selects a forward node for itself according to received rate adjustment values from $f(i)$. The node i selects the one as a forward node which received rate

value from it is max. Then node i send notification to selected forward node. For increasing the throughput, the other forward nodes of node i which is not selected as a forward node of this node adjust the new rates for their other backward nodes.

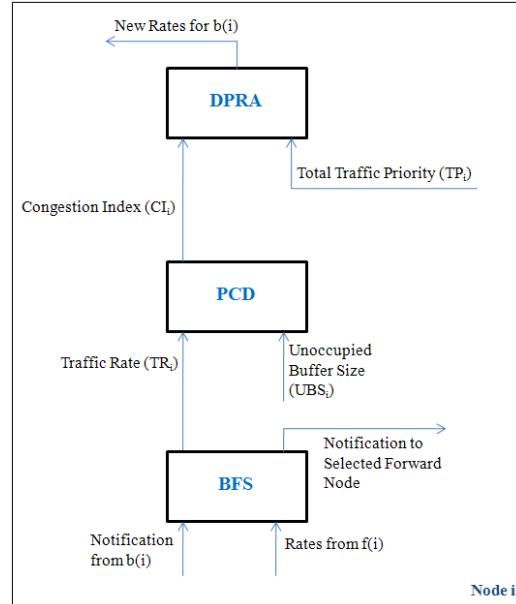


Fig. 5 Congestion control scheme in node i

4.2 Predictive Congestion Detection

Congestion index (CI_i) reflecting the current congestion level at each sensor node i is determined on its unoccupied buffer size (UBS_i) and traffic rate (TR_i) at MAC layer as follows:

$$CI_i = UBS_i - TR_i \quad (1)$$

$$UBS_i = MBS_i - OBS_i \quad (2)$$

$$TR_i = (r_s^i + \sum_{j \in b(i)} r_{ji} - \sum_{k \in f(i)} r_{ik}) \times T \quad (3)$$

Here, MBS_i and OBS_i are defined as the maximal buffer size and current queue length of node i . r_{ji} and r_{ik} denote the average upstream input traffic rate from node j to i and downstream output traffic rate from node i to k , respectively.

On the other hand, r_{ji} and r_{ik} are updated periodically at each time interval T as follows:

$$r_l^{new} = [(1 - \omega) \times r_l^{old}] + [\omega \times \frac{n_T}{T}], \forall l = i, ji, ik \quad (4)$$

Here, n_T denotes number of the new arriving packets during the time period T , and ω is a constant satisfying $0 < \omega < 1$.

If $CI_i < 0$, it means that congestion may occur in the node i with this traffic rate. In this state, DPRA component must adjust the traffic rates of backward nodes to avoid congestion.

4.3 Dynamic Priority-based Rate Adjustment

Total traffic priority (TP_i) in each sensor node i is calculated as follows:

$$TP_i = \sum_{j \in b(i)} TP_j + SP_i \quad (5)$$

Here, SP_i and TP_j are defined as local source traffic priority of node i and total traffic priority of node j which is the member of $b(i)$, respectively. Traffic priority ratio of node i (TPR_i) and its backward nodes (TPR_{ji} , $j \in b(i)$) in one hop are obtained as follows:

$$TPR_i = \frac{SP_i}{TP_i} \quad (6)$$

$$TPR_{ji} = \frac{TP_j}{TP_i} \quad (7)$$

According to Eq. (6) and Eq. (7), source traffic rate of node i and each transit traffic rate of this node can be allocated with the traffic priority as follows:

$$r_{s_new}^i = TPR_i \times CI_i \times \frac{1}{T} \quad (8)$$

$$r_{ji_new} = TPR_{ji} \times CI_i \times \frac{1}{T} \quad (9)$$

5. Performance Evaluation

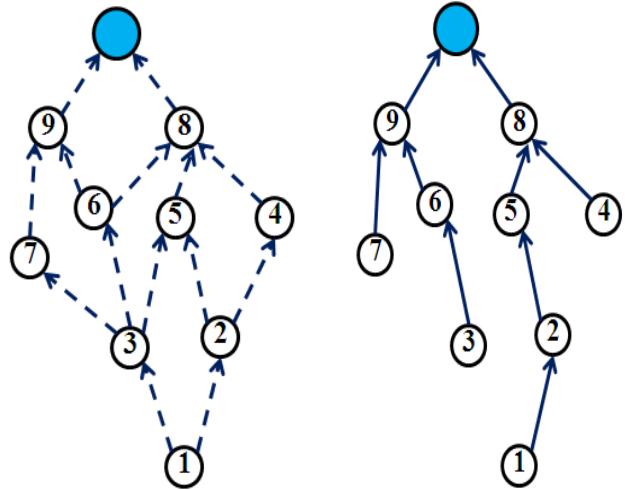
In this section, performances of the DPCC protocol are shown by simulation. The scenario is similar to Fig. 6 (a), where 9 nodes send traffic packets to the sink with different SP and form many-to-one upstream traffic. The network stack of each node consists of IEEE 802.11 MAC layer. Simulation parameters are listed in Table 1.

Table 1: Simulation parameters

| Parameters | Value |
|-------------------|------------|
| Data rate | 1 Mbps |
| Buffer size | 20 packets |
| Number of sensors | 9 |
| Initial energy | 2J |
| Time interval T | 1 sec |
| ω | 0.2 |
| Simulation time | 80 sec |
| Data packet size | 512 bits |

Performance comparisons of the DPCC with PCCP protocol on throughput and fairness are provided as follows.

The network model in this scenario is assumed similar to Fig. 6 (b) for PCCP protocol.



a) Network model in DPCC b) Network model in PCCP
 Fig. 6 Network model in this scenario

5.1 Normalized Throughput

We study the impact of changing the source traffic rate on throughput. We change the traffic rate at the each source node from 5 to 40 packets/sec. We assume that priority of all nodes is same in this evaluation.

Normalized network throughputs of DPCC and PCCP are shown in Fig. 7.

As it can be seen, proposed protocol has performance better than MCMP in network throughput especially when that traffic rate at the source node is increased.

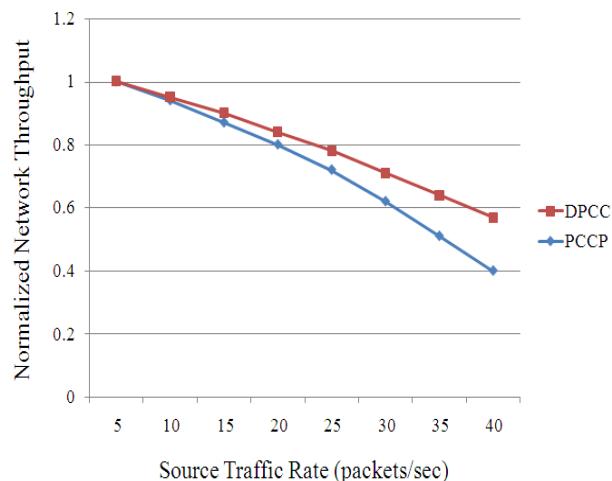


Fig. 7 Normalized network throughput

5.2 Priority-based Fairness

In this case, we use the same topology as in Fig. 6, but the nodes will be configured with different source traffic priority index (SP_i) as follows: node 4 with source traffic priority index 3, node 5 with source traffic priority index of 2 and all other nodes with source traffic priority index of 1.

It is assumed that node 5 only remains active in the time interval [20 sec, 60 sec] and node 6 only remains active in the time interval [30 sec, 50 sec] and generate traffic packets based on source traffic priority SP_5, SP_6 .

By the Fig. 8 it can be seen that priority-based fairness has been achieved.

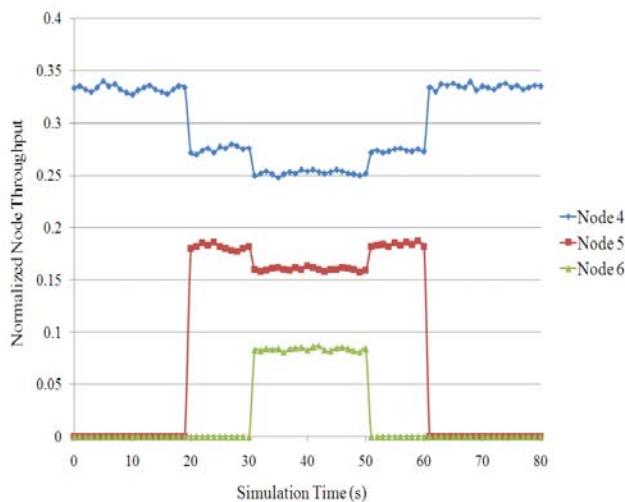


Fig. 8 Normalized node throughput in DPCC

6. Conclusion

In this paper, we propose a dynamic predictive congestion control (DPCC) algorithm. The DPCC can predict congestion in the node and will broadcast traffic on the entire network fairly and dynamically. Simulation results show that the proposed protocol is more efficient than previous algorithms especially in network throughput evaluation.

References

- [1] T. Bokareva, W. Hu, S. Kanhere, B. Ristic, N. Gordon, T. Bessell, M. Rutten and S. Jha, "Wireless Sensor Networks for Battlefield Surveillance", In roceedings of The Land Warfare Conference (LWC)- October 24 – 27, 2006, Brisbane, Australia.
- [2] A. Mainwaring, J. Polastre, R. Szewczyk, D. Culler, and J. Anderson, "Wireless Sensor Networks for Habitat Monitoring," in the Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications (ACM-WSNA), Pages: 88-97, September 28 - 28, 2002, Atlanta, Georgia, USA.
- [3] N. Xu, S. Rangwala, K. Chintalapudi, D. Ganesan, A. Broad, R. Govindan, and D. Estrin, "A Wireless Sensor Network for structural Monitoring," in Proc. ACM SenSys Conf., Nov.2004.
- [4] M. Hefeeda, M. Bagheri, "Wireless Sensor Networks for Early Detection of Forest Fires", in the proceedings of IEEE International Conference on Mobile Adhoc and Sensor Systems, 2007. MASS 2007. Volume , Issue , 8-11 Oct. 2007 Page(s):1 – 6, Pisa, Italy.
- [5] K. Srinivasan, M. Ndoh, H. Nie, H. Xia, K. Kaluri, and D. Ingraham, "Wireless Technologies for Condition-Based Maintenance (CBM) in Petroleum Plants," Proc. of DCOSS'05 (Poster Session), 2005.
- [6] Bashir Yahya, Jalel Ben-Othman, "Towards a classification of energy aware MAC protocols for wireless sensor networks", Journal of Wireless Communications and Mobile Computing, Wiley.
- [7] B. Hull, K. Jamieson, and H. Balakrishnan, "Mitigating congestion in wireless sensor networks, in Proc. ACM Sensys'04.
- [8] Chonggang Wang, Kazem Sohraby, Victor Lawrence, Bo Li, Yueming Hu, "Priority-based Congestion Control in Wireless Sensor Networks", IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing , Vol 1 (SUTC'06), 2006, pp. 22-31.
- [9] C. T. Ee and R. Bajcsy, "Congestion control and fairness for many-to-one routing in sensor networks," in Proc. Of ACM Sensys'04.
- [10] B. Hull, K. Jamieson, and H. Balakrishnan, "Mitigating congestion in wireless sensor networks," in Proc. ACM Sensys'04.
- [11] C. Y. Wan, S. B. Eisenman, and A. T. Campbell, "CODA: Congestion detection and avoidance in sensor networks," in Proc. of ACM Sensys'03.
- [12] A. Woo and D. C. Culler, "A transmission control scheme for media access in sensor networks," in Proc. Of ACM Mobicom'01.
- [13] Yogesh S., O.B.Akan, Ian F.Akyildiz, "ESRT: Event-to-Sink Reliable Transport in Wireless Sensor Networks", in Proc. of Mobi- Hoc03, Annapolis, Maryland, USA, June, 2003.
- [14] Chonggang Wang, Kazem Sohraby, Victor Lawrence, Bo Li, Yueming Hu, "Priority-based Congestion Control in Wireless Sensor Networks", IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing , Vol 1 (SUTC'06), 2006, pp. 22-31.
- [15] B. Hull, K. Jamieson, and H. Balakrishnan, "Mitigating Congestion in Wireless Sensor Networks," in Proc. of ACM SenSys '04.

Off-Line Handwritten Signature Identification Using Rotated Complex Wavelet Filters

M.S. Shirdhonkar¹ and Manesh Kokare²

¹ Department of Computer Science and Engineering, BLDEA's CET, Bijapur, Karnataka, India

² 2Department of Electronics and Telecommunication, SGGS IOT, Nanded, Maharashtra, India

Abstract

In this paper, a new method for handwritten signature identification based on rotated complex wavelet filters is proposed. We have proposed to use the rotated complex wavelet filters (RCWF) and dual tree complex wavelet transform(DT-CWT) together to derive signature feature extraction, which captures information in twelve different directions. In identification phase, Canberra distance measure is used. The proposed method is compared with discrete wavelet transform (DWT). From experimental results it is found that signature identification rate of proposed method is superior over DWT

Keywords: Signature identification, rotated complex wavelet filters, discrete wavelet transform person's identification.

1. Introduction

1.1 Motivation

Authentication and affirmation of statements, documents, scripts etc, from times immemorial has been done through signatures. Even today, where every thing has gone digital, signature plays a vital role. They appear on many types of documents such as bank cheques, credit cheques, governmental documents, wills over assets of a person and many other documents of greater importance. But this type of authentication is also subject to mal practices and crimes. Forgery and imitation of signatures of other person may help anyone to gain access to his/her valuable assets or can lead to undesirable consequences. Identification of signatures by human eye and study may be error prone and manipulative, thus an automated document processing system that can analyze and identify a signature serves as an effective, useful and less error prone and non manipulative tool.

Signature's validity confirmation for different documents is an important problem domain in automatic document processing. An area where signature identification finds application is in banking, user login in computers or PDA (Personal digital assistant), for access control, to check for authentication of official documents etc. There are two modes for signature identification and

verification: Static or off-line and Dynamic or on-line. In static mode, the input of system is a 2D image of signature. Contrary to this, in dynamic mode, the input is signature trace in time domain. In the Dynamic mode, the person puts his signature on an electronic tablet through an electronic pen. His/ her obtained signature is sampled with each sample having three attributes: The 2-dimensional co-ordinates; x and y and time of sample occurrence, t. The time attribute of each sample is used to extract useful information such as start and stop points, velocity and acceleration of the signature stroke. Some electronic tablets in addition to time sampling can digitize the pressure. Such additional information in the dynamic mode increases identification rate as compared to the static mode. But Dynamic mode has a greater disadvantage: it is on-line, hence it requires presence of the person whose signature is needed and that too has been taken digitally. Thus it cannot be applied to other important cases where there is absence of person whose sign is needed and cases where analysis and identification needs to be carried out of existing documents with signature marks. Thus off-line signature verification becomes an inevitable choice and finds universal application.

1.2 Related works

Signature verification contain two areas: off-line signature verification, where signature samples are scanned into image representation and on-line signature verification, where signature samples are collected from a digitizing tablet which is capable of pen movements during the writing. In 2009, Ghandali and Moghaddam have proposed an off-line Persian signature identification and verification based on Image registration, DWT (Discrete Wavelet Transform) and fusion. They used DWT for features extraction and Euclidean distance for comparing features. It is language dependent method [1]. In 2008, Larkins and Mayo have introduced a person dependent off-line signature verification method that is based on Adaptive Feature Threshold (AFT) [2]. AFT enhances the method of converting a simple feature of signature to binary feature vector to improve its

representative similarity with training signatures. They have used combination of spatial pyramid and equimass sampling grids to improve representation of a signature based on gradient direction. In classification phase, they used DWT and graph matching methods. In another work, Ramachandra et al [3], have proposed cross-validation for graph matching based off-line signature verification (CSMOSV) algorithm in which graph matching compares signatures and the Euclidean distance measures the dissimilarity between signatures.

In 2007, Kovari et.al [4] presented an approach for off-line signature verification, which was able to preserve and take usage of semantic information. They, used position and direction of endpoints in features extraction phase. Porwik [5] introduced a three stages method for offline signature recognition. In this approach the Hough transform ,center of gravity and horizontal-vertical signature histogram have been employed, using both static and dynamic features that were processed by DWT has been addressed in[6].The verification phase of this method is based on fuzzy net using the enhanced version of the MDF(Modified Direction feature)extractor has been presented by Armand et.al [7].The different neural classifier such as Resilient Back Propagation(RBP) neural network and Radial Basis Function(RBF)network have been used in verification phase of this method. In 2005, Chen and Srihari [8] described an approach that obtains an exterior contour of the image to define pseudo writing path. To match two signatures a dynamic time wrapping (DTW) method has been employed to segment signature into curves.

The main contribution of this paper is that, we have proposed an off-line handwritten signature identification using rotated complex wavelet filters and dual tree complex wavelet transform, which captures information in twelve different directions for identification. In identification phases Canberra distance measure is used. The experimental results of proposed method were satisfactory and found that it gives better results as compared with earlier approach. The rest of paper is organized as follows. In section 2, discusses the feature extraction phase. The signature identification approaches is presented in section 3. In section 4, the experimental results and the selection of training samples are presented, and finally section 5 concludes the work.

2. Feature Extraction Phase

The major task of feature extraction is to reduce image data to much smaller in size which represents the important characteristic of the image. In signature identification, edge information is very important in characterizing signature properties. Therefore we proposed the use of DT-CWT and DT-RCWF jointly, which

captures the information in twelve different directions. The performance of the system is compared with standard discrete wavelet transform which captures information in only three directions.

2.1 Discrete Wavelet Transform Features

The multi resolution wavelet transform decomposes a signal into low pass and high pass information. The low pass information represents a smoothed version and the main body of the original data. The high pass information represents data of sharper variations and details. Discrete Wavelet Transform decomposes the image into four sub-images when one level of decomposing is used. One of these sub-images is a smoothed version of the original image corresponding to the low pass information and the other three ones are high pass information that represents the horizontal, vertical and diagonal edges of the image respectively. When two images are similar, their difference would be existed in high-frequency information. A DWT with N decomposition levels has $3N+1$ frequency bands with $3N$ high-frequency bands [9]. The impulse response associated with 2-D discrete wavelet transform are illustrated in Fig. 1 as gray-scale image.

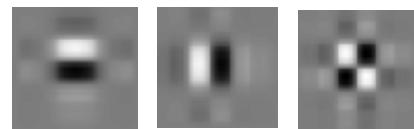


Fig.1. Impulse response of 0° , 90° and $\pm 45^{\circ}$ of DWT

2.2 Dual Tree Rotated Complex Wavelet Filters

Drawbacks of the DWT are overcome by the complex wavelet transform (CWT).By introducing limited redundancy into the transform. But still it suffer from problem like no perfect reconstruction is possible using CWT decomposition beyond level 1, when input to each level becomes complex. To overcome this, Kingsbury [11] proposed a new transform, which provides perfect reconstruction along with providing the other advantages of complex wavelet, which is DT-CWT. The DT-CWT uses a dual tree of real part of wavelet transform instead using complex coefficients. This introduces a limited amount of redundancy and provides perfect reconstruction along with providing the other advantages of complex wavelets. The DT-CWT is implemented using separable transforms and by combining subband signals appropriately. Even though it is non-separable yet it inherits the computational efficiency of separable transforms. Specifically, the 1-D DT-CWT is implemented using two filter banks in parallel, operating on the same data. For d-dimensional input, a L scale DT-CWT outputs an array of real scaling coefficients corresponding to the low pass subbands in each dimension. The total

redundancy of the transform is 2^d and independent of L . The mechanism of the DT-CWT is not covered here. See [10], [12-13] for a comprehensive explanation of the transform and details of filter design for the trees. A complex valued $\psi(t)$ can be obtained as

$$\psi(x) = \psi_h(x) + j\psi_g(x) \quad (1)$$

Where $\psi_h(x)$ and $\psi_g(x)$ are both real-valued wavelets. The impulse responses of six wavelets associated with 2-D dual tree complex wavelet transform are illustrated in Fig. 2.

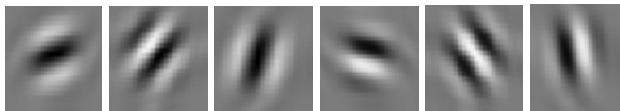


Fig. 2 Impulse response of six wavelet filters $+15^\circ, +45^\circ, +75^\circ, -15^\circ, -45^\circ$ and -75° of complex wavelet.

2.3 Dual Tree Rotated Complex Wavelet Filters

Directional 2D RCWF are obtained by rotating the directional 2D DT-CWT filters by 45° so that decomposition is performed along new direction, which are apart from decomposition 45° directions of CWT[10]. The size of a filter is $(2N-1)X(2N-1)$, where N is the length of the 1-D filter. The decomposition of input image with 2-D RCWF followed by 2-D down sampling operation is performed up to the desired level. The computational complexity associated with RCWF decomposition is the same as that of standard 2-D CWT, if both are implemented in the 2-D frequency domain. The set of RCWFs retains the orthogonal property. The six sub bands of 2D DT-RCWF gives information strongly oriented at $(30^\circ, -30^\circ, 60^\circ, 90^\circ, 120^\circ)$. The mechanism of the DT-RCWF is not covered here. See [10],[12-13] for a comprehensive explanation of the transform and details of filter design for the trees. Thus, the 2D DT-CWT and RCWF provide us with more directional selectivity in the direction

$$\left\{ \begin{array}{l} (+15^\circ, +45^\circ, +75^\circ, -15^\circ, -45^\circ, -75^\circ), \\ (0^\circ, +30^\circ, +60^\circ, +90^\circ, 120^\circ, -30^\circ) \end{array} \right\} \text{ than}$$

the DWT whose directional sensitivity is in only three directions $\{0^\circ, \pm 45^\circ, 90^\circ\}$. The six wavelets associated with rotated complex wavelet transform are illustrated in Fig.3.

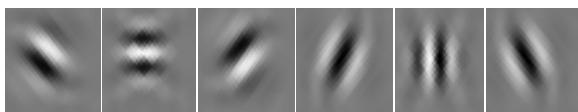


Fig. 3 Impulse response of $-30^\circ, 0^\circ, +30^\circ, +60^\circ, 90^\circ$ and 120° of rotated complex wavelet

2.4 Feature Database Creation

To conduct the experiments, we were computed two different feature sets using *algorithm 1* and *algorithm 2*, which uses DWT and combined DT-CWT and DT-RCWF respectively. To construct the feature vectors of each signature in the database, we decomposed each signature using DT-CWT and DT-RCWF up to 6th level. The Energy and Standard Deviation (STD) were computed separately on each sub band and the feature vector was formed using these two parameter values. The Energy E_k and Standard Deviation σ_k of kth sub band is computed as follows

$$E_k = \frac{1}{M \times N} \sum_{i=1}^M \sum_{j=1}^N |W_k(i, j)| \quad (2)$$

$$\sigma_k = \left[\frac{1}{M \times N} \sum_{i=1}^N \sum_{j=1}^M (W_k(i, j) - \mu_k)^2 \right]^{\frac{1}{2}} \quad (3)$$

Where $W_k(i, j)$ is the kth wavelet-decomposed sub band, $M \times N$ is the size of wavelet decomposed sub band, and μ_k is the mean of the kth sub band. The resulting feature vector using energy and standard deviation are

$$\bar{f}_E = [E_1 \ E_2 \ \dots \ E_n] \quad \text{and} \\ \bar{f}_\sigma = [\sigma_1 \ \sigma_2 \ \dots \ \sigma_n] \quad \text{respectively. So combined feature vector is}$$

$$\bar{f}_{\sigma\mu} = [\sigma_1 \ \sigma_2 \ \dots \ \sigma_n \ E_1 \ E_2 \ \dots \ E_n] \quad (4)$$

The step by step procedure for feature database creation using discrete wavelet transform and combined DT-CWT and DT-RCWF are explained in algorithm 1 and algorithm 2 respectively.

Algorithm 1: Feature database creation using DWT

Input:

Signature image Database: DB

1D filters : LF, HF

Handwritten Signature : S_i

Output: Feature database FV

Begin

For each S_i in DB **do**

Decompose the S_i by applying low pass LF and high pass HF filters up to 6th level

Calculate energy E and standard deviation SD for each subband using (2) and (3) respectively in each level

Feature vector f= [E U SD]

FV=FV U f

End for

End

Algorithm 2: Feature database creation using DT-CWT and DT-RCWF

Input:

Signature image Database: DB
 2D DT-CWT filters : F
 Handwritten Signature : S_i

Output:

Feature database : FV

Begin

If DT-RCWF

Rotate 2D filters F by 45^0

End if

For each S_i in DB do

Decompose the S_i by applying 2D filters F up to 6th Level. Calculate energy E and standard deviation SD for each subband using (2) and (3) respectively in each level

Feature vector $f = [E \ U \ SD]$

$FV = FV \cup f$

End for

End

3. Signature Identification Phase

There are several ways to work out the distance between two points in multidimensional space. We have used Canberra distance metric as distance measure. If x and y are the feature vectors of the database and query signature, respectively, and have dimension d, then the Canberra distance is given by

$$\text{Canb}(x, y) = \sum_{i=1}^d \frac{|x_i - y_i|}{|x_i| + |y_i|} \quad (5)$$

The step by step procedure of identification is as follows,

Algorithm 3: Handwritten Signature Identification

Input: Test signature: St

Feature database: FV

Output: Distance vector: Dist

Handwritten signature identification

Begin

Calculate feature vector of test signature St using algorithm 1

For each fv in FV do

Dist= Calculate distance between test signature and fv using (5)

End for

Display the minimum distance signature from distance vector.

End

4. Experimental Results

4.1. Image Database

The signatures were collected using either black or blue ink (No pen brands were taken into consideration), on a white A4 sheet of paper, with eight signature per page. Signatures were scanned subsequently to digitize individual with a resolution in 256 grey levels. Images were obtained in rectangular areas of size 256x256 pixels. Sample signature image database is shown in Fig.3. A group of 52 persons are selected for 16 specimen signatures which make the total of $52 \times 16 = 832$ signature database.



Fig.3. Sample Signature Images Database

4.2. Identification Performance

For each person 12 signatures for training and 4 signatures for testing are used. This makes the total of $4 \times 52 = 208$ signature. The identification rate is 90.6% using proposed method and 61.45 % using DWT. Fig.4 shows comparison between DWT and proposed method. From Fig. 4, we observed that signature identification rate of proposed method is superior over DWT.

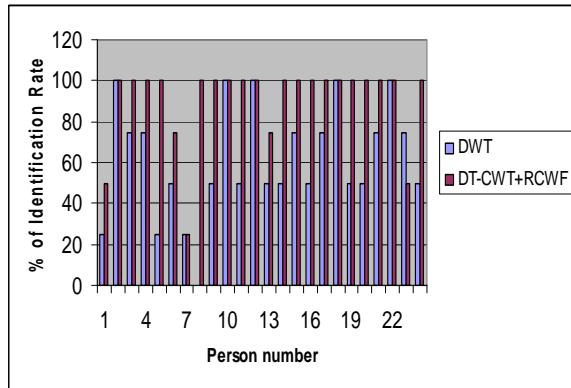


Fig.4. Comparison between DWT and RCWF

5. Conclusions

In this paper, we introduced new approach for identification of off-line signatures. The proposed approach uses RCWF and DT-CWT jointly for extracting details in twelve different directions and Canberra distance for comparing features. The experimental results we found that signature identification rate for proposed method is superior over DWT.

Acknowledgments

The authors would like to appreciate all participants who gave permission to use their handwritten signatures in this study.

References

- [1] Samanesh Ghandali and Mohsen Ebrahimi Moghaddam, "Off-Line Persian Signature Identification and Verification based on Image Registration and Fusion" In: Journal of Multimedia, volume 4, 2009, pp. 137-144.
- [2] Larkins, R. Mayo, M., "Adaptive Feature Thresholding for Off-Line Signature Verification", In: Image and vision computing New Zealand, 2008, pp. 1-6.
- [3] Ramachandra, A.C. Pavitra, K. and Yashasvini, K. and Raja, K.B. and Venugopal, K.R. and Patnaik, L.M., "Cross-Validation for Graph Matching based Off-Line Signature Verification", In IDICON 2008, India, 2008, pages: 17-22.
- [4] Kovari, B. Kertesz, Z. and Major, a., "Off-Line Signature Verification Based on Feature Matching: In: Intelligent Engineering Systems, 2007, pp. 93-97.
- [5] Porwik P., "The Compact Three Stages Method of the Signatures Recognition", 6 th International Conference on Computer Information Systems and Industrial Management Applications, 2007, pp. 282-287.
- [6] Wei Tian Yizheng Qiao Zhiqiang Ma, "A New Scheme for Off-Line Signature Verification uses DWT and Fuzzy net", In: Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2007, pages: 30-35.
- [7] Armand S., Blumenstein, M., Muthukumarasamy V. "Off-Line Signature and Neural Based Classification", In: Neural Networks, 2006 IJCNN, pp.: 684-691.
- [8] Chen S., Srihari S., "Use of Exterior Contours and Shape Features in Off-Line Signature Verification", In: Eighth International Conference Document Analysis and Recognition, 2005, pp. 1280-1284.
- [9] Gongalo Pajares, Jesus, Mahuel de la Cruz, "A wavelet-based image fusion Tutorial", Pattern Recognition Volume 37, Issue 9, September 2004, Elsevier Science Inc, pp. 1855-1872.
- [10] Manesh Kokare, P.K. Biswas, and B.N. Chatterji, "Texture Image retrieval using New Rotated Complex Wavelet Filters," IEEE Trans. on systems, man, and Cybernetics-Part B: Cybernetics, vol. 35, no.6, Dec. 2005
- [11] N.G. Kingsbury, "Image processing with complex wavelet," Phil. Trans. Roy. Soc.
- [12] N. G. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," J.App. Comput. Harmon. Anal., vol. 10, no.3, pp.234-253, May 2001.
- [13] JI. Selesnick, R. Baraniuk, and N. Kingsbury, "The dual-tree complex wavelet transform," IEEE Signal Process. Mag., vol.22, no. 06, pp.123-151, Nov. 2005.

M. S. Shirdhonkar completed his B. E. and M.E. from the Department of Computer Science and Engineering, Shivaji University, Kolhapur, India in the years 1994, 2005 respectively. From 1997-1999, he was worked as lecturer in Computer Science Department at JCE, Institute of Technology, Junnar, Maharashtra, India. In 2000, he joined as a lecturer in the Department of Computer Science at B. L. D. E's. Dr. V. P. P.G.H. College of Engineering and Technology, Bijapur, Karnataka, India, where he is presently holding



position of Assistant Professor and doing PhD at S.R.T.M. University, Nanded, Maharashtra, India. His research interests include image processing, pattern recognition, and document image retrieval. He is a life member of Indian Society for Technical Education and Institute of Engineers.

Manesh Kokare (S'04) was born in Pune, India, in Aug 1972. He received the Diploma in Industrial Electronics Engineering from Board of Technical Examination, Maharashtra, India, in 1990, and B.E. and M. E. Degree in Electronics from Shri Guru Gobind Singhji Institute of Engineering and Technology Nanded, Maharashtra, India, in 1993 and 1999 respectively, and Ph.D. from the Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur, India, in 2005. Since June 1993 to Oct 1995, he worked with Industry. From Oct 1995, he started



his carrier in academics as a lecturer in the Department of Electronics and Telecommunication Engineering at S. G. G. S. Institute of Engineering and Technology, Nanded, where he is presently holding position of Assistant Professor. His research interests include wavelets, image processing, pattern recognition, and Content Based Image Retrieval.

Ternary Tree and Memory-Efficient Huffman Decoding Algorithm

Pushpa R. Suri¹ and Madhu Goel²

¹Department of Computer Science and Applications,
Kurukshetra University, Haryana, India

²Department of Computer Science and Engineering,
Kurukshetra Institute of Technology and Management, KUK

Abstract

In this study, the focus was on the use of ternary tree over binary tree. Here, a new one pass Algorithm for Decoding adaptive Huffman ternary tree codes was implemented. To reduce the memory size and fasten the process of searching for a symbol in a Huffman tree, we exploited the property of the encoded symbols and proposed a memory efficient data structure to represent the codeword length of Huffman ternary tree. In this algorithm we tried to find out the starting and ending address of the code to know the length of the code. And then in second algorithm we tried to decode the ternary tree code using binary search method. In this algorithm we tried to find out the starting and ending address of the code to know the length of the code. And then in second algorithm we tried to decode the ternary tree code using binary search method.

Key words: Ternary tree, Huffman's algorithm, adaptive Huffman coding, Huffman decoding, prefix codes, binary search

1. INTRODUCTION

Ternary tree or 3-ary tree is a tree in which each node has either 0 or 3 children (labeled as LEFT child, MID child, RIGHT child).

Huffman coding is divided in to two categories:-

1. Static Huffman coding
2. Adaptive Huffman coding

Static Huffman coding suffers from the fact that the uncompressed need have some knowledge of the probabilities of the symbol in the compressed files. This can need more bits to encode the file. If this information is unavailable, compressing the file requires two passes.

FIRST PASS finds the frequency of each symbol and constructs the Huffman tree. SECOND PASS is used to compress the file. We already use the concept of static Huffman coding [12] using ternary tree And we conclude that representation of Static Huffman Tree [12] using Ternary Tree is more beneficial than representation of Huffman Tree using Binary Tree in terms of number of internal nodes, Path length [8], height of the tree, in memory representation, in fast searching and in error detection & error correction. Static Huffman coding methods have several disadvantages.

Therefore we go for adaptive Huffman coding.

Adaptive Huffman coding calculates the frequencies dynamically based on recent actual frequencies in the source string. Adaptive Huffman coding which is also called dynamic Huffman coding is an adaptive coding technique based on Huffman coding building the code as the symbols are being transmitted that allows one-pass encoding and adaptation to changing conditions in data. The benefits of one-pass procedure is that the source can be encoded in real time, through it becomes more sensitive to transmission errors, since just a single loss ruins the whole code.

Implementations of adaptive Huffman coding: -

There are number of implementations of this method, the most notable are

1. FGK (Faller Gallager Knuth) Algorithm
2. Vitter Algorithm

We already use the concept of FGK Huffman coding [13] using ternary tree And we conclude that representation of FGK Huffman Tree using Ternary Tree is more beneficial than representation of Huffman Tree using Binary Tree in terms of number of internal nodes, Path length [12], height of the tree, in memory representation, in fast searching and in error detection & error correction.

We also already use the concept of Vitter Huffman coding [14] using ternary tree And we conclude that representation of algorithm V Huffman Tree using Ternary Tree is more beneficial than representation of Huffman Tree using Binary Tree in terms of number of internal nodes, Path length [8], height of the tree, in memory representation, in fast searching and in error detection & error correction.

All of these methods are defined- word schemes that determine the mapping from source messages to codewords on the basis of a running estimate of the source message probabilities. The code is adaptive, changing so as to remain optimal for the current estimates. In this way, the adaptive Huffman codes responds to locality, in essence, the encoder is learning the characteristics of the source. The decoder must learn along with the encoder by continually updating the Huffman tree so as to stay in synchronization with the encoder. Here we are given the concept of error detection and error correction. And the main point is that, this thing is only beneficial in TERNARY TREE neither in binary tree nor in other possible trees.

Now here we try to use the concept of adaptive Huffman decoding algorithm using ternary tree.

In 1951, David Huffman [2] and his MIT information theory classmates gave the choice of a term paper or a final exam. Huffman hit upon the idea of using a frequency-sorted binary tree and quickly proved this method the most efficient. In doing so, the student out did his professor, who had worked with information theory inventor Claude Shannon to develop a similar code. Huffman built the tree from the bottom up instead of from the top down.

Huffman codes are widely used in the area of data compression and telecommunications. Some applications include JPEG [3] picture compression and MPEG video and audio compression. Huffman codes are of variable word length, which means that the individual symbols used to compose a message are represented (encoded) each by a distinct bit sequence of distinct length. This characteristic of the codeword helps to decrease the amount of redundancy in message data, i.e., it makes data compression possible.

The use of Huffman codes [7] affords compression, because distinct symbols have distinct probabilities of incidence. This property is used to advantage by tailoring the code lengths corresponding to those symbols in accordance with their respective probabilities of occurrence. Symbols with higher probabilities of incidence are coded with shorter codeword, while symbols with lower probabilities are coded with longer codeword.

However, longer codeword still show up, but tend to be less frequent and hence the overall code length of all codeword in a typical bit string tends to be smaller due to the Huffman coding.

A basic difficulty in decoding Huffman codes is that the decoder cannot know at first the length of an incoming codeword. As previously explained, Huffman codes are of variable length codes. Huffman codes can be detected extremely fast by dedicating enormous amounts of memory. For a set of Huffman code words with a maximum word length of N bits, 2^N memory locations are needed, because N incoming bits are used as an address into the lookup table to find the corresponding code words.

A technique requiring less memory is currently performed using bit-by-bit decoding, which proceeds as follows. One bit is taken and compared to all the possible codes with a word length of one. If a match is not found, another bit is shifted in to try to find the bit pair from among all the code words with word length of two. This is continued until a match is found. Although this approach is very memory-efficient, it is very slow, especially if the codeword being decoded is long.

Another technique is the binary tree search method. In this implementation technique, Huffman tables used should be converted in the form of binary trees. A binary tree is a finite set of elements that is either empty or partitioned into three disjoint subsets. The first subset contains a single element called the root of the tree. The other two subsets are referred to as left and right sub trees of the original tree. Each element of a binary tree is called a node of the tree. A branch connects two nodes. Nodes without any branches are called leaves. Huffman decoding for a symbol search begins at the root of a binary tree and ends at any of the leaves; one bit for each node is extracted from bit-stream while traversing the binary tree [1]. This method is a compromise between memory requirement and the number of Huffman code searches as compared to the above two methods. In addition, the coding speed of this technique will be down by a factor related to maximum length of Huffman code.

Another technique currently used to decode Huffman codes is to use canonical Huffman codes. The canonical Huffman codes are of special interest since they make decoding easier. They are generally used in multimedia and telecommunications. They reduce memory and decoding complexity. However, most of these techniques use a special tree structure in the Huffman codeword tables for encoding and hence are suitable only for a special class of Huffman codes and are generally not suitable for decoding a generic class of Huffman codes.

As indicated in the above examples, a problem with using variable codeword lengths is the difficulty in achieving balance between speed and reasonable memory usage.

Huffman is a fairly standard compression algorithm, and it is still commonly used. In order to do this you need a very simple tree. The nodes need a char and a number of occurrences (I used an unsigned short in mine). The tree does not need any of the standard BST methods, but you will need to be able to create a tree by merging two existing trees. All data is stored in the leaf nodes, frequency information is stored in every node in the tree.

- The message “go eagles” requires 144 bits in Unicode but only 38 using Huffman coding
- A Huffman tree is a binary tree [10] used to store a code that facilitates file compression

There are basically two concepts in Huffman coding

- Huffman Encoding
- Huffman Decoding

1.1 HUFFMAN ENCODING:-

This is a two pass problem. The first pass is to collect the letter frequencies. You need to use that information to create the Huffman tree. Note that char values range from -128 to 127, so you will need to cast them. I stored the data as unsigned chars to solve for this problem, and then the range is 0 to 255.

Open the output file and write the frequency table to it. Open the input file, read characters from it, gets the codes, and writes the encoding into the output file.

Once a Huffman code has been generated, data may be encoded simply by replacing each symbol with its code.

1.2 HUFFMAN DECODING:-

This can be done in one pass. Open the encoded file and read the frequency data out of it. Create the Huffman tree [14] base on that information (The total number of encoded bytes is the frequency at the root of the Huffman tree.). Read data out of the file and search the tree to find the correct char to decode (a 0 bit means go left, 1 go right for binary tree and 00 bit means go left, 01 bit means go mid, 10 bit means go right in case of ternary tree) This gets tricky since you read in 8 bit blocks, but the codes can be shorter or longer than that and there are no separators.

If you know the Huffman code for some encoded data, decoding may be accomplished by reading the encoded data one bit at a time. Once the bits read match a code for symbol, write out the symbol and start collecting bits again.

1.3 Huffman codes to binary data

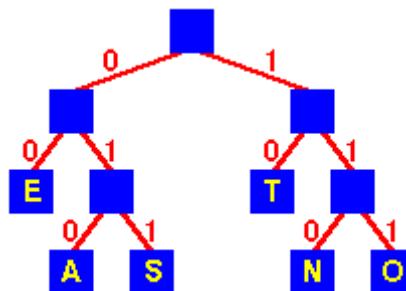
Since they are arbitrary in length, Huffman codes can be difficult to represent. The string data type has major advantages, the length [15] can be changed, and characters can be appended to them, or removed from them at either end. While you will probably use strings to represent the codes, you are not going to write a string of ones and zeros to the file. That would defeat the point of the program, which is file compression. You will need to convert from a string of length 8 to a char value which can be written to the file, and do the reverse process as well. This problem is that of finding the minimum length bit string which can be used to encode a string of symbols.

One application is text compression:

What's the smallest number of bits (hence the minimum size of file) we can use to store an arbitrary piece of text?

Huffman's scheme uses a table of frequency of occurrence for each symbol (or character) in the input. This table may be derived from the input itself or from data which is representative of the input. For instance, the frequency of occurrence of letters in normal English might be derived from processing a large number of text documents and then used for encoding all text documents. We then need to assign a variable-length bit string to each character that unambiguously represents that character. This means that the encoding for each character must have a unique prefix. If the characters to be encoded are arranged in a binary tree:

Encoding tree for ETASNO



An encoding for each character is found by following the tree from the route to the character in the leaf: the encoding is the string of symbols on each branch followed.

For example:

| String | Encoding |
|--------|------------|
| TEA | 10 00 010 |
| SEA | 011 00 010 |
| TEN | 10 00 110 |

We already used the concept of Huffman encoding using ternary tree. Further I tried to use the concept of Huffman decoding [13] using ternary tree we now implemented an algorithm which is used for Huffman decoding.

Huffman codes are widely used and very effective techniques for compressing data. Huffman's algorithm uses a table of the frequencies of occurrence of each character to build up an optimal way of representing each character as a binary string (i.e. a codeword). The running time of Huffman algorithm on a set of n characters is $O(\log n)$.

Hasemian presented an algorithm [6] to speed up the search process for a symbol in a Huffman tree and to reduce the memory size. He used a tree clustering algorithm[13] to avoid high sparsity of the Huffman tree. However, finding the optimal solution of the clustering problem is still open. Moreover, the codeword of a single side growing [16] Huffman tree is different from the codeword of the original Huffman tree. Later, Chung gave a memory efficient data structure, which needs the memory size $2n-3$, to represent the Huffman tree. In this paper, we shall propose a more efficient algorithm to save memory space.

Now I try to introduce new concept of memory efficient Huffman decoding using binary search method.

Huffman code has been widely used in text, image and video compression. For example, it is used to compress the result of quantization stage in JPEG (Hashemain, 1995). The simplest data structure used in the Huffman decoding is the Huffman tree. Array data structure (Chen *et al.*, 2005) has been used to implement the corresponding complete ternary tree for the Huffman tree. However, the sparsity in the Huffman tree causes a huge waste of memory space for array implementation (Chen *et al.*, 2005) which needs $O(2^d) = O(n \log n)$ memory, where n is the number of source symbols and d is the depth of the Huffman tree. The number of nodes in the Huffman tree is $2n-1$ and the search time is $O(d)$ (Chen *et al.*, 2005).

Huffman decoding can be done in one pass. Create the Huffman tree (Pushpa and Goel, 2008) base on that information (The total number of encoded bytes is the frequency at the root of the Huffman tree.).

Here we are first presents a new array data structure to represent the Huffman tree using Ternary Tree. The memory required in the proposed data structure is $nd = O(n)$, which is less than the previous ones. We then address an efficient Huffman decoding algorithm based on the proposed data structure; given a Huffman code, the search time for finding the source symbol is $O(\log n)$.

2 MATERIALS AND METHODS

The proposed data structure: Consider a set of source symbols $S = \{S_0, S_1, \dots, S_{n-1}\}$ with frequencies $W = \{w_0, w_1, \dots, w_{n-1}\}$ for $w_0 \geq w_1 \geq \dots \geq w_{n-1}$, where the symbol

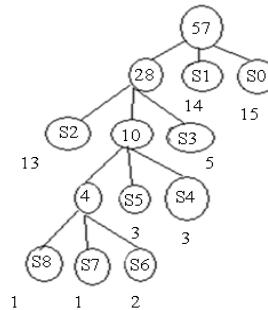


Fig. 1: An example of the Huffman tree

Table 1: An example of Huffman encoding

| Symbol | Weight | Code words |
|--------|--------|------------|
| S0 | 15 | 10 |
| S1 | 14 | 01 |
| S2 | 13 | 0000 |
| S3 | 5 | 0010 |
| S4 | 3 | 000110 |
| S5 | 3 | 000101 |
| S6 | 2 | 00010010 |
| S7 | 1 | 00010001 |
| S8 | 1 | 00010000 |

Table 2: Interval representation of symbols in Table 1

| Interval address | Symbol | Starting |
|------------------|--------|----------|
| Int2 | S2 | 00000000 |
| Int8 | S8 | 00010000 |
| Int7 | S7 | 00010001 |
| Int6 | S6 | 00010010 |
| Int5 | S5 | 00010100 |
| Int4 | S4 | 00011000 |
| Int3 | S3 | 00100000 |
| Int1 | S1 | 01000000 |
| Int0 | S0 | 10000000 |

S_i has frequency w_i . Using the Huffman algorithm to construct the Huffman tree T, the codeword c_i , $0 \leq i \leq n-1$, for symbol S_i can then be determined by traversing the path from the root to the leaf node associated with the symbol S_i , where the left branch is corresponding to "00", mid branch is corresponding to "01" and the right

branch is corresponding to “10”. Let the level of the root be zero and the level of the other node is equal to summing up its parent’s level and one. Codeword length l_i for s_i can be known as the level of S_i . Then the weighted external path length $\sum_{i=0}^{n-1} w_i l_i$ is minimum.

For example, the Huffman tree corresponding to the source symbols $\{s_0, s_1, \dots, s_8\}$ with the frequencies $\{15, 14, 13, 5, 3, 3, 2, 1, 1\}$ is shown in Fig. 1. The codeword set $C = \{c_0, c_1, \dots, c_8\}$ is derived as $\{10, 01, 0000, 0010, 000110, 000101, 00010010, 00010001, 00010000\}$, where the length set $L = \{l_0, l_1, \dots, l_8\}$ is $\{2, 2, 4, 4, 6, 6, 8, 8, 8\}$ is given in Table 1. here $d = 4$ is the depth of the Huffman tree.

The codeword generated from the Huffman tree could be treated as prefixes, which obey the prefix property, i.e., no codeword is the prefix, or start of the code for another codeword. Each prefix could be expressed as an interval with d -bit starting/ending addresses by appending 0’s/1’s to the prefix. For example, c_0 could be treated as an interval from address 10000000-1111111 and c_5 starts from 00010100- 00010111. Since there is no empty branch in the Huffman tree, each address is occupied by exactly one interval of the prefix.

3 RESULTS & DISCUSSION

Lemma 1: Each address is occupied by exactly one interval of the codeword prefix.

Proof: If some address is not occupied by any interval, this means that there is undefined code in the compressed data and will result in error decoding. Otherwise, if some address is occupied by at least two intervals, this also means that the code corresponding to this address can be decoded as two code words and violates the prefix property.

Consequently, we could represent the Huffman tree by the set of intervals. As shown in Table 2. The range of each interval int_i (i.e., ending address-starting address+1) can derive the prefix length of the corresponding symbol s_i . For example, the range of int_1 is $64 = 01111111-01000000+1$ (or 2^6), hence the height of S_i equals $(2d-6 =) 2$. Since each symbol corresponds to a leaf in the Huffman tree, each Huffman tree could be expressed as a unique set of intervals.

The data structure [13] of the intervals could be further improved by exploiting two properties: Ascending order and contiguity. Observing the leaves of the Huffman tree, the starting addresses corresponding to the leaves’ intervals is in an ascending order from left to mid and mid to right. Thus the intervals with ascending order could be derived by reading the symbols from left to mid and mid to right (or by depth first search). In addition, these intervals are contiguous, i.e., $\forall i, 0 \leq i \leq n-1$, int_i whose

starting address equals to the ending address of int_j plus one. Therefore, the successive intervals could be expressed by merely their starting addresses, as shown in Table 3.

Table 3: Contiguous interval representation of symbols in Table 1

| Interval address | Symbol | Start add | End add |
|------------------|--------|-----------|----------|
| Int0 | S0 | 10000000 | 10111111 |
| Int1 | S1 | 01000000 | 01111111 |
| Int2 | S2 | 00000000 | 11111111 |
| Int3 | S3 | 00100000 | 00101111 |
| Int4 | S4 | 00011000 | 00110000 |
| Int5 | S5 | 00010100 | 00010111 |
| Int6 | S6 | 00010010 | 00010010 |
| Int7 | S7 | 00010001 | 00010001 |
| Int8 | S8 | 00010000 | 00010000 |

1. Interval generation algorithm:
2. Input: The constructed d -level Huffman tree with n symbols
3. Output: N sorted mutually exclusive intervals
4. Generate (root, level, count, address) BEGIN
5. IF (root → leaf! = true) BEGIN
6. Count1 = Generate (root → left, level+1, count, address);
7. count1 = Generate (root → mid, level+1, count, address+ $2^{d-level-1}$);
8. count = Generate (root → right, level+1, count1, address+ $2^{d-level-1}$);
9. return count;
10. END
11. ELSE BEGIN
12. Int [count]. address = address;
13. Int [count]. symbol = root → symbol;
14. Return count+1;
15. END
16. END

In the following, the detailed algorithms to generate the intervals are presented. The algorithm is based on the depth first search. By traversing the Huffman tree, the begin address for each symbol is derived. The generated interval is then appended to the entry $int[i]$, $0 \leq i \leq n-1$, of the interval array. Since each node is traversed exactly once, the time complexity for the interval generation is $O(n)$.

For each Huffman tree, the required storage for the interval representation is n entries. Each entry contains two fields: Address and symbol. The length of address is d bits and the storage complexity is $O(n)$.

Decoding algorithm: With the array representation of Huffman tree, we can accomplish the decoding procedure by simple binary search. Given an input code ‘00100000’, the fifth $((1+9)/2 = 5)$ entry ‘00010100’ is tested. Since ‘00010100’ is smaller than ‘00100000’, the third $((6+9)/2 = 8)$ entry ‘01000000’ is evaluated. Since the input code is smaller than the 8th entry, the next entry compared is the fourth $((6+7)/2 = 7)$ one. Since the input code is equal to the 7th entry, the seventh entry corresponding to S_3 is the result. In the following, the length of the codeword must

be derived to decide the starting bit of the next input code. The length can be calculated by subtracting the next interval to the matching one. If the input is totally equal to matching one, then no need to calculate next input code. Otherwise we have to repeat our algorithm and again calculate the starting address of the input bit. This extra calculation can eliminate the storage for the length of the codeword.

The detailed algorithm is listed below.

1. Decoding algorithm:
2. Input: The constructed n-entry interval array and the input code
3. Output: The symbol of the matching interval
4. Decode (code, start, end) BEGIN
5. Mid = [start+end/2];
6. IF (int [mid]>code) BEGIN
7. IF (mid = start+1)
8. Return start
9. Decode (code, start, mid-1)
10. END
11. ELSE BEGIN
12. IF (end = mid+1)
13. Return mid
14. Decode (code, mid, end)
15. END
16. END

4 CONCLUSIONS

The main contribution of this study is exploiting the property implied in the Huffman tree to simplify the representation of the Huffman tree and the decoding procedure. Moreover our algorithm can also be parallelized easily. We already showed that representation of Huffman Tree using Ternary tree is more beneficial than representation of Huffman Tree using Binary tree. The proposed data structure does not need any branch or pointer, thus is very compact as compared with the existing schemes.

ACKNOWLEDGEMENTS

The author Madhu Goel would like to thank Kurukshetra University Kurukshetra for providing me University Research Scholarship & support of Kurukshetra Institute of Technology & Management (KITM).

REFERENCES

1. BENTLEY, J. L., SLEATOR, D. D., TARJAN, R. E., AND WEI, V. K. A locally adaptive data compression scheme. Commun. ACM 29,4 (Apr. 1986), 320-330.
2. Chen, H.C., Y.L. Wang and Y.F. Lan, 2005, "A memory efficient and fast Huffman decoding algorithm." Proceedings of the 19th International Conference 2005 IEEE. 69: 119-122.scialert.net/fulltext/?doi=itj.2007.776.779
3. DAVID A. HUFFMAN, Sept. 1991, profile Background story: Scientific American, pp. 54-58
4. ELIAS, P. Interval and regency-rank source coding: Two online adaptive variable-length schemes. IEEE Trans. Inj Theory. To be published.
5. FALLER, N. An adaptive system for data compression. In Record of the 7th Asilomar Conference on Circuits, Systems, and Computers. 1913, pp. 593-591.
6. GALLAGER, R. G. Variations on a theme by Huffman. IEEE Trans. Inj Theory IT-24, 6 (Nov.1978), 668-674.
7. Hashemain, "memory efficient and high-speed search Huffman Coding" IEEE Trans. Communication 43(1995) pp. 2576-2581.
8. Hu, Y.C. and Chang, C.C., "A new losseless compression scheme based on Huffman coding scheme for image compression".
9. KNUTH, D. E, 1997. The Art of Computer Programming, Vol. 1: Fundamental Algorithms, 3rd edition. Reading, MA: Addison-Wesley, pp. 402-406
10. KNUTH, D. E. Dynamic Huffman coding. J. Algorithms 6 (1985), 163-180.
11. MacKay, D.J.C., *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003.
12. MCMASTER, C. L. Documentation of the compact command. In UNIX User's Manual, 4.2Berkeley Software Distribution, Virtual VAX- I Version, Univ. of California, Berkeley, Berkeley,
13. Calif, Mar. 1984. ,
14. PUSHPA R. SURI & MADHU GOEL, Ternary Tree & A Coding Technique, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.9, September 2008
15. PUSHPA R. SURI & MADHU GOEL, Ternary Tree & FGK Huffman Coding Technique, IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.1, January 2009
16. PUSHPA R. SURI & MADHU GOEL, A NEW APPROACH TO HUFFMAN CODING, Journal of Computer Science. VOL.4 ISSUE 4 Feb. 2010 .
17. ROLF KLEIN, DERICK WOOD, 1987, on the path length of Binary Trees, Albert-Lapwings University at Freeburg.
18. ROLF KLEIN, DERICK WOOD, 1988, On the Maximum Path Length of AVL Trees, Proceedings of the 13th Colloquium on the Trees in Algebra and Programming, p. 16-27, March 21-24.
19. SCHWARTZ, E. S. An Optimum Encoding with Minimum Longest Code and Total Number of Digits. If: Control 7, 1 (Mar. 1964), and 37-44.
20. TATA MCGRaw HILL, 2002 theory and problems of data structures, Seymour lipshutz, tata McGraw hill edition, pp 249-255

21. THOMAS H. CORMEN, 2001 Charles e. leiserson, Ronald l. rivest, and clifford stein.
22. Thomas H.Cormen Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. Introduction to algorithms, Second Edition. MIT Press and McGraw-Hill, 2001. Section 16.3, pp. 385–392.

Dr. Pushpa Suri is a reader in the department of computer science and applications at Kurukshetra University Haryana India. She has supervised a number of Ph.D. students. She has published a number of research papers in national and international journals and conference proceedings.



Mrs. Madhu Goel has Master's degree (University Topper) in Computer Science. At present, she is pursuing her Ph.D. and working as Lecturer in Kurukshetra Institute of Technology & Management (KITM), Kurukshetra University Kurukshetra. Her area of research is Algorithms and Data Structure where she is working on Ternary tree structures. She has published a number of research papers in national and international journals.

IJCSI CALL FOR PAPERS MAY 2011 ISSUE

Volume 8, Issue 3

The topics suggested by this issue can be discussed in term of concepts, surveys, state of the art, research, standards, implementations, running experiments, applications, and industrial case studies. Authors are invited to submit complete unpublished papers, which are not under review in any other conference or journal in the following, but not limited to, topic areas. See authors guide for manuscript preparation and submission guidelines.

Accepted papers will be published online and indexed by Google Scholar, Cornell's University Library, DBLP, ScientificCommons, CiteSeerX, Bielefeld Academic Search Engine (BASE), SCIRUS, EBSCO, ProQuest and more.

Deadline: 31st March 2011

Notification: 30th April 2011

Revision: 10th May 2011

Online Publication: 31st May 2011

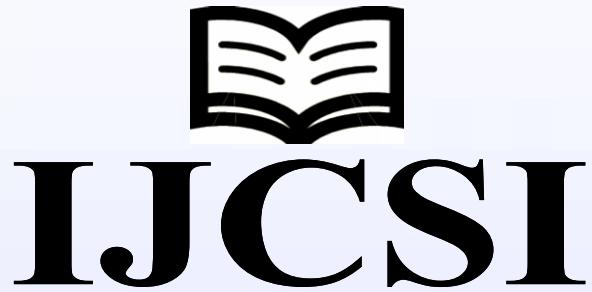
- Evolutionary computation
- Industrial systems
- Evolutionary computation
- Autonomic and autonomous systems
- Bio-technologies
- Knowledge data systems
- Mobile and distance education
- Intelligent techniques, logics, and systems
- Knowledge processing
- Information technologies
- Internet and web technologies
- Digital information processing
- Cognitive science and knowledge agent-based systems
- Mobility and multimedia systems
- Systems performance
- Networking and telecommunications
- Software development and deployment
- Knowledge virtualization
- Systems and networks on the chip
- Context-aware systems
- Networking technologies
- Security in network, systems, and applications
- Knowledge for global defense
- Information Systems [IS]
- IPv6 Today - Technology and deployment
- Modeling
- Optimization
- Complexity
- Natural Language Processing
- Speech Synthesis
- Data Mining

For more topics, please see <http://www.ijcsi.org/call-for-papers.php>

All submitted papers will be judged based on their quality by the technical committee and reviewers. Papers that describe on-going research and experimentation are encouraged.

All paper submissions will be handled electronically and detailed instructions on submission procedure are available on IJCSI website (www.IJCSI.org).

For more information, please visit the journal website (www.IJCSI.org)



The International Journal of Computer Science Issues (IJCSI) is a well-established and notable venue for publishing high quality research papers as recognized by various universities and international professional bodies. IJCSI is a refereed open access international journal for publishing scientific papers in all areas of computer science research. The purpose of establishing IJCSI is to provide assistance in the development of science, fast operative publication and storage of materials and results of scientific researches and representation of the scientific conception of the society.

It also provides a venue for researchers, students and professionals to submit ongoing research and developments in these areas. Authors are encouraged to contribute to the journal by submitting articles that illustrate new research results, projects, surveying works and industrial experiences that describe significant advances in field of computer science.

Indexing of IJCSI

1. Google Scholar
2. Bielefeld Academic Search Engine (BASE)
3. CiteSeerX
4. SCIRUS
5. Docstoc
6. Scribd
7. Cornell's University Library
8. SciRate
9. ScientificCommons
10. DBLP
11. EBSCO
12. ProQuest