IJCSIS Vol. 15 No. 11 (I), November 2017
ISSN 1947-5500

International Journal of Computer Science & Information Security

Pennsylvania, USA

Indexed and technically co-sponsored by :

IJCSIS

ISSN (online): 1947-5500

Please consider to contribute to and/or forward to the appropriate groups the following opportunity to submit and publish original scientific results.

CALL FOR PAPERS
International Journal of Computer Science and Information Security (IJCSIS)
January-December 2017 Issues

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 Authors are invited to submit complete unpublished papers, which are not under review in any other conference or journal in the following, but not limited to, topic areas.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.

Indexed by Google Scholar, DBLP, CiteSeerX, Directory for Open Access Journal (DOAJ), Bielefeld Academic Search Engine (BASE), SCIRUS, Scopus Database, Cornell University Library, ScientificCommons, ProQuest, EBSCO and more.

Deadline: see web site
Notification: see web site
Revision: see web site
Publication: see web site

Context-aware systems
Networking technologies
Security in network, systems, and applications
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
Knowledge for global defense
Information Systems [IS]
IPv6 Today - Technology and deployment
Modeling
Software Engineering
Optimization
Complexity
Natural Language Processing
Speech Synthesis
Data Mining

For more topics, please see web site https://sites.google.com/site/ijcsis/

For more information, please visit the journal website (https://sites.google.com/site/ijcsis/)

Editorial
Message from Editorial Board

*It is our great pleasure to present the **November 2017 issue** (Volume 15 Number 11, Part I & II) of the **International Journal of Computer Science and Information Security** (IJCSIS). High quality research, survey & review articles are proposed from experts in the field, promoting insight and understanding of the state of the art, and trends in computer science and technology. It especially provides a platform for high-caliber academics, practitioners and PhD/Doctoral graduates to publish completed work and latest research outcomes. According to Google Scholar, up to now papers published in IJCSIS have been cited over 9775 times and the number is quickly increasing. This statistics shows that IJCSIS has established the first step to be an international and prestigious journal in the field of Computer Science and Information Security. There have been many improvements to the processing of papers; we have also witnessed a significant growth in interest through a higher number of submissions as well as through the breadth and quality of those submissions. IJCSIS is indexed in major academic/scientific databases and important repositories, such as: Google Scholar, Thomson Reuters, ArXiv, CiteSeerX, Cornell's University Library, Ei Compendex, ISI Scopus, DBLP, DOAJ, ProQuest, ResearchGate, Academia.edu and EBSCO among others.*

On behalf of IJCSIS community and the sponsors, we congratulate the authors and thank the reviewers for their outstanding efforts to review and recommend high quality papers for publication. In particular, we would like to thank the international academia and researchers for continued support by citing papers published in IJCSIS. Without their sustained and unselfish commitments, IJCSIS would not have achieved its current premier status.

*"We support researchers to succeed by providing high visibility & impact value, prestige and excellence in research publication." For further questions or other suggestions please do not hesitate to contact us at **ijcsiseditor@gmail.com**.*

A complete list of journals can be found at:
http://sites.google.com/site/ijcsis/

IJCSIS Vol. 15, No. 11, November 2017 Edition

ISSN 1947-5500 © IJCSIS, USA.

Journal Indexed by (among others):

Bibliographic Information
ISSN: 1947-5500
Monthly publication (Regular Special Issues)
Commenced Publication since May 2009

Editorial / Paper Submissions:
IJCSIS Managing Editor
(ijcsiseditor@gmail.com)
Pennsylvania, USA
Tel: +1 412 390 5159

IJCSIS EDITORIAL BOARD

IJCSIS Editorial Board	IJCSIS Guest Editors / Associate Editors
Dr. Shimon K. Modi [Profile] Director of Research BSPA Labs, Purdue University, USA	**Dr Riktesh Srivastava [Profile]** Associate Professor, Information Systems, Skyline University College, Sharjah, PO 1797, UAE
Professor Ying Yang, PhD. [Profile] Computer Science Department, Yale University, USA	**Dr. Jianguo Ding [Profile]** Norwegian University of Science and Technology (NTNU), Norway
Professor Hamid Reza Naji, PhD. [Profile] Department of Computer Enigneering, Shahid Beheshti University, Tehran, Iran	**Dr. Naseer Alquraishi [Profile]** University of Wasit, Iraq
Professor Yong Li, PhD. [Profile] School of Electronic and Information Engineering, Beijing Jiaotong University, P. R. China	**Dr. Kai Cong [Profile]** Intel Corporation, & Computer Science Department, Portland State University, USA
Professor Mokhtar Beldjehem, PhD. [Profile] Sainte-Anne University, Halifax, NS, Canada	**Dr. Omar A. Alzubi [Profile]** Al-Balqa Applied University (BAU), Jordan
Professor Yousef Farhaoui, PhD. Department of Computer Science, Moulay Ismail University, Morocco	**Dr. Jorge A. Ruiz-Vanoye [Profile]** Universidad Autónoma del Estado de Morelos, Mexico
Dr. Alex Pappachen James [Profile] Queensland Micro-nanotechnology center, Griffith University, Australia	**Prof. Ning Xu,** Wuhan University of Technology, China
Professor Sanjay Jasola [Profile] Gautam Buddha University	**Dr . Bilal Alatas [Profile]** Department of Software Engineering, Firat University, Turkey
Dr. Siddhivinayak Kulkarni [Profile] University of Ballarat, Ballarat, Victoria, Australia	**Dr. Ioannis V. Koskosas,** University of Western Macedonia, Greece
Dr. Reza Ebrahimi Atani [Profile] University of Guilan, Iran	**Dr Venu Kuthadi [Profile]** University of Johannesburg, Johannesburg, RSA
Dr. Dong Zhang [Profile] University of Central Florida, USA	**Dr. Zhihan lv [Profile]** Chinese Academy of Science, China
Dr. Vahid Esmaeelzadeh [Profile] Iran University of Science and Technology	**Prof. Ghulam Qasim [Profile]** University of Engineering and Technology, Peshawar, Pakistan
Dr. Jiliang Zhang [Profile] Northeastern University, China	**Prof. Dr. Maqbool Uddin Shaikh [Profile]** Preston University, Islamabad, Pakistan
Dr. Jacek M. Czerniak [Profile] Casimir the Great University in Bydgoszcz, Poland	**Dr. Musa Peker [Profile]** Faculty of Technology, Mugla Sitki Kocman University, Turkey
Dr. Binh P. Nguyen [Profile] National University of Singapore	**Dr. Wencan Luo [Profile]** University of Pittsburgh, US
Professor Seifeidne Kadry [Profile] American University of the Middle East, Kuwait	**Dr. Ijaz Ali Shoukat [Profile]** King Saud University, Saudi Arabia
Dr. Riccardo Colella [Profile] University of Salento, Italy	**Dr. Yilun Shang [Profile]** Tongji University, Shanghai, China
Dr. Sedat Akleylek [Profile] Ondokuz Mayis University, Turkey	**Dr. Sachin Kumar [Profile]** Indian Institute of Technology (IIT) Roorkee

ISSN 1947 5500

TABLE OF CONTENTS

1. PaperID 31101704: Extended Multi Queue Job Scheduling in Cloud (pp. 1-8)

Neeraj Kumar Pandey, Research Scholar, Uttaranchal University, Dehradun Uttarakhand
Sumit Chaudhary, Research Scholar, Uttaranchal University, Dehradun Uttarakhand
N. K. Joshi, Director, Uttaranchal Institute of Technology, Uttaranchal University, Dehradun Uttarakhand

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

2. PaperID 31101705: Analyzing and Predicting Hash tags Popularity on Twitter (pp. 9-16)

Katarla Krishna Vasu (1) & Dr. Vedula Venkateswara Rao (2)
(1) Dept of CSE, Sri Vasavi Engineering College Pedatadepalli, Tadepalligudem, A.P, India
(2) Associate Professor, Dept of CSE, Sri Vasavi Engineering College Pedatadepalli, Tadepalligudem, A.P, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

3. PaperID 31101706: Invisible Digital Watermarking Methodology for Image Validation (pp. 17-22)

Dr. V. Kakulapati, SNIST, Hyderabad, India
Dr. SR Kattamuri, SNIST, Hyderabad, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

4. PaperID 31101707: Designing an Efficient Distributed Algorithm for Big Data Analytics: Issues and Challenges (pp. 23-28)

Mohammed S. Al-kahtani, Dept. of Computer Engg., Prince Sattam bin Abdulaziz University, Saudi Arabia
Lutful Karim, School of ICT, Seneca College of Applied Arts & Technology, Toronto, Canada

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

5. PaperID 31101711: A Survey on Memory Virtualization in Cloud (pp. 29-34)

P. V. S. S. Gangadhar, Scientist-D, Research Scholar, NIC, MEITY, Govt of India, Gitam University, Visakhapatnam, Andhra Pradesh, India
Dr. Ashok Kumar Hota, Scientist-F, NIC, MEITY, Govt of India
Dr. Mandapati Venkateswara Rao, Professor, Dept of IT, Gitam Institute of Technology, Gitam University, Visakhapatnam, Andhra Pradesh, India
Dr. Vedula Venkateswara Rao, Professor, Dept of CSE, Sri Vasavi Engineering College, Tadepalligudem, Andhra Pradesh, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

6. PaperID 31101715: A Comparative Study on MAC Protocols for Wireless Sensor Networks on Energy Reduction (pp. 35-40)

Rajan Sharma, Research Scholar, Department of Electronics and Communication Engineering, I.K. Gujral Punjab Technical University, Jalandhar, Punjab, India
Balwinder Singh Sohi, Department of Electronics and Communication, CGC Group of Colleges, Punjab, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

7. PaperID 31101717: Performance Analysis of Adaptive Routing Protocol for Cognitive Radio Wireless Sensor Networks using Bio-inspired Methods (pp. 41-49)

(1) Amit N. Thakare, (2) Dr. Latesh Malik (Bhagat), (3) Mrs. Achamma Thomas
(1, 3) Department of Computer Science & Engineering, G.H. Raisoni College of Engineering, Nagpur, MS, India
(2) Department of Computer Science & Engineering, Government College of Engineering, Nagpur, MS, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

8. PaperID 31101722: Application of the ORBAC Model in the Context of Physical Access Control (pp. 50-63)

Belbergui Chaimaa, STIC Laboratory, Chouaib Doukkali University, El jadida, Morocco
Elkamoun Najib, STIC Laboratory, Chouaib Doukkali University, El jadida, Morocco
Rachid Hilal, STIC Laboratory, Chouaib Doukkali University, El jadida, Morocco

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

9. PaperID 31101723: Clustering Algorithms: A Review (pp. 64-76)

*Nikitha Johnsirani Venkatesan & Dong Ryeol Shin**
School of Electrical and Computer Engineering, Sungkyunkwan UniversitySuwon, South Korea

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

10. PaperID 31101725: Implementation Challenges in Campus Network Security (pp. 77-81)

Zaka Ullah, Department of Computer Science, Lahore Garrison University, Lahore, Pakistan
Muhammad Zulkifl Hasan, Department of Computer Science, Lahore Garrison University, Lahore, Pakistan

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

11. PaperID 31101728: Improving the Data Quality in the Research Information Systems (pp. 82-86)

Otmane Azeroual (abc),
(a) German Center for Higher Education Research and Science Studies (DZHW), Schützenstraße 6a, Berlin, 10117, Germany
(b) Otto-von-Guericke-University Magdeburg, Department of Computer Science, Institute for Technical and Business Information Systems Database Research Group, P.O. Box 4120; 39106 Magdeburg, Germany
(c) University of Applied Sciences HTW Berlin, Study Program Computational Science and Engineering, Wilhelminenhofstraße 75A, 12459 Berlin, Germany

Mohammad Abuosba (c),
(c) University of Applied Sciences HTW Berlin, Study Program Computational Science and Engineering, Wilhelminenhofstraße 75A, 12459 Berlin, Germany

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

12. PaperID 31101733: One Hop Forwarding Technique for QOS Routing with an Improved Fault Tolerant Clustering Mechanism for Multimedia Sensor Networks (pp. 87-100)

(1) R. Guru, (2) Dr. Siddaraju, (3) Ananda babu
(1) Research scholar, (2) Professor, (3) Assistant Professor
(1) Sri Jayachamarajendra College of Engineering, Mysuru, India.
(2) Dr. Ambedkar Institute of Technology, Bangalore, India
(3) Kalpataru Institute of Engineering, Tiptur, India.

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

13. PaperID 31101736: Constructions of Design Elements from Software Requirement Specifications (SRS) (pp. 101-105)

Syed Naimatullah Hussain, CIT, Taif University, Taif, KSA

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

14. PaperID 31101737: An Efficient Way of Medical Image Encryption using Cat Map and Chaotic Logistic Function (pp. 106-115)

Ranu Gupta (1), Rahul Pachauri (2), Ashutosh K. Singh (3)*
(1, 2) Jaypee University of Engineering and Technology, Raghogarh, Guna (M.P.) India, 473226
(3) Thapar Institute of Engineering and Technology University, Patiala, 147004, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

15. PaperID 31101741: The Methodology of Data Collecting and the Real Time Synchronization in ETL (pp. 116-122)

Zijadin Krasniqi (1), Enver Ahmeti (2), Adriana Gjonaj (3)
(1) Ph.D. in Information Systems, Information Technology, Pristina, Kosovo.
(2) M.sc in Information Technology, Pristina, Kosovo.
(3) Professor emeritus, European University of Tirana, Albania

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

16. PaperID 31101752: Conceptual Model of Intelligent Collaborative Educational System: Possible Solutions (pp. 123-128)

Sabina Katalnikova, Faculty of Computer Science and Information Technology, Riga Technical University, Riga, Latvia
Leonids Novickis, Faculty of Computer Science and Information Technology, Riga Technical University, Riga, Latvia

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

17. PaperID 31101754: Develop a Distributed Intrusion Detection System based on Cloud Computing (pp. 129-132)

Khoi Nguyen, Datic Lab, DaNang University, DaNang, Viet Nam
Trang Dang Thi, IT Faculty, PhamVanDong University, QuangNgai, VietNam

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

18. PaperID 31101755: A comprehensive approach on RLE and ECC (Elliptical Cure Cryptography) using Mean Square Error (MSE) feature (pp. 133-138)

Ayushi Mathur, Department of Computer Engineering, Government Women Engineering College, Ajmer, India
Dr. Varun Prakash Saxena, Department of Computer Engineering, Government Women Engineering College, Ajmer, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

19. PaperID 31101757: Enhanced Multi-Modal Biometric Based Security Scheme With Feature Based Machine Learning Approach (pp. 139-143)

Rakhi Choudhary, Er. Krishan Kumar, Dr. Himanshu Monga
Department of Computer Science Engineering, Jan Nayak Chaudhary Devi Lal Memorial College of Engineering

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

20. PaperID 31101760: Evaluating Similarity using near Set Theory for Plastic Surgery Face Images (pp. 144-149)

Sujata G. Bhele, Electronics Engg, Priyadarshini College of Engg, Nagpur, India
Vijay H. Mankar, Electronics and Telecommunication Engg, Government Polytechnic, Nagpur, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

21. PaperID 31101761: Face Image Recognition Based on Linear Discernment Analysis and Cuckoo Search Optimization with SVM (pp. 150-165)

Jamal Mustafa AL-Tuwaijari & Suhad Ibrahim Mohammed
Department of Computer Science - College of Science - University of Diyala – Iraq

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

22. PaperID 31101764: Generic Synthesis System Based on Agile Methodology for Multimedia Mobile Web Learning Modules (pp. 166-171)

Linda lafta Gashim, Department of Computer Science, College of Science / Mustansiriyah University, Baghdad, Iraq
Assist. Prof. Dr. Karim Qasim Hussein, Department of Computer Science, College of Science / Mustansiriyah University, Baghdad, Iraq

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

23. PaperID 31101765: Data Mining Methods for Worm Detection Using Variable Length Instruction Sequences (pp. 172-183)

Muazzam Ahmed Siddiqui & Safaa Alkatheri
Department of Information Systems, Faculty of Computing & Information Technology, King Abdulaziz University

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

24. PaperID 31101768: Tuning of Canny Image Edge Detection (pp. 184-187)

Jamil A. M. Saif,
(1) Hodeidah University, Hodeidah, Yemen
(2) University of Bisha, Bisha KSA

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

25. PaperID 31101770: The Necessity of Developing a Standard for Exchanging a Chain of Custody of Digital Evidence Data (pp. 188-191)

Jasmin Cosic, IT Section of Police Administration, MoI of Una-sana canton, 502.V.bbr, Bihac, Bosnia and Herzegovina
Miroslav Baca, Faculty of Organization and Informatics, University of Zagreb, Pavlinska 2, 42 000 Varaždin, Croatia
Petra Grd, Faculty of Organization and Informatics, University of Zagreb, Pavlinska 2, 42 000 Varaždin, Croatia

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

26. PaperID 31101772: Survey of Research Challenges in Cyber Physical Systems (pp. 192-199)

Swati Nikam, Research Scholar, DIT, Pune, India
Dr. Rajesh Ingle, Sr. Member, IEEE

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

27. PaperID 31101773: Detecting Money Laundering in a Financial System Based on Genetic Algorithm (pp. 200-208)

Ramadan Mahmood Ramo, Department of Management Information systems, University of Mosul
Prof. Dr. Khalil Ibrahim Alsaif, Department of Computer, University of Mosul

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

28. PaperID 31101778: Design of a New Small Antenna for Passive UHF RFID Tags (pp. 209-214)

Ines Frigui, Mohamed Salah Karoui, Hamadi Ghariani, Mongi Lahiani
Laboratory of Electronics and Technologies of Information (LETI)
National School of Engineers of Sfax (ENIS), Sfax, Tunisia

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

29. PaperID 31101780: Efficient Multi-Level Authentication for Cloud API based on RestPL (pp. 215-222)

M. J. Balachandran & Dr. P. Sujatha
School of Computing Sciences, Vels University, Chennai, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

30. PaperID 31101781: Bitmap Indexes for Faster Query Execution (pp. 223-228)

Samadhi P. Kumarasinghe, Nishadi D. Kirielle, Malmee Weerasinghe, Sasini Madhumali, Amal S. Perera, Sachini Herath, Amila De Silva, Lasantha Fernando
Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

31. PaperID 31101783: Optical Flow for Robot Navigation (pp. 229-237)

Nixon Adu-Boahen, Department of Computer Science, KNUST, Kumasi, Ghana
J. B. Hayfron-Acquah, Department of Computer Science, KNUST, Kumasi, Ghana
J. K. Panford, Department of Computer Science, KNUST, Kumasi, Ghana

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

32. PaperID 31101789: Dataset of Graphs and Sub-graphs: Storage Representation and its Graphical form (pp. 238-245)

Rachna Somkunwar (1) & Dr. Vinod M. Vaze (2)
(1) JJT University, Research scholar, Rajasthan, India
(2) JJT University, Assistant Professor, Rajasthan, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

33. PaperID 31101790: Fingerprint Image Retrieval using Statistical Methods (pp. 246-249)

Sudhir Vegad, Department of Information Technology, A D Patel Institute of Technology, New V V Nagar, India
Dr. Tanmay Pawar, Electronics Department, BVM Engineering College, Vallabh Vidyanagar, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

34. PaperID 31101791: Proactive Detection of Catastrophe Trends for Rescheduling Real-Time Systems with Scenario Shift (pp. 250-268)

A. Christy Persya (1), T. R. Gopalakrishnan Nair (2)
(1) Advanced Real-Time Computing Group, RRGI Research Centre, Bangalore, India.
(2) Advanced Real-Time Computing Group, RRGI Research Centre, Rector RRGI & Visiting Professor, NIAS, Bangalore, India.

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

35. PaperID 31101792: Big Data Analytics Framework for Peer-To-Peer Botnet Detection Using Random Forest and Deep Learning (pp. 269-277)

Saraladevi D., PG Student, Department of Computer Science and Engineering, Pondicherry Engineering College
Sathiyamurthy K., Associate Professor, Department of Computer Science and Engineering, Pondicherry Engineering College
Vijayaprabakaran K., PhD Scholar, Department of Computer Science and Engineering, Pondicherry Engineering College

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

Extended Multi Queue Job Scheduling in Cloud

Neeraj Kumar Pandey
Research Scholar, Uttaranchal University, Dehradun Uttarakhand
Email: dun.neeraj@gmail.com

Sumit Chaudhary
Research Scholar, Uttaranchal University, Dehradun Uttarakhand
Email:iimtsumit@gmail.com

N.K.Joshi
Director, Uttaranchal Institute of Technology, Uttaranchal University,
Dehradun Uttarakhand
Email: nkjoshi2001@yahoo.com

Abstract:

Cloud computing refers to a 'Utility Computing' where the services can be used as per the requirements and the demand can be expanded and contracted as per the business needs. One of the major features provides the facility to pay as per the use. Scheduling is a major challenge in any cloud environment where multiple tasks are arriving at same time and CSP has to manage them pretending the unlimited resources for any request. While cloud service provider have limited amount of resources to fulfill all request. Min min algorithm used to minimize the make span time of task execution. Each algorithm is designed in such a way that by using it CSP must achieve user satisfaction. In this paper an improved economy based scheduling strategy is proposed after analyzing traditional algorithm which is based on task length, priority and cost of execution. These parameters are considered before the allocation of the resources so this strategy shows the considerable improvement in the existing strategies.

Keywords: Cloud Service Provider (CSP), Resource Allocation Strategy, Cloud computing, Task priority, Task length.

Introduction: A cloud is a technology where many users are connected to many systems through a network. A scalability property allows the user to demand any service at any time without bothering about the availability of resource to fulfill the corresponding request. So this situation creates a great challenge of managing the resources so before the CSP in such a way so that the entire request must be obeyed within time constraints.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Resource Scheduling is a one of the major challenge in cloud computing. There are numerous strategies used by the CSP depending upon the situation or demand of the user task. When a task is assigned to execute the data center will take care of task that how it will execute. The following figure will illustrate the basic architecture of assigning any task to data center. The tasks are submitted by different user to the datacenter broker. The datacenter broker acts like a mediator between the user and data center.

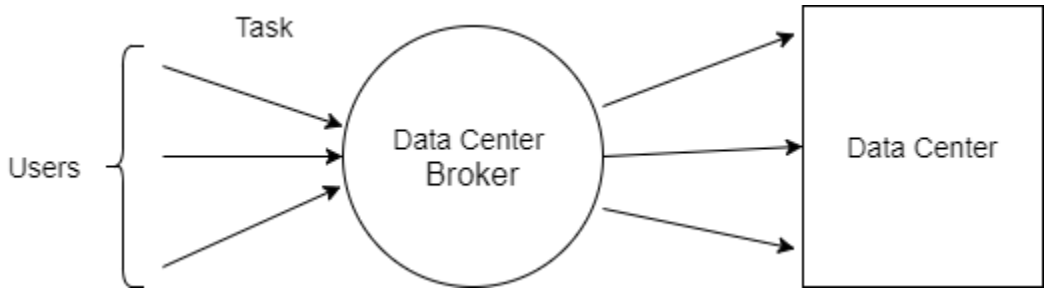


Figure.1

Min min strategy considers the task length as the level for executing the assignments. An ideal condition happens when undertaking with least length is executed first. With a specific end goal to accomplish this ideal condition datacenter intermediary ought to consider this factor. But on the other hand the order of execution of the task may differ if the other parameter like priority and the cost of the demanded resource to execute the task is considered. The main idea about the proposed approach is to consider the both criteria i.e. priority of the task and resource cost to execute the task.

2. Related work:

A lot of resource scheduling algorithm is available and used now days. The selection of scheduling algorithm is depends on the parameters it works. But the performance is always an issue to analyze. Three parameters are considered in the proposed approach i.e size of the task, priority of the task and cost of execution of task. Many researchers are working with the study area multi queue scheduling algorithm [2], OCRP [3], Activity based costing [4][5], Simulated Annealing (SA) [6]. A lot of survey [7][8] have been done to identify the characteristic of different scheduling strategies so that the optimized scheduling method can be found.

2.1 Priority based: priority is a major concern in job scheduling algorithms. There are numerous algorithms which performs job scheduling by considering priority as a major factor. Analytical hierarchy process(AHP)[9] uses multi criteria decision making and multi-attribute decision making models to evaluate the priority of the task.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

2.2 Cost Based Resource Scheduling: this strategy is based on the cost of the allocation of resource to any request received by the CSP. CSP calculates the cost by considering the resource cost and the performance of the requested resource cost. If the execution time is increases by allocating the resource then it will increase the execution cost which will be combined with the resource cost. Then finally the cost of allocation of resource gets increased. Paper [10] calculates the cost by reservation and demand strategy in which resources are reserved as per the demand generated in the future.CSP can change the resource before it is consumed by the user.

2.3 Energy efficiency Based: A decentralized architecture of energy aware resource management is proposed in the paper [15]. A problem is defined to minimize the energy consumption of resource allocation policy in the cloud data center. It will produce the environment friendly strategy to manage the energy efficient data center development concept.

2.4 Non Preemption Based: to implement the non-preemptive (Parallel) Scheduling Naphele scheduler has been introduced in the paper [12]. Naphele calculated the critical time (maximum time allowed to take by any process) so that it the execution reaches the critical time it will be suspended and added to the waiting queue. Naphele [11] provided a framework for the parallel data processing in the cloud which completely exploits the dynamic resource allocation.

2.5 Heuristic Based: in this paper [13] a heuristic based cost optimization algorithm is proposed. Heuristic provides optimal solution of resource allocation problem under some restriction. Although resource provisioning is a NP hard problem but heuristic based solution provides optimum result in various conditions.

2.6 Reliability Based: Minimum failure resource allocation (MFRA)[14] problem have been discussed for requested VMs residing in the multiple data centers located in different geographical locations. MFRA is a NP complete problem but ILP framework is proposed to provide solution of small scale problem of distributed cloud. Two heuristic algorithms are proposed as a large scale topology which gives a optimal result in a small scale topology.

2.7 Replication Based: in this strategy cloud is divided into sub cloud and submitted job is replicated to all sub cloud following different resource provision algorithm. Each sub cloud executed the same replicated job. If any sub cloud finishes the job it will broadcast the finish message to all other sub cloud

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

to stop execution. It is not a feasible solution of resource provisioning because of wastage of time and resource use in the partial execution of the job by sub cloud.

3. Proposed Approach

Jobs having equal priority or no economic urgency will be divided into two queues based on burst time. These queues will execute the jobs from both queues one by one as seen in figure 2[16][8]. But if the jobs come with different priority then queue selector will put them into a priority queue which will execute the jobs according their priority (preemptive method). Dynamic job selector will select the job dynamically between the two different types of queue for the cloud environment. The order of execution of jobs is:

1) For equal priority jobs

There are two types of queue one for smaller burst time i.e.70% of total job and other is large burst time i.e.30% of total jobs. So the order of execution is

P1→P2→P3→P4→P5→P6→P7→P8

2) For different priority jobs:

P2→p1→p3→p4→p8→p7→p5→p6

Pseudo code for priority scheduling:

1. For all T_s(Task Submitted)
2. Find the priority of submitted task.
3. Maintain the ready queue based on newly arrived job
4. For all newly arrived job

   If (priority of new jobT_n> Priority of job in execution T_e)

   $T_e=T_n$
5. Run the job queue using priority scheduling
6. Else
7. Divide the jobs according their burst time and use traditional Scheduling

Input Data

Job ID	P1	P2	P3	P4	P5	P6	P7	P8
Burst Time(ms)	15	88	85	42	35	20	92	72

Table.1(FCFS)

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Job ID	P1	P6	P5	P4	P8	P3	P2	P7
Burst Time(ms)	15	20	35	42	72	85	88	92

Table.2 (SJF)

Job ID	P1	P4	P5	P2	P3	P6	P8	P7
Burst Time(ms)	15	42	35	88	85	20	72	92

Table.3 (CBA)

Job ID	P1	P8	P6	P3	P5	P2	P4	P7
Burst Time(ms)	15	72	20	85	35	88	42	72

Table.4 (MQS)

The priority based queue will give result based on their priority. The order of execution will be P1→P8→P6→P3→P5→P2→P4→P7

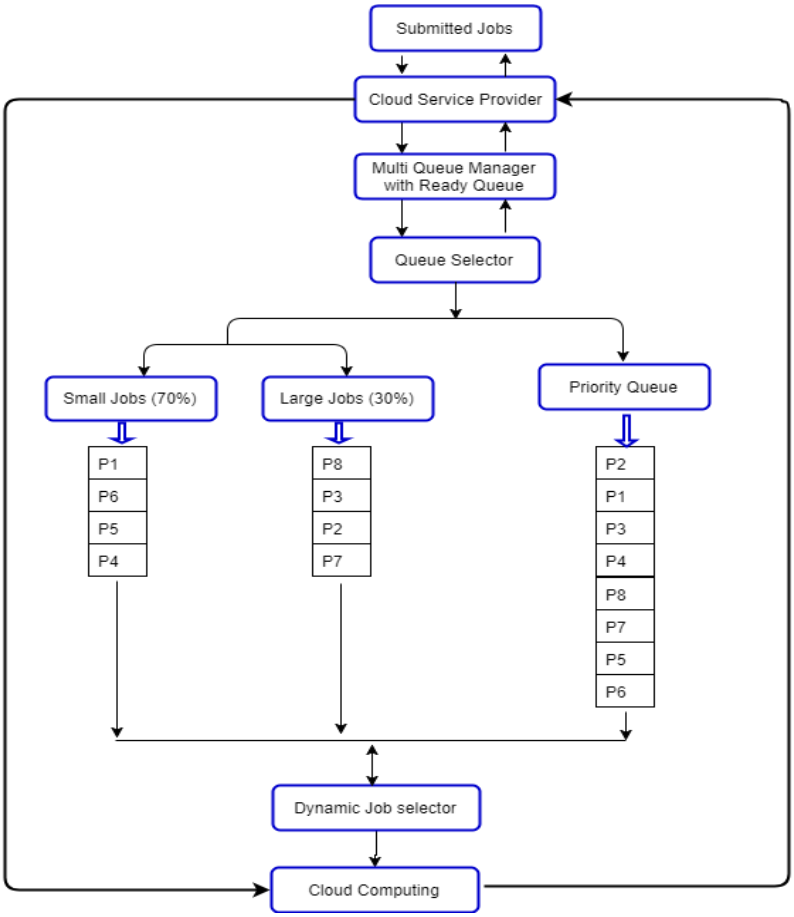
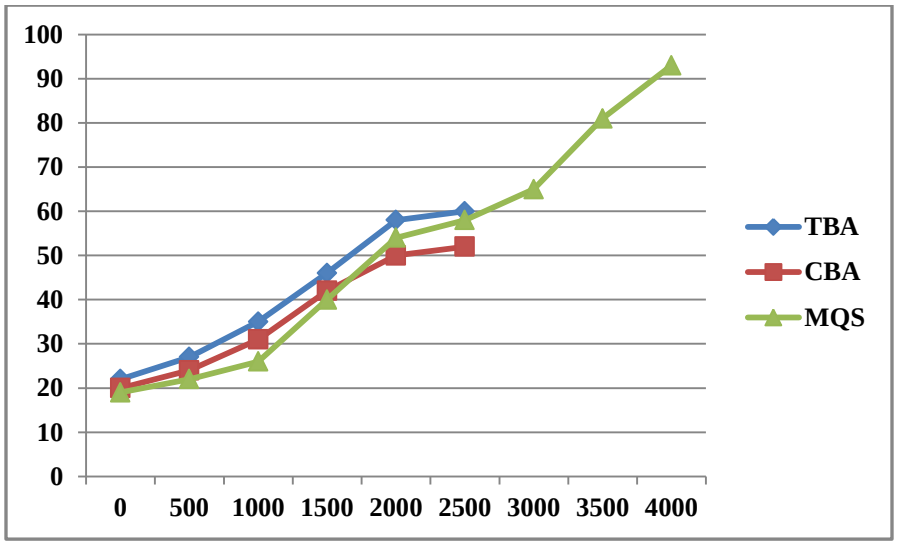


Figure.2

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Result and Discussion

The result of proposed multi queue scheduling strategy gives optimized result when compared to some well-known job scheduling strategy. The proposed approach divides the job into two parts based on their priority i.e. multi queue for equal priority queue for different priority process. It gives better result in real time systems. The proposed strategy provides optimum result for cloud environment and enhances the resource utilization in cloud environment. Cloud simulator is used to implement the proposed strategy which simulates the allocation of processing element (PE) which is shared to the cloudlet job. Cloudlet attains the PE depends on the number of jobs in the ready queue. When the finishes the execution it is removed from the queue (Execution Set). Cloudsim schedules all jobs depend on the current position of the job present in ready queue. The simulation of MQS and cloudlet execution into small queue P1, P6, P5, P4 and P8, P3, P2, P7 are queued. The scheduler takes one job from queue in every iteration for equal priority jobs and if priority is different the execution is maintained in different queue in which priority of job is calculated first. If any high priority jobs arrive then current executing job is preempted. It follows the priority with preemption policy for current execution set. It provides optimized result in cloud environment.



Graph 1: comparison of the proposed method with existing method

Conclusion and future scope

The scheduling methodology simply defining the efficient way of execution of job according the demand of the user. If users give some priority to the job then it is handled by priority scheduler but if it does not have any priority then it is considered as zero priority and put into a queue as per the burst time which is handled by traditional scheduler. The future of this method can add economy based queue after calculating the cost of the execution.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Reference:

[1] Antony Thomas, Krishnalal G, Jagathy Raj V P "Credit Based Scheduling Algorithm in Cloud Computing Environment" International Conference on Information and Communication Technologies (ICICT 2014) pp. 913-920.[2] AV.Karthick, Dr.E.Ramaraj, R.Ganapathy Subramanian "An efficient multi queue job scheduling for cloud computing" IEEE World Congress on Computing and Communication Technologies 2014 pp.164-166.

[3] Abhishek Vichare, Zenia P. Gomes, Noella Fernandes and Flavin Cardoza "Cloud computing using OCRP and virtual machines for dynamic allocation of resources" IEEE International Conference on Technologies for Sustainable Development (ICTSD-2015), Feb. 04 – 06, 2015, Mumbai, India.

[4] Qi cao,Zhi-bo Wei,Wen-mao Gong. An Optimized Algorithm for Task Scheduling Based On Activity Based Costing in Cloud Computing.

[5] J Blythe, S Jain, E Deelman, Y Gil, K Vahi. Task scheduling strategies for workflow-based applications in grids. Cluster Computing and the Grid; 2005.

[6] Diptangshu Pandit, Samiran Chattopadhyay, Matangini Chattopadhyay "Resource allocation in cloud using simulated annealing" IEEE Applications and Innovations in Mobile Computing (AIMoC) 2014).

[7] T. a1, A. Jaya Lakshmi 2, Vuyyuru K. Reddy "Resource Allocation Methods in Cloud Computing : Survey"international journal on emerging trend in technology ISSN: 2350 – 0808 September 2015 Volume 2 Issue 2 pp 416-419.

[8] Neeraj Kumar pandey, Sumit Chaudhary,N.K.Joshi "Resource allocation strategies used in cloud computing: A critical analysis"IEEE 2nd International Conference on Communication Control and Intelligent Systems (CCIS), 2016.

[9] Swachil J. Patel, Upendra R. Bhoi "Improved priority based job scheduling algorithm in cloud computing using iterative method" IEEE Fourth International Conference on Advances in Computing and Communications 2014 pp.199-202.

[10] Abhishek Vichare, Zenia P. Gomes, Noella Fernandes and Flavin Cardoza "Cloud computing using OCRP and virtual machines for dynamic allocation of resources" IEEE International Conference on Technologies for Sustainable Development (ICTSD-2015), Feb. 04 – 06, 2015, Mumbai, India.

[11] Warneke, Daniel, and Odej Kao. "Nephele: efficient parallel data processing in the cloud." Proceedings of the 2nd workshop on many-task computing on grids and supercomputers. ACM, 2009.

[12] Gandhali Upadhye,Truspti Dange "Cloud resource allocation as nonpreemptive approach" IEEE 2nd International Conference on Current Trends in Engineering and Technology(ICCTET)2014 pp.352-356.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[13] Sunirmal Khatua, Preetam Kumar Sur, Rajib Kumar Das,and Nandini Mukherjee “Heuristic-Based Resource Reservation Strategies for Public Cloud” IEEE transactions on cloud computing, vol. 4, no. 4, october-december 2016.

[14] Yi Zhu, Yan Liang, Qiong Zhang, Xi Wang, Paparao Palacharla and otoyoshi Sekiya “Reliable Resource Allocation with Weighted SRGs for Optically Interconnected Clouds” IEEE Global Communications Conference 2014 pp. 2186-2191.

[15] Anton Beloglazov and Rajkumar Buyya “Energy Efficient Resource Management in Virtualized Cloud Data Centers” 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing 2010 pp. 826-831.

[16] AV.Karthick, Dr.E.Ramaraj, R.Ganapathy Subramanian “An efficient multi queue job scheduling for cloud computing” IEEE World Congress on Computing and Communication Technologies 2014 pp.164-166.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Analyzing and Predicting Hash tags Popularity on Twitter

Krishna Vasu Katarla
M.Tech Final Year, Dept of CSE,
Sri Vasavi Engineering College
Pedatadepalli, Tadepalligudem, A.P, India
vasu.krish2@gmail.com

Dr.Vedula Venkateswara Rao
Associate Professor, Dept of CSE,
Sri Vasavi Engineering College
Pedatadepalli, Tadepalligudem, A.P, India
Venkatvedula2017@gmail.com

***Abstract*— Twitter and micro blogging services have became in dispensable sources of information on today's web. Understanding the main factors that make certain pieces of information spread quickly on these platforms can be deceived for the analysis of opinion formation. *Twitter* Data is one of the large amounts of sized data, because it is having a millions of tweets every day. It is one of the largest social media site. We are using this twitter data for the business purpose and industrial or social purpose according to our data requirement and processing the data. It is very large amount of sized data increasing every second that is known as big data. In today's world twitter is not only a social networks but also an increasingly interesting news important information broadcasting mechanisms and new media's that have many followers are not important information sources. Thus of becomes very important to understand news related propagation from news media called as super nodes. A useful practice in social media network analysis is to predict future popularity of event good platform to perform such analysis with twitters topic structure on mind, the problems can be stated as : "knowing current and previous tweet activity in future or can we predict if it will become more popular and if so how much. This understanding prediction benefits many applications such as social management advertisement and interaction optimization between media and followers.**

Keywords- micro blog, twitter, retreat, hash tag, super node, registration analysis, predictions

I. INTRODUCTION

Twitter, with its public discussion model, is a good platform to predict future popularity of a topic or event. Knowing current and previous tweet activity for a hash-tag (#) , we can predict if it became more prominent and trendy in the future and if yes by how much. In the times of information age, the magnitude of online social media activity has reached unprecedented level. Hundreds of millions of users spend hours online everyday to stay connected and communicate with the rest of world. Millions of users participate in these social networks of Social awareness streams. [19] People generate huge amount of data everyday on various social media networks, which in aggregate indicate the interests and current attention of the local and global communities. There are many events and topics discussed on Twitter. Some topics may get a lot attention and some may not. Some of these topics become very popular and focus of interests for large number of people. The connections and the nature of social network let information disseminate to large number of other people, a phenomena known as going "viral". These popular topics of discussions are also called "trends" in the social network. These trends are very dynamic and temporal in nature which exposes the expose the aggregate interests and attention of global and local communities. Trends in social networks are of high significance and a major point of interest in both the industry [19] and the research community [8, 15]. Many applications on web and business can be immensely benefitted from knowing what is currently \trending", which represents an answer to the age-old query what are people talking about? [9]. From stock exchange making real time decision to search engines delivering more updated, relevant search results.

Figures 1.1, 1.2 Twitter is one of the most popular social networking and micro-blogging service, which had more than 200 Million registered users by 2013, producing 400 Millions tweets everyday [17].

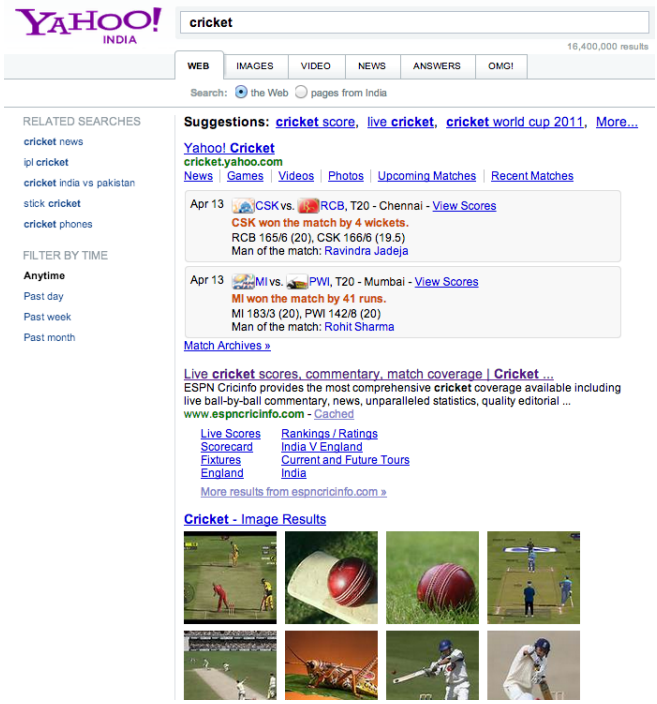


Fig 1.1 Trending topic \Cricket"

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

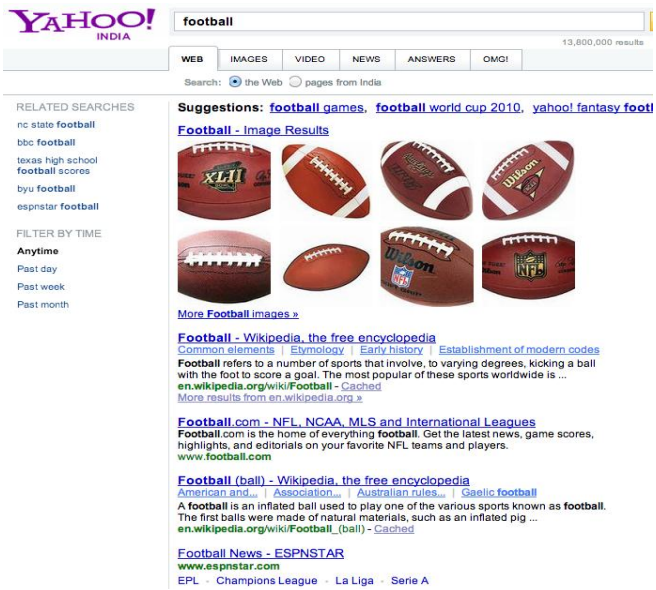


Fig 1.2 Non-trending topic "Football"

A. *Motivation*

In twitter more than 19% of the tweets are about organizations or product brands, less than 20% of which are shown to have significant sentiment. Predicting the tweets which are likely to stimulate users' interests can improve the sale and marketing of different products and brands. Online advertisements could use such predicted messages to efficiently target the locations of networks which are visited the most. Moreover successful predictions can also increase user satisfaction by providing them with more attractive contents. Media companies could learn how to effectively generate buzzes for new films and shows. In political campaigning, groups could learn who they should target in order to successfully spread their message. Predicting the popularity of content in twitter is also quite important for several other purposes such as viral marketing, popular news detection, personalized message recommendation and trend analysis. Users with many connections can suffer from information overload. It is quite important to filter information flow for the end users and to provide them with important tweets. Popularity prediction is also helpful in personalizing the content and finding the right tweets for end users. On the other hand, understanding how and why a tweet becomes popular, can help to gain a better insight into how the information is dispersed over the network. In the case of marketing, predicting popular tweets is quite useful for determining what are the trending topics and products.

In this work we developed an automatic learning-based approach to predict the popularity of content in tweeter. Automatic prediction with machine power has much lower costs compared to human-based work

The problem of popularity prediction in twitter has been studied in some previous works. In most of the recent works the popularity of tweets is defined as the number of retweets since retweeting is potentially the most effective way to disseminate messages due to its viral nature.

B. *Problem Statement*

Our problem statement can be formally defined as follows. Create an automated system, which takes in continuous stream of raw tweets, processes these tweets to filter out \noise" and get relevant informative tweets. Further mine these \filtered" tweets to detect and predict evolving \trending" topics in their very early stage.

There are three key aspects to this problem:

- Creating a model to process real time tweets feed to filter out the \noise" and output only relevant tweets.
- Creating a system to store the social graph of twitter users such that it can be efficiently used as a data structure for answering various user queries and get relations for a given user.
- Creating a model to distinguish a \trending" topic from \non-trending" to make

II. Related Work

A. *Introduction*

The problem of popularity prediction in social networks has always been widely studied. This problem has not only been studied in conjunction with twitter, but also in connection with other social networks. In this section we provide an overview of the existing approaches to popularity prediction in social networks, we discuss the related work and we elaborate on the advantages and limitations of existing methods.

B. *Information Dissemination in Social Networks*

A growing line of research has been followed on information dissemination through social networks. These studies propose that network cascades can play an important role as mediums for the dissemination of various information. These studies tend to be based on the idea that the information is spread by various infection mechanisms Under the same category, Kempe et al. (2003) studied a combinatorial optimization problem sometimes known as the influence maximization problem. The problem involves finding a small set of seed nodes in social networks to target initial activation so that the largest expected spread of information can be yielded. However, the exact computation of information cascades is an NP-hard problem (Chen et al., 2010). Information diffusion has been studied in several online social networks, such as Flicker.

C. *Predicting the Popularity of Content in Social Networks*

Due to the advent of web 2.0, user-generated content has increased dramatically. There are various types of contents that can be generated by users, such as comments and reviews on photos, movies and products. Most of these web 2.0 services connect the user with other users through social network, thus producing a social graph. For instance, in micro blogging services such as Twitter this social graph is called a follower network. Any content generated from a user becomes

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

visible to all of his/her followers and each of these contents has the chance to be re-posted by these followers who subsequently disperse the content over the social network. Re-posting, commonly known as retweeting, gives post the chance to become popular. The problem of popularity prediction in social networks has been widely studied.

D. Popularity Prediction in Twitter

Due to the popularity of the twitter micro blogging service there have been many studies on twitter. A great amount of work has been done to predict the popularity of tweets in this network. In this section we first justify why users do retweeting in twitter by reviewing the related literature and then we briefly explain the related work on popularity prediction on twitter social network. Understanding how users tweet and their motivations for tweeting is potentially important for predicting whether a tweet will be popular or not. In fact discovering what contents users choose to retweet can help to explain why a particular tweet becomes popular. The motivations for the act of retweeting are well explored in the study done by Boyd et al. (2010). They highlighted the mains reasons for retweeting as given by users. They introduced 10 different motivations for retweeting such as commenting on tweets, propagating tweets to new audiences, to inform specific persons or groups and to save tweets for future personal access.

E. Popularity Predictions and Recommendations

The problems of popularity predictions and recommendations are similar in some aspects. Both problems try to identify influential contents. While in popularity prediction problems the focus is more on the popularity of content, in recommender systems the focus is more on the user, the goal being to recommend the items to a user which satisfy him the most. Predicting the popularity of contents can be quite useful in connection with making recommendations.

F. Social Network Prediction Applications

Predictive models analyze past information to assess how likely it is than an event will occur in the future. Although human experts could have greater accuracy they are not scalable and do not work properly in cases when events have very low or high probability and they are definitely more expensive compared to the computer-based approach.

III. Data Collection – Data Set

Twitter data is collected by querying popular hash-tags related to the 2015 Super Bowl spanning a period starting from 2 weeks before the game to a week after the game. We use this data to train a regression model and then use the model to making predictions for other hash-tags. The test data consists of tweets containing a hash-tag in a specified time window, and we have then used our model to predict number of tweets containing the hash-tag posted within one hour immediately following the given time window. Twitter is a social networking and micro blogging service that allows users to post real time messages, called tweets.

Tweets are short messages, restricted to 140 characters in Length. Due to the nature of this micro blogging service, people use acronyms; make spelling mistakes, use emoticons and other characters that express special meanings. Following is a brief terminology associated with tweets.

- Emoticons: These are facial expressions: pictorially represented using punctuation and letters; they express the user's mood
- Target: Users of Twitter use the @ symbol to refer to other users on the micro blog. Referring to other users in this manner automatically alerts them.
- Hash tags: Users usually use hash tags to mark topics. This is primarily done to increase the visibility of their tweets.

IV. Preprocessing

A. Tokenization

After downloading the tweets using the tweet id's provided in the dataset, we first tokenize the tweets. This is done using the Tweet-NLP developed by ARK Social Media Search. This tool tokenizes the tweet and returns the POS tags of the tweet along with the confidence score. It is important to note that this is a twitter specific tagger in the sense it tags the twitter specific entries like Emoticons, Hash tag and Mentions too. After obtaining the tokenized and tagged tweet we move to the next step of preprocessing.

B. Remove Non-English Tweets

Twitter allows more than 60 languages. However, in our work currently focuses on English tweets only.

C. Replacing Emoticons play an important role in determining the sentiment of the tweet. Hence we replace the emoticons by their sentiment polarity by looking up in the Emoticon Dictionary.

D. Remove Url

The url's which are present in the tweet are shortened using TinyUrl due to the limitation on the tweet text. These shortened url's did not carry much information regarding the sentiment of the tweet. Thus these are removed.

E. Remove Target

The target mentions in a tweet done using '@' are usually the twitter handle of people or organization. The information is also not needed to determine the sentiment of the tweet. Hence they are removed.

F. Replace Negative Mentions

Tweets consist of various notions of negation. In general, words ending with 'nt' are appended with a not. Before we remove the stop words 'not' is replaced by the word 'negation'. Negation play a very important role in determining the sentiment of the tweet. This is discussed later in detail.

G. Hash tags

Hash tags are basically summarizer of the tweet and hence are very critical. In order to capture the relevant information

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

from hash tags, all special characters and punctuations are removed before using it as a feature.

H. Sequence of Repeated Characters

Twitter provides a platform for users to express their opinion in an informal way. Tweets are written in random form, without any focus given to correct structure and spelling. Spell correction is an important part in sentiment analysis of user-generated content. People use words like 'coooool' and 'hunnnnngry' in order to emphasise the emotion. In order to capture such expressions, we replace the sequence of more than three similar characters by three characters. For example, wooooow is replaced by wooow. We replace by three characters so as to distinguish words like 'cool' and 'cooooool'.

I. Numbers

Numbers are of no use when measuring sentiment. Thus, numbers which are obtained as tokenized unit from the tokenizer are removed in order to refine the tweet content.

J. Nouns and Prepositions: Given a tweet token, we identify the word as a Noun word by looking at its part of speech tag given by the tokenizer. If the majority sense (most commonly used sense) of that word is Noun, we discard the word. Noun words don't carry sentiment and thus are of no use in our experiments.

K. Stop-word Removal

Stop words play a negative role in the task of sentiment classification. Stop words occur in both positive and negative training set, thus adding more ambiguity in the model formation. And also, stop words don't carry any sentiment information and thus are of no use to us. We create a list of stop words like he, she, at, on, a, the, etc. and ignore them while scoring the sentiment.

V. ARCHETECTURE

In this section we build feature base model to perform classification using a combination of both these models. Our approach can be divided into various steps. Each of these steps are independent of the other but important at the same time. Figure 1 represent the approach for training and explains in detail about different steps in training process.

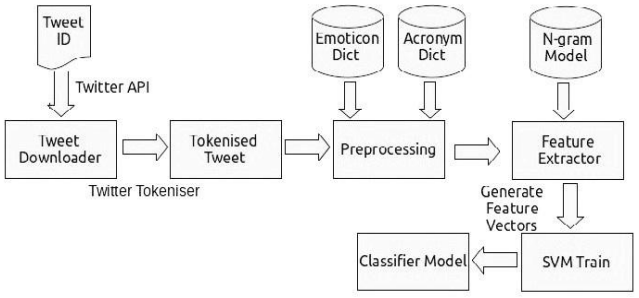


Figure 4.1 Flow Diagram of Training Data (A Hybrid Model)

Figure 4.2 represents testing the model where it describes various steps for testing the application with varying data sets.

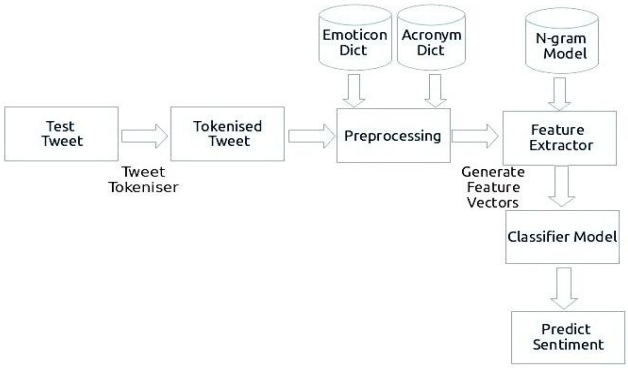


Figure 4.2 Flow diagram of testing (A Hybrid Model)

VI. MODEL TO MINE TWEETS

A. Linear Regression

A linear regression model was created using 5 features to predict number of tweets in the next hour, with features extracted from the tweet data in the previous hour. The features used to create the model were,

1. Numbers of Tweets (Class Variable)
2. Total Number of Re-tweets - (metrics/citations/total)
3. Sum of the number of followers of the users - (authors/followers)
4. Maximum number of followers of the users posting the hash tag
5. Time of the data - Obtained using the post-time of the tweet

The same hour-window approach was employed to calculate all the features. The output variable for each hour-window was the tweet count for the next hour-window. The model was trained using the OLS stats model library.

B. Regression Model with Extra Features

A new regression model was created using custom extra features (including original features considered in Question 2.) that were considered based on various papers and observation of the data. The new features considered were as follows,

1. Numbers of Tweets (Class Variable)
2. Total Number of Re-tweets - (metrics/citations/total)
3. Sum of the number of followers of the users - (authors/followers)
4. Maximum number of followers of the users posting the hash tag
5. Time of the data - Obtained using the post-time of the tweet
6. Ranking Score - (metrics/ranking_score)
7. Impression Count - (metrics/impression) - Measures the number of times a user is served a Promoted

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Tweet either in time-line or on search

8. Favorite Count - (tweet/favorite_count) - Number of tweets favorite's by users

9. Number of Users per hour - (tweet/user/id) - Counted number of users posting per hour

10. Number of Long Tweets per hour - (title) - Counted the number of tweets with length > 100 characters.

C. Cross Validation

This function gives internal and cross-validation measures of predictive accuracy for ordinary linear regression. The data are randomly assigned to a number of `folds'. Each fold is removed, in turn, while the remaining data is used to re-fit the regression model and to predict at the deleted observations. Cross validation is an approach to partition the sample data into a training (or model-building) set**,** which we can use to develop the model, and a validation (or prediction) set, which is used to evaluate the predictive ability of the model.

VII. Experimental Results and Analysis

A. Tweet Data Statistics

The training tweet data was loaded and statistics for each hash-tag was calculated in this question. In order to keep track of the hour count we have employed a hour-window approach. Since the tweets are all in sorted order of their posting time (firstpost_date). The first tweet is considered and the 1st hour-window is created using the formula

end_time = start_time + 3600 (1)

We loop through each tweet in the file and compare the post-time of the tweet with the end time of the present window. If it lies within the window we increase the hour-count if it doesn't we create a new window by using eqn:(1) and adding 3600 again to the end-time. At the same time a count is kept for the number of followers of users (author/followers) and the number of re-tweets (metrics/citations/total) for each tweet. The statistics calculated using the above procedure is listed below.

Table 7.1: Statistics for Each Hash tag

Hash tag	Total Tweets	Avg. # Tweets/hr	Avg. # of Followers of Users	Avg. # of Retweets
#gohawks	**188135**	**193.5438**	**1596.443**	**2.0146**
#gopatriots	**26231**	**38.3832**	**1292.2031**	**1.4001**
#gopatriots	**259019**	**279.5503**	**4394.2539**	**1.5385**
#patriots	**489710**	**499.4200**	**1607.4407**	**1.7828**
#sb49	**826905**	**1419.886**	**2229.694**	**2.5111**
#superbowl	**1348766**	**1401.2445**	**3675.3394**	**2.3882**

Analysis of the Statastics

1. Most Tweeted Hash tags per hour: #sb49 and #superbowl
2. Most Followers of Users for Hash tag : #nfl and #superbowl
3. All of the tweet data collected comprise of tweets that are not re-tweeted or are re-tweeted by very few users hence making the average re-tweet count _ 2.

In order to visualize the number of tweets in an hour a histogram was plotted for #SuperBowl and #NFL. A steep-rise can be seen for both the graphs at the same time which indicates the hour of the event.

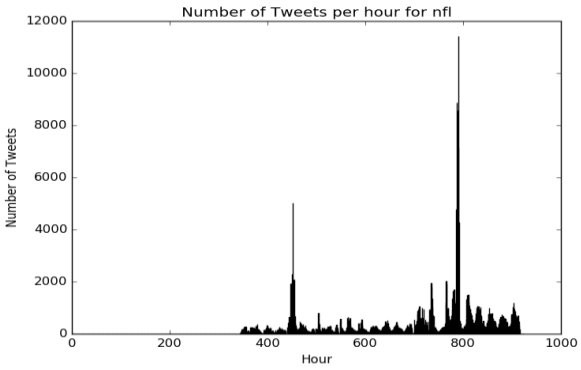
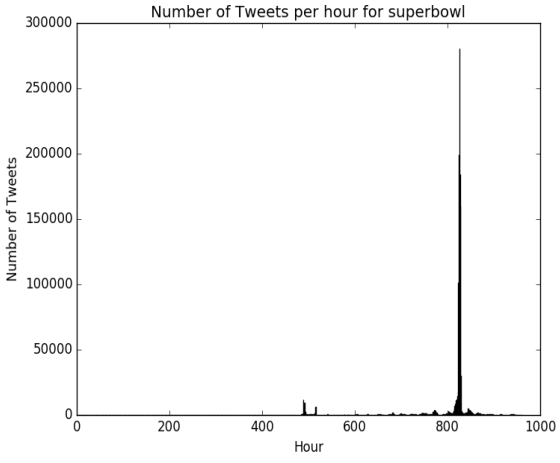


Figure 7.1: Number of tweets in hour : #NFL

Figure 7.2: Number of tweets in hour : #SuperBowl



B. Linear Regression

The same hour-window approach was employed to calculate all the features. The output variable for each hour-window was the tweet count for the next hour-window. The model was trained using the OLS stats model library. The results obtained for each of the hash tag are as follows,

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 7.2: Model Accuracy for each Hash tag

Hash tag	Accuracy
#gohawks	41.78
#gopatriots	43.15
#gopatriots	54.69
#patriots	43.72
#sb49	58.54
#superbowl	66.13

The (p-value [1], t-value [2]) for each attribute was recorded as well, the results are as follows

Table 7.3: p-value & t-value for Model Parameters

Hash tag	#of Retweets	∑of # of followers of users	Max. # of followers	Time of the data
#gohawks	(2.115*10^{-5},4.273)	(1.066*10$^{-7,}$,5.355)	(1.290*10^{-6},-4.871)	(4.230*10^{-3},2.867)
#gopatriots	(8.732*10^{-27},11.247)	(3.517*10$^{-16,}$,-8.386)	(1.048*10^{-12}, 7.278)	(8.643*10^{-1},-0.170)
#nfl	(3.525*10^{-16},8.266)	(7.424*10$^{-2,}$,1.787)	(4.185*10^{--},-0.809)	(2.314*10^{-5},4.235)
#patriots	(4.432*10^{-63},18.053)	(6.807*10$^{-14,}$,-7.602)	(2.209*10^{-5},4.263)	(4.776*10^{-3},0.710)
#sb49	(4.681*10^{-56},17.647)	(2.329*10$^{-31,}$,-12.347)	(1.320*10^{-15},8.222)	(6.402*10^{-2},-1.855)
#superbowl	(4.897*10^{-149},31.256)	(1.612*10$^{-7,}$,-26.502)	(4.853*10^{-52},-6.145)	(7.136*10^{-2},1.805)

Analysis of Results

- According to the definition of p-value and t-value it can be seen that the most contributing feature towards the regression model in all hash-tag files is the Number of Re-tweets posting a hash-tag.
- A fairly low accuracy is obtained for most of the hash-tag. This can be attributed to the window-size of one-hour as in the initial hours the average number of tweets are pretty low and creating a model for these sparse features is more difficult.

C. Regression Model with Extra Features

A total of 9 features were used to create the new regression model and after employing the same methodology of Question-2, features were collected using one-hour window. Since the last hour window cannot predict a tweet-count value it has been removed while creating the model. The model was tested and the results obtained were as follows

Hash tag	Accuracy
#gohawks	78.384
#gopatriots	53.118
#nfl	nfl 64.840
#patriots	58.793
#sb49	70.623
#superbowl	77.089

Table 7.4: Model Accuracy for each Hash-tag

As seen from the above results we have a significant increase in the accuracy of the model for each of the hash-tag, this can be attributed to features that are not sparse and have a well defined distribution through-out the period of the SuperBowl. Metrics employed in the tweet-data have been used to model the importance of the tweet for a given window frame thereby increasing the accuracy. In order to better visualize the contribution of the features in the model a scatter plot was created of the Top 3 features for each hash-tag. Since the initial hours have less number of tweets, all of the graphs exhibit clustering of values near low of tweets/hour.

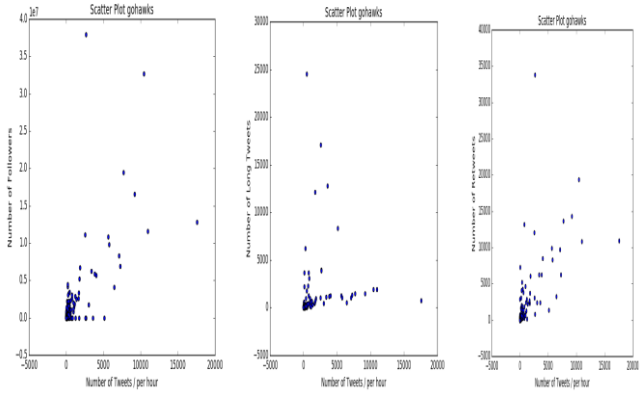


Figure 7.3Top 3 feature for #gohawks (# of Followers, # of Re-tweets, # of Long Tweets)

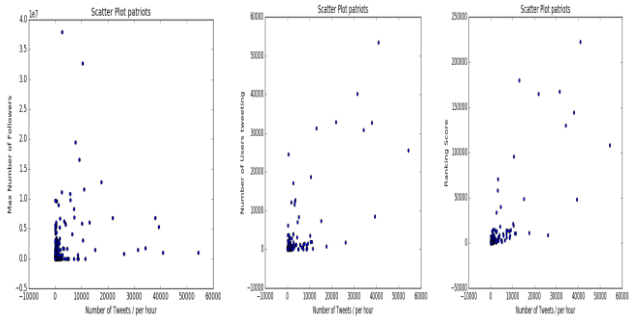


Figure 7.4: Top 3 feature for #gohawks (Max # of Followers, # of Users Tweeting, Ranking Score)

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

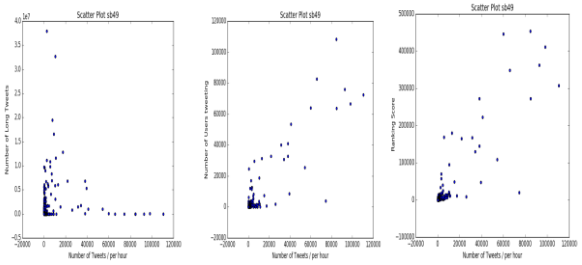


Figure 7.5: Top 3 feature for #gohawks (# of Long Tweets, # of Users Tweeting, Ranking Score)

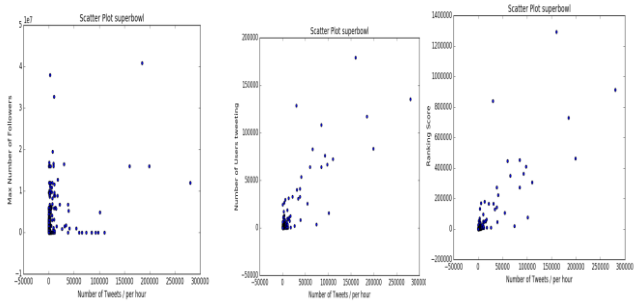


Figure 7.6: Top 3 feature for #gohawks (Max # of Followers, # of Users Tweeting, Ranking Score)

Analysis of Results

Table 7.5: Analysis of Top 3 Features Scatter Plot

Hash tag	Analysis
#gohawks	A linear proportionality can be seen in the scatter plots signifying a good relationship between all the 3 features
#gopatriots	Almost identical scatter plots with clustering towards the region of the origin
#nfl	A constant relationship can be seen for features (Favorite Count, # of Long Tweets) while a linear relationship is visible for # of Retweets
#patriots	Constant relationship for Max # of Followers feature while other two show linear proportionality
#sb49	Similar analysis to #patriots
#superbowl	Clustered regions with a very small linear deviation. Large number of instances fits a better regression model hence the higher accuracy

D. Cross Validation

This also requires the usage of same features used in Linear Regression with extra features to perform 10-fold Cross Validation across data. The accuracy results obtained across various hash-tags and over every fold given below.

Table 7.6: Average Error of 10 Fold Cross Validation

Fold Number	#gopatriots	#gohawks	#nfl	#patriots	#sb49	#superbowl
(1)	7.782	20.127	23.921	180.855	31.417	229.980
(2)	8.438	46.514	1.376	84.489	61.089	255.881
(3)	10.145	4.814	3.181	31.927	99.079	337.870
(4)	204.985	2.245	28.109	52.189	64.583	397.136
(5)	15.497	117.978	185.833	265.855	124.529	361.339
(6)	41.759	629.267	133.980	997.125	301.904	2506.928
(7)	19.302	147.079	93.183	687.341	881.058	1168.849
(8)	18.391	171.120	194.827	466.046	2854.875	2756.248
(9)	30.380	850.131	524.838	2046.537	1032.974	19664.687
(10)	247.476	5.099	137.612	176.498	321.142	1661.469
Average Error	60.415	199.437	132.686	498.886	577.265	2934.039

Analysis of Results

- We can see that there is a relationship between the number of tweets for a hash-tag and the average error of cross validation. Greater the number of tweets leads to a higher absolute average error for the hash-tag.
- In particular it is seen that for each hash-tag the error of one of the cross-validation fold is too high due to the uneven distribution of the data-set. A fold might consider a split wherein the test-data has all high values for the class (tweets during the time of the SuperBowl) and training-data has all low values for the class (tweets before and after the SuperBowl), hence producing a high error value for that fold (e.g. Fold 9 for #gopatriots).

E. Cross Validation with Time Periods

The second part of Question-4 deals with analysis of regression models created for different time-periods during the SuperBowl. Three different time-periods were considered to create the regression models,

1. Before Feb. 1, 8:00 a.m.
2. Between Feb. 1, 8:00 a.m. and 8:00 p.m.
3. After Feb. 1, 8:00 p.m.

Each tweet was segregated based on the time it was posted and split into windows of one-hour. The models were tested using 10-fold Cross Validation and the average errors for all folds obtained were as follows

Table 7.7: Average Error of 10 Fold Cross Validation for each Time-Period

Hash tag	Before	Between	After
#gohawks	167.189	7022.163	2607.692
#gopatriots	16.217	238.102	1760.682
#nfl	75.919	753.944	533.593
#patriots	190.869	93528.077	9745.065
#sb49	39.833	51166.878	12012.449
#superbowl	203.754	12861.877	11834.395

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Analysis of Results

- It can be clearly seen that due to the between time-period having only 12 one-hour window the number of instances in this time-period to create a model is very low. Hence the model created is giving very high average error values.
- Since the before time-period has a greater number of instances a better model is created hence giving low average error values.

VIII. Conclusion

As proposed an Event-Sequencing and Ad-Celeb Popularity/Comparison Checker was implemented and the results were presented above. The scope of the problem can be further being spread into areas of analytics for advertising agencies and for the celebrity PR teams. Sentiment analysis of the tweets collected can further represent the feelings of an advertisement or a celebrities´ performance during the SuperBowl.

References

[1] Bo Wu, Haiying Shen, "Analyzing and predicting news popularity on Twitter" in International Journal of Information Management, 35 (2015) 702–711.
[2] Shing H. Doong, "Predicting Twitter Hash tags Popularity Level", in 2016 49th Hawaii International Conference on System Sciences, 1530-1605/16 $31.00 © 2016 IEEE DOI 10.1109/HICSS.2016.247, pp:1959-1968
[3] Shoubin Kong, Qiaozhu Mei, Ling Feng, Fei Ye, Zhe Zhao, "Predicting Bursts and Popularity of Hashtags in Real-Time ", in *SIGIR'14,* July 06 - 11 2014, Gold Coast, QLD, Australia. Copyright 2014 ACM 978-1-4503-2257-7/14/07 http://dx.doi.org/10.1145/2600428.2609476.
[4] L.Jaba Sheela," A Review of Sentiment Analysis in Twitter Data Using Hadoop ",in International Journal of Database Theory and Application Vol.9, No.1 (2016), pp.77-86 http://dx.doi.org/10.14257/ijdta.2016.9.1.07
[5] Anisha P. Rodrigues, Niranjan N. Chiplunkar," Sentiment Analysis of Social Media Data using Hadoop Framework: A Survey", in *International Journal of Computer Applications (0975 – 8887) Volume 151 – No.6, October 2016*
[6] Maximilian Jenders, Gjergji Kasneci "**Analyzing and Predicting Viral Tweets" in** WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.ACM 978-1-4503-2038-2/13/05.
[7] Twitter nlp. 2012.
[8] Graph databases, http://www.neo4j.org/learn/graphdatabase. 2013.
[9] neo4j http://www.neo4j.org. 2013.
[10] Twitter api. 2013.
[11] Y. Altshuler, W. Pan, and A. S. Pentland. Trends prediction using social di_usion models. In Proceedings of the 5th international conference on Social Computing, Behavioral-Cultural Modeling and Prediction, SBP'12, pages 97{104, Berlin, Hei-delberg, 2012. Springer-Verlag.
[12] S. Ardon, A. Bagchi, A. Mahanti, A. Ruhela, A. Seth, R. M. Tripathy, and S. Triukose. Spatio-temporal analysis of topic popularity in twitter. CoRR,abs/1111.2904, 2011.
[13] L. Barbosa and J. Feng. Robust sentiment detection on twitter from biased and noisy data. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, COLING '10, pages 36{44, Stroudsburg, PA, USA, 2010.Association for Computational Linguistics.
[14] L. M. L. Delcambre and G. Giuliano, editors. Proceedings of the 2005 National Conference on Digital Government Research, DG.O 2005, Atlanta, Georgia, USA,May 15-18, 2005, volume 89 of ACM International Conference Proceeding Series.Digital Government Research Center, 2005.
[15] N. P. Fang Fang1. Detecting Twitter Trends in Real-Time. WITS 2011.
[16] M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, SIGMOD '10, pages 1155{1158, New York, NY, USA, 2010.ACM.
[17] Wei Chen, Chi Wang, and Yajun Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, (January):1029–1038,2010.URLhttp://portal.acm.org/citation.cfm?doid=1835804.1835934.
[18] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, Tweet, retweet: Conversational Aspects of retweeting on Twitter, volume 0, pages 1–10. IEEE,2010.URL http://www.computer.org/portal/web/csdl/doi/10.1109/HICSS.2010.412.
[1] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[2] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

AUTHORS PROFILE

Krishna Vasu Katarla is a M.Tech student in the Department of Computer Science Engineering at Srivasavi Engineering College, tadepalligudem, India. His research interests include Cloud Computing and Distributed Systems, Data Mining, Big Data Analytics

Dr.Vedula Venkateswara Rao is Professor in the Department of Computer Science Engineering at Srivasavi Engineering College, tadepalligudem, India. He received Matsers Degree in Computer Science Engineering from JawaharLalNehru Technological University Kakinada, Masters Degree In Information Technology from Punjabi University, Patiayala, India and PhD from Gitam University. His research interests include Cloud Computing and Distributed Systems, Data Mining, Big Data Analytics and Image Processing.He published several papers in International conferences and journals.

Invisible Digital Watermarking Methodology for Image Validation

[1] Dr. V.Kakulapati, SNIST, Hyderabad, India, vldms@yahoo.com
[2]Dr. SR Kattamuri, SNIST, Hyderabad, India, sreeram1203@gmail.com

Abstract— Watermark is the pattern of bits inserted into a digital image that identified the file's copyright information. The word watermarking is originated from faintly visible marks imprinted on organizational stationery. We propose robust and novel strategic in-visible approaches for insertion-extraction of a digital watermark in color images are presented. Contrasting printed watermarks, which are intended to be somewhat visible; the digital watermarks are designed to be completely invisible: the invisible insertion of the watermark is performed in the most significant region of the host image such that tampering of that portion for an intention to remove or destroy will degrade the esthetic quality and value of the image. One feature of the algorithm is that this user defined characters are used as a region of interest for the water marking process and eliminates the changes of watermark removal. Specifically dithering techniques are developed and intended to embed color water marking into the color image. A new technique is proposed and implemented using dithering techniques in various color spaces like RGB, HSV, and CMY. An attempt is made to develop full color water marking the scheme using those techniques.

Key words: - Watermarking; RGB; HSV; CMY**;** Dither

I. INTRODUCTION

Many researchers has allows digitally representation of water marking to ascertain itself as a prospective solution for n fortification of rights and regulate information piracy of images . Development of images by using Water marking techniques are classified into two types one is visible and another one is invisible approaches [20]. The visible methods provide indicates for overt assertion of rights with logos and the second methods are presenting covert protection of these rights. Images in first type of methods cannot protected the intellectual property due to easy process of any image in trendy graphic software collections. Whereas, digital water marking come into sight as a tool for secure the multimedia data from copyright violation [8]. Digital water mark will include in turn about the holder and in consequence protect it aligned with illegal use. As the water mark embed with a secret key is utilized to, no one other than the owner will be able to detect this watermark.

Now–a-days. Provision of validity is becoming increasingly significant as more of the world's information is accumulated as enthusiastically moveable bits. This is type of watermarking has been extensively useful to solve copyright issues of digital media concerning to illegitimate utilization or circulation. Numerous gray level image and water marking methods have been suggested [19], but the use of full color watermarks is still not well studied and a major weakness of water marking techniques is that they do not work if the image needs printing .No invisible watermark survives the distortions introduced by printers and any subsequent scanning does not allow recovery of originally embedded watermark. Hence, the color image watermarking in printed image is focused here. In the present work, we explore the techniques to embed invisible watermark in color images that can survive printed distortions. Printers use a technique called dithering to render images on paper [1]. Dithering involves printing pure color dots on paper in specific patterns determined by the algorithms and dither masks used. Specifically, we investigate dithering techniques to embed color watermark into color image. Here the dithering techniques are proposed and implemented in RGB, HSV, and CMY color spaces.

II. RGB AND HSV COLOR SPACES

In RGB color, space dithering is performed by mapping the colors in the source image colors in RGB set i.e. Red, Green and Blue. A dither mask is issued to map the colors to RGB color sets .The mask is constructing by randomly distributing red, green and blue colors and threshold values are taken by dividing the color range based on the number of tiles for each color inside the mask. This mask is tiled for each color inside the mask .This mask is tiled over the image and dither algorithm is applied .If the image patch value exceeds the threshold value, then the primary color of that particular tile (mask) is retained , otherwise the pixel is made white .The following is the dither mask containing RGB colors .We have chosen the 4X4 mask containing 7 red tiles , 5 green and 4 blue. To avoid patterns in the dithered image, all the three color's tiles are randomly dispersed in the mask. The threshold values in the mask are generated by dividing the color range(0-255) separately for each color channel .Since there are 7 red tiles in the mask, the color range (0 to 255) is divided into 7 part ie. (37,74,110,150,185,220,255) and are randomly placed in the red cells , similarly the color range (0-255) is divided into 5 parts (50,100,150,200,250) for green and 4 parts for blue (64,128,190,255) according to the figure 1.

Figure.1: Dither Mask

Algorithm:

1. The compute RGB values of each pixel in the image.
2. Construct a Dither mask (4X4 block) containing randomly distributed RGB values.
3. The Dither mask is tiled over the image and one to one comparison of the corresponding points is done, i.e., pixel values are compared with the threshold values of the mask.
4. If the image pixel value exceeds the threshold value retain the primary color of the otherwise make the pixel white.
5. This process is applied selectively on the channels ie., the threshold values in red tiles are applied on red channels only. Repeat the step 3 till the whole image dithered .The sample images are as shown below in figure 2,3and 4

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Figure 2: Image

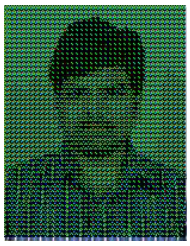


Figure 3: Dithered image

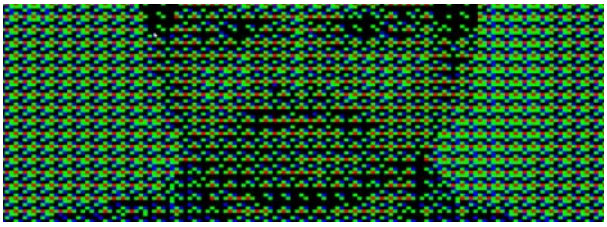


Figure 4: The dithered image zoomed

A. Dithering on the highest Values of RGB

In each pixel of the image, the maximum value of RGB is computed and is compared with threshold value in the mask .If that value is greater than the threshold of the maximum color then the component of that cell is retained. If we consider an example with pixels as Pixel (30, 50,220), Threshold = 150 then the Max (R, G, B) =blue (220) is calculated and Max (RGB)>threshold. Hence blue is retained in the output image as shown in figure 5

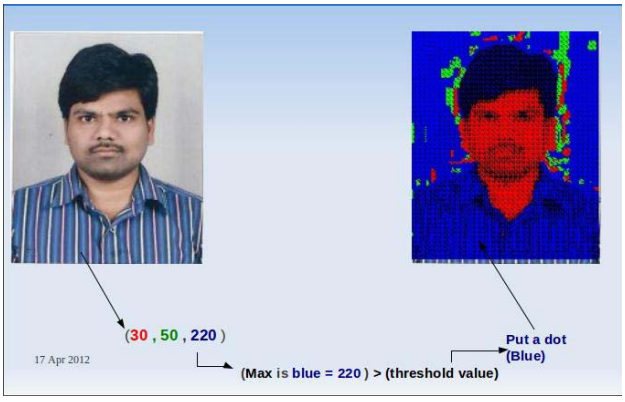


Figure 5 : Dithered to maximum value from primary colors

B .Dithering in HSV space

Hue Dithering

One of the important properties of color is Hue. It refers to dominant wavelength of a specific color. By dithering the image based on 'hue' parameter, the color content of the image can be changed [22].

Algorithm

1) Consider an image .Convert each pixel in the RGB image into a HSV.
2) Construct a dither mask (4X4) containing randomly distributed hue values.
3) Tile, the mask over the image.
4) Perform one of the comparisons of threshold value with the hue value of each pixel of the image.
5) If the hue value of the image exceeds the threshold value, then hue is set as 120 degrees .Otherwise, it is taken as '0'.
6) Repeat step 3 till the whole image dithered.
7) Convert HSV values into RGB primary colors.

Sample images dithered based on hue are shown below in figure 6,7, and 8

Figure 6: Image

Figure 7: Dithered image

Figure 8: Dithered image containing red (hue=0) and green (hue=120)

In the above figure, hue values are either '0' or '120'. The colors appeared in the dithered image are red and green .The sample image to hue at 120 and 240 are shown below in figure 9, 10 and 11

Figure 9: Image

Figure 10: Dithered image

Figure 11: Zoomed blue (hue=240) and green (hue=120)

C. Dithering on Dispersion

Dithering on Saturation refers to the purity of color. By dithering on saturation the purity of the color can be varied .Hereby considering a dither mask containing random values of saturation 16 parameters, dithering is performed on saturation parameter only. The sample images are as shown below in figure 12 and 13.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Figure 12: Image

Figure 13: Dithered image on 'S'parameters

D. Dithering on 'Values'

In geometry, the central vertical axis is called value and contains the unbiased, achromatic, or gray colors, value of black and white ranging from 0 and 1. Figure 14 and 15

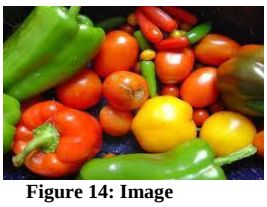


Figure 14: Image

Figure 15: Dithered image on 'V' Parameter

E. Dithering in CMYK color space

Printing presses and some ink-jet printers utilize four colors (cyan, magenta, yellow and black).Dithering is the most common means of reducing the color range of images .It is one of the principles behind printing technology. The colors are printing by mapping the colors in the source image to CMYK.

Figure 16 : CMYK mask

The color palette contains CMY colors .The figure ….above shows the mask that is designed by using 6 cyan tiles, 5 magenta and 5 yellow. The threshold values for cyan are taken by splitting the color range into 6 parts (1,32,100,150,200,255) as shown in figure 16 .Similarly the color range is divided into 5 parts each for yellow and magenta .To map the colors to the CMYK mask , the mask is tiled over the color image and one to one comparison of pixel values of corresponding cells is performed .If the pixel value exceed threshold value, then the color of the corresponding cell is retained .Otherwise white color is retained.

Results :

Image

Dithered image

Color image

Dithered to CMY

F. Implanting Secret Information

Invisible Digital Image Watermarking

Our algorithm inserts watermark into the color image and utilizes dithering technique to embed watermark in the color image .To make the watermark survive against the distortion created by printers, the image dithered in CMYK color space .To embed secret information two different types of dither masks are using by dithering the image in CMYK color space. For this, a secret code can be embedded in the image.

Embedding Digital watermarking into image

Algorithm

1. For an image converts the primary RGB values of each pixel into subtractive primary CMY colors.
2. Generate two different types of masks (mask-1, mask containing randomly distributed CMY colors.
3. The characters that are to be embedded are converted into ASCII equivalents, the ASCII values are further computed to its bit equivalent.
4. The region where secret information is embedded dithered using two kinds of masks (mask-1 to encode bit 0, mask-2 to encode bit 1.The remaining region of the image dithered using a single mask.
5. If the bit value is '0' mask-1 is tiled over the image .The image is dither with other masks if the bit value is '1'.

Dithering masks

In the embedding the secret code in the image two kinds of masks are used. These masks contain randomly distributing CMY colors and threshold values are generate by dividing the color range separately for each color channel .The following are the dither mask that contains the threshold values of three color's cyan, magenta and yellow. The whole image dithered with a single mask, the place where secret information as to be inserted by dithered using two kinds of masks as shown in the figure17 and 18.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Figure17: Mask-1

Figure 18: Mask-2

Experimental results:

Figure 18: Cover image

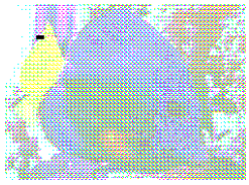


Figure 19: Watermarked image

The black spot in the watermarked image in the above figure 19 is calibrated mark. A secret message is embedded just after the calibrated mark.

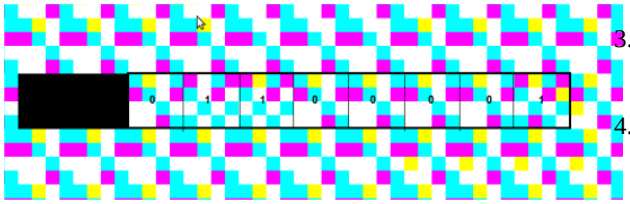
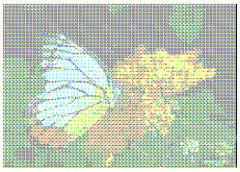


Figure 20: Encoded bits in watermarked image (zoomed)

The figure 20 above shows the embedded bits of character 'a' .The image is calibrated to find the encoded region. A single character 'a' is embedded in the image .The ASCII equivalent of 'a' is computed which is '97' and is further converted into bit equivalent i.e. 01100001.These bits encoded in the image using two kinds of masks as shown in figures 20.

Results:

Cover image



Single character encoded image

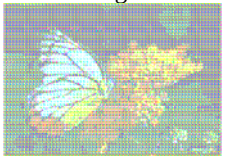


Image watermarked with 3 characters

Deciphered information from the scanned image

This Algorithm retrieves watermark from the scanned image. The fundamental proposal of dithering is to reduce the color range of the image. The same concept is used in retrieving secret information .The embedded information can be decoded by dithering the scanned image to CMYK color set.

The Watermarked image Printing and Scanning

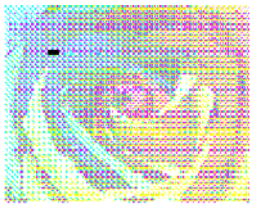
The Calibration mark

Initially the watermarked image is calibrated with a black mark near watermarked region, so that embedded information can be easily located. Here we used 4X8 block of black pixels that act as a calibration mark .The watermarked image is printed at different resolutions and scanned at different dpi. The following decoding algorithm is applied on the scanned images to retrieve the secret information.

Algorithm

1. For every pixel in the scanned image compute the mean square distance to each color in CMY color set.
2. Compute the minimum distance of each pixel in the scanned image with the colors in CMYK color space.
3. The scanned image dithered by replacing each pixel in the scanning image to the neighboring color in the CMYK color set.
4. After mapping all colors to CMYK color space, the image is examined near the calibrated mark to find whether the embedded masks are retrieved.

III. EXPERIMENTAL RESULTS



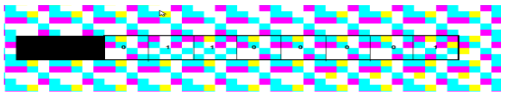
Watermarked the image with the calibrated mark

Mask -1

Mask-2

As shown in the figure …. The above, Mask-1 is used to encode '0' and mask-2 to encode '1'.Single character 'a' is encoded in the image, its bits equivalent can be observed in the figure above.

Step 1: Encoding the secret characters



https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Encoded bits (enlarged)

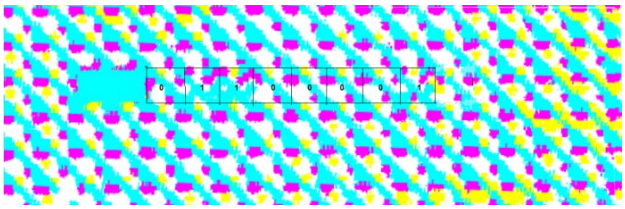
Step 2: Scanning the watermarked image

Scanned image at 600dpi

Step 3: The scanned image colors are mapped with CMYK colors.

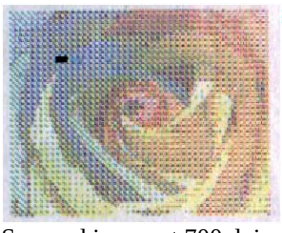
Scanned the image dithered to CMYK set

Step 4: Examine the region near calibration mark to decode secret bits

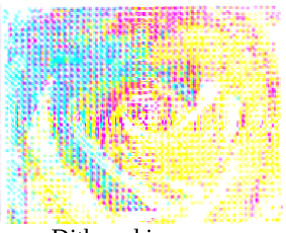


Retrieved watermark (decoded bits)

Image at 75 Dpi



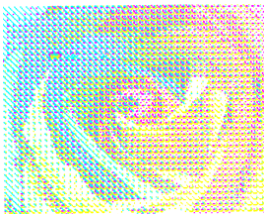
Scanned image at 700 dpi



Dithered image.

Image at 1200 Dpi

Scanned image at 1200 dpi



Dithered image

IV. CONCLUSION

In this work, a new method was proposed for embedding secret information invisibly into an image printed on a page. The secret embedding is in the form of dither masks. The pattern of colors printed is unique to the mask and we showed that the pattern is recoverable if scanned at sufficiently high resolution. In our experiment, we got the best results when the scanned resolution is the more than four times the print resolution .This technique has the potential to greatly enhance our concept of secure documents.

Bibliography

[1] V.Ostromoukhov ,P.Emmel , N.Rudaz, I.Amidror, R.D Herch, Multilevel color.halftoning algorithms .Symposium on Advanced Imaging and Network Technologies Berlin ,SPIE Vol 2949 pp. 332-340, oct 1996.

[2] Nirav Patel ,Binghampton University .Voronoi Diagrams Robust and Efficient implementation Thomas J.Watson School of engineering and Applied Sciences 2005.

[3] Victor Ostromoukhov Pseudo-Random Halftone screening for color and black and white printing .The 9th International congress in Non-Impact Printing Technologies, Yoohoma ,japan pp .579-582,1993.

[4] Dr. Chakravarthy Bhagvati,color image processing manual.

[5] Piyu Tsaia,Yu-Chen Hub ; Chin-chen Changa A color image watermarking scheme based on color quantization .A Department of Computer Science and Information Engineering .National Cung Cheng University ,Chiayi ,Taiwan pages 1754-1759,2002

[6] Gary Bradski , Adrian Kaebler learning open cv manual Computer vision with open cv library

[7] Joachim J. Eggers, Bernd Girod Quantization effects on Digital Watermarks Telecommunications Laboratory ,University of Erlangen –Nuremberg Stanford University 2000.

[8] R.G.Van Schyndel A.Z .Tirkel and C.F Osborne , A digital watermark IEEE. Int. Conf.Image Processing vol 2 1994 .pp 86-90.

[9] W.R. Bender D.Gruhl and N. Morimoto , Techniques for data hiding and Storage and Retrieval of image and video Databases , vol 2420,1995, pp 164-173

[10] G.K.Wallace, the JPEG still picture compression standard ACM , vol 34 1991 pp 30-44

[11] F.Hartung B. Girod , Fast public-key watermarking of compressed video in : Proceddings of the IEEE

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

International conference on Image Processing Santa Barbara ,CA, USA, October 1997.

[12] Dariusz Bogumi An assymetric image watermarking scheme resistant against geometrical distortions Institute of Computer , Warsaw University Of technology .Nowoweiejska 15/19,00-665 Warszawa,Poland June 2005

[13] M.Kutter,F.Jordan,F.Bossen,Digital signature of color image using amplitude modulation ,in: I. K. Sethi R.Jain(Eds.)Storge and Retrival for Image and Video Databases V,Vol.3022,SPIE,SanJose,CA,1997,pp.518-526.

[14] A.Piva,F.Bartonlini,V.Cappellini, M.Bamni, Exploiting the cross-correlation of RGB-channels for robust watermarking of color image.processdings of the IEEE International Conference on Image Processing,Vol. 1,

[15] Campbell,Alastair.The Designer's Lexicon, Chronicle, SanFrancisco,2000

[16] C.A.Bouman Ditigal Image Processing January 9,2012.

[17] Danny Pascale A Review of RGB Color Space from CMY to RGB The Bable Color Company 5700 Hector Desloges Montreal(Quebec) Canada H1T3Z6.2003.

[18] Scott Cravr, Nasir Memon, Boon-Lock yeo. Resolving Rightful ownership with In-Visible watermarking techniques: limitation, attacks and Implications IEEE may 1998.

[19] Stephane Derrode, Robust and Efficient Fourier-Mellin Transform Approximations for Gray-Level Image Reconstruction and Complete Invariant Description deBretagne,Technople,Brest-Iroise.BP832,292858 Brest Cedex,France,31

[20] Saraju P. Mohanty ,A Dual watermarking technique for imagess Dept. of Electrical. Engg. Indian institute of science ,Bangalore 1999.

[21] Victor Ostromukhov. Color in art and science2002.

[22] Wen chen Yun Qshi,Guoriong Xuan, Identifying computer graphics using HSV color model and statistical moments of characteristic functions. Dept. of Computer Science -China.

[23] Damin Cardani Adventures in HSV Space dcardani@buena.com.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Designing an Efficient Distributed Algorithm for Big Data Analytics: Issues and Challenges

Mohammed S. Al-kahtani

Dept. of Computer Engg., Prince Sattam bin Abdulaziz University, Saudi Arabia
Email: alkahtani@psau.edu.sa

Lutful Karim

School of ICT, Seneca College of Applied Arts & Technology, Toronto, Canada
Email: lutful.karim@senecacollege.ca

Abstract - As designing computationally efficient distributed algorithms is very important for analyzing big data this paper presents the current state-of-the-art research in designing distributed algorithms of big data analytics. More specifically, this papers presents a comprehensive survey on existing distributed algorithms of big data analytics – present working principle of existing algorithms, their advantages, limitations and also compare these algorithms in terms of several features. This paper also presents issues and research challenges that may arise to design an efficient distributed algorithm for big data analytics and proposes some solutions to address such challenges. Our research find that these algorithms support parallel processing and designed based on the MapReduce paradigm of big data for a particular application.

***Keywords*:** Big Data, Distributed Algorithm, MapReduce, DBMS, Commodity Hardware.

I. INTRODUCTION

As a large amount of data are being produced from social media, cloud, computer networks, content delivery networks and other emerging technologies and transmitted through Internet Big Data analytics have achieved a widespread popularity. At the same time, analyzing big data has proven to be challenging as traditional computing systems are not able to handle them. Three attributes of big data, velocity, volume and variety, namely 3V also reflects such challenges. Velocity deals with the issue of how quickly large amounts of data are being sent in. Volume deals with the size and amount of data being stored and processed. Variety represents different formats of data, which are mostly unstructured. A good example of 3V attributes of Big Data would be Instagram, which has over 400 million active users with 80 million upload a day on an average. These uploads include multiple formats of pictures and videos [1].

To cope with the challenges big data has emerged, horizontal scaling of a computing system is much more important than the vertical scaling [2]. This means that one computer is not built to be more powerful but the work is spread out over many more less powerful machines. This necessitates the use of distributed algorithms to process big data. Hence, designing efficient distributed algorithms for processing big data is significantly important to achieve computational efficiency.

Several distributed algorithms [1-14] have been designed to process big data. These algorithms are mostly designed to process data of a specific application. These algorithms are also designed based on the MapReduce Paradigm. The MapReduce runs on a cluster of commodity machines and thus, can be used in Hadoop operations and functionalities for large scale data processing. Among many, the Parallel IdeaGraph Algorithm uses MapReduce paradigm to handle big data challenges by implementing a parallel distribution of IdeaGraph [2]. Another Algorithm is Probabilistic Latent Semantic analysis (PLSA) [3] that implements a parallel method to train PLSA under the MapReduce framework. This algorithm addresses the scalability issues in PLSA. Item-based Collaborative-filtering algorithm [4] is a very effective and computationally efficient algorithm in MapReduce Paradigm that uses "hotweight" as the weight [4]. Evolutionary algorithm called the Grouping Genetic Algorithm solves the problem of grouping and is found in the schema optimization in HBase[1]. In addition to these algorithms, Locality-aware Scheduling Algorithm (LaSA) performs data locality assignment in Hadoop scheduler to enhance the performance of big data applications [5]. Parallel Two-Pass MDL (PTP-MDL) algorithm, Scalable Nearest Neighbor and Convex Optimization are other algorithms that reduce the computational, storage, and communications bottlenecks [6]. However, these algorithms are application dependent and hence, cannot be used for processing data of all applications. These algorithms only works in commodity machines and cannot be used in low powered wireless nodes such as sensors.

This paper provides a comprehensive survey on distributed algorithms of big data i.e., present working principle of existing algorithms, their advantages, limitations and also compare these algorithms in terms of several features. This paper also presents several challenges and issues to design an efficient distributed algorithm and proposes some solutions for future improvement. The rest of the paper is organized as follow.

Section 2 defines some terminologies and presents MapReduce Paradigm. Section 3 presents a comprehensive review on distributed algorithms of big data. Section 4 analyzes the algorithms, classifies and compares them. Section 5 identifies research challenges to design an efficient distributed algorithm and proposal for improvement. Sections 6 summarizes our research work in this paper.

II. PRELIMINARIES

To understand the working principle of distributed algorithms, the MapReduce framework of Big Data, which has been popularized by Google, is presented first. MapReduce is a scalable and fault-tolerant data processing tool that processes a large amount of data in

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

parallel with a number of low-end computing nodes [7]. It breaks a large file into a number of small chunks or task and distributes the small unit of works into a number of processing nodes to be executed in parallel. As programs written for MapReduce paradigm are automatically parallelized and can be executed on a large cluster of commodity computers.

The schedule of tasks execution is not predefined in MapReduce. Tasks are scheduled or assigned to nodes during runtime. Moreover, the underlying distributed file system ensures data locality and availability [7]. A few benefits of MapReduce among many are listed below.

- Simple but efficient for query processing in database management system (DBMS).
- It is flexible and easy to use.
- Fault tolerant and highly scalable.
- Can provide it functionalities inside DBMS to support user-defined functions in the relational DBMS.

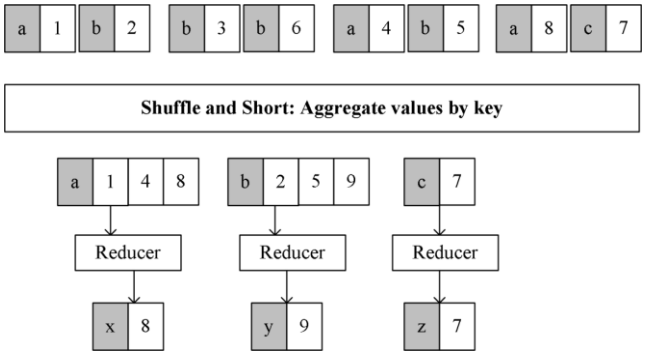


Figure 1. A simplified view of MapReduce

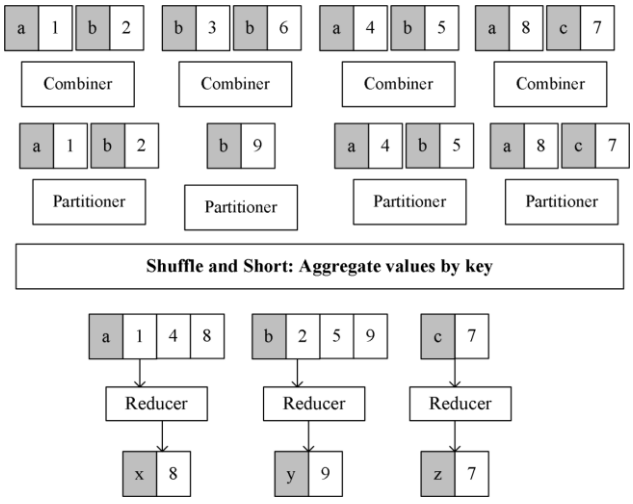


Figure 2. MapReduce with combiners and partitioners

MapReduce uses clustering mechanism on big data analytics. The simplified view of MapReduce is illustrated in Figures 1 and 2. A large file or database gets split into a number of small files when it is stored into the distributed file system of MapReduce [7]. The mapper works onto the every key-value pair and generates intermediate key-value pair. The reducers reduces the number of values by working on the same intermediate key associated with a number values. Thus, the reducer generates output key value pairs. The general concept is that distributed algorithms based on MapReduce are designed to run on more than one job. For instance, nine jobs run (one at a time) in MapReduce for IdeaGraph algorithm [2]. Figure 3 also illustrates the map and reduces operation process.

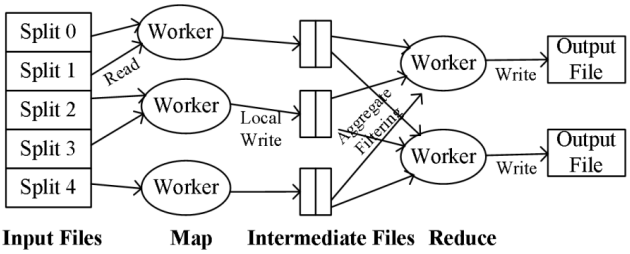


Figure 3. Map and Reduce operation process

III. DISTRIBUTED ALGORITHM IN LITERATURE

Analyzing huge amount of data collected/captured by social media, wireless networks, cloud, and organizations become computationally inefficient unless we design an efficient distributed algorithm to do so. In the following sections, we present several distributed algorithms for big data analysis.

A. IdeaGraph

IdeaGraph is a parallel distributed algorithm that uses MapReduce paradigm and detects chances[2] where, a chance represents a rare event. The chance discovery (CD) is very significant in human-computer interaction process for quick decision making. IdeaGraph [2] performs well in searching chances. The results generated by IdeaGraph are represented using scenario graph, which can be easily interpreted. IdeaGraph can be both parallel and sequential. The sequential IdeaGraph performs well for small datasets but is not suitable for big data. Though running sequential IdeaGraph on a powerful computer can mitigate the problem to some extent it cannot complete eliminate the problem. This is because a single process has limited computational resources. Thus, the implementation of parallel IdeaGraph that runs into a distributed system such as Apache Hadoop is greatly required.

Thus, the parallel IdeaGraph based on MapReduce is used to improve the performance. The parallel IdeaGraph introduced in [2] is able to discover rare chance from large datasets and also reduces the processing time for big data. The parallel IdeaGraph implements a MapReduce version of IdeaGraph to maintain the accuracy of the original IdeaGraph and achieve better performance. In this implementation of the algorithm, nine jobs run one after another to get the final results. After the initial tests have been run, the authors come to three conclusions. First, the proposed implementation is valid to identify the possibility to get the same results as the sequential IdeaGraph [2]. Second, the parallel implementation consists of several MapReduce jobs and intermediate results are stored into the MapReduce file system (not into the memory). The parallel

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

implementation does not work well for small data collections as compared to the sequential implementation. Finally, the parallel IdeaGraph implementation is more efficient than the sequential IdeaGraph implementations in terms of processing time and memory cost for large data sets.

B. Probabilistic Latent Semantic Analysis (PLSA)

Probabilistic Latent Semantic Analysis (PLSA) is a powerful statistical technique to analyze relation between co-occurrence data. It has widespread applicability in automated information processing tasks [3]. However, it is often difficult and time consuming to train large datasets. This scalability problem of PLSA can be addressed by implementing a parallel method to train PLSA under the MapReduce computing framework. P2LSA and P2LSA plus are two types of parallel P2LSA that are designed based on MapReduce paradigm in [3]. The P2LSA executes E-step and M-step similar to the Map and Reduce functions, respectively. A large amount of data is transferred between E and M-steps to train PLSA models and perform the operations (i.e., read and process the input records). This overloads the network and increases the overall processing time. The P2LSA plus performs two different jobs to finish the whole task, but that reduces the degree of parallelism. The work done in [3] introduces an algorithm that works differently from these two MapReduce algorithms. In this approach, the authors implement the EM algorithm and then parallelize it. Thus, this algorithm is more straightforward and applicable to many areas.

C. Recommendation Algorithm

The idea of recommendation system was first introduced in 1994 by a research team who launched a system called GroupLens and set up a collaborative filtering engine [4]. Recently, the recommendation systems that use recommendation algorithms have achieved widespread applicability in commercial video, audio streaming and e-commerce applications such as Youtube, Netflix video recommendation system, Amazon, ebay. Hence, the recommendation systems also require processing huge amount of data and information as a part of big data analytics. Thus, the traditional operating mode of recommendation system does not work well because of the limited resources to process big data [4]. Hence, Hadoop is being used as the key to solve this problem as the distributed infrastructure of Hadoop can fully utilize the storage and computing ability. This leads to another algorithm, which is a collaborative filtering recommendation algorithm running on a Hadoop platform using “hotweight” [12] as the weight. The “hotweight” is the average rating of each item taken from historical data of users. This greatly improves the efficiency of a system when dealing with huge amounts of data. Traditional algorithms cannot recommend an item if a user is added to the system (as item vectors are all zeros). However, collaborative algorithms that use hotweight can recommend items for new users, which are basically the bestselling products. The computation of hotweight uses MapReduce computation model, i.e., distributed framework available in Hadoop. The process starts by collecting key-value pairs with the same key. By using this distribute system, the computational efficiency is signifacantly improved [4]. The hotweight calculation process is shown in Figure 4.

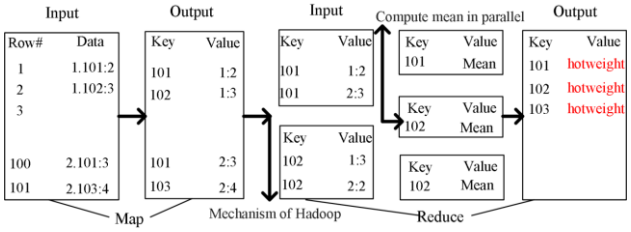


Figure 4. MapReduce process of computing hotweight

D. Locality-aware Scheduling Algorithm (LaSA)

Many data intensive computing applications such as MapReduce, HBase, and Hadoop [5] have been built for big data processing. Data access, placement and computations pose a great challenge in big data framework. These problems can be alleviated by increasing data locality in distributed environment [5]. MapReduce uses Hadoop as a distributed file system, process scheduler, execution environment and framework. However, the process scheduler fails to account for data locality issues. Thus, the work done in [5] introduces Locality-aware Scheduling Algorithm (LaSA) algorithm that enhances data locality assignment in Hadoop scheduler and also increases the performance of data-intensive computing applications. The end point of LaSA is to accomplish locality-aware resource assignment in order to decrease the bottleneck of transmission by following the weight of data interference. The concept of data interference weight is made known to MapReduce framework by the LaSA as well as the locality-aware scheduler is introduced in JobTracker. The data interference weight of each node that depends on resource assignment is calculated by the LaSA. Then, the job tracker selects a node with the least weight in order to execute the task, which in turn reduces the congestion of network transmissions.

E. Parallel Two-Pass Minimum Description Length (PTP- MDL)

In data storage and communications, distributed cloud computing and big data frameworks introduce new challenges. For instance, data streaming uses distributed computing. Data is processed remotely in clusters and results are streamed through a network to the user. The use of data streaming in big data framework makes it imperative to use fast lossless data compression algorithms whose primary features include good compression quality and high throughput [8]. The PTP-MDL algorithm is used for such data compression. It possesses the random access property. Hence, if a part of a file is compressed using PTP-MDL that part can be decompressed without decompressing the whole file. Also the PTP-MDL algorithm has numerical results that provide a better trade-off between compression and throughput. Thus, this algorithm is very useful and efficient for big data problems.

The Parallel Two-Pass MDL (PTP-MDL) can compress certain data files or part of files while decompresses

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

others. The block diagram of the PTP-MDL encoder is displayed in Figure 5 as well the block diagram of the PTP-MDL decoder in Figure 6.

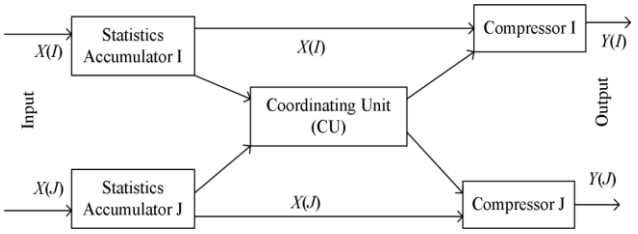


Figure 5. PTP-MDL encoder

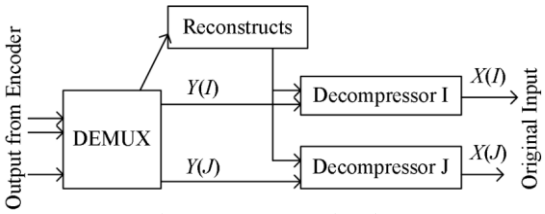


Figure 6. PTP-MDL decoder

F. Scalable Nearest Neighbor Algorithm

Many computer vision algorithms are computationally inefficient as they search for the most similar matches to high-dimensional vectors. This is also known as nearest neighbor matching. Scalable nearest neighbor algorithms [9] are efficient as they can find the nearest neighbors in large data sets very fast. However, the use of large training sets is significantly important to achieve better performance in these algorithms [9]. Though many applications improve processing time they are subject to settle for an approximate search i.e., after performing the search the neighbors returned as search result are not all exact. Although it is an approximation they are still very close to the exact neighbors. It provides almost 100% correct neighbors. Moreover, it is much faster than linear search (two or more orders of magnitude faster).

The Scalable Nearest Neighbor algorithms are able to match the nearest neighbor very fast in a large data set and thus, it improves the overall performance of network bandwidth. The priority search *k*-means tree algorithm is one of the very fast nearest neighbor algorithms. The search tree is built using *k*-mean clustering. This clustering algorithm partitions the data points at each level into *k* individual clusters/regions and recursively applies the same approach to the points in each region. The recursion terminates when the number of points in a region becomes smaller than *k*.

G. Convex Optimization Algorithm

The main focus of convex optimization algorithms for big data [6, 11] is to reduce the computational, storage, and communication bottlenecks. Convexity in signal processing dates back to the dawn of the field, with problems like least squares (LS) being ubiquitous across nearly all subareas [6]. The importance of convex formulations and optimizations has increased even more dramatically in the last decade because of the rise of new theory for structured sparsity and rank minimization, and successful statistical learning models such as support vector machines [6]. Due to the popularity of convex optimization, the algorithms are considered greatly to accommodate large data sets and also to solve problems in unprecedented dimensions.

H. Other Algorithms

The Item-Based collaborative filtering recommendation algorithm (IBCF) computes the product of co-occurence matrix and user item vector (which also estimates the complexity of this algorithm). Tradtional algorithms require huge storage for large matrix. The method in [4] is designed to save space to store sparse matrix that is used by IBCF. This method is just storing the nonzero data with the coordinates. In matrix mutltiplcation, the output matrix $C[m \times k]$ is computed by two input matrices $A[m \times n]$ and $B[n \times k]$. The computation of each element in matrix C is independent of each other. Hence, in Map phase, all the elements from matrix A and matrix B should be collected together under the same key to calculate the element of matrix C. Then the Reduce function calculates every element of matrix C in parallel [4]. The matrix multiplication process is shown below in Figure 7.

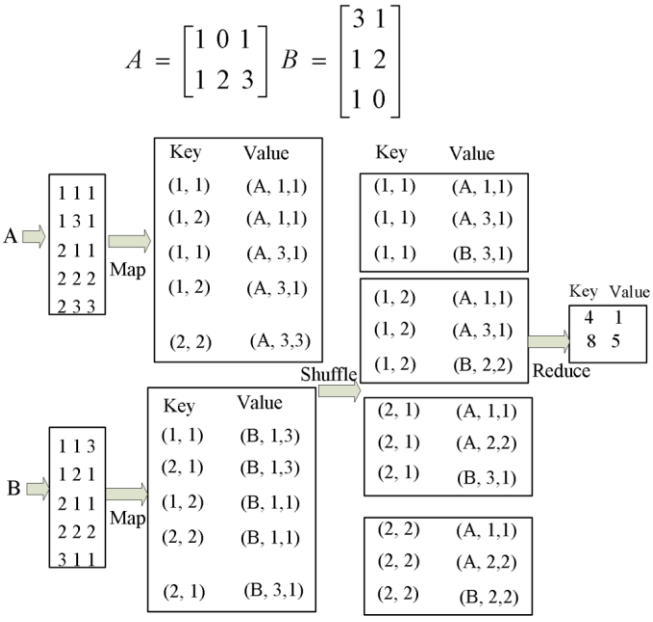


Figure 7. MapReduce process of matrix multiplication

Operations of an evolutionary algorithm are very important, which conducts the direction of evolution to optimized solutions. An algorithm with efficient operations can let solutions jump from a local area to search in the global space, which leads the evolution to a better and faster convergence point [1]. For this HBase optimizing problem, genetic operations together with validity check for evolution process use mutation and crossover. A mutation operator adds or removes one or more columns inside one CF. Redundant columns can be removed to make a CF more effective for query types that do not need those columns. Moreover, some other

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 1. Comparison of existing distributed algorithms

Features	Idea-Graph	PLSA	Recommen-dation Algorithm	LaSA	PTP-MDL	Nearest Neigbor	Convex Optimization
Parallel processing	√	√	√	√	√	√	√
Sequential processing	√	X	X	X	X	X	√
Uses MapReduce paradigm	√	√	√	√	√	√	X
Statistical Model	X	√	X	X	X	X	√
Use high throughput lossless data compression	X	X	X	X	√	X	X
Use collaborative filtering engine	X	X	√	X	X	X	X
Used in computer vision algorithms for searching the most similar matches to high-dimensional vectors	X	X	X	X	X	√	X
Support locality-aware resource assignment	X	X	X	√	X	X	X
Used in structured sparsity and rank minimization	X	X	X	X	X	X	√
Able to match the nearest neighbor very fast in a large data set	X	X	X	X	X	√	X

columns can be added so that the CF can match more query types [1]. The crossover operator exchanges columns between two CFs inside an HBase schema. Such crossover mixes good sub-solutions of CFs without any disruption of the partitions. The split up operator is designed to split a CF into two smaller CFs inside a solution. A diving point is chosen by the operator, and the columns before diving point are kept in the current CF. The columns after the dividing point form a new CF. Among other algorithms, the work done in [13] introduces distributed structured prediction learning algorithm that reduces processing time and storage requirements and the work done in [14] presents a comprehensive survey on in-memory big data management and processing such as fault tolerance and consistency of in memory environment.

IV. PERFORMANCE ANALYSIS

In section 3, we presented a number of distributed algorithms for processing big data. Most of these algorithms are based on MapReduce paradigm as it provides scalability and fault tolerance for big data. However, MapReduce cannot substitute DBMS rather it can complement DBMS with scalable and flexible parallel processing[3]. However, input and output costs of MapReduce still require to be addressed for successful implications.

Parallel IdeaGraph based on MapReduce framework can handle big data challenges. It has been found that parallel IdeaGraph is better than sequential counterpart for processing huge volume of data. On the other hand, sequential IdeaGraph is much better for processing small data set. In general, the parallel IdeaGraph implementation performs better than the sequential IdeaGraph implementations in terms of data processing time and cost of memory [2].

The Probabilistic Latent Semantic Analysis (PLSA) parallel algorithm is another popular distributed algorithm of big data. The performance of the PLSA is evaluated in terms of speedup which is defined as the ratio of standalone running time to the Hadoop cluster running time. Three benchmarked datasets: 20Newsgroups dataset, its subset mini-Newsgroups, and a crawled document dataset [3] were used to evaluate its performance. The speedup is high when the dataset is small. However, the speed up does not increase linearly for small dataset with the increasing number of processors because of the communication costs between the Map phase and the Reduce Phase. In addition, the initialization and the partition processes also take some time. The speedup increases almost linearly as the number of the processors for large datasets. Results on performance evaluation of PLSA [3] conclude that the PLSA training method on MapReduce can deal with large datasets efficiently and hence provides a practical solution to big data analysis applications.

The LaSA is a job scheduler algorithm for Hadoop-MapReduce. It improves the data locality problem of Job Tracker in the original MapReduce framework [5]. The LaSA arranges the resource assignment to avoid required data missing by using the weight of data interference concept. On the other hand, the compression ratio of the PTP-MDL algorithm for real-world data is comparable to existing universal data compressors. Moreover, the throughput of PTP-MDL scales well with the number of parallel units.

The scalable nearest neighbor algorithms can search for the nearest neighbor in a high dimensional vector space very fast. Thought this makes the algorithms computationally inefficient they are widely used in many computer vision and machine learning algorithms [8]. As distributing the search into multiple machines requires less memory access/overhead their performance gets better.

An Evolutionary Algorithm was also presented that finds the optimum column family schema for a given set of user queries. The reading performance of the optimized column family schema was evaluated using a real dataset, which contains 2.6 million rows of aggregated tracking data and 1.3 million user queries [1]. It is found that the reading performance of HBase is improved using optimized column family schema. User queries from a testing set shows that the average response time is reduced by up to 72% compared to un-optimized column family schemas [1]. Table 1 compares several distributed

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

algorithms that we presented in this paper in terms of a number of features.

V. RESEARCH CHALLENGE AND PROPOSED SOLUTIONS

After analyzing existing distributed algorithms of big data that are presented in previous sections, the following challenges/limitations have been identified.

- They are mostly designed based on MapReduce paradigm for a specific application or context and thus, cannot efficiently be used for other applications.
- They are not dynamic. They are mostly designed for large amount of data. Hence, it is not efficient for small amount of data.
- They are designed considering that processing is distributed into a large number of commodity computers. However, they do not consider distributed processing at data collecting devices such as sensors.
- They support either sequential or parallel implementations. They do not support both.
- Statistical analysis is a great tool for big data analysis and hence, can be considered an integral part of an efficient distributed algorithm of big data analytics. Most of them do not support statistical models.
- They cannot extract rare but important events (for making decisions) from huge data unless they are designed to do so.
- Moreover, partitioning highly correlated data to maximize data locality and reduce communication cost is a great challenge.

These research challenges can be alleviated by introducing a distributed algorithm for big data framework that is application independent and can be applied to any big data application. The proposed algorithm is scalable, dynamic and fault tolerant, i.e., it supports (i) as many nodes as needed, (ii) both parallel and sequential implementation based on the amount of data to process (i.e., if the amount of data exceed a certain threshold it works in parallel mode; otherwise, it should work in sequential mode to reduce processing overhead) (iii) continuous operation even if any node fails. Moreover, the algorithm should work in different types of nodes ranging from low powered RFID tags and sensors to high speed/powered computers. The proposed algorithm also integrates a statistical model to validate the results.

VI. CONCLUSION

We presented a literature survey on the existing distributed algorithms for big data analytics. We compared these algorithms based on different features and identified their limitations. We also identified research challenges to design an efficient distributed algorithm for big data analytics and proposed some solutions. This research work will led us to design an efficient distributed algorithm for big data analysis in terms of processing time, memory requirements in future.

REFERENCES

[1] Yang F, Milosevic D, Cao J. An Evolutionary Algorithm for Column Family Schema Optimization in HBase. *IEEE First International Conference on Big Data Computing Service and Applications (BigDataService),* Redwood City, CA, 2015, pp. 439-445.

[2] Wang Q. , Wang H, Zhang C, Wang W, Chen and F. Xu. A Parallel Implementation of Idea Graph to Extract Rare Chances from Big Data. 2014 *IEEE International Conference on Data Mining Workshop*, Shenzhen, 2014, pp. 503-510.

[3] Liang Z, Li W and Li Y. A Parallel Probabilistic Latent Semantic Analysis method on MapReduce Platform. *2013 IEEE International Conference on Information and Automation* (*ICIA*, Yinchuan, 2013, pp. 1017-1022.

[4] Lu F, Hong L, Changfeng L. The improvement and implementation of distributed item-based collaborative filtering algorithm on Hadoop. *2015 34th Chinese Control Conference* (*CCC*), Hangzhou, 2015, pp. 9078-9083.

[5] Chen TY, Wei HW, Wei MF, Chen YJ, Hsu TS and Shih WK. LaSA: A locality-aware scheduling algorithm for Hadoop-MapReduce resource assignment. *2013 International Conf. on Collaboration Technologies and Systems* (*CTS*), San Diego, CA, 2013, pp. 342-346.

[6] Cevher V, Becker S, Schmidt M. Convex Optimization for Big Data: Scalable, randomized, and parallel algorithms for big data analytics. In *IEEE Signal Processing Magazine*, Sept. 2014, 31(5), pp. 32-43.

[7] Maitrey S. Jha CK. Handling Big Data Efficiently by Using Map Reduce Technique. 2015 *IEEE International Conference on Computational Intelligence & Communication Technology* (*CICT*), Ghaziabad, 2015, pp. 703-708.

[8] Krishnan N. Baron D. A Universal Parallel Two-Pass MDL Context Tree Compression Algorithm. In *IEEE Journal of Selected Topics in Signal Processing*. June 2015, 9(4), pp. 741-748.

[9] Muja M. Lowe DG. Scalable Nearest Neighbor Algorithms for High Dimensional Data. in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Nov 2014, 36(11).

[10] Facchinei F, Scutari G, Sagratella S. Parallel Selective Algorithms for Nonconvex Big Data Optimization. In *IEEE Transactions on Signal Processing*, April 2015, 63(7), pp. 1874-1889.

[11] Daneshmand A, Facchinei F, Kungurtsev V, Scutari G. Hybrid Random/Deterministic Parallel Algorithms for Convex and Nonconvex Big Data Optimization. In *IEEE Transactions on Signal Processing*, August 2015, 63(15), pp. 3914-3929.

[12] Ghuli P, Ghosh A, Shettar R. A collaborative filtering recommendation engine in a distributed environment. 2014 *International Conference on Contemporary Computing and Informatics* (*IC3I*), Mysore, 2014, pp. 568-574.

[13] Schwing AG, Hazan T, Pollefeys M, Urtasun R. Distributed Structured Prediction for Big Data. *NIPS Workshop on Big Learning*, Nevada, USA, Dec 2012.

[14] Zhang H, Chen G, Ooi BC, Tan KL, Zhang M. In-Memory Big Data Management and Processing: A Survey. In *IEEE Transactions on Knowledge and Data Engineering*, 2015, 27(7), pp. 1920-1948.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

A Survey on Memory Virtualization in Cloud

PVSS Gangadhar
Scientist-D& Research Scholar,
NIC, Meity, Govt of India,
Dept. of IT, GIT,
Gitam University
Visakhapatnam,
Andhra Pradesh, India
pvss.gangadhar@nic.in

Dr.Ashok Kumar Hota,
Scientist-F,
NIC, MEITY,
Odisha state Unit,
Bhubaneswar
Govt of India
ak.hota@nic.in

Dr.M.Venkateswara Rao
Professor, Dept of IT, GIT,
Gitam University,
Visakhapatnam,
Andhra Pradesh, India
mandapti_venkat@yahoo.com

Dr.V.Venkateswara Rao
Professor, Dept. of CSE,
Sri Vasavi Engg College
Tadepalligudem,
Andhra Pradesh, India
venkatvedula2017@gmail.com

***Abstract*—the Cloud Computing Infrastructure-as-a-Service (IaaS) layer offers a service for on demand virtual machine images (VMIs) operation. This service offers a elastic platform for cloud users to build up, install, and test their applications. Virtual machine environments are becoming more universal due to the augmented performance of service hardware and the materialization of cloud computing for huge scale applications. Because of the reason that virtual machines continues to grow, performance critical applications will need proficient methods to realize their tasks. The operation of a VMI classically engages booting the image, installing and configuring the software packages. In the conventional approach, when a cloud user requests a new platform, the cloud provider chooses a suitable template image for cloning and deploying on the cloud nodes. The template image encloses pre-installed software packages. If it does not fit the requirements, then it will be tailored or the new one will be formed from scratch to fit the request. In the context of cloud service management, the customary approach features the difficult issues of handling the complexity of interdependency between software packages, scaling and maintaining the deployed image at runtime. The cloud providers would like to automate this process to improve the performance of the VMIs provisioning process, and to give the cloud users more flexibility for selecting or creating the appropriate images while maximizing the benefits for provider's intern of time, resources and operational cost. The increasing demand for storage and computation has driven the growth of large data centers–the massive server farms that run many of today's Internet and business applications.**

A data center can comprise many thousands of servers and can use as much energy as a small city. The massive amounts of computation power contained in these systems results in many interesting distributed systems and resource management problems. In this paper we focus to investigate challenges related to data centers, with a particular emphasis on how new virtualization technologies can be used to simplify deployment, improve resource efficiency, and reduce the cost of reliability, all in application agnostic ways. We first study problems that relate to the initial capacity planning required when deploying applications into a virtualized data center, issues related to memory utilization among virtual machines and performance metrics for memory virtualization..

Keywords-Cloud Computing; IaaS; Virtual Machine Image(VMI); Data Center; Memory Virtualization; Capacity plannin;

I. INTRODUCTION

All Cloud Computing is a model for enabling service users to have ubiquitous, convenient and on- demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services), that can be rapidly provisioned and released with minimal management effort or service-provider interaction. The Cloud will provide IT similar to public utilities providing electricity, gas, and water. There is no need to have to own the hardware & the staff. There will be multiple public cloud providers. The generalized architecture of cloud computing is given in figure-1 below.

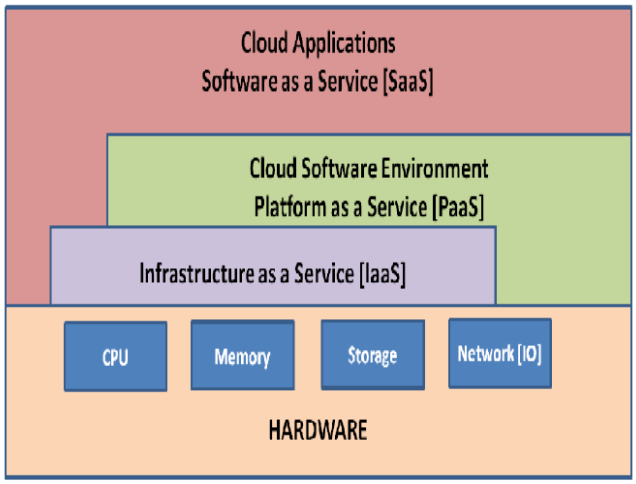


Figure1: Cloud computing Architecture

Infrastructure as a service refers to the sharing of hardware resources for executing services, typically using virtualization technology. With this so called Infrastructure as a Service (IaaS) approach, potentially multiple users use existing resources. The resources can easily be scaled up when demand increases, and are and are typically charged for on a per -pay-use basis. In the Platform as a Service (PaaS).approach, the offering also includes a software execution environment, such as an application server. In the Software as a Service approach (SaaS), complete applications are hosted on the Internet. Virtualization is the process of decoupling the hardware from the operating system on a physical machine. Virtualization can be thought of essentially as a computer within a computer, implemented in software. This is true all the way down to the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

emulation of certain types of devices, such as sound cards, CPUs, memory, and physical storage. The Figure2 below describes how the physical resources like servers, storage and network are separated from the operating systems and the applications using the virtual infrastructure.

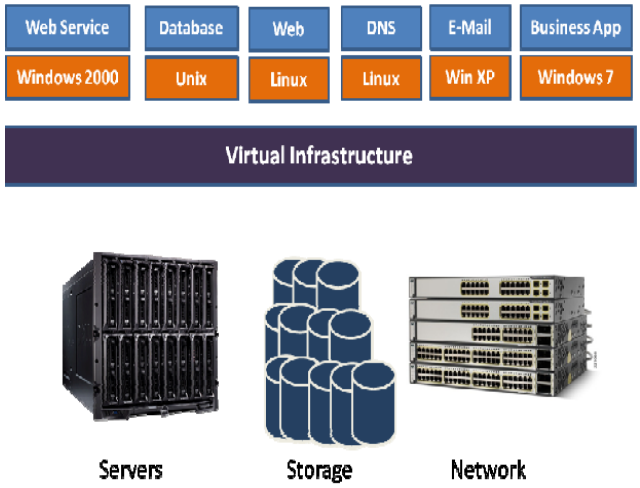


Figure2: Hardware / Software Separation in a Virtualized Environment

Cloud Computing takes Virtualization to the Next Step. Virtual Machines & services can be rented as needed from a cloud service provider. However for the management of such a process requires that the resources be managed in an efficient way. Resource Management is this process of managing the Physical Resources like CPU, Memory, Networketc across various Virtual Machines (VM) based on policies. The figure-3 below gives a frame work of the cloud infrastructure with resource orchestration.

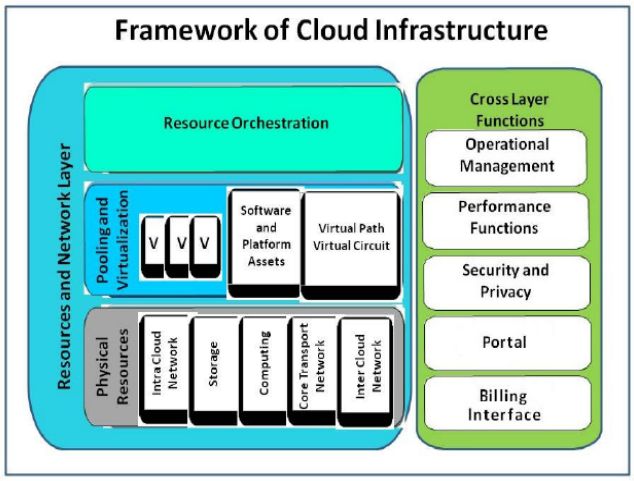


Figure3: Framework of cloud infrastructure with resource orchestration.

II. VIRTUALIZATION

Virtualization is the process of decoupling the hardware from the operating system on a physical machine. An instance of an operating system running in a virtualized environment is known as a virtual machine. Figure-4 below gives architecture of the virtualization with multiple virtual machines sharing the same hardware resources

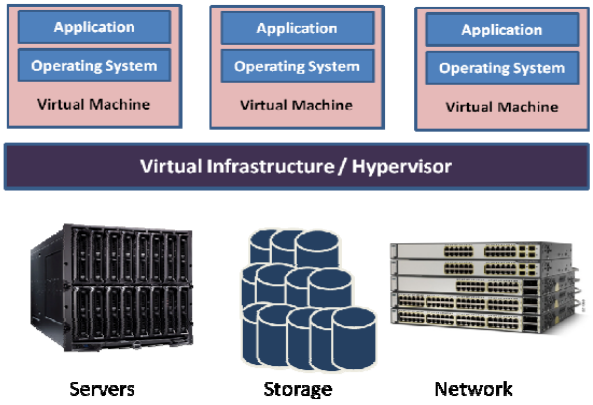


Figure 4 Virtualization Architecture

Virtualization technologies allow multiple virtual machines, with heterogeneous operating systems to run side by side and in isolation on the same physical machine. By emulating a complete hardware system, from processor to network card, each virtual machine can share a common set of hardware unaware that this hardware may also be being used by another virtual machine at the same time. The operating system running in the virtual machine sees a consistent, normalized set of hardware regardless of the actual physical hardware components.

Virtualization is adopted in cloud architecture because of the following reasons

- Hardware independence: The guest VM Sees the same Hardware regardless of the host Hardware
- Isolation –VM's operating system is isolated from the host operating system
- Encapsulation–Entire VM encapsulated into a single file
- Simplified administration because of Hardware
- Increased hardware utilization, Server consolidation, Decreased provisioning, independence & portability times and Improved security
- The other reasons include reduced capital expenditure, reduced operating expenditure, reduced risks of data outage and reduced energy consumption

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

A. Virtualization terminology:

Host Machine: A host machine is the physical machine running the virtualization software. It contains the physical resources, such as memory, hard disk space, and CPU, and other resources, such as network access, that the virtual machines utilize.

Virtual Machine: The virtual machine is the virtualized representation of a physical machine that is run and maintained by the virtualization software. Each virtual machine, implemented as a single file or a small collection of files in a single folder on the host system, behaves as if it is running on an individual, physical, non-virtualized PC.

Virtualization Software: Virtualization software is a generic term denoting software that allows a user to run virtual machines on a host machine.

B. Components of virtualization:

Guest OS: A guest OS is an operating system that runs in a virtual environment. A guest OS may be a client desktop, physical server or any other operating system that runs independently of dedicated hardware resources. Instead, the guest OS uses hardware resources allocated dynamically through a hypervisor or similar intermediary software.

Hypervisor or virtual machine manager (VMM): A hypervisor also called a virtual machine manager (VMM), which is a program that allows multiple operating systems to share a single hardware host. Each operating system appears to have the host‘s processor, memory, and other resources all to itself. The task of this hypervisor is to handle resource and memory allocation for the virtual machines, ensuring they cannot disrupt each other, in addition to providing interfaces for higher level administration and monitoring tools [8].

A virtual machine monitor monitors a system of virtual machines (sometimes called hardware virtual machines), which allow the sharing of the underlying physical machine resources between different virtual machines, each running its own operating system. A virtual machine monitor monitors the software layer providing the virtualization, which is called a virtual machine. The VMM is the control system at the core of virtualization. It acts as the control and translation system between the VMs and the hard**ware.**

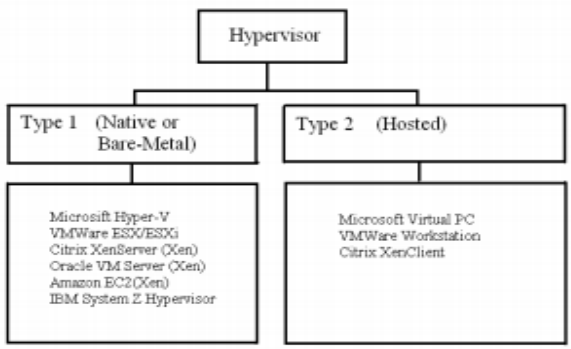


Figure 5 hypervisor types

III. Resource Management

Resource Management is the process of managing the Physical Resources like CPU, Memory, Network etc across various Virtual Machines (VM) based on policies. The figure-6 below gives positioning of resource management in the management and distributed services in virtualization architecture.

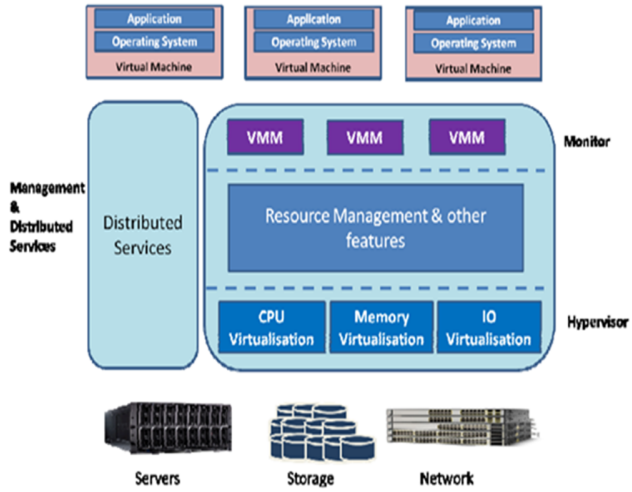


Figure 6: Management & Distributed Services in Virtualization Architecture

The Goal of resource management is three fold.

- Performance isolation: This prevent VMs from monopolizing resources and Guarantees predictable service rates
- Efficient utilization: This is achieved through exploiting under committed resources and over commit with graceful degradation
- Support flexible policies: Absolute service-level agreements are met and relative importance of VMs is controlled efficiently.

A. Capacity monitoring

The key metrics monitored for capacity planning are:

(i) Server utilization: Peak/average server resource utilization – memory /CPU/resource, server bottlenecks and correlation with a number of users/VMs.

(ii) Memory usage: Memory utilization on each server, capacity bottlenecks and relationship with number of users/VMs and with different cloud services.

(iii) Network usage: Peak/average network utilization, capacity/bandwidth bottlenecks and relationship with a number of users/VMs and with different cloud services.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

(iv) Storage utilization: Overall storage capacity metrics, VM/virtual disk utilization, I/O performance metrics, snapshot monitoring and correlation with a number of users/VMs and with different cloud services.

IV. MEMORY VIRTUALIZATION

It divides the physical memory, allocate memory for virtual machine instances when starting up, and release memory from virtual machines when shutting down. Every running instance of OS sees a continuous memory space and is isolated from the memory space of other instances. The hypervisors are capable of memory address conversion from the guest instance physical memory address to the machine physical address. The operating system of a running instance maps the application virtual memory to guest instance physical memory. It can increase the memory allocation to a guest OS at a later stage. The virtualization software has commit feature for the virtual machines.
The operating system maps the virtual page numbers to physical page numbers that are stored in page tables. All modern x86 CPUs contain a memory management unit (MMU) and a translation look aside buffer (TLB) for improvement virtual memory. In order to run multiple virtual machines on a system, another level of memory virtualization is essential. Thus, one has to virtualized the MMU to support the guest OS. The guest OS controls the mapping of virtual addresses to the guest memory physical addresses, but the guest OS cannot have access to the actual machine memory. The TLB hardware is used by the VMM to map virtual memory directly to the machine memory in order to avoid the two levels of translation on every access.

A. Memory Management

Memory manager (MM) such as Hypervisor or ESXi in VMware uses five memory management mechanisms. Page sharing, ballooning, memory compression, swap to host cache, and regular swapping to dynamically reduce the amount of machine PM required for each VM. Virtualization causes an increase in the amount of PM required due to the extra memory needed by MM for its own code and for data structures.

This additional memory requirement can be separated into two components

- A system-wide memory space overhead for the VM kernel and various hosts.
- An additional memory space overhead for each VM.

The amounts of memory reserved for these purposes depend on a variety of factors, including the number of virtual CPUs, the configured memory for the guest operating system, whether the guest operating system is 32-bit or 64-bit, and which features are enabled for the virtual machine.

B. Re-Claiming unused Memory

The memory manager of the hypervisor detects whether the virtual memory is actually used by the guest OS or not. If not, the hypervisor shall be able to assign the unused part of the memory to another guest OS, so that the memory can be shared among the guest OS. Hence this feature is required for memory over-commitment.

This is achieved through traditional method of adding transparent swap layer or using an implicit co-operation. Ballooning is a method where Guest OS manages memory implicit cooperation by paging in / out of the virtual disk. In Page Sharing, multiple VMs running same OS de-duplicate redundant copies of code, data etc.

C. Non Uniform Memory Access scheduling

Periodic rebalancing of the memory usage computes VM entitlements & memory locality, assign “home” node for each VM and migrate VMs and pages across nodes VM migration is to move all VCPUs and threads associated with VM and migrate to balance load and improve locality Page migration allocates new pages from home node and carries migration and replication.

D. Implementations of Memory Virtualization

Application-level integration – Applications executing on connected computers openly connect to the memory pool through an Application Programming Interface(API)

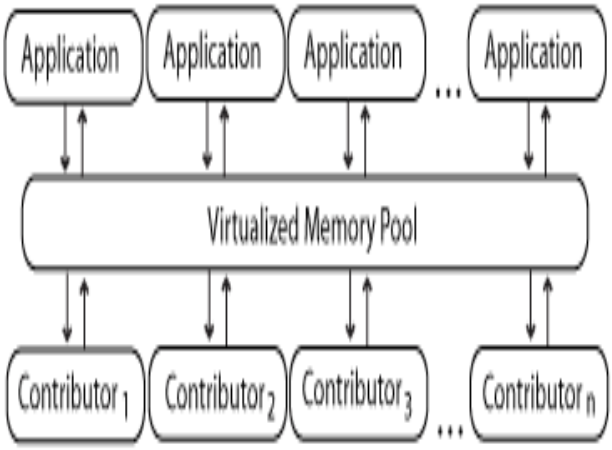


Figure 7: Memory Virtualization through Application Level Integration

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Operating System Level Integration – The operating system first connects to the memory pool, and makes that pooled memory available to applications.

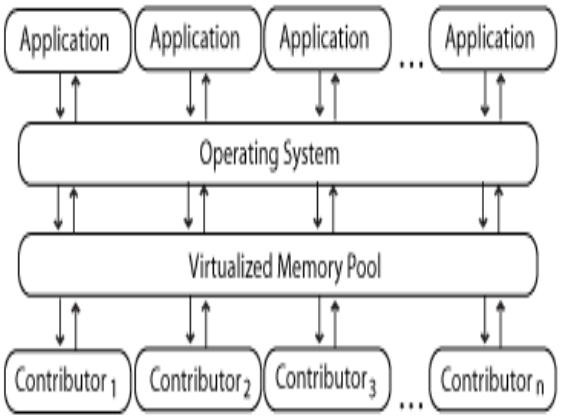


Figure 8: Memory Virtualization through Operating System Level Integration

V. CONCLUSION

Data center and desktop computing successfully use virtualization for better utilization of computing capacity, to balance computing load, manage complexity and parallelism and improve security by isolation. However Virtualization may not work well for Resource-intensive applications where VMs may have RAM/CPU/SMP limitations or situations where custom hardware devices are required. Some hardware architectures or features are impossible to be virtualized as certain registers or states are not exposed. Mobile and embedded computing currently lag behind virtualization since most hypervisors only support the x86 platform, require large memories, have poor real-time support and are inefficient with microkernel Operating System's.

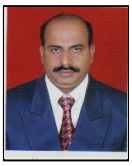
REFERENCES

1. Q. Ali, "Scaling web 2.0 applications using Docker containers on vSphere 6.0," http://blogs.vmware.com/performance/2015/04/scaling-web-2-0-applicationsusing-docker-containers-vsphere-6-0.html, 2015, (Accessed on 03/21/2016).
2. J. Almeida, V. Almeida, D. Ardagna, C. Francalanci, and M. Trubian, "Resource Management in the Autonomic Service-Oriented Architecture," in Proceedings of the 2006 IEEE International Conference on Autonomic Computing (ICAC 2006), June 2006, pp. 84–92.
3. J. Anselmi, E. Amaldi, and P. Cremonesi, "Service Consolidation with End-to-End Response Time Constraints," in Proceedings of 34th Euromicro Conference on Software Engineering and Advanced Applications (SEAA 2008.), September 2008, pp. 345–352.
4. M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica et al., "A view of cloud computing," Communications of the ACM, vol. 53, no. 4, pp. 50–58, 2010.
5. B. Arnold, S. A. Baset, P. Dettori, M. Kalantar, I. I. Mohomed, S. J. Nadgowda, M. Sabath, S. R. Seelam, M. Steinder, M. Spreitzer, and A. S. Youssef, "Building the ibm containers cloud service," IBM Journal of Research and Development, vol. 60, no. 2-3, pp. 9:1–9:12, March 2016.
6. M. Assuncao, M. Netto, B. Peterson, L. Renganarayana, J. Rofrano, C. Ward, and C. Young, "CloudAffinity: A framework for matching servers to cloudmates," in Proceedings of the 2012 IEEE Network Operations and Management Symposium (NOMS 2012), April 2012, pp. 213–220.
7. J. Banks, J. S. Carson, B. L. Nelson, and D. M. Nicol, Discrete-event system simulation. Prentice Hall, 2010.
8. P. Barham et al., "Xen and the art of virtualization," in Proceedings of the 19th ACM Symposium on Operating Systems Principles (SOSP 2003), October 2003, pp. 164–177.
9. L. A. Barroso and U. H¨ olzle, "The case for energy-proportional computing," Computer, vol. 40, no. 12, pp. 33–37, Dec. 2007.
10. C. L. Belady and D. Beaty, "Roadmap for datacom cooling," ASHRAE journal, vol. 47, no. 12, p. 52, 2005.
11. A. Beloglazov and R. Buyya, "Energy efficient allocation of virtual machines in cloud data centers," in Proceedings of the 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing (CCGrid 2010), May 2010, pp. 577–578.
12. "Adaptive threshold-based approach for energy-efficient consolidation of virtual machines in cloud data centers," in Proceedings of the 8th International Workshop on Middleware for Grids, Clouds and e-Science, December 2010, pp. 4:1–4:6.
13. F. Caglar, S. Shekhar, and A. Gokhale, "A performance Interference-aware virtual machine placement strategy for supporting soft realtime applications in the cloud," Institute for Software Integrated Systems, Vanderbilt University, Nashville, TN, USA, Tech. Rep. ISIS-13-105, 2013.
14. R. Calheiros and R. Buyya, "Energy-efficient scheduling of urgent Bag-of-Tasks applications in clouds through DVFS," in Proceedings of the 6th IEEE International Conference on Cloud Computing Technology and Science (CloudCom 2014), December 2014, pp. 342–349.
15. R. N. Calheiros, M. A. Netto, C. A. De Rose, and R. Buyya, "EMUSIM: an integrated emulation and simulation environment for modeling, evaluation, and validation of performance of cloud computing applications," Software: Practice and Experience, vol. 43, no. 5, pp. 595–612, 2013.
16. M. Chen, H. Zhang, Y.-Y. Su, X. Wang, G. Jiang, and K. Yoshihira, "Effective VM sizing in virtualized data centers," in Proceedings of the 2011 IFIP/IEEE International Symposium on Integrated Network Management (IM 2011), May 2011, pp. 594–601.
17. Y. Chen, S. Alspaugh, D. Borthakur, and R. Katz, "Energy efficiency for large-scale MapReduce workloads with significant interactive analysis," in Proceedings of the 7th ACM European Conference on Computer Systems, April 2012, pp. 43–56.
18. Y. Chen, A. S. Ganapathi, R. Griffith, and R. H. Katz, "Analysis and lessons from a publicly available Google cluster trace," University of California, Berkeley, Tech. Rep., 2010.
19. P. DELFORGE, "Energy efficiency, data centers—NRDC," http://www.nrdc.org/ energy/data-center-efficiency-assessment.asp, (Accessed on 02/18/2016).
20. B. H. Li, X. Chai, and L. Zhang, "New advances of the research on cloud simulation," in *Advanced Methods, Techniques, andApplications inModeling and Simulation*, vol. 4 of *Proceedings in Information and Communications Technology*, pp. 144–163, 2012.
21. S. Jafer, Q. Liu, and G. Wainer, "Synchronization methods in parallel and distributed discrete-event simulation," *Simulation Modelling Practice andTheory*, vol. 30, pp. 54–73, 2013.
22. R. Fujimoto, A. Malik, and A. Park, "Parallel and distributed simulation in the cloud," *SCS Modeling and Simulation Magazine*, pp. 1–10, 2010.
23. A. W. Malik, A. J. Park, and R. M. Fujimoto, "An optimistic parallel simulation protocol for cloud computing environments," *SCS M&S Magazine*, vol. 4, 2010.
24. A. J´avor and A. Fur, "Simulation on the Web with distributed models and intelligent agents," *Simulation*, vol. 88, no. 9, pp. 1080–1092, 2012.
25. IEEE Std 1516.1-2010, *IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA)*, Framework and Rules Specification, 2010.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

26. IEEE Std 1516.2-2010, *IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA)*, Object Model Template (OMT) Specification, 2010.
27. IEEE Standard, *1516.1-2010—IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA)—Federate Interface Specification*, 2010.
28. Google, "Google App Engine", (2012), [online]. Available: cloud.google.com [Nov 1, 2012].
29. Amazon, "Amazon Elastic Compute Cloud (Amazon EC2)", (2012), [online]. Available: aws.amazon.com/ec2/ [Nov 1, 2012].
30. Microsoft, "Windows Azure.", (2012), [online]. Available:windowsazure.com [Nov 1, 2012].
31. IBM, "SmartCloud." (2012), [online]. Available: ibm.com/cloudcomputing [Nov 1, 2012].
32. P. Mell and T. Grance, "The NIST definition of cloud computing(draft)," *NIST special publication*, vol. 800, p. 145.
33. A. Desai, "Virtual Machine." (2012), [online]. Available: http://searchservervirtualization.techtarget.com/definition/virtualmachin
34. VMWare, "vSphere ESX and ESXi Info Center.", (2012), [online]. Available: vmware.com/products/vsphere/esxi-and-esx [Nov 1, 2012].
35. Microsoft, "Windows Virtual PC.", (2012), [online]. Available: http://www.microsoft.com/windows/virtual-pc/ [Nov 1, 2012].
36. Xen, "Xen Hypervisor.", (2012), [online]. Available: http://www.xen.org/products/xenhyp.html [Nov 1, 2012].
37. Microsoft, "Hyper-V Server 2012.", (2012), [online]. Available: microsoft.com/server-cloud/hyper-v-server/ [Nov 1, 2012].
38. KVM, "Kernel-based Virtual Machine.", (2012), [online]. Available: linux-kvm.org [Nov 1, 2012].
39. Oracle, "VirtualBox.", (2012), [online]. Available: virtualbox.org [Nov1, 2012]

Author Biographies



P.V.S.S.GANGADHAR is Scientist-D in NIC & Ph.D Scholar. Presently studying Ph.D, Department of Information Technology at Gitam Institute of Technology, Gitam University, Vishakapatnam, and AndhraPradesh, India. His Research interests e-governance, Cloud computing, fuzzy logic and Data mining.

Dr. Ashok Kumar Hota is Scientist-F at NIC, MEITY,OSU, Bhubaneswar, Govt of India. His research interests include e-governance, Tribal informatics, Cloud Computing, Data Mining, and Big Data Analytics. He published several papers in International conferences and journals

Dr. Mandapati Venkateswara Rao is Professor in Department of Information Technology at Gitam Institute of Technology, Gitam University, Vishakhapatnam, India. He has received M.Tech in CST and PhD in Robotics from Andhra University. His Research Interests includes Robotics, Cloud Computing and Image processing. He published several papers in International conferences and journals.

Dr.Vedula Venkateswara Rao is Professor in the Department of Computer Science Engineering at Srivasavi Engineering College, tadepalligudem, India. He received Matsers Degree in Computer Science Engineering from JawaharLalNehru Technological University Kakinada, Masters Degreen In Information Technology from Punjabi University, Patiayala, India and PhD from Gitam University. His research interests include Cloud Computing and Distributed Systems, Data Mining, Big Data Analytics and Image Processing. He published several papers in International conferences and journals.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

A Comparative Study on MAC Protocols for Wireless Sensor Networks on Energy Reduction

Rajan Sharma,Research Scholar,
Department of Electronics and Communication Engineering
K. Gujral Punjab Technical University, Jalandhar
Punjab, India
E-mail: rajansharma.ece@gmail.com

Balwinder Singh Sohi
Department of Electronics and Communication Engineering
CGC Group of Colleges
Punjab, India
E-mail: bssohi@yahoo.com

***Abstract*—Wireless Sensor Networks (WSNs) leads to a most versatile solution in numerous applications such as smart building, tracking the targets and many more. Basically, WSNs comprises of high count of inexpensive sensor nodes those are scattered in target area for gathering the desired information. Medium Access Control (MAC) Protocols plays key role for energy efficient operation in WSNs as it manages radio communication on the shared medium. WSN is multidisciplinary area of research and has increased acceptance because of its application. Due to lower sensing range and changing topology there is needed to achieve an efficient medium access for energy efficiency. Variety of MAC protocols have been proposed for WSNs, based upon different objectives like energy efficiency, delay, throughput and packet loss. In recent years, WSNs has gained special attention in research community. This paper outlines the properties of WSNs that plays and important role for designing of MAC layer protocol highlighting pros and cons. Finally, we investigate on various MAC protocols designs for WSN.**

Keywords-Wireless Sensor Networks; MAC; Energy Reduction; Energy Efficiency

I. INTRODUCTION

Wireless Sensor Networks (WSNs) is an amalgamation of large number of sensing nodes with one or more base stations. In sensor network, sensing nodes are deployed randomly to monitor the physical parameters such as temperature, sound, pressure, etc. Sensor nodes are autonomous petite devices that comprises of components such as sensing module, processing module, radio unit and power unit as presented in Figure 1. Sensor nodes capture, gather analyze data and communicate with other sensor nodes through radios. The gateways exchange gathered data with applications [1]. These nodes basically sense, compute and communicate the data to other nodes or base station. It preserves the association to the outer world where the activities such as data management, data evaluation, and data grant providence to the effective approaches.

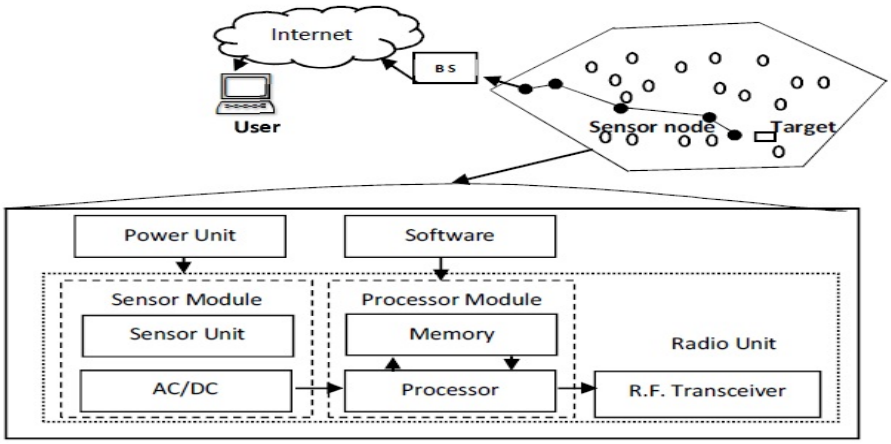


FIGURE 1: BASIC STRUCTURE OF WIRELESS SENSOR NETWOR

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Wireless Sensor Network (WSN) provides applications for various fields like defense, environment monitoring, disaster relief, target tracking and detection etc.

The data link layer operates between physical layer and network layer. The data link layer is further segmented into Logic Link Control (LLC) sub layer and Medium Access Control (MAC) sub layer. The Control Mechanism for channel access at MAC layer influence significantly on energy efficiency, performance and lifetime of the network as it is responsible for fair and efficient access of shared medium by nodes, resolving conflicts between competitor nodes and correcting errors. MAC sub layer provides various other functions like dentation, addressing, frame recognition and delimiting, gathering information of source station and transferring the data [2]. According to 802.3 functions MAC needs to fulfill are: padding, checking errors in the frame sequence, transferring and receiving of frames, appending and checking frame check sequence and adding and removing preamble [3].

WSNs Medium Access Control (MAC) protocol design requirements are entirely different from the conventional wireless networks. The prime design requirement in WSNs is to achieve higher energy efficiency but in the case of conventional wireless computer networks, higher throughput is needed. Since only source of energy for sensors nodes are batteries which are inconvenient to replace so saving of energy is primary goal for design of MAC layer protocols. As sensor nodes are battery powered, these nodes face energy depletion issues because of which packets are delayed or abort.MAC protocols are designed with an aim to

reduce power consumption and to prolong network lifetime.

Hence, we need the operations to be energy efficient so that the sensors lifetime can be extended. Out of sources of power consumption (sensing, radio operation and computation of data) the radio operation consumes maximum power of sensor nodes [4]. Reducing energy consumption with a aim to achieve higher energy efficiency are the significant aspect that can increase the battery life of mobile terminals which is limited [5]. There are many protocols which have been designed for wireless sensor networks. The protocols which are related to the wireless sensors should have an awareness related to the energy to increase the life of the network [6]. The sources of energy wastage in WSNs that occurs due to the following reasons:

1. *Overhearing:* Sometimes nodes gather the packets that were supposed to be intended to other nodes on the shared medium which wastes the energy; later leads to energy depletion [7]. These sensor nodes may run out of power while they need to be in active state.
2. *Packet Overheads:* The size of control packets (headers and trailers) associated with data should be short and exchanging handshaking packets should be limited as they consume lot of energy.
3. *Collision:* Collision of packets either due to retransmissions or discarding of packet requires more

   energy of the nodes in comparison to the normal transmission.
4. *Idle Listening:* Sensor nodes sometimes waste energy when they are active in accessing channel but they have no packets to listen.

The major goal in designing MAC protocol is to minimize these issues to save power and to achieve higher energy efficiency [8]. Section II discuss the Prior work for various types of MAC protocol used in WSN that provide solution for power consumption reduction with other constraints of sensor nodes. Section III shows the Valuation of MAC protocols discovered, followed by providing the comparisons and conclusion in Section IV.

● – *USA* □- *Singapore* ○ - *California* ◊ -*Switzerland* *- *Italy* # - *Saudi Arabia*

Parameters	Late 90-2000	2000-2005	2005-2010	2010-2015	2015-2017
Packet Size		◊		◊ ●	
Energy		◊	◊	● ◊ # □	●
Throughput	○		●	● *	
HTP			○	◊	
Latency			◊	● □ *	
Synchronization			● ○	●	
Efficiency			● ○ □	◊ #	●

FIGURE 2: MAJOR CONTRIBUTION BY VARIOUS COUNTRIES IN THE FIELD

II. PRIOR WORK

The work in the respective field has always been on its track. Thus, the analysis of the major contributed countries in the golden years of the field are discussed in Figure 2.

Kulkarni [9] explains, the major challenges faced by the sensor nodes, those are the problem of early sleeping, synchronization and overheads of packets. In some of the protocols energy consumption was improved yet need to be compromise on other metrics like packet delay and throughput of the network. IEEE 802.15.4 provides preamble based protocol which can save the energy but face the problem related to acknowledgement of preamble.

W. Ye et al. [10] provides sensor MAC (S-MAC), protocol that manages periodic sleep and listen synchronization. The message passing concept where packet is divided into frames before sending reduces communication overheads, which further saves the energy. The sleep schedule helps in saving energy of idle listening. The variable traffic load creates problem in Listen and sleep period which leads to a decrease in efficiency of algorithm. Overhearing issues are still there because of the collision probability which was also faced in Halkes [17].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

El-Hoiydi [11] proposed Spatial TDMA and CSMA with preamble sampling protocol where the nodes use different channel for data and control. TDMA method is used for data channel whereas CSMA method is used for control channel. The preamble along with data is sent with an aim to reduce the power consumption by finding active and in-active state of nodes but the size of preamble may be the disadvantage.

A.El-Hoiydi et al. [12] gave Wise MAC, protocol that is quite similar to CSMA but use a single channel to reduce the idle listening. Dynamic preamble is used to reduce the power consumption over fixed length preamble. This helps to manage the variable traffic conditions. Hidden terminal Problem creates issue in efficiency of the algorithm. This algorithm suffers from higher latency and increase in power consumption because of buffering of broadcasting packets for finding active and in active sensor nodes.

Rajendran et al. [13], proposed a TDMA based TRAMA (Traffic Adaptive MAC), that build the vitality effectiveness by wiping out concealed terminal issue and guarantees that the hubs display at one jump from the transmitter will get the bundles without the impacts. Time is categorized into schedule access and random-access periods. Less collision probability and more sleep time helps in achieving better results but the timing slots creates issue of synchronization.

Bao et al. [14] proposed, node activation multiple access (NAMA) where time slots are divided and nodes present at two hop from the transmitter is elected to avoid the collision. This technique faces problem of delay due to more sleep time. Tay Y.C et al. [15], proposed event driven sensing environment. SIFT is MAC based protocol that uses probability distribution function that is non-uniform. The latency and energy consumption was taken into consideration but idle listening was still there because of slot listening while sending even the overhearing problem was forced. So, this technique increases complexity.

Alazzawi and Elkateeb [16], focuses on features of WSN are dense deployment multi-hop routing, self-organizing and dynamic topology. The algorithm divides time into slots and assigns those to the sensing nodes depending upon data gathering tree. Though the latency is reduced but collision could be avoided.

Halkes et al. [17] proposed T-MAC to handle variable traffic load. Time out MAC (T-MAC) provided better result as the nodes sleep when periods end listens, but synchronization can create problem of early sleeping.

Lin et al. [18] proposed Dynamic Sensor MAC (DS-MAC), that include feature of dynamic duty cycle that helps in reducing latency. In synchronization periods, the sensing nodes have one hop latency value. When this value is high it reduces twice so that neighboring nodes are not affected which hence saves the power consumption.

Safwat et al. [19], provides integrated MAC layer protocol with routing protocol. Cui et al. [20] integrates MAC protocol with routing protocol at physical layer where the TDMA scheme of variable length helps in energy utilization but the node distance creates the complexity. Ding et al. [21] proposed MINA that integrates routing protocol with MAC in multi hop infrastructure network architecture; the sensing nodes are grouped under base stations which are at one hop distance. This head of cluster will decide the schedule for data transmission, but this technique is in tolerant to the failure.

Zorzi et al. [22] proposed Geographic Random Forwarding (GeRaF) which combines routing protocol with MAC protocol using CSMA/CA technique. Rugin et. al. [23] improved by reducing up to one channel system, but still the latency issue was there due to transmission request during awake time.

Roy and Sharma [24] proposed AEEMAC that is Adaptive energy efficient MAC protocol that uses duty cycle to avoid idle listening and reusing the channel by adaptive sleeping. This protocol by protocol with MAC protocol using CSMA/CA technique.

Protocol	Technique	Communication pattern support	Adaptively to change	Time needed for synchronization
S-MAC	CSMA	All	Good	No
T-MAC	CSMA	All	Good	No
DS-MAC	CSMA	All	Good	No
Wise-MAC	Np-CSMA	All	Good	No
D-MAC	TDMA/ Slotted ALOHA	Coverage based	Good	No
TRAMA	TDMA	All	Good	Yes
SIFT	CSMA/ CA	All	Good	No

TABLE 1 ASSESSMENT OF MAC PROTOCOLS

This protocol shows better results than ESMAC that was proposed by S. Hayat [25], but still complexities in data and acknowledgment packets where there. Kwan-WuChin et al. [26] provides E2MAC that focus on providing better results for packet delivery rate and delay in Packets.

Bhuiyan et. al [27] proposed Intelligent MAC (I-MAC) that works on the basis of intelligent sleep and wakeup schedule and is inspired from 802.11 DCF. Cano at al. [28], proposed MAC protocol in which preamble of IEEE 802.15.4 is used to find the sleep time for the sensor node. This process creates packet overhead reducing the efficiency of the protocol.

N. Javaid [29] reduced the ideal listening time with proposed protocol by extending MAC header. The sensor nodes periodically check the channel so that they could increase their sleep time which provides energy efficiency.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Later Escolar et al.[30] discussed that due to dynamic change of topology the energy saving is the major issue in WSN. A lot of research has been done to provide protocols for this purpose. Most of the protocols aim to achieve energy reduction by focusing on data transmission. This provides reduction in packet loss as well as packet delay. Figure 3 depicts the complete parametric contribution by various researchers in the golden era of the field.

FIGURE 3: PARAMETRIC CONTRIBUTION BY VARIOUS RESEARCHERS IN THE MAJOR YEARS OF RESEARCH

Parameters						
	Packet size		[28]		[19][12]	[28]
	Energy Consumption		[10]	[15]	[9][10][12][18]	[19]
	Throughput			[21]	[9][16][22][23]	
	HTP	[24]	[27]	[13]	[12]	[29]
	Latency		[26]	[15]	[12][18][10]	[30]
	Synchronization	[25]		[13][17]		
	Efficiency		[10]	[13][17]	[10]	[19]

III. THE ASSESSMENT OF MAC PROTOCOLS EXPLORED

Table 2. Provides comparison of various MAC protocol based upon parameters like communication pattern, time needed for synchronization and adaptability to the changes. The Table 2 and Figure 4 provides the comparison between ESAC, E2MAC, IMAC and AEEMAC protocols on the basis of data and acknowledgement packets which are sent during the time slots.

Protocol	Data packets	Acknowledgement packets	Total	Total/min
ES-MAC	490	490	980	16
E2MAC	240	240	480	8
AEEMAC	100	100	800	16
I-MAC	80	80	170	2

TABLE2. COMPARISON OF DIFFERENT MAC PROTOCOL

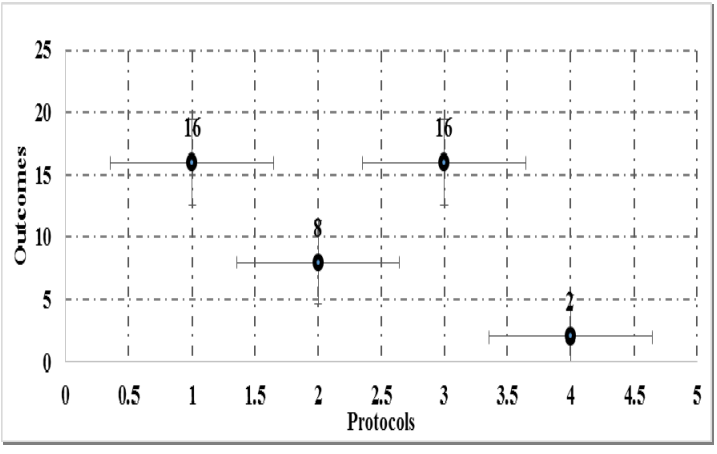


FIGURE 4. COMPARISON OF MAC PROTOCOLS BASED UPON NUMBER OF PACKETS

IV. CONCLUSION

Wireless Sensor Networks (WSNs) have become popular due to its exceptional capabilities & numerous applications in many fields such as military, industries, health sector and surveillance etc. Various MAC protocols available in the literature have been reviewed in this paper. Power saving of WSNs is considered as the prime aspect in designing MAC protocols because it is not feasible to change the batteries of sensor nodes in many applications wherever there is less human intervention. The responsibility of the MAC protocol is to manage radio communication fairly and efficiently on shared medium. In this survey, we found that the main target is to keep the sensor nodes in power saving mode to the maximum extent and to reduce the parameters which are source of energy inefficiency such as re-transmission due to collision or congestion, idle channel sensing, overhearing, and overhead due to control messages so that lifetime of network can be increased. This review will be beneficial as it highlights to the possibilities of the development of some new energy efficient MAC approaches to extend lifetime of the WSNs.

ABBREVIATIONS

WSN	Wireless Sensor Network
MAC	Medium Access Control
LLC	Logic Link Control
HTTP	Hyper Text Transfer Protocol
SMAC	Sensor Medium Access Control
TDMA	Time Division Multiple Access
CSMA	Carrier Sense Multiple Access
TRAMA	Traffic Adaptive Medium Access
NAMA	Node Activation Multiple Access
T-MAC	Time Out Medium Access Control
AEEMAC	Adaptive Energy Efficient Energy Protocol
I-MAC	Intelligent Medium Access Control
E2MAC	Energy Efficient Medium Access Control
ES-MAC	Energy Efficient Sensor Medium Access Control

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

REFERENCES

[1] S. Kurt, H. U. Yildiz, M. Yigit, B. Tavli, and V. C. Gungor, "Packet Size Optimization in Wireless Sensor Networks for Smart Grid Applications," *IEEE Trans. Ind. Electron.*, vol. 64, no. 3, pp. 2392–2401, 2017.

[2] N. Zaman, L. T. Jung, and M. M. Yasin, "Enhancing Energy Efficiency of Wireless Sensor Network through the Design of Energy Efficient Routing Protocol," vol. 2016, 2016.

[3] S. Zhuo, Z. Wang, Y. Q. Song, Z. Wang, and L. Almeida, "A Traffic Adaptive Multi-Channel MAC Protocol with Dynamic Slot Allocation for WSNs," *IEEE Trans. Mob. Comput.*, vol. 15, no. 7, pp. 1600–1613, 2016.

[4] K. Pattamasiriwat and C. Jaikaeo, "Evaluation of Low Power Listening MAC Protocol on Network Monitoring in Wireless Sensor Networks," pp. 0–5, 2017.

[5] M. Dong, K. Ota, and A. Liu, "RMER: Reliable and Energy-Efficient Data Collection for Large-Scale Wireless Sensor Networks," *IEEE Internet Things J.*, vol. 3, no. 4, pp. 511–519, 2016.

[6] M. Dong *et al.*, "Evaluation of Low Power Listening MAC Protocol on Network Monitoring in Wireless Sensor Networks," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 0–5, 2016.

[7] J.-D. Decotignie, C. Enz, V. Peiris, and M Hübner, "WiseNET: An Ultra Low-Power Concept for Wireless Sensor Networks," *Tech. Mess. - Platf. für Methoden, Syst. und Anwendungen der Messtechnik*, vol. 77, no. 2, pp. 107–112, 2010.

[8] V. K. Sachan, "Energy Efficient Wireless Sensor Networks using Co- operative MIMO : A Technical Review," vol. 135, no. 11, pp. 20–27, 2016.

[9] S. S. Kulkarni, "TDMA service for sensor networks," in *24th International Conference on Distributed Computing Systems Workshops, 2004. Proceedings.*, 2004, pp. 604–609.

[10] W. Ye, J. Heidemann, and D. Estrin, "Sleeping for Wireless Sensor Networks," vol. 12, no. 3, pp. 493–506, 2004.

[11] A. El-Hoiydi, "Spatial TDMA and CSMA with preamble sampling for low power ad hoc wireless sensor networks," in *Proceedings - IEEE Symposium on Computers and Communications*, 2002, pp. 685–692.

[12] A. El-Hoiydi and J. D. Decotignie, "WiseMAC: An ultra low power MAC protocol for multi-hop wireless sensor networks," *Algorithmic Asp. Wirel. Sens. Networks*, no. 5005, pp. 18–31, 2004.

[13] V. Rajendran, "Energy-efficient medium access control for wireless sensor networks," *Wirel. Networks*, pp. 445–452, 2006.

[14] L. Bao and J. J. Garcia-Luna-Aceves, "A New Approach to Channel Access Scheduling for Ad Hoc Networks," *Comput. Eng.*, pp. 210–220, 2001.

[15] Y. C. Tay, K. Jamieson, and H. Balakrishnan, "Collision-minimizing CSMA and its applications to wireless sensor networks," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 6, pp. 1048–1057, 2004.

[16] L. K. Alazzawi, A. M. a. M. Elkateeb, and A. Ramesh, "Scalability Analysis for wireless sensor networks Routing Protocols," *22nd Int. Conf. Adv. Inf. Netw. Appl.*, pp. 139–144, 2008.

[17] G. P. Halkes, T. V. A. N. Dam, and K. G. Langendoen, "Comparing Energy-Saving MAC Protocols for Wireless Sensor Networks," pp. 783–784, 2005.

[18] P. Lin, C. Qiao, and X. Wang, "Medium Access Control With A Dynamic Duty Cycle For Sensor Networks," in *Proc. {IEEE} Wireless Commun. and Networking Conf. (WCNC)*, 2004, vol. 3, pp. 1534–1539.

[19] A. Safwat, H. Hassanein, and H. Mouftah, "ECPS and E2LA : New Paradigms for Energy Efficiency in Wireless Ad hoc and Sensor Networks," *GLOBECOM '03. IEEE Glob. Telecommun. Conf. (IEEE Cat. No.03CH37489)*, pp. 3547–3552, 2003.

[20] S. Cui, R. Madan, A. J. Goldsmith, and S. Lall, "Joint Routing, {MAC}, and Link-Layer Optimization in Sensor Networks with Energy Constraints," in *Proc. IEEE International Conference on Communications (ICC'05)*, 2005, pp. 725–729.

[21] D. Jin, S. Krishna, R. Kashyapa, and C. Lu Jian, "A multi-layered architecture and protocols for large-scale wireless sensor networks," *Veh. Technol. Conf. 2003. VTC 2003-Fall. 2003 IEEE 58th*, vol. 3, p. 1443–1447 Vol.3, 2003.

[22] M. Zorzi, "A new contention-based MAC protocol for geographic forwarding in ad hoc and sensor networks," *IEEE Int. Conf. Commun.*, vol. 6, no. c, pp. 3481–3485, 2004.

[23] R. Rugin and G. Mazzini, "A simple and efficient MAC-routing integrated algorithm for sensor network," *IEEE Int. Conf. Commun. (IEEE Cat. No.04CH37577)*, vol. 0, no. c, p. 3499–3503 Vol.6, 2004.

[24] A. Roy and N. Sarma, "AEEMAC: Adaptive energy efficient MAC protocol for wireless sensor networks," *2011 Annu. IEEE India Conf.*, pp. 1–6, 2011.

[25] S. Hayat, N. Javaid, Z. A. Khan, A. Shareef, A. Mahmood, and S. H. Bouk, "Energy Efficient MAC Protocols," pp. 1–8.

[26] K.-W. Chin and D. Klair, "E2MAC : An energy efficient MAC for RFID-enhanced wireless sensor networks," *Pervasive Mob. Comput.*, vol. 7, no. 2, pp. 241–255, 2011.

[27] M. M. Bhuiyan, I. Gondal, and J. Kamruzzaman, "I-MAC : Energy Efficient Intelligent MAC Protocol for Wireless Sensor Networks," 2011 17th Asia-Pacific Conf. Commun. 2nd – 5th Oct. 2011 | Sutera Harb. Resort, Kota Kinabalu, Sabah, Malaysia, no. October, pp. 133–[1].

[28] C. Cano, B. Bellalta, J. Barceló, and A. Sfairopoulou, "A novel mAC protocol for event-based wireless sensor networks: Improving the collective QoS," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5546 LNCS, pp. 1–12, 2009.

[29] N. Javaid, A. Ahmad, A. Rahim, Z. A. Khan, M. Ishfaq, and U. Qasim, "Adaptive medium access control protocol for wireless body area networks," Int. J. Distrib. Sens. Networks, 2014.

[30] S. Escolar, S. Chessa, and J. Carretero, "Energy-neutral networked wireless sensors," Simul. Model. Pract. Theory, vol. 43, pp. 1–15, 2014.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

AUTHORS PROFILE

Rajan Sharma has received his degrees of B.Tech & M. Tech from Punjab Technical University, Jalandhar, Punjab, and pursuing Ph.D. in the field of Electronics and Communication from the same University. He is having experience of 13 years in teaching. His research areas of interest are wireless communication, computer networks and wireless sensor networks.

Dr. B.S. Sohi has received his degrees of B.Sc. Engineering, Master of Engineering & Ph.D. in Electronics in years 1971, 1981, 1992 from Panjab University, Chandigarh. He is the Ex- Director of UIET, Panjab University, Chandigarh and presently he is working as Professor in the Department of Electronics and Communication Engineering at CGC Group of Colleges, Mohali, Punjab. He is having experience of 35 years in teaching and administration. He has 105 research publications in various fields. His areas of research are wireless networking, computer networking and wireless sensor networks.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Performance Analysis of adaptive routing protocol for cognitive radio wireless sensor networks using bio-inspired methods

[1]Amit N. Thakare, [2]Dr. Latesh Malik(Bhagat), [3]Mrs. Achamma Thomas
[1,3]Department of Computer Science & Engineering, G.H. Raisoni College of Engineering, Nagpur, MS, India
[2]Department of Computer Science & Engineering, Government College of Engineering, Nagpur, MS, India

***Abstract*:** Wireless sensor network is the need of today's and next generation requirement to provide the optimal solution for routing techniques in telecommunication. The many different challenges and issues are workout by various authors. We consider the fault-tolerance issues with multi-sink and multi-channel probabilistic approach. The sensor nodes are deployed in the region in an unattended fashion for data gathering from different places. The swarm optimization techniques as artificial intelligence are used to find out the optimal solution. In the communication network, there is a need for designing a fault-tolerance routing protocol. The node energy depletion as the nodes is in active mode. The node gets exhausted as it is used continuously for routing the packets. We extend our work in designing the fault-tolerance issues using the multi-sink approach as probabilistic techniques inspired from ant colony optimization for cognitive radio wireless sensor network (CR-WSN) using spectrum sensing, sharing techniques applying on AODV and DSR routing protocol. In this paper we try to analyze the adaptive routing protocol with existing bio-inspired AntHocNet routing protocol.

Keywords: fault-tolerance, Ant Colony Optimization approach, CRWSN, Multi-sink, Multichannel, performance analysis.

I. Introduction

In the wireless sensor network (WSN) are having a wide variety of applications to solve the multiple issues in wildlife tracking, military applications, environmental monitoring, traffic surveillance, health care, etc. Like the ad hoc network wireless sensor network to acquire the knowledge of cognitive network using spectrum frequency utilization with 2.4 GHz. The next generation communication network using the cognitive radio sensor network as a self-organized network that consists of a large number of low cost and low powered sensor devices called as sensor nodes with cognitive capabilities. Each sensor node is equipped with sensing unit for capturing, the wireless transceiver is used to capture events and reach to the base station. The analogy to this technique is an ant colony optimization designed by M. Darigo provide the routing protocol as novel approach such as AntNet, AntHocNet, BeeHocNet, BeeSensor etc. are developed to provide optimal solutions. The cognitive radio sensor network is a self-organized network as we prove earlier in previous publications. The primary user (PU) is used the licensed spectrum band for data sensing, data communication, and data process. If the PU is unused the secondary user (SU) is used as unlicensed spectrum band as PU-SU activity. The PU arrival time and departure time is maintaining the proper utilization of spectrum band. As soon as PU departure time processed the SU starts the arrival process as spectrum holes utilization.

Here the network communication layer is considered for fault tolerance issues in wireless communication links. All the sensor nodes with cognitive capabilities, select the neighbor.
A cluster using nearest neighbor techniques and select the one of node select cluster head to forward the data packet from one node to other node using multiple channels and multi-sink approach using the radio interference ranges of sensor nodes.

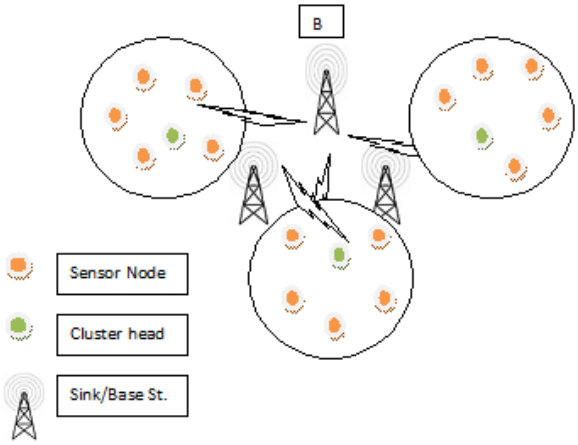


Figure 1. Architecture of multi-sink cognitive radio sensor networks

The retransmission occurs if the failure of data transfer causes a delay, sometimes link failure, and sometimes only using single sink the nodes get exhausted and energy depletion occurs. To solve this problem we are using two sink nodes to avoid the fault occurrences and solve the fault-tolerance issues as similar as inspired ant colony optimization as ANTs starts food searching from source to destination using multiple paths using hops-to-hops to communication as probabilistic approach, either select one path or other as hypothetical approach {0, 1} that it selects either one of the paths or the other path and reach the destination by depositing chemical substances called the pheromone over the all complete path as other ants follow the same path to using pheromone value. The pheromone gets also evaporate as time going on. The obstacle in all complete paths can avoid by ants reach to the destination called forward ants and return back using the same or other bath called backward ants as ACO meta-heuristic.

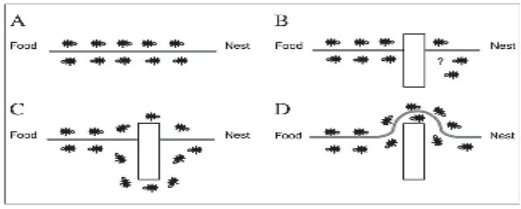


Figure 2. Ant Colony Optimization

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

In ant colonies, the ant has less intelligence (Caro and Dorigo, 1998) and collective behavior has more intelligence with respect to performing any task during the searching food. They divide the number of the working ants as labor to perform the individual task and this is spectrum sharing in cognitive WSNs.

The many researchers carried out experimentation for WSN and biologically inspired methods. We have gone through the designed protocols for adaptive routing protocols based on WSNs. The next generation novel approach is Cognitive radio sensor networks (E. Fadel, I. F. Akyildiz et al., 2017). Here the utilization of unused spectrum band as unlicensed spectrum band for the other applications, used as to sense the band as spectrum sensing and the spectrum sharing with accessing by the other channel is utilized for the proper flow of data communications between the numbers of users exchanging from node to node and collectively provide to base station. Thus, the analogies between the cognitive radio sensor networks as with Ant Colony Optimization (ACO) provides better outcomes for finding the optimized solution for routing techniques in a telecommunication network.

The rest of the section is as follows, Section 2 describes the literature review; section 3 design issues; section 4 provides the proposed protocol; section 5 observed the Performance analysis; section 6 Concluded the paper.

II. Literature Review

Hai Liu et al.,[2009] proposed a how fault-tolerance addresses in different applications of WSN. They consider the five different parameters that are node replacement of WSN, topology control, target & event detection, data gathering and aggregation. N.Fami-Tafreshi et al., [2016] proposed the idea of exploiting the mobility of data collection points (sinks) for an increasing network lifetime of WSN. The proposed method is based on ant colony optimization algorithm. Lei Cao, et al.,[2010] proposed multiple sink cluster WSN schemes which combine two solutions, also proposed an efficient transmission power control scheme for sink centric cluster routing protocol in multiple sink WSN that is distributed, scalable, self-organized, adaptive system. Athar Ali Khan et al., [2017] evaluate the requirement and key design challenges for routing and MAC protocols in the CR-based smart grid & provide research and review of research carried out for CR based routing. E. Fadal et al., [2017] proposed honey bee mating optimization based routing and cooperative channel assignment algorithms. The development of framework decreases the probability of packet loss and preserve high link quality among sensor nodes in the harsh smart grid environment and performance evaluations is done based on the packet delivery ratio, transmission delay, and residual energy. A. Ozan Bicen, Ozgur B. Akan [2011] point out the need for a novel reliability and congestion control mechanism for CRSN with transport layer simulated & demonstrate performance of protocol in CRSN and much more.

III. Design issues of routing protocols

Wireless sensor network has different challenges and issues to overcome the constraints of WSN's and solve the design application issues. The challenges and design issues in WSN are limited energy consumptions, fault tolerance, scalability, productive cost, data aggregation, load balancing, congestion, security, self-organization.

Self-organization: A WSN is expected to remain operational for an extended period of time. Here the new node added in this network, maybe the other nodes fail because of failures or exhaust their batteries becomes un-operational. A routing protocol must resilient to such dynamic & generally unpredictable variations sustain the long-term availability of network services.

Fault-tolerance: CR-WSNs should have self-forming, self-configuration and self-healing properties. In other words, whenever some nodes or links fail, an alternative path that avoids the faulty node or link must be derived. In CR-WSNs, faults can occur for a variety of reasons, such as hardware or software malfunctioning, or natural calamities, e.g., fire, floods, earthquakes, volcanic eruptions, or tsunamis *etc.* A CR-WSN should always be prepared to deal with such situations. The fault tolerance or reliability of a wireless sensor node using the Poisson distribution within the time interval (0,t) occurs.

Scalability: Sensor node deployment in WSNs is application dependent and affects the performance of the routing protocol. A large number of nodes is deployed in the region having short communication range and high failure rates. The routing protocols have effectively acceptable for those challenges.

IV. Proposed adaptive routing protocol

The objective of the proposed routing protocol to solve the fault-tolerance issue as adaptive in nature in cognitive radio sensor network, that is the next generation wireless communication. The WSN is equipped with cognitive capabilities of spectrum sensing, spectrum sharing, spectrum decision with low power battery and energy level was slightly exhausting after transmission of data from multiple routes with multiple sink and multiple channels. As the number of channels increases the variation in output parameters are reflected in the evolution. The neighboring nodes discover the route and coordinate to transmission and sharing of data collectively and form the cluster. After forming the cluster as per the energy level they all node in cluster chose the cluster head.

The all techniques are inspired from ANT colony optimization for finding the food source by trailing the pheromone. Which was followed further by rest of ANTs to find the shortest route and food source called forward ANTs and reverse came back to their nest. The hurdles in between the route are avoided and find out the alternate path to reach the destination. As they also use multiple paths shortest route to achieve the target. Inspired from biological organism, we try to develop the routing protocol with solving some issues in communication technology. The Pseudo code is given below.

If a sink node fails to send the data to the base station, the solution for improvement for solving the issues are using multi-sink, multichannel operation in network. Ant –based clustering is a probabilistic approach with multi-hops techniques. We introduced such a probabilistic based routing protocol for fault-tolerance issues. [Awwad et al., 2010] Here if the one of the sink node fails the solution is to use multi-sink base station. The process of working in the figure given below that there is deployment of sensor nodes in the region and then started the neighboring techniques; the nodes started finding the neighbor in the probabilistic hops manner. Then there is cluster is formed and chose the cluster head (CH) as per energy level of individual node. As per bio-inspired techniques the Poisson distribution, the arrival and departure rate of PU (primary user) provides the performance analysis and solves the issues.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

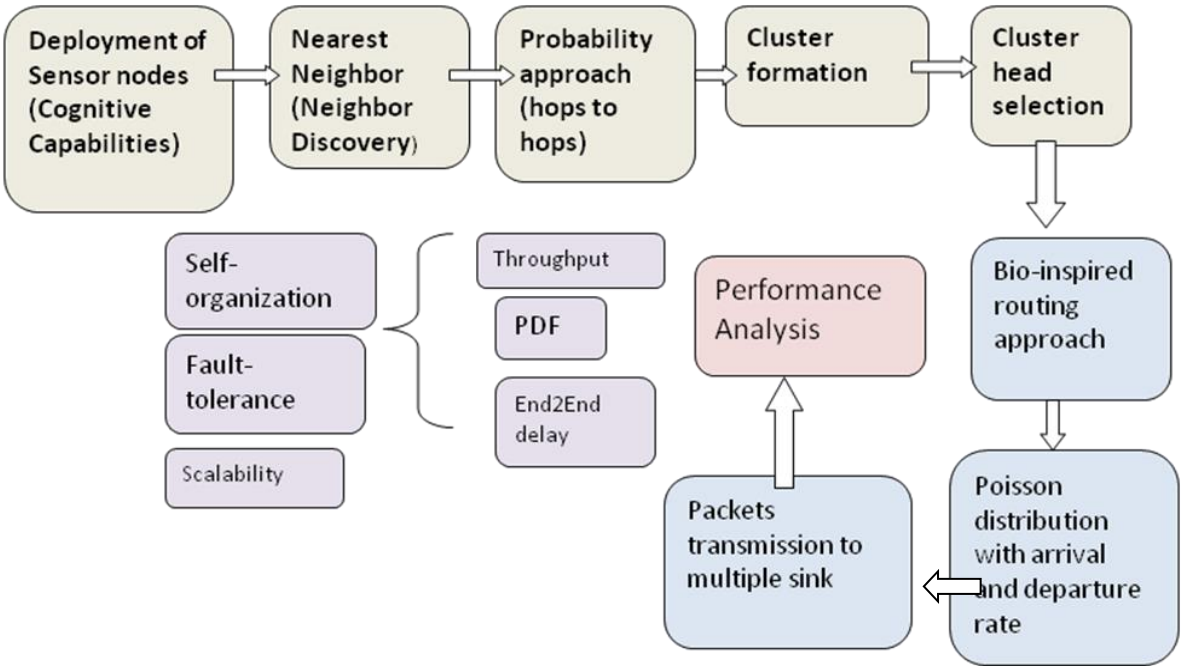


Figure 3. Block diagram of actual working of proposed work

4.2 Pseudo code for routing in WSN

Let consider the CRSN sensor node i and j as edge (i, j) ant start searching food from the best path.
Cluster Head Selection Algorithm

Setup Phase

for *each path (i , j)* **do**
perform spectrum sensing;
end for
finding each hops from i to j as best path (shortest)
for *each ant k1 & k2* **do**
Compute probability for choosing multi- path;
if [*node j nearest neighbor to node I, ant k1 & k2 checks both path & take decision]* ← *chose path*, **then**
Calculate length ;
Compute as shortest path and form a cluster (all nearer nodes);
else longest path
ant k computes best path (i , j) by dropping pheromone trail;
end
pheromone value = 1;
if *node with greater pheromone value then others* **then**
ant k select path(i ,j) in multi-hop fashion;
else if *node which have deeper pheromone value choose as CH* **then**
Advertise as cluster head (CH);
End
Transfer data from CH to Master CH (as sink node);
for each path (i, j) do
If one sink exhaust (because of energy level) the other channel and sink node chosen
End

Steady Phase: Multi –Hop routing algorithm

for *each edge* **do**
Set initial pheromone value 1;
end for
for each CH do compute the visibility
end for
While *not stop* **do**
for *each ant k* **do**
Select CH;
for i=1 to n do
Compute probability Pij and select next CH j with;
Probability pij;
end for
end for
for *each path* **do**
Update the pheromone value;
end for
end while
select shortest path ij &
Choose the CH as a leader for each CH to base station
do
sends packet under leader control;
end for

Communication inside cluster

for each cluster do
All nodes sends data to CH
CH check the availability of network
if (Network is busy) then
perform spectrum sensing;
if a spectrum band is available then
utilize primary network;
transfer the data through primary networks(Pus);
wait for selection of CH and master cluster Head(in which all CH sends data);
if CH is elected in transferring data then
send the data with max packet size;
else
send the data with restricted packet size;
end
else
utilize secondary network(SUs);
Here access SU over CSMA protocol;
end
else
Utilize the secondary network;
end
end

In this paper, we consider the three conditions as channel selection, mobility as a static sink. Initially, we consider the static sink.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The initial assumption in the proposed protocol is considered in the following form.

1. All sensor nodes and sink node (Base Station) are stationary after deployment.
2. The sensor nodes are uniformly distributed with random deployment.
3. The sensor nodes with heterogeneous configuration.
4. All nodes have limited energy.
5. Cross layer MAC layer and Network layer jointly work with TCP transport protocol.

In proposing work we consider two methods for our demonstration of the adaptive routing protocol for solving the fault-tolerance issues in communication.

a. Single sink and multiple channels.
b. Multiple sink and multiple channels.

Now consider multiple sinks, multiple channels and demonstrate and find out performance. Here while finding the shortest path among routing in communication the wireless sensor network having the capabilities of cognitive radio spectrum utilization. Here the spectrum senses the available spectrum band search for a certain duration of time for neighboring node and jumps on the vacant band, identify the primarily used band if free serve the channel and if busy, it went back for spectrum sensing. The served channel take a decision and senses the carrier which sends the data to the cognitive sensor node, a user (CR) as forwarding ANT send the information to next coming ants through pheromone trails either send back to carrier sensing for the other vacant band.

The communication protocols are not sufficiently flexible regarding environmental changes. These environmental changes and control on each layer in wireless sensor network architecture operate on widely different timescales.

MAC layer supports for one hops communication where data transmission takes a few milliseconds in most sensor networks. Energy efficiency MAC protocol with sleep scheduling for prolonging for a lifetime is assumed in sensor networks.

Whereas routing layer has deals with topological changes to realize source to destination communications. In static sensor nodes manage network topology by using HELLO message every several tens of seconds.

The timescale operation external control of self-organization is longer than routing layer. It is insufficient to discuss the robustness within one layer. We consider two layer combinations as cross layer approach, MAC and Network layer. We have designed for the fault-tolerance based routing protocol for the condition of channel selection, mobility and node failure.

In MAC layer the sleep control is expected, so that power saving option is successful. The MAC protocol with the sleep control allows the node to sleep in every ten milliseconds. So that each node can communicate with each node only when it's awake. The cycle of sleep control means the minimum unit time of one-hop data transmission. In MAC layer the selection of next hop node when it is in sleep mode the data is to hold for certain period of time. There is condition comes to probabilistic channel selection for communication. The channel selection from spectrum sensing, sharing utilizing the available spectrum band called as cognitive radios in nature, self-organization for sensor node in wireless sensor networks, which is to be analogous to biologically inspired methods, i.e., availability of different paths for ANTs and BEEs for searching the food among different path to reach the destination i.e., sharing resources.

Consider the insect colony as cognitive radios networks. The insects are as a cognitive radio for spectrum utilization. Task allocation is available to channel and task associated stimuli as a permissible power to channel.

1) Selecting the channel that has the minimum channel selection probability (appropriate channels) with avoiding the interference at the same time.(select the paths by ants and bees as availability of different paths)
2) Manage the number of sensor nodes, which are nearer to each other as a neighboring node forming the groups called clusters (the node that has the fewest radiuses). Gathering the data as a response to sink node for the HELLO message as a multi-hop communication and accomplish the task of joining regarding to pheromone levels.
3) If a sink node fails to send the data to the base station. To improve the problem at that moment the multi-sink operation is useful.

A. Markov Chain model for probabilistic approach

The development in a cognitive radio sensor network based on partially observable on Markov chain decision[19]. In different applications the probabilistic approach plays a vital role in solving the number of problems with optimum solution based on Markov chain model.

Markov Chain is a class of characterizing stochastically the process by Markov property. The meaning is the next process takes depends on the current state and not on the previous state. To prove this for discrete time as given as follows and stated by [23]. Here P is the probability of selecting the behavior of next-hope sensor node and X_t is the condition probability.

$$P[X_t+1=x \mid X_0=x_0,.....,X_t]=P[X_t+1=x \mid X_t=x_t]$$

This process called as a Markov Chain model.

The Markov chain model for sensor node is based on (ON/OFF) condition of channel selection of channel by primary user and secondary user, also the probability of next hope sensor is either {0, 1} as shown in the figure given below.

In the following condition state that as equivalent to Ants search food stochastically as random walk probability, same as for channel selection by the secondary user after primary user.

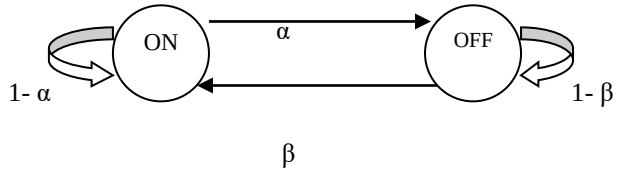


Figure 4. Model for behavior of next-hope sensor node analogous to channel sensing state as ON/OFF.

In random sensing techniques the secondary users cooperate to sense the licensed channels. Each secondary users stochastically chooses a licensed channels for sensing. Su is considered as secondary user as CR node. We can develop the Markov chain model to calculate the probability mass function (PMF) of secondary users and denoted as Probability of secondary user denoted as Pr {Su} for number of secondary users S = 0, 1, 2, 3 ….. n. Each one select n channels as each sensor nodes select the nodes uniformly with 1/n model is

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

defined for n + 1 states. The transition probability from states i to j is P (i, j).

$$\Pr(i, j) = \Pr(Su+1 = j \mid Su = i) = \begin{cases} i/n & j = i; \\ 1 - i/n & i = i+1; \\ 0, & otherwise \end{cases}$$

The above equation, prove that the condition of a number of states from state to another state with probability from one user that is the initial state to another node as the last state with respect to transition from 0 to n and the middle one state is n-1 transition with number of probabilities to reach upto j state.

V. Performance results and discussions

To evaluate the performance of the proposed scheme against the recent routing schemes, we considered the three different parameters such as packet delivery fraction, end to end delay and throughput. We evaluated the CRSNs as next generation WSNs. The simulation was done by extending in the open source software NS2 version NS-2.34 with multi-sink & multi-channel extensions to enable dynamic spectrum access and cognitive radio sensor networks. The CRSN network consists of 25 nodes, which are arranged in a grid topology of 1,000m x 1,000m. The licensed arrival and departure time are considered with packet size variation of 100,200,300,400,500 with a queue size of 50. It is assumed that the network utilizes an unlicensed channel which is shared with other legacy networks. This condition causes network interference, which is represented by the parameter (i.e., PU channel). Here we consider the 11 channels for our experimentation in CRSN network. We compare the performance [AODV-Cogns], [DSR-Cogns], AntHocNet. In this simulation, we used Mac/Cogmac standard as the MAC layer protocol. Here we consider the following parameters.

A. Simulation Analysis

The simulation analysis is done on the three parameters, packet delivery factorial, throughput, and delay. We require the high packet delivery factorial, High throughput, low delay. We consider the parameters with their respective values as mention in table 1.

We apply the novel technique on two protocols AODV and DSR with MAC layer protocol as Cognitive radio sensor network. We analyze on two techniques.

1. Packet Variation
2. Pause Time variation
3. Variation on Packet arrival and departure (α, β) respectively.

The packet variation in the novel protocol when applying with the fixed number of nodes deployed in the specific region is given below.

a) Throughput: The parameter evaluation in wireless sensor networks with the considering the above mentioned parameters in NS2 was evaluated. Depends on analysis of throughput with respect to packet size shown in the screenshot above. In the simulation throughput of the network is decreased with respect to increasing the number of packet size. The throughput is calculated as.

Table 1. Parameter Consideration

Simulation Parameters	Values
Channel Type	Channel/WirelessChannel
Radio-propagation model	Propagation/TwoRayGround
Network Interface Type	Phy/WirelessPhy
MAC type	MAC/cogmac/802_11
Interface queue Type	Queue/DropTail/PriQueue
Antenna Model	Antenna/OmniAntenna
Max packet in ifq	50
Number of Mobile nodes	25
Routing Protocol	/AODV/DSR /AntHocNet
Topography	1000mx1000m
Energy Model(Initial)	100 Joule
Data transmission rate	2 sec
Radio transmission range	200
Pause time	0,30,60,90,120,150 sec.
Packet Size	100,200,300,400,500 bytes
Constant Bit Rate (CBR)	256 kb
Channel Number Type/radio	11
Time of simulation	150

Throughput: Total number of bits received /Receiving time of last packet –Sending time of first packet.

***b) Average End-to-End Delay*:** The same as throughput the average end to end delay for the 25 number of nodes with respect to packet size with some statistical values mention in each graphical representation. Here the average end to end delay is less at starting state and increasing as packet size increased.

The average end to end delay = total receiving time of ith We observed that as the variation of packets size though we utilize the multiple sink and multiple channel for shortest route to packets transmission betwwen source to destination in communication by gathering data collectively.Each number of nodes discover the route and by hops-to-hops communication with probabilisitic approach. The figure 5 shows that as we increase the packet size the End 2 End delay decreases and some are increases. Though the as compare to AODV(Cogns) , DSR(Cogns) and Anthocnet, the DSR shows less End 2 End Delay as compare to other two, means the DSR with CogNS capabilistic provides less delay for packet transmission, so it is best proposed protocol with cognitive capabilistics which solves and provides high performance for issues such as fault-tolerance.

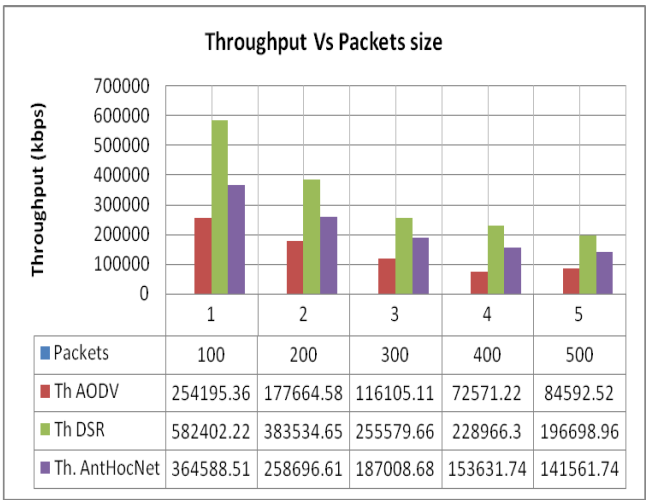
packet of destination – total sending time of ith packet to source / Total number of received packets by destination nodes.

***c) Packet Delivery factorial*:** The PDF is another parameter in the sensor networks also depends on the total number of sending and receiving packets from any source to destination with respect to time. Here initially it is less as the packets size increases, it also increases.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

5.1.1 Packet time Variation

a) Throughput Vs packets size

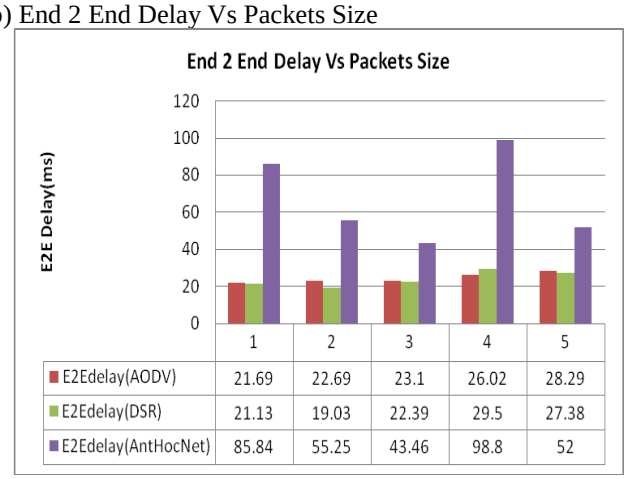


Packets	100	200	300	400	500
Th AODV	254195.36	177664.58	116105.11	72571.22	84592.52
Th DSR	582402.22	383534.65	255579.66	228966.3	196698.96
Th. AntHocNet	364588.51	258696.61	187008.68	153631.74	141561.74

Figure 5. Variation in packets with Throughput analysis

We observed that as the variation of packets size though we utilize the multiple sinks and multiple channel for the shortest route to packets transmission between source to destination in communication by aggregating the data. Each number of nodes discover the route and by hops-to-hops communication with the probabilisitic approach. The figure 5 shows that as we increase the packet size the throughput(kbps) decreases. Though the as compared to AODV(Cogns) , DSR(Cogns) and Anthocnet, the DSR shows High Throughput as compare to other two, means the DSR with CogNS capability provides high throughput for packet transmission.

b) End 2 End Delay Vs Packets Size

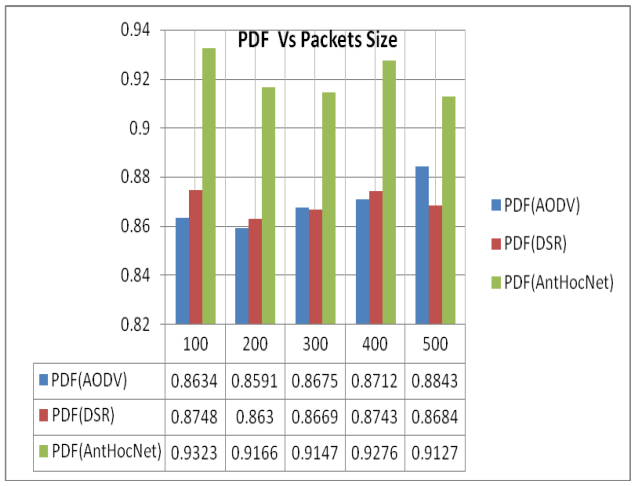


E2Edelay(AODV)	21.69	22.69	23.1	26.02	28.29
E2Edelay(DSR)	21.13	19.03	22.39	29.5	27.38
E2Edelay(AntHocNet)	85.84	55.25	43.46	98.8	52

Figure 6. Variation in packets with End 2 End Delay analysis

We observed that as the variation of packets size though we utilize the multiple sink and multiple channel for shortest route to packets transmission betwwen source to destination in communication by gathering data collectively.Each number of nodes discover the route and by hops-to-hops communication with probabilisitic approach. The figure 6 shows that as we increase the packet size the End 2 End delay decreases and some are increases. Though the as compare to AODV(Cogns) , DSR(Cogns) and Anthocnet, the DSR shows less End 2 End Delay as compare to other two, means the DSR with CogNS capabilistic provides less delay for packet transmission, so it is best proposed protocol with cognitive capabilistics which solves and provides high performance for issues such as fault-tolerance.

a) PDF Vs Packets Size



PDF(AODV)	0.8634	0.8591	0.8675	0.8712	0.8843
PDF(DSR)	0.8748	0.863	0.8669	0.8743	0.8684
PDF(AntHocNet)	0.9323	0.9166	0.9147	0.9276	0.9127

Figure 7. Variation in packets with End 2 End Delay analysis

The figure 7 shows that as we increase the packet size the PDF(packet delivery fraction) increases. Though the as compare to AODV(Cogns) , DSR(Cogns) and Anthocnet, In this case AntHocnet increases in PDF as compare to other two. Means the DSR with CogNS capabilistic still intermediate averagely between AODV and AntHocNet.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Pause Time Variation

a) Throughput Vs Pause Time

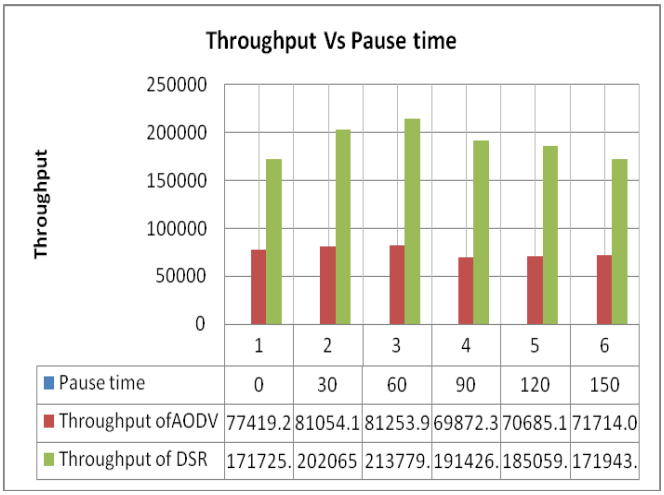


Figure 8. Variation in Pause time with Throughput

The figure 8 shows that as we varies the pause time of 0,30,60,90,120,150. The pause time is defined as the parameter in which the node moves randomly choose path towards destination. When it hits destination pause for some time and changed the direction. where the packet size of 500, it is fixed for all the pausetime. The throughput of DSR with CogNS have high throughput as compare to AODV with CogNS. Here variation in pausetime in AntHocNet never shows any variation.

a) End 2 End Delay Vs Pause Time

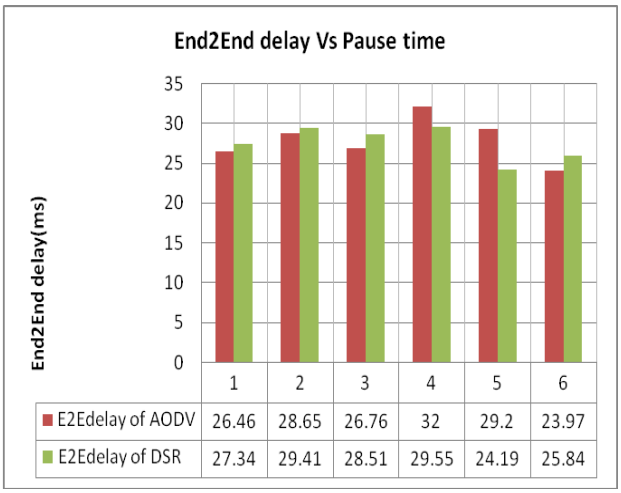


Figure 9. Variation in Pause time with End2End delay

The figure 9 shows that as we varies the pause time of 0,30,60,90,120,150 which we for the packet transmission of 500 in size , fixed for all the pause time. The end2end delay of DSR with CogNS have less delay as compared to AODV with CogNS. Here variation in pause time in AntHocNet do not shows any varaition.

a) PDF Vs Pause Time

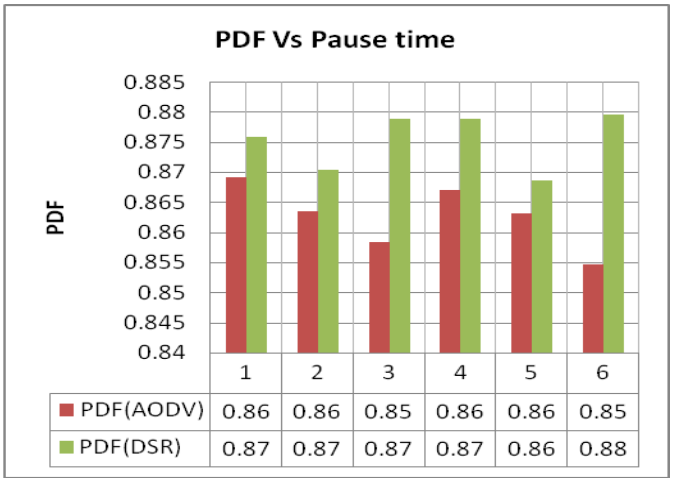


Figure 10. Variation in Pause time with PDF

The figure 10 shows that as we varies the pause time of 0,30,60,90,120,150 which we for the packet transmission of 500 in size , fixed for all the pause time. The end2end delay of DSR with CogNS have high packet delivery ratio as compare to AODV with CogNS. Here variation in pause time in AntHocNet do not shows any varaition.

5.1.3 Variation scenarios of primary user arrival and departure.

Arrival = α Departure = β
Scenario 1: α=1, β= 1
Scenario 2: α=1, β= 2
Scenario 3: α=1, β= 3
Scenario 4: α=2, β= 1
Scenario 5: α=3, β= 1

The Primary User has the arrival rate and departure rate applied on different scenarios as mention above. The protocol has multi-sink, multi-channel fashion among different clusters.

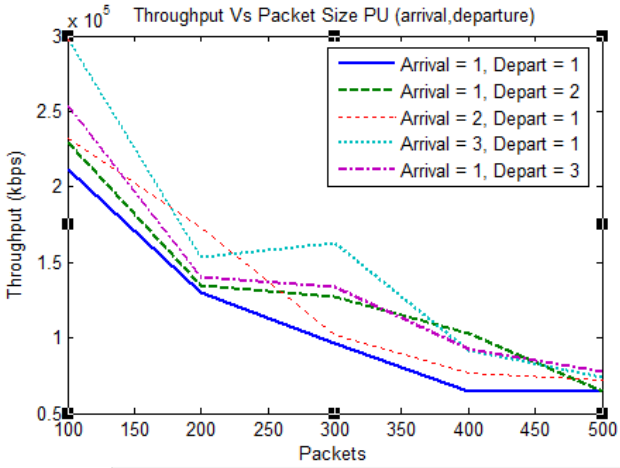


Figure 11. Variation in arrival and departure time with Throughput

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Here arrival 3, and departure 1, shows better output of throughput as compare to rest of the transmission of arrival and departure of primary user based on ACO techniques.

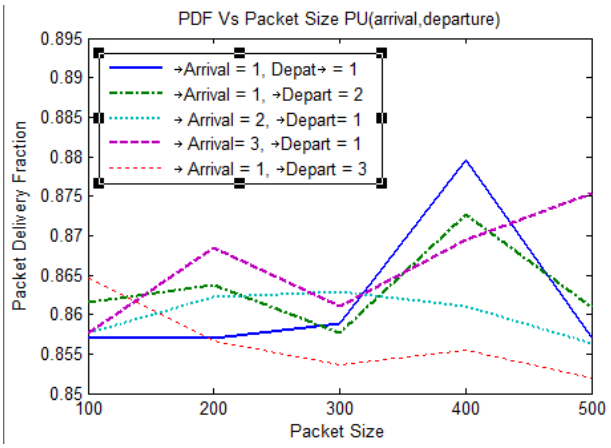


Figure 12. Variation of PU activity with packet delivery fraction

Here arrival 3, and departure 1, shows PDF with high PDF after 300 packet size it gradually increases but arrival with 1, and departure with 1, increases after 300 packet size shows accurate as compare to rest of the transmission of arrival and departure of primary user based on ACO techniques.

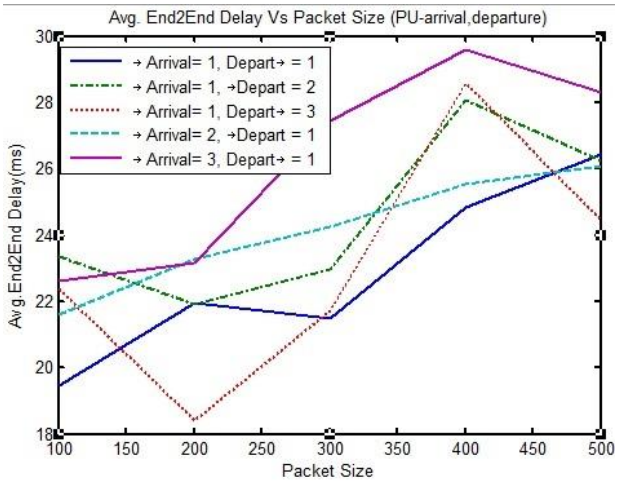


Figure 13. Variation of PU activity with End 2 End delay

Here arrival 3, and departure 1, shows end to end delay after 300 packet size it gradually increases but arrival with 3, and departure with 1, increases after 300 packet size shows accuracy as compare to rest of the transmission of arrival and departure of primary user based on ACO techniques.

Thus the based on three parameters the DSR with Cognitive capabilities based on bio-inspired techniques of Ant Colony optimization provides good accuracy for data transmission. The arrival and departure of PU activity give a spectrum hole for utilization of SU user as a shortest method to route the packet solving the fault-tolerance issue.

VI Conclusion

The proposed bio-inspired cognitive capable routing algorithm with AODV, DSR and AntHocNet routing protocol is analyzed. It is observed that the proposed method of multi-sinks, multi-channel routing techniques are implemented on AODV and DSR, the findings and observations from above graphs are that the proposed bio-inspired DSR (cogns) gives overall accuracy and reliability with highest throughput and high packet delivery fraction and less end to end delay as compare to AODV (cogns) and AntHocNet with maintaining the energy level. Thus for next generation routing protocol for wireless sensor network which are prone to failure of nodes even though the energy of individual nodes exhaust, finout the alternate solution for the same with using available spectrum bands as secondary users. In future work we have to solve the scalability issues. In future work we solve the scalability. Issue for CRSN.

Acknowledgment

I would like to thanks to my guide Dr. Latesh Malik for their regular guidance and also thanks to all the researcher for their valuable research contribution in my research.

References

[1] Hang Su, Xi Zhang, "Cross – layer based opportunistic MAC Protocols for QoS Provisioning Over Cognitive Radio Wireless Networks", IEEE Journal on Selecting areas in Communications, Vol. 26, No. 1, January 2008.

[2] D. G. Reina, S. L. Toral Marin, E. Asimako poulou, F. Barrero, N. Bessis, The role of congestion in probabilistic broadcasting for ubiquitous wireless multi-hop networks through mediation Analysis.

[3] G. Bolch, S. Greiner, H. de Meer, K. S. Trivedi, and K. S. Trivedi, "Queueing Networks and Markov Chains: Modeling and Performance Evaluation With Computer Science Applications", Wiley-Inderscience, 1998.

[4] V. Esmaeelzadeh, R. Berangi, S. M. Sebt, E. S. Hosseini, and M. Parsinia, "CogNS: A Simulation Framework for Cognitive Radio Networks", Wireless Personal Communications, vol. 72, no. 4, pp. 2849 - 2865, Oct. 2013.

[5] M. D. Felice, K. R. Chowdhury, " Search: A routing protocol for mobile cognitive radio ad-hoc networks", Elsevier , Computer Communications 32(2009) 1983-1997.

[6] Zhangliang Liang et al., "Delay Performance Analysis for supporting real-time traffic in a cognitive radio sensor network", IEEE Transactions on Wireless Communications, Vol. 10, No. 1, January 2011.

[7] Ozgur B. Akan et al., "Cognitive Radio Sensor Network", IEEE Network, Vol.23, Issue 4, July-August 2009.

[8] G. P. Joshi et al., "Cognitive radio wireless sensor networks: applications, challenges and research trends," *Sensors*, vol. 13, no. 9, pp. 11196–11228, 2013.

[9] B. Chandrashekaran, "Survey of Network Traffic Models" http://www.cse.wustl.edu/jain/cse56706/ftp/trafficmodels3.pdf/

[10] Le The Dung, et al., "Simulation modeling and analysis of one hop count distribution in cognitive radio ad-hoc networks with shadow fading", Elsevier, simulation, Modeling, practice and theory, 69 (2016), 43-54.

[11] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "Next generation/dynamic spectrum access/cognitive radio wireless

networks: A survey," Computer Networks, vol.50, no. 13, pp. 2127-2159, Sep. 2006.

[12] Muhammad Saleem, Gianni A. Di Caro, Muddassar Farooq, The survey of Swarm Intelligence based routing protocol for wireless sensor networks, *International Journal of Elsevier on Information Science, 2010*.

[13] H. Liu, A.Nayak, et al, "Fault-tolerance algorithm/protocols in wireless sensor networks," in Guide to wireless sensor Networks, S.Misra, Eds. New York: Springer-verlag, 2009, pp. 261-292

[14] N. Fami-Tefreshi and M Davoudi-Monfared, 2016. Ant Colony Based Mobile Sink Routing Algorithm for Wireless Sensor Network. Asian Journal of information Technology, 15: 2981-2987.

[15] Cao L, Xu C, Shao W, et al., Distributed Power Allocation for Sink-Centric Clusters in Multiple Sink Wireless Sensor Networks. Sensors (Basel, Switzerland). 2010, 10(3): 2003-2026.doi:10.339./s 100302003.

[16] Athar Ali Khan, et al., "Cognitive –Radio-based Internet of Things: Applications, Architecture, Spectrum relate Functionalities and future research directions",

[17] I. F. Akyilidiz et al., "NeXt generation /dynamic spectrum access/cognitive radio wireless networks: A survey," Computer Network Journal (Elsevier), Sept. 2006.

[18] Awwad, S.A.B. et al. (2010) 'Cluster based routing protocol for mobile nodes in wireless sensor network', 23 May, *Wireless Press Commun.*, Vol. 61, No. 2, pp.251–281.

[19] Wayne Yang, (2013), "Markov Chains and Ant Colony Optimization".

[20] NS2 official website, http://www.isi.edu/nsnam/ns/

AUTHORS PROFILE

He is pursuing the PhD.(Computer Science & Engineering)from G. H. Raisoni College of Engineering, Nagpur, M.Tech (Computer Science & Engineering)in 2008 from MGMs College of Engineering,Shree Ramanand Teerth Marathwada University, Nanded, B.E.(Computer Technology) in 2002, from Manoharbhai Patel Institute of Engineering & Technology, Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur (Maharashtra), India, He is life member of ISTE, Member of CSI and IEI,Paper presented and Published more than 15 papers.

She has completed Ph.D. (Computer Science & Engineering) from Visveyaraya National Institute of Technology in 2010, M.Tech. (Computer Science & Engineering) from Banasthali Vidyapith, Rajasthan, India and B.E. (Computer Engineering) from University of Rajasthan, India . She is gold medalist in B.E. and M.Tech. She is currently working as Associate Professor & Head of Department in Department of Computer Science & Engineering at Government College of Engineering, Nagpur, MS, India. She has teaching experience of 19 years. Mrs. Latesh Bhagat (Malik) is life member of ISTE, CSI, ACM and presented 46 papers in international journal and 57 papers in international conference. She is recipient of 2 RPS and 1 MODROBs by AICTE. She guided 31 PG projects and 8 Ph.D. students are registered under RTM Nagpur University.

A.Thomas received M.Tech(Computer Science & Engineeirng) in 2013. She is currently the Head of Computer Science Department at G.H.Raisoni College of Engineering, Nagpur. She has completed M.Phil(Computer Science) in the year 2011.Her area of specialization is soft Computing and Data mining.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

APPLICATION OF THE ORBAC MODEL IN THE CONTEXT OF PHYSICAL ACCESS CONTROL

Belbergui Chaimaa
STIC Laboratory
Chouaib Doukkali University
El jadida, Morocco
belbergui.c@ucd.ac.ma

Elkamoun Najib
STIC Laboratory
Chouaib Doukkali University
El jadida, Morocco
elkamoun.n@ucd.ac.ma

Rachid Hilal
STIC Laboratory
Chouaib Doukkali University
El jadida, Morocco
hilal.r@ucd.ac.ma

***Abstract*—Controlling logical access to servers, applications and data is not sufficient in an academic institution such as a faculty. The control of physical access to the premises of this one is also paramount. For the reason that, unauthorized access to the faculty computer room as an example can cause the unavailability of all services. Unfortunately, this aspect of security has almost never been taken into account in research works.**

We demonstrate, by the present work, the utility of the use of OrBAC in the context of physical access control, as a solution of all the problems presented above. This, is accomplished through the development of a case study of physical access control within a faculty. The results of tests approve that the defined security policy is dynamic. The permissions of access are activated if the context is enabled, and deactivated if not. During simulation, when the time of the access request is modified or when the location of the access requester change some access rules are deactivated. There are other contexts that were tested.

Keywords-physical access; access control; security policy modeling; OrBAC; MotOrBAC.

I. INTRODUCTION

A faculty consists of premises assigned to a public service, but that are not characterized by the public place nature, it welcomes only specific users. The traditional solutions for controlling physical access to the premises of an institution were based on the use of keys. This method is not secured since it is easy to make a key copy.

The risk increases if the keys are lost or stolen, as this involves access to premises by unauthorized people, and so put the security of the premises at risk.

In the absence of academic solutions to this kind of problems, private companies have taken over and offered physical access solutions using badges.

These solutions offer the advantage of reliability compared to the traditional access management. The loss of badges does not affect the security of the premises since it can be immediately blocked by the administrator. They thus limit the risk of theft and ensure the traceability of events. Simplicity is also a strong point of these solutions resulting in the use of a single badge as a means of access to all premises instead of needing a set of keys.

These solutions also have limits, such as the policy complexity. The definition of access rules is made per person, not per role, also the redefinition of the whole policy is made if modification of access rules is needed, and the definition of contextual permissions is not possible. Another fault exists, it is the lack of means of; modeling, and testing coherence and conformity to the specifications.

The objective of this work is to control physical access within a faculty as an important pillar of security. This, is done by limiting the access to the authorized persons, in order to avoid the intrusions in the premises of this one.

The solution is based on the technology of badges, which is associated with the use of the formal access control model "Organization based Access Control" (OrBAC) as a remedy for its inadequacies.

The article is organized as follows: Section 2 discusses the overview of the research works dealing with the issue of access control. Section 3 presents the used methods, Section 4 explains the concept of the system and the access control requirements, Section 5 models the access control policy of the faculty with OrBAC, Section 6 presents the simulation and discusses the results, and Section 7 concludes the work.

II. AN OVERVIEW OF THE RESEARCH WORKS DEALING WITH THE ISSUE OF ACCESS CONTROL

The physical access control problematic has almost never been addressed in the research works. Researchers are more interested in logical access within several kinds of systems and networks. Distributed, collaborative and multi-agent systems, web services, social networks, cloud computing and others.

Article [6] provides a language to control access in distributed systems. The article [7] introduces the notion of Team-based Access Control as an approach to collaborative environments. As for paper [8], a new access control model is developed to meet all collaboration requirements. It assumes that users interact with a collaborative application by concurrently editing its data structures. Some solutions have also been proposed for multi-agent systems [9] and [10].
The work [9] describes a typing system for a distributed π-calculus, which guarantees that distributed agents cannot access the resources of a system without first being granted the capability to do so. The paper [10] presents a mobile agent structure which supports authentication, security management and access control for mobile agents.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The paper [11] concerns Web services. It proposes the design of an access control scheme that addresses access control issues, and proposes also an extended, trust-enhanced version of the XML-based Role Based Access Control (X-RBAC) framework that incorporates trust and context into access control.

Several researches are interested in access control for social networks [12], [13] and [14]. The article [12] adopts a rule-based approach for specifying access policies on the resources owned by network participants, with constraints of the type, depth, and trust level of the relationships between nodes. For the article [13], it shows the gap between access control management options offered by Facebook and the real requirements of users in the same context. The paper [14] proposes a new access control solution for Facebook. It is an extension of the model "Organization based Access Control" that is specially dedicated to the social network. This proposal meets the needs of users and covers the deficiencies of the policy currently used by Facebook.

The access control issues in the cloud computing environments were also discussed in many works. The paper [15] proposes an access control model to meet the identified cloud access control requirements. The article [16] proposes a novel distributed architecture embedded principles of computing security. A dynamic access control for securing data in the cloud is also presented in [17]. As for the work [18], it leads to the design of attribute based access control for cloud computing. An onto-ACM (ontology-based access control model) [19] was also proposed as a solution to the difference in service providers and users. Using encryption concept in the context of Cloud is also proposed in [20] and [21]. And finally, the paper [22] proposes a novel access control model for Cloud computing systems inspired by OrBAC model, based on trust evaluation, and introduced a Trusted Third party controlling interactions between access requestor and the system that hosts requested resources.

This observation motivates us to explain the importance of taking physical access control into account in any security policy. Thus, to show the usefulness of modeling the physical access control policy within any organization and the benefits to make it using the OrBAC concepts, such as the dynamism. In this work, the case study is carried out within a faculty.

III. USED METHODS

In this part, the methods used in this work are presented below.

A. Physical access control

1) Definitions

An access control can be logical or physical. It therefore allows to control the access to logical resources such as applications and data, or physical ones like an organization premises.

Physical access control is a technique that consists in subjecting the entrance of a premises, organizations, buildings or places to an access authorization. The aim of this access authorization is to protect persons or properties. The access control policy may concern, for example, access by staff to sensitive areas such as computer rooms, offices, etc.

2) Physical access control by badges

Access control by badges is a widely used solution. It is used by several kinds of organizations to restrict access, to sensitive premises, to specific persons.

This solution allows not only to manage access permissions to the different premises, but also to keep traceability of all accesses of participants.

The access control platform by badges consists of these elements:

- A main console: Generally, it is a computer configured to control the elements of the badge reader device, to analyze and to keep track of the different accesses.
- Applications: These are applications that allow to define access rules such as the configuration of the door opening. It is also the dashboards allowing the supervision and the traceability of the events collected thanks to the badge reader.Using these applications allow to manage users and control units, and centralize user-related items such as access rights, alarms, and more.
- The badge reader: This is a device used for authenticating an individual requesting access to a given premises.

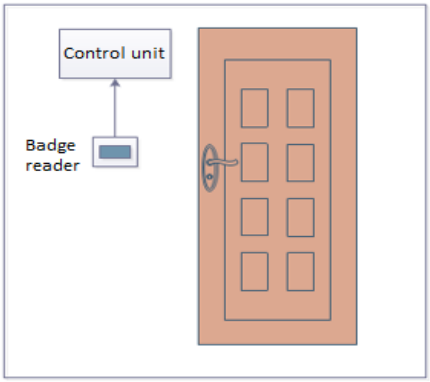


Figure 1. The access control solution by badges

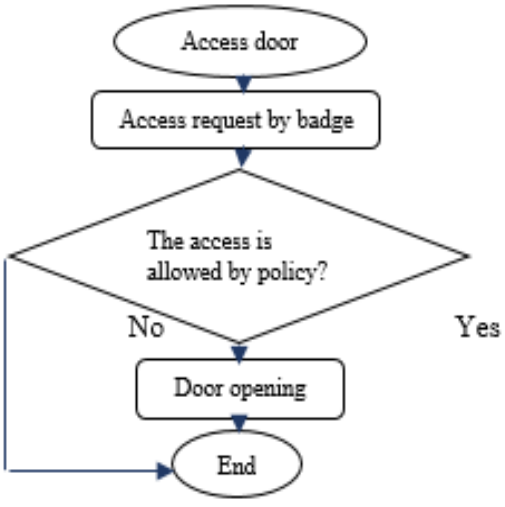


Figure 2. The principal

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

B. Organization Based Access Control

1) Concept

Traditional access control models are based on the three entities: subject, action, and object. Therefore, access permission specifies whether a subject (s) is authorized to perform an action (a) on an object (o). While, OrBAC adds three abstract entities; role (r), activity (a), and view (v).

Access rules are specified between those entities at the abstract level, and then the concrete access rules can be derived from these ones [3]. The advantage is the independence of the implementation. We present below the abstract and concrete entities, as well as the relations between them, and the definition of contexts.

a) The subject and role entities, and the relationship Empower ()

The role corresponds to the mission of the subject within the organization. Thus, the concept of role models how the organization (org) empowers a subject [4]. Thanks to this concept, the addition of a new subject does not require the modification of the whole policy. It is enough to empower the new subject in the corresponding role since the access rights are already defined.

The relationship between the subject (s) and the role (r) is defined as: Empower (org, s, r). For example, Empower (faculty, Chloe, professor) means that the faculty empowers the subject Chloe in the role professor.

b) The object and view entities, and the relationship Use ()

A view allows modeling and structuring objects. It characterizes how objects are used in the organization (org), and corresponds to a set of objects that satisfy the same property.

The relationship between object (o) and view (v) [4] is defined as follows: Use (org, o, v). For example, Use (faculty, file.doc, administrative document) means that the faculty uses the object file.doc in the view administrative document.

c) The action and activity entities, and the relation Consider ()

The concept of activity allows to model in what way the organization performs actions. The same activity can be implemented in different ways in the same organization [4].

The relationship between action (a) and activity (a) is defined as follows: Consider (org, a, a). For example, Consider (faculty, read, consult) means that the faculty considers the action read as the activity consult.

d) The context and the relationship Define ()

The context is an essential element of OrBAC. The context represents an access constraint [4] which can have different forms; Temporal, spatial, etc. When the constraint is validated, the subject (s) will be allowed to perform the action (a) on the object (o).

The definition of the context is done at the concrete level as follows: Define (org, s, a, o, context). Consider the example: Define (Faculty, Chloe, read, file.doc, working hours). This means that the professor can access the administrative document only during working hours.

2) Expression of security policies in the OrBAC model

Thanks to OrBAC, it is possible to express access rules of different types; Permissions, prohibitions, obligations and recommendations. The notation of an access rule at the abstract level is as follows:

Permission (org, r, a, v, c) means that the organization (org) allows the role (r) to perform the activity (a) on the view (v) if the context (c) is validated.

The notation of the access permission at the concrete level is as follows: Is_permitted (s, α, o). It means that the subject (s) is allowed to perform the action (a) on the object (o). The derivation of access rights at the concrete level is done automatically after defining the relationships between abstract and concrete entities, the context and the access rules at the abstract level as follows:

org ∈ Organizations, s ∈ Subjects, α ∈ Actions, o ∈ Objects, r ∈ Roles, a ∈ Activities, v ∈ Views, c ∈ Contexts,

Permission (org, r, a, v, c) ∧
Empower (org, s, r) ∧
Consider (org, α, a) ∧
Use (org, o, v) ∧
Define (org, s, α, o, c)
→ Is_permitted (s, α, o)

The notation of prohibitions, obligations and recommendations at the abstract level, as well as the transition to the concrete level are similar to permission definitions.

3) The MotOrBAC simulator

There is a security policy design tool developed with OrBAC. The tool MotOrBAC [5] allows to simulate the access control policy and to test its proper functioning; by checking incoherencies and proposing solutions to resolve it.

IV. CONCEPT AND ACCESS REQUIREMENTS

The concept and the requirements of physical access control within the faculty are presented in what follows.

A. Concept

Every faculty receives a significant number of people per day, they are mostly students, but also professors, researchers, administrative and technical staff, as well as many visitors.

Since intrusions originate in most cases from the interior, and following the problems of disappearance of various objects, it is necessary to restrict the access to the premises (administration, departments, Amphis, etc.) to the authorized persons, by means of automatic opening and closing doors. The opening of each of these doors is controlled by a badge reader placed in the proximity.

Badges are issued to persons who must access the protected premises. Access rules should be defined in such a way as to allow these persons to carry out their tasks within the faculty.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The access rules can be defined for a single door or for a group of doors. Similarly, it may be assigned to one person or to a group of persons. Similarly, a person may have more than a role, so his access rights should correspond to the union of these roles' access rights.

During this work, we are interested in controlling physical access within a faculty. The premises of the faculty are shown in the figure below.

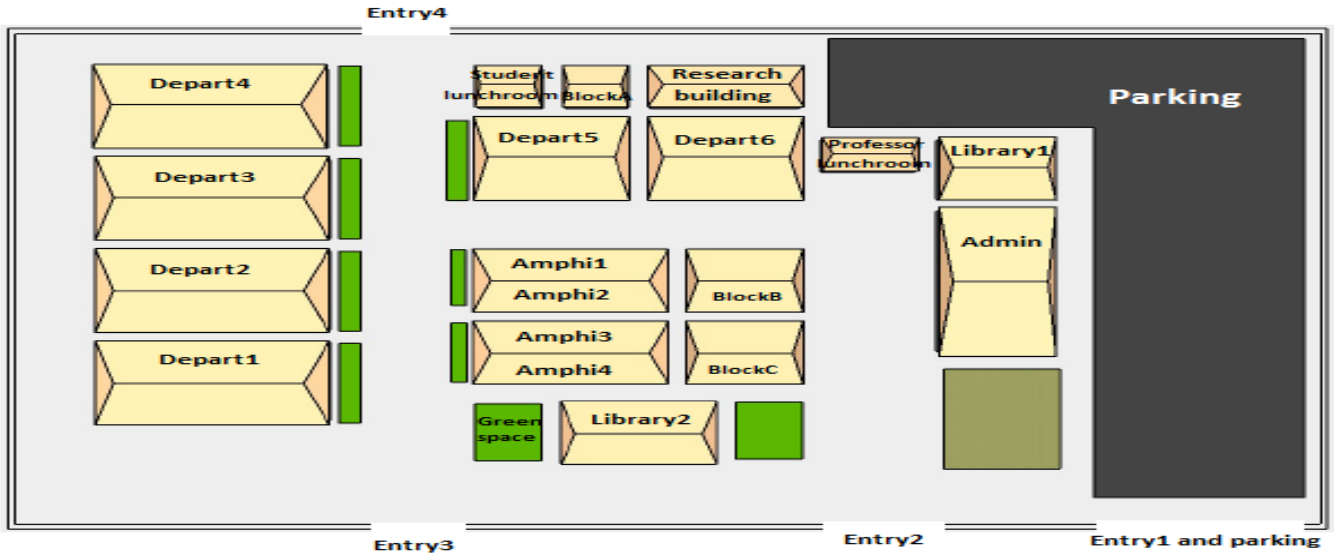


Figure 3. The premises of the faculty

B. Access control requirements

The definition of access rules is made by describing for each person or group of persons the entries authorized to access. Access rules must depend on different constraints; temporal, spatial, and other ones. It is specified in an annual calendar which describes the access requirements, week by week, taking into consideration weekends and holidays. Below is a summary of the recommended access policy.

There are two types of entries to the faculty; an entry that is reserved for professors and staff (Door_for_professors), and three other entries for students and everyone (Doors_for_students). All entries are accessible to professors, staff, and researchers, regardless of the context, even during weekends and holidays. However, students and visitors can access only through the three doors (Doors_for_students), and during working days and hours. The parking is reserved for the exclusive use of professors and staff regardless of the context.

At the administration, administrative staff (Administrative_staff) can access the administration door (Admin_door), has access to the administration offices (Admin_offices) and cannot access the computer room (computer_room), by default. The technical staff (Technical_staff) can access the administration door and has access to the computer room, by default. On the other hand, he can access the administration offices and all the buildings, rooms, offices and amphis only during working days and hours or in case of emergency.

Housekeepers are allowed to access all the buildings of the institution through the main doors during working days, between 6 and 8 am or throughout the day, while access to the offices, rooms and amphis may be authorized only between 6 and 8 hours.

The remainder of participants can access the administration door only during working days and hours, and the administration offices in the same context but with the presence of the administrative staff. As for access to the computer room, it is strictly prohibited, by default.

In departments, the department door (Depart_door) is always open during working days and hours, and therefore can be accessed by everyone. Heads and secretaries of departments, professors and researchers are allowed to access it also outside working days and hours in case of belonging to the department.

The office of the head of department (office_head_depart) is accessible by him whatever the context. On the other hand, it is accessible by the other persons only during working days and hours with his presence in the office. The same principle is applied to the secretary of department office (office_secret_depart) and to the professors' offices (office_prof_course1, etc.).

Access to classrooms, tutorial rooms (T_room), practical work rooms (PW_room) and research laboratories (researchLab_depart) is prohibited by default for administrative staff and visitors.

However, it is authorized for heads and secretaries of departments, during the working days and hours with the constraint of belonging to the department.

Professors are allowed to access the classrooms, the tutorial rooms and the practical work rooms respectively, during the time, of course, of tutorial or of practical work. As well as for

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

students in the same context, but with the presence of the professor.

Research laboratories are accessible to researchers, but with the constraint of belonging to the department.

Any Amphi is accessible in two cases; or when an event is organized in it, the constraints are the participation in the event and the time of the event. Either, by the professor, of course, at the course time, and by researchers and students in the same context with the presence of the professor.

Access to tutorial blocks (Block) through the main door is allowed to everyone during working days and hours. On the other hand, access to the block rooms is only allowed to professors of tutorials during the time of tutorials, as well as to students in the same context with the presence of the professor.

The research building is accessible by everyone during working days and hours. However, the laboratories are only accessible by researchers.

Libraries are accessible to everyone during working days and hours.

Two lunchrooms exist; that of the students, but whose access is allowed at the same time to everyone. As well as that reserved for the exclusive use of professors and staff (including the heads and secretaries of departments) and whose access is prohibited for any other person.

Access permissions to classrooms, tutorial and practical work rooms, as well as to amphis, will be made according to the timetable presented in TABLE I.

TABLE I. TIMETABLE OF THE WEEK

	08h 09h 10h 11h 12h 13h 14h 15h 16h 17h 18h
Monday 15 April	Course 1 classroom 1 | Course 2 classroom 2 | Tutorial 1 Block A room 1 | Course 3 classroom 3
Tuesday 16 April	Course 4 classroom 4 | Course 5 Amphi B | Tutorial 2 Block A room 1
Wednesday 17 April	Course 1 classroom 1 | Conference Amphi B | Course 6 Amphi A | Tutorial 3 Tutorial room 1
Thursday 18 April	PW 1 PW room1 | **Holiday** | PW 2 PW room2
Friday 19 April	PW 3 PW room3 | Course 5 Amphi B | Course 2 classroom 2 | Tutorial 4 Tutorial room 2

V. MODELING OF THE STUDIED SYSTEM WITH ORBAC

The modeling of the studied system with OrBAC is made by firstly specify the organization that needs to secure access to its premises. After that, abstract and concrete entities, also the contexts should be defined. Finally, access rules can be developed.

A. The organization

In this case study, the physical access control policy will be modelled for a faculty. In the OrBAC sense, the central entity is the organization "faculty" (Fac), bringing together a group of professors, students, etc.

B. Roles and subjects

The faculty (Fac) welcomes researchers, as well as students so that professors teach them collectively. Professors can provide them theoretical but also practical courses, so there are three types of professors; professors of courses, professors of tutorials and professors of practical works. It also welcomes the administrative, technical, and housekeeping staff, one head and one secretary per department, and others. As well as visitors who may be students' parents, information seekers, etc.

Roles and subjects are presented below (Fig. 4).

The faculty empowers subjects in the various roles. Jack in the role of Dean, Tracy and Amber in the role of administrative staff, and others. The relationship between these subjects and roles is modeled according to OrBAC as follows:

-Empower (Fac, Jack, dean),
-Empower (Fac, Tracy, administrative_staff),
-Empower (Fac, Amber, administrative_staff), etc.

In addition, we use three functions:

-The function Office (subject): It returns the subject office.

-The function geo_loc (subject): It returns the geographical location of the subject during a time t.

-The function Depart (subject): It returns the department to which the subject belongs.

C. Views and objects

The spaces to be protected are divided into several buildings; The administration building, a building per department (there are six departments: Biology, Geology, Chemistry, Maths, Computer Science and Physics), one building per amphi (there are four amphis), one building per tutorial block (there are three tutorial blocks), one research building, one building per library (there are two libraries), one building per lunchroom (there are two lunchrooms, one for professor and one for students).

The entrances to the institution must also be protected (four entries exist, one for professors and staff, and staff and three for all persons). The parking is reserved for professors and staff and must also be protected from unauthorized access.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

In the administration, access from the administration door, offices and computer room must be controlled.

Within the department, access from the main door, access to the offices of the head and secretary of the department, the professors' offices, the classrooms, the Tutorial and the practical work rooms, the Research laboratory and the detergent storage room should be restrained.

By space constraints, we cannot enter in the details of each view, some examples are presented (Fig. 6).

The faculty (Fac) considers several objects belonging to the different views which was presented above. The relationship between these views and objects is modeled according to OrBAC as below. The Fac considers the entry1 as an entrance reserved for professors and staff. On the other hand, entries 2, 3 and 4 are accessible by students as well as by other people.

- Use (Fac, entry1, door_for_professors),
- Use (Fac, entry2, door_for student),
- Use (Fac, entry3, door_for student) and
- Use (Fac, entry4, door_for student).

D. The activity "access" and the action "enter"

The main and unique activity in our case study is the "access" to the various premises of faculty. This activity corresponds to the action "enter". The relation between the activity and action is modeled according to OrBAC as follows:

- Consider (Fac, Access, Enter).

E. The contexts

In this physical access management, the modeling with OrBAC allowed to define the access constraints expressed in the section "requirements". We have specified several contexts (spatial, temporal, etc.) that can be cumulated according to the security requirements.

The definition of contexts is done according to the timetable (TABLE I).

1) Temporal contexts

All temporal needed contexts are defined below.

a) The default context, it is defined as follows:

∀s∀o∀a (define (Fac, s, a, o, default) ↔ true): The default context is always true between subject s, action a, and object o.

b) Working days, it is defined as follows:

∀s∀o∀a (define (Fac, s, a, o, working_days) ↔ ((D≥15 ∧ D ≤17) ∨ (D=19)) ∧ (MO=4) ∧ (Y = 2017)): In Fac, the context is true between s, a, and o, when the day (D) is between 15 and 17 or equal to 19. The month (MO) is April and the year (Y) is 2017.

c) Working hours, it is defined as follows:

∀s∀o∀a (define (Fac, s, a, o, working_hours) ↔ (h≥7 ∧ h≤ 19)): In Fac, the context is true between s, a, and o, when the time (h) is between 7 and 19 hours.

d) Time of event, it is defined as follows:

∀s∀o∀a (define (Fac, s, a, o, time_event) ↔ (h≥10 ∧ h≤12) ∧ (D=17) ∧ (MO=4) ∧ (Y = 2017)): In Fac, the context is true between s, a, and o, when h is between 10 and 12 o'clock on April 17, 2017.

e) Time of courses

- Time of course1, it is defined as follows:

∀s∀o∀a (define (Fac, s, a, o, time_course1) ↔ (h≥8 ∧ h≤10) ∧ ((D=15) ∨ (D=17)) ∧ (MO=4) ∧ (Y = 2017)): In Fac, the context is true between s, a, and o, when h is between 8 and 10 of the day 15 or 17 April 2017.

Times of courses 2, 3, 4, 5 and 6 are defined in the same way:

- Time of course2

∀s∀o∀a (define (Fac, s, a, o, time_course2) ↔ (((h>=10 ∧ h<=12) ∧ (D=15)) ∨ ((h>=14 ∧ h<=16) ∧ (D=19))) ∧ (MO=4) ∧ (Y = 2017)))

- Time of course3

∀s∀o∀a (define (Fac, s, a, o, time_course3) ↔ ((h>=16 ∧ h<=18) ∧ (D=15) ∧ (MO=4) ∧ (Y = 2017)))

- Time of course4

∀s∀o∀a (define (Fac, s, a, o, time_course4) ↔ ((h>=9 ∧ h<=11) ∧ (D=16) ∧ (MO=4) ∧ (Y = 2017)))

- Time of course5

- ∀s∀o∀a (define (Fac, s, a, o, time_course5) ↔ ((h>=11 ∧ h<=12) ∧ ((D=16) ∨ (D=19)) ∧ (MO=4) ∧ (Y = 2017)))

- Time of course6

∀s∀o∀a (define (Fac, s, a, o, time_course6) ↔ ((h>=14 ∧ h<=16) ∧ (D=17) ∧ (MO=4) ∧ (Y = 2017))).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

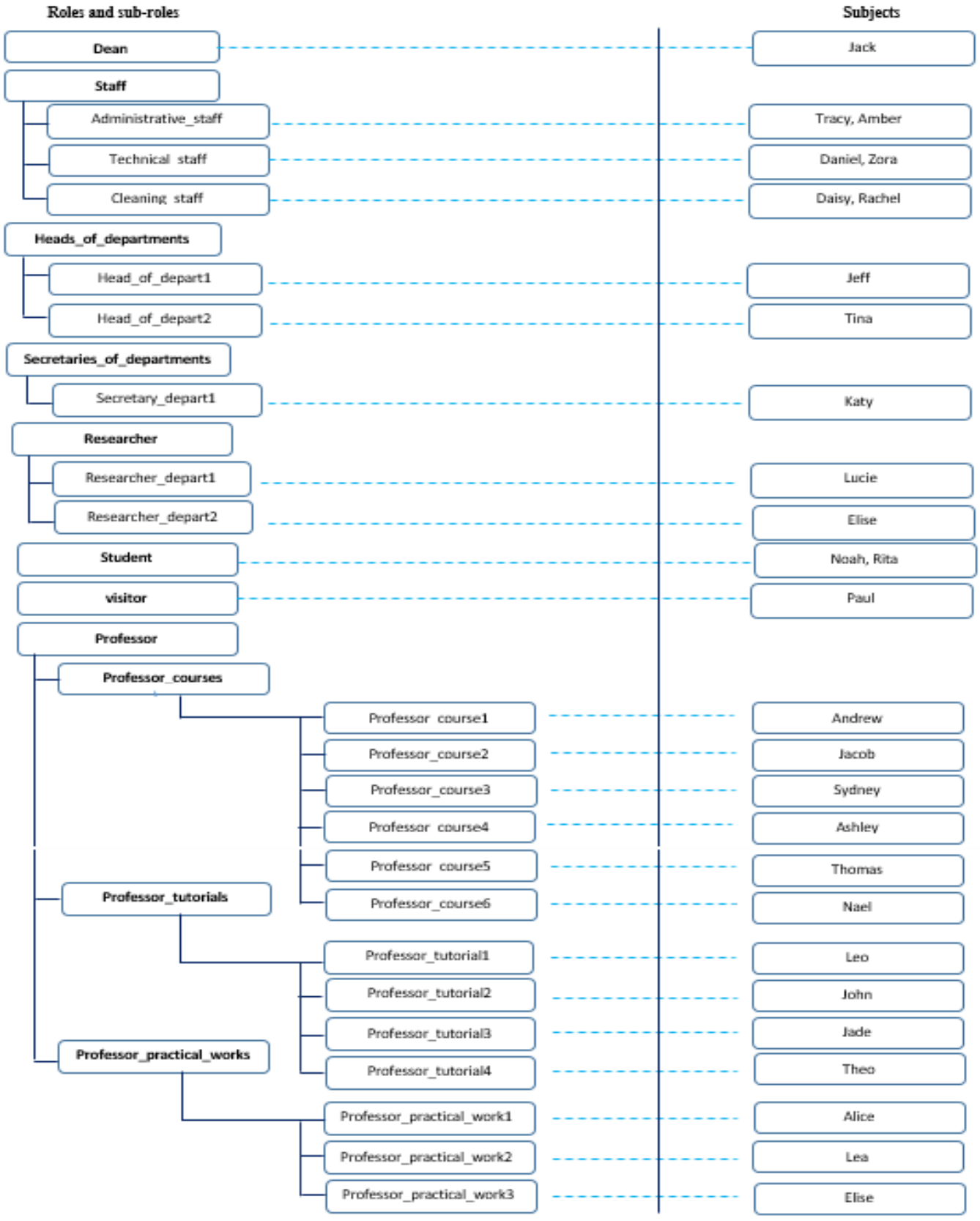


Figure 4. Roles and subjects

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

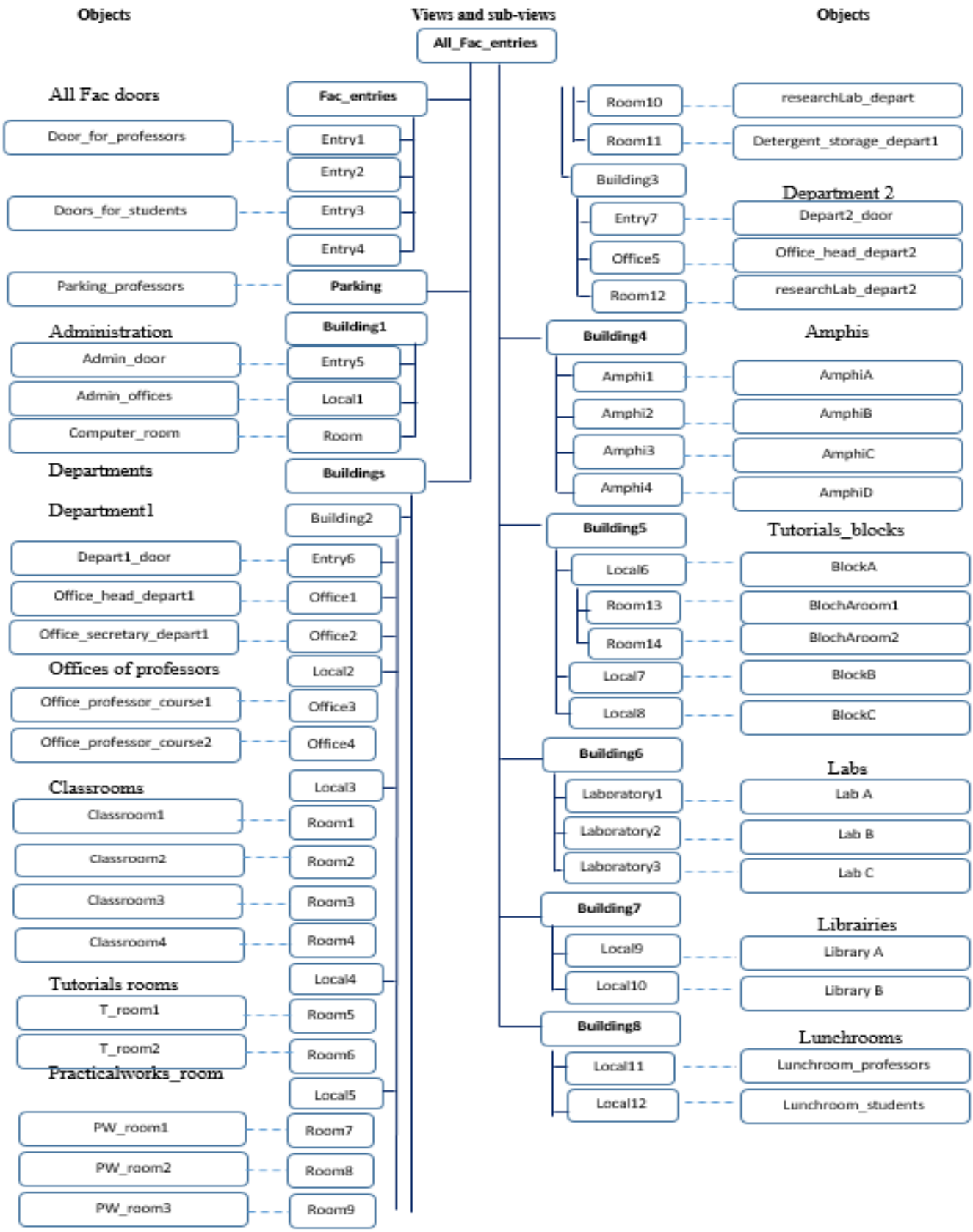


Figure 5. Views and objects

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

f) Time of tutorials

- Time of tutorial1, it is defined as follows:

∀s∀o∀a (define (Fac, s, a, o, time_tutorial1) ↔ ((h>=14 ∧ h<=16) ∧ (D=15) ∧ (MO=4) ∧ (Y = 2017))): In Fac, the context is true between s, a, and o, when h is between 14 and 16 h of the day April 15, 2017.

Times of tutorials 2, 3 and 4 are defined in the same way:

- Time of tutorial2

∀s∀o∀a (define (Fac, s, a, o, time_tutorial2) ↔ ((h>=14 ∧ h<=17) ∧ (D=16) ∧ (MO=4) ∧ (Y = 2017)))

- Time of tutorial3

∀s∀o∀a (define (Fac, s, a, o, time_tutorial3) ↔ ((h>=16 ∧ h<=18) ∧ (D=17) ∧ (MO=4) ∧ (Y = 2017)))

- Time of tutorial4

∀s∀o∀a (define (Fac, s, a, o, time_tutorial4) ↔ ((h>=16 ∧ h<=18) ∧ (D=19) ∧ (MO=4) ∧ (Y = 2017))).

g) Time of practical works (PW)

- Time of PW1, it is defined as follows:

∀s∀o∀a (define (Fac, s, a, o, time_PW1) ↔ ((h>=8 ∧ h<=10) ∧ (D=18) ∧ (MO=4) ∧ (Y = 2017))): In Fac, the context is true between s, a, and o, when h is between 8 and 10 of the day April 18, 2017.

The times of PW 2 and 3 are defined in the same way:

- Time of PW2

∀s∀o∀a (define (Fac, s, a, o, time_PW2) ↔ ((h>=14 ∧ h<=17) ∧ (D=18) ∧ (MO=4) ∧ (Y = 2017)))

- Time of PW3

∀s∀o∀a (define (Fac, s, a, o, time_PW3) ↔ ((h>=8 ∧ h<=11) ∧ (D=19) ∧ (MO=4) ∧ (Y = 2017))).

2) Spatial contexts

All spatial needed contexts are defined below.

a) Presence

- Presence of administrative staff (pres_admin_staff), it is defined as follows:

∀s∀o∀a (define (Fac, s, a, o, pres_admin_staff) ↔ ($subject.geo_loc= admin_offices)): In Fac, the context is true when the administrative staff is present in his office, the geographical location of the administrative staff should be "admin_offices".

The contexts below are defined in the same way:

- Presence of head of department (pres_resp_depart)
- Presence of secretary (pres.secret)
- Presence of course professor (pres_course_prof)
- Presence of tutorials professor (pres_tuto_prof)
- Presence of practical work professor (pres_pw_prof).

3) The contexts; emergency, belongs to depart and participant

a) Emergency, it is defined as follows:

∀s∀o∀a (define (Fac, s, a, o, emergency) ↔ true): In Fac, the emergency context is true when a fault is detected at the equipments or when an alarm is triggered.

b) Belongs to depart, it is defined as follows:

∀s∀o∀a (define (Fac, s, a, o, Belongs_to_depart) ↔ (s ϵ depart.members)): In Fac, the context is true between s, a, and o, when the subject is a member of the department.

c) Participant, it is defined as follows:

∀s∀o∀a (define (Fac, s, a, o, Participant) ↔ (s ϵ event.participants)): In Fac, the context is true between s, a, and o, when the subject appears in the list of the event participants.

4) The composed contexts

Several composed contexts are used in this work:

- Working hours and working days, it is defined as follows:

∀s∀o∀a (define (Fac, s, a, o, workingdaysANDhours) ↔ (working_hours ∧ working_days)): In Fac, the context is true between s, a, and o, when the two contexts working_hours and working_days are true at the same time.

- Time of technical works, it is defined as follows:

∀s∀o∀a (define (Fac, s, a, o, time_technicalworks) ↔ workingdaysANDhours v emergency): In Fac, the context is true between s, a, and o, in case of emergency or when the time of the access request respects the context workingdaysANDhours.

- Working hours, working_days and presence of administrative staff

∀s∀o∀a (define (Fac, s, a, o, workingdaysANDhoursANDpres_admin_staff) ↔ workingdaysANDhours ∧ pres_admin_staff): In Fac, the context is true between s, a, and o, when the administrative staff is present in his office and when the time of the access request respects the context workingdaysANDhours.

- Time of course1 and presence of course1 professor

∀s∀o∀a (define (Fac, s, a, o, time_course1ANDpres_prof_course1) ↔ time_course1 ∧ pres_prof_course1): In Fac, the context is true between s, a, and o, when the course1 professor is present in the classrom1and when the time of the access request respects the context time_course1.

Other composed contexts, of time of course and presence of professor of course, time of tutorial and presence of tutorial professor, and time of practical works and presence of practical work professor, are defined in the same way.

- Participant and time of event

∀s∀o∀a (define (Fac, s, a, o, participantANDtime_event) ↔ participant ∧ time_event): In Fac, the context is true between s, a, and o, when the access requester is a participant in the event and the time of the access request respects the context time_event.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

• Participant and time of event, or time of course5 and presence of course5 prof

∀s∀o∀a (define (Fac, s, a, o, ParticipantANDtime_eventORtime_course5ANDpres_prof_course5) ↔ (Participant ∧ time_event) v (time_course5 ∧ pres_prof_course5)): In Fac, the context is true between s, a and o, either when the access requestor is a participant in the event and the time of the access request, respects the time_event context, either when the course5 professor is present in the amphi and when the time of the access request respects the context time_course5.

F. The security policy modeled with OrBAC

The access control policy is defined at two levels, abstract and concrete one.

1) Policy at the abstract level

By space constraint, we will not be able to present all the permissions of policy. We will only present some examples of those that will be tested in the simulation section.

Access permissions are expressed in OrBAC as follows:

• Permission (Fac, technical_staff, access, local2, time_technicalworks)

• Permission (Fac, professor_course1, access, room1, time_of_course1)

• Permission (Fac, student, access, room13, time_of_tutorial2 & pres_prof_tutorial2)

The first permission indicates that the technical staff is permitted to access the local2 during the time of technical works. The second one indicates that even the professor of course1 cannot access classroom1 at any time. Access is only allowed during the time of course1. The third one indicates that the access of the student to the blockAroom1 is only allowed during the time of the tutorial2 with, the presence of the professor of this tutorial.

2) Policy at the conrete level

OrBAC, concrete permissions and prohibitions are derived automatically after defining the relationships between all abstract and concrete entities, the contexts and the permissions at the abstract level. In the example below, the Fac allows technical staff to access local2 during the time of technical work. It empowers Daniel in the role of technical staff, considers the action access as the activity enter, uses the office of course1 professor in local2 and defines the context time of technical works as an access constraint.
This, implies that when the context is verified, the subject Daniel is allowed to enter the office the course1 professor. It is expressed as follows:

Permission (Fac, technical_staff, access, local2, time_technicalworks) ∧

Empower (Fac, Daniel, technical_staff) ∧

Consider (Fac, enter, access) ∧

Use (Fac, office_professor_course1, local2) ∧

Define (Fac, Daniel, enter, office_professor_course1, time_technicalworks)

→ Is_permitted (Daniel, enter, office_professor_course1).

VI. SIMULATION AND DISCUSSION

We simulate, in this part, the physical access control policy that was modeled in the previous section. We will illustrate the effectiveness of the policy, modeled with OrBAC, through examples.

We will modify the various entries (time of the access request, presence professors, etc.), in each test, in order to verify the effectiveness of the physical access control policy within the faculty, which was modeled by OrBAC in the previous section.

The MotOrBAC software is used in this simulation. The MotOrBAC simulation window allows to view the permissions and prohibitions at the concrete level corresponding to the access rules of the abstract level. These are activated if the context is enabled, and deactivated if not. The color of a permission is green. However, the color of a prohibition is red. When the access rule is activated, the color is clear. If not, it is dark.

Due to the space constraint, we will only present examples of tests. We will try to vary the roles, views, and contexts during the tests.

A. Example 1

The technical staff is permitted to access to the local2 during the time of technical works. The permission at the concrete level is as follows:

• Permission (Fac, technical_staff, access, local2, time_technicalworks)

The “time_technicalworks” is composed context. Its definition is: (working_hours & working_days) | Emergency, with:

- working_hours is equivalent to ((D>=15 & D<=17) | (D=19)) & (MO=4) & (Y = 2017),

- working_days is equivalent to (h>=7 & h<=19),

- Emergency is true if a fault is detected in the faculty equipments or if an alarm is triggered, and false if not.

When either the context (working_hours & working_days) or the emergency context is activated, the permission is granted at the concrete level:

• Is_permitted (Daniel/Zora, enter, office_professor_course 1/2)

Test1: The emergency context is false (Fig. 6), but the “workingdays & hours” context is enabled (Fig. 7). Access permission is activated (Fig. 7).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

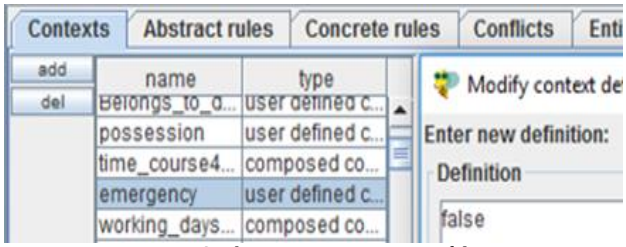


Figure 6. The context emergency is false

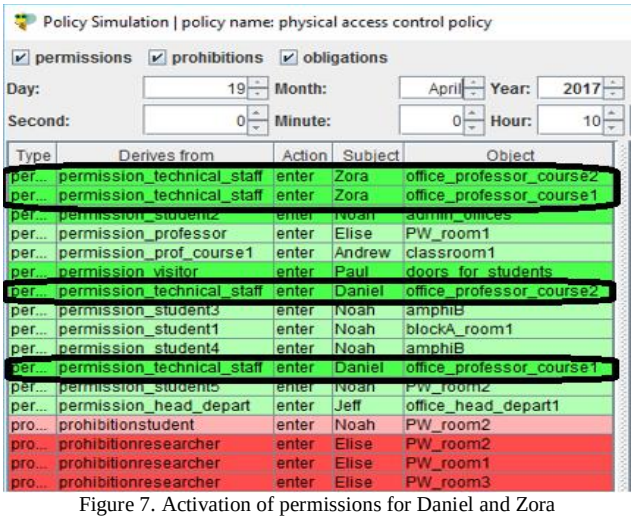


Figure 7. Activation of permissions for Daniel and Zora

Test2: The context "emergency" is true (Fig. 8). Therefore, even if the context "workingdays & hours" is not validated because h = 6 (Fig. 9), the access permissions are activated (Fig. 9).

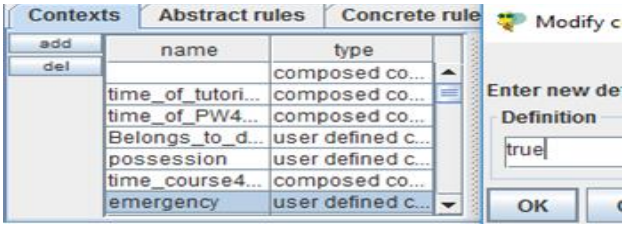


Figure 8. The context emergency is true

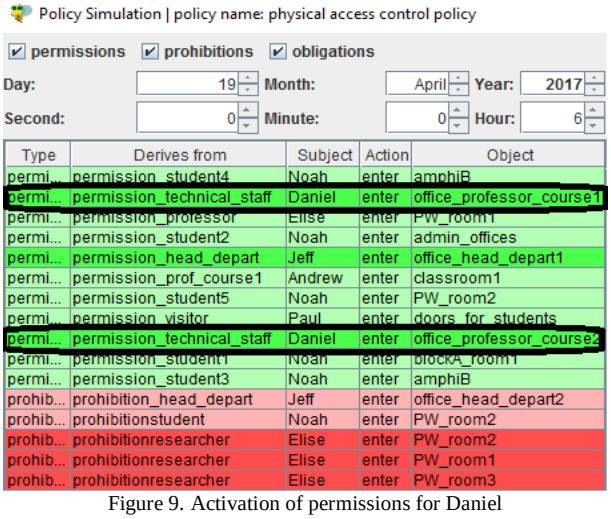


Figure 9. Activation of permissions for Daniel

B. Example

Even the professor of course1 cannot access classroom1 at any time. Access is only allowed during the time of course1. Permission at the abstract level is as follows:

- Permission (Fac, professor_course1, access, room1, time_of_course1), with:

- Time_of_course1 is equivalent to ((h>=8 & h<=10) & ((D=15) | (D=17)) & (MO=4) & (Y = 2017))

When the context time of course1 is validated, permission is activated at the concrete level:

- Is_permitted (Andrew, enter, classroom1)

Test 1: We check the professor's access during the time of course1 (Fig. 10). Access is granted (Fig. 10).

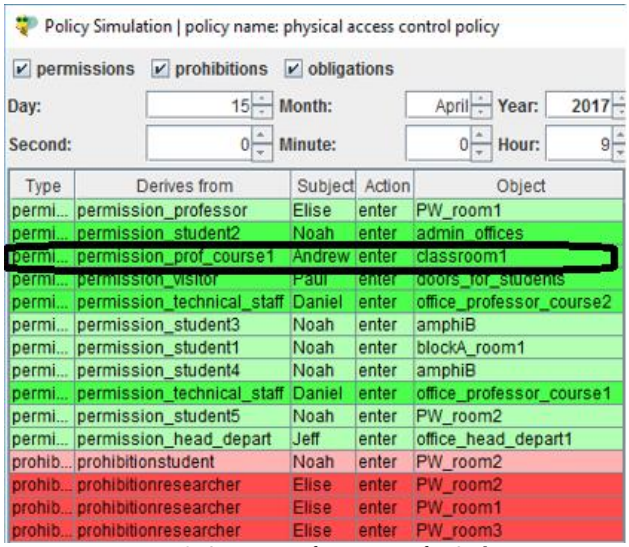


Figure 10. Activation of permission for Andrew

Test 2: We check the professor's access during non-course1 time (Fig. 11). So, access is not authorized (Fig. 11).

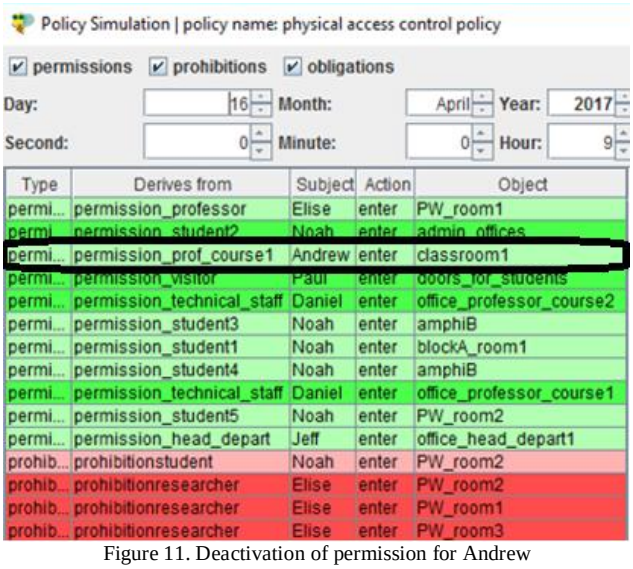


Figure 11. Deactivation of permission for Andrew

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

C. Example 3

The access of the student to the blockAroom1 is only allowed during the time of the tutorial2 with, the presence of the professor of this tutorial. Permission at the abstract level is as follows:

• Permission (Fac, student, access, room13, time_of_tutorial2 & pres_prof_tutorial2), with:

- Time_of_tutorial2 is equivalent to ((h>=14 & h<=17) & (D=16) & (MO=4) & (Y = 2017))
- Pres_prof_tutorial2 is true when $John.geo_loc = BlockAroom1.

When both contexts are activated, access is allowed at the concrete level:

• Is_permitted (Noah, enter, BlockA_room1)

Test1: We show that student Noah cannot access the BlochA_room1 during the non-tutorial2 time (h=13). Permission is disabled (Fig. 12).

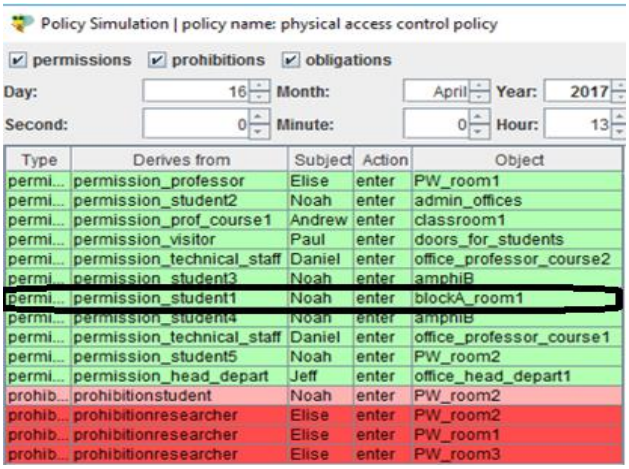


Figure 12. Deactivation of permission for Noah

Test2: Even if it is the time of tutorial2, the access of students to the BlockAroom1 is refused if the professor of Tutorial2 (John) is not present (Fig. 13). Permission is disabled (Fig. 14).

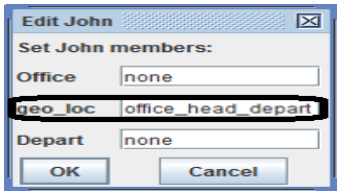


Figure 13. Geographical location of John

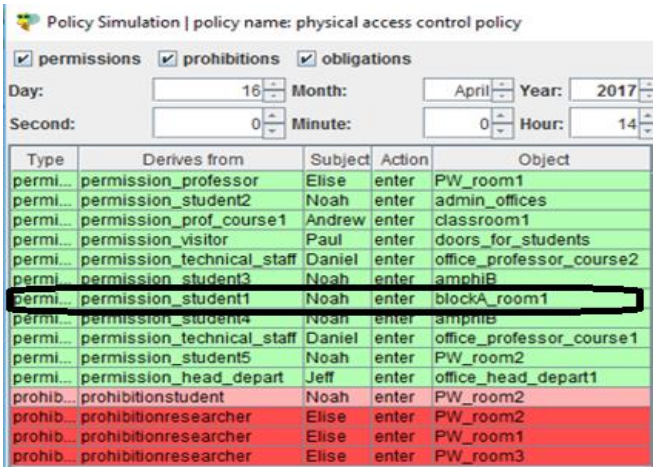


Figure 14. Deactivation of the permission for Noah

Test3: When access is required at the time of tutorial2 and the professor is present (Fig. 15). The permission is enabled (Fig. 16).

Figure 15. Geographical location of John

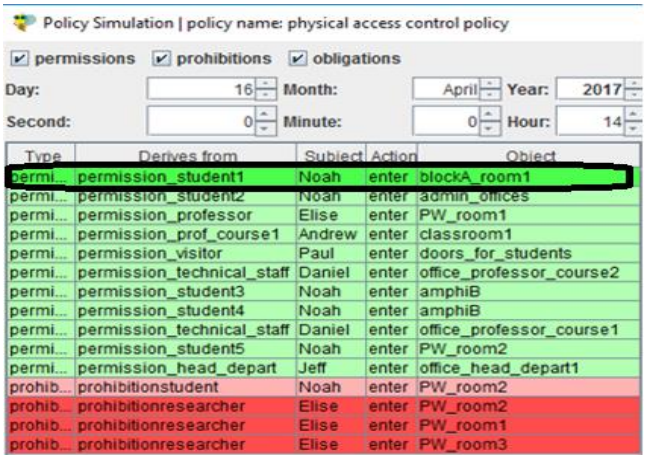


Figure 16. Activation of permission for Noah

D. Example 4

A head of department must access the office of the head of department 1 only if he belongs to this department. In other departments, he can access the head department offices only if the head of this department is present in his office. Permission at the abstract level is defined as follows:

• Permission (Fac, head_of-depart, access, office1, belongs_depart1), with:

- Belongs_depart1 is true if $subject.depart = depart1

When context is enabled, permission at the concrete level is activated:

• Is_permitted (Jeff, enter office_head_depart1)

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Test1: The head of department 1 has a permission to access the office of department 1 (Fig. 17) because it is his office (Fig. 18), but access to the office of the head of department 2 is prohibited (Fig. 17) because he does not belong to this department (Fig. 17), and the head of department 2 (Tina) is not present in his office (Fig. 19).

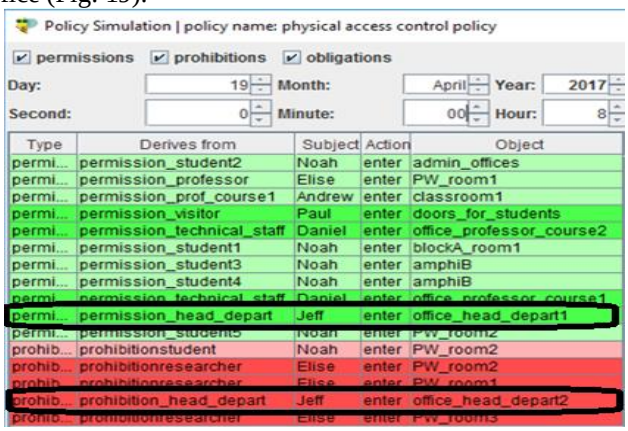


Figure 17. Access rights of Jeff

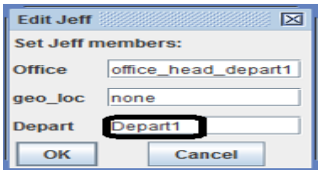


Figure 18. Jeff infos

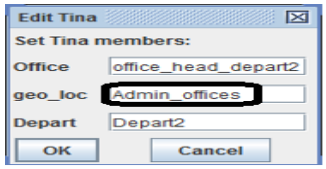


Figure 19. Tina infos

Test2: The head of department 1 has a permission to access the office of the head of department 1 (Fig. 20) because it is his office (Fig. 18). Access to the office of the head of department 2 is also authorized (Fig. 20) because the head of department 2 (Tina) is present in his office (Fig. 21).

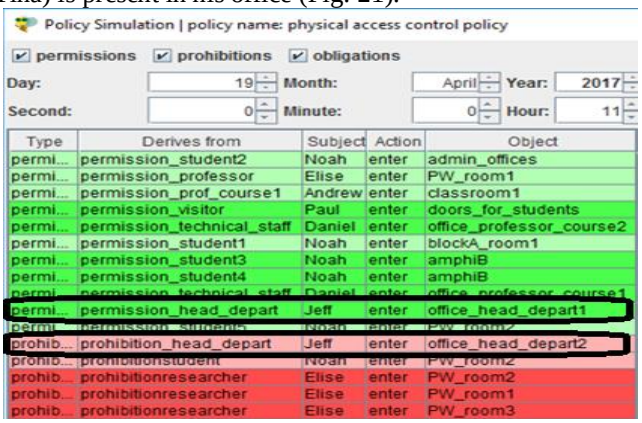


Figure 20. Jeff access rights

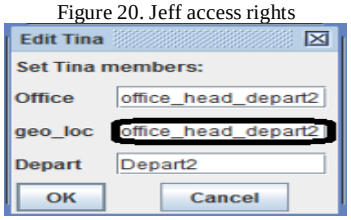


Figure 21. Tina infos

E. Example 5

A visitor can only access the Fac through the doors that are not reserved for professors (entries for student). This access is allowed only during working hours and days. Permission at the abstract level is defined as follows:

- Permission (Fac, visitor, access, entry2, workinghoursANDworkingdays), with:

- workingdays is equivalent to ((D>=15 & D<=17) | (D=19)) & (MO=4) & (Y = 2017) and
- workinghours is equivalent to (h>=7 & h<=19)

When both contexts are validated, access is allowed:

- Is_permitted (Paul, enter, doors_for_students).

Test 1: Access is allowed because the temporal context is validated (Fig. 22).

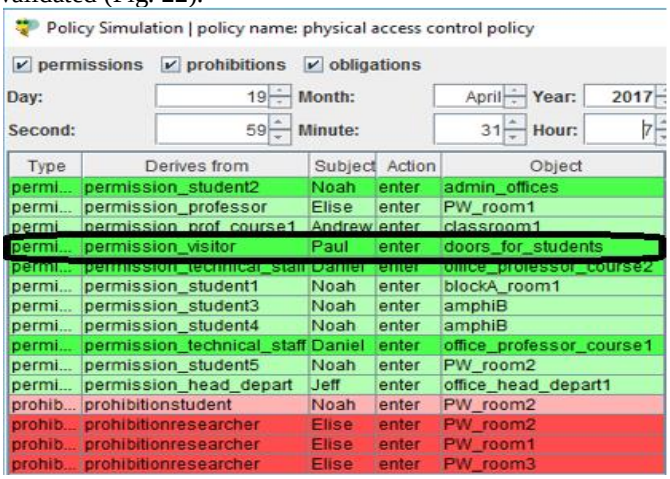


Figure 22. The permission is activated for Paul

Test 2: The context “working days” is validated. On the other hand, 20 h does not belong to the time range "working hours", so this context is not validated. The access permission is deactivated (Fig. 23).

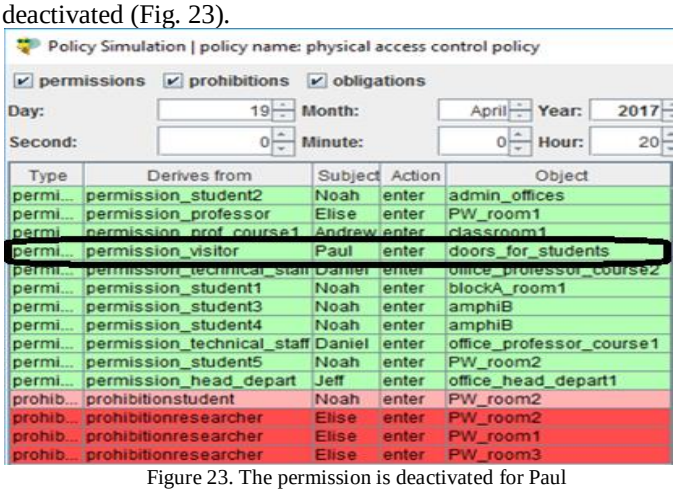


Figure 23. The permission is deactivated for Paul

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

VII. Conclusion

In this work, we deal with a serious matter that has almost never been addressed in the research works, to my knowledge, it is the physical access control. We have explained the importance of controlling not just logical access, but also the physical one within any organization. This interest motivated us to model a physical access policy within an academic institution as an example. This, by using the OrBAC model, for all the benefits that it offers; the expression of contextual permissions, the independence of implementation, and others.

We have demonstrated the ease of modeling with OrBAC through excerpts from the access control policy. Subsequently, we have made a test of coherences so as not to have conflicts during the real management of accesses. We have also shown the interest of defining contextual access rules, by modifying the entries; the time of the access request, etc. The access rights assigned to the same role change automatically. Through examples, we have also demonstrated the good functioning of the policy.

Acknowledgment

This research is supported by the National Center for Scientific and Technical Research of Morocco.

References

[1] A. A. E. Kalam et al., "Organization based access control," in Policies for Distributed Systems and Networks, 2003. Proceedings. POLICY 2003. IEEE 4th International Workshop on, 2003, pp. 120–131.

[2] A. Baïna, "Controle d'accès pour les grandes infrastructures critiques. Application au réseau d'énergie électrique.," INSA de Toulouse, 2009.

[3] R. Thion, "STRUCTURATION RELATIONNELLE DES POLITIQUES DE CONTRÔLE D'ACCÈS REPRÉSENTATION, RAISONNEMENT ET VÉRIFICATION LOGIQUES," Université Paul Sabatier Toulouse, 2008.

[4] A. A. El Kalam et al., "Or-BAC: un modele de contrôle d'acces basé sur les organisations," Cahiers francophones de la recherche en sécurité de l'information, vol. 1, pp. 30–43, 2003.

[5] F. Cuppens, N. Cuppens-Boulahia, & C. Coma, MotOrBAC: "un outil d'administration et de simulation de politiques de sécurité". In Security in Network Architectures (SAR) and Security of Information Systems (SSI), First Joint Conference on June 2006, pp. 6-9.

[6] M. Abadi, M. Burrows, B. Lampson, and G. Plotkin, "A calculus for access control in distributed systems," ACM Transactions on Programming Languages and Systems (TOPLAS), vol. 15, no. 4, pp. 706–734, 1993.

[7] R. K. Thomas, "Team-based access control (TMAC): a primitive for applying role-based access controls in collaborative environments," in Proceedings of the second ACM workshop on Role-based access control, 1997, pp. 13–19.

[8] H.Shen, & P. Dewan, "Access control for collaborative environments", in Proceedings of the 1992 ACM conference on Computer-supported cooperative work (pp. 51-58). ACM.

[9] M. Hennessy and J. Riely, "Resource Access Control in Systems of Mobile Agents," Information and Computation, vol. 173, no. 1, pp. 82–120, Feb. 2002.

[10] V. Roth and M. Jalali-Sohi, "Access control and key management for mobile agents," Computers & Graphics, vol. 22, no. 4, pp. 457–461, 1998.

[11] R. Bhatti, E. Bertino, and A. Ghafoor, "A trust-based context-aware access control model for web-services," in Web Services, 2004. Proceedings. IEEE International Conference on, 2004, pp. 184–191.

[12] B. Carminati, E. Ferrari, and A. Perego, "Enforcing access control in Web-based social networks," ACM Transactions on Information and System Security, vol. 13, no. 1, pp. 1–38, Oct. 2009.

[13] C. Belbergui, N. Elkamoun, and R. Hilal, "Modeling Access Control Policy of a Social Network," International Journal of Advanced Computer Science and Applications, vol. 7, no. 6, 2016.

[14] C. Belbergui, N. Elkamoun, and R. Hilal, "An access control model for a social network: Case of Facebook," International Journal of Computer Science and Information Security (IJCSIS), vol. 14, no. 11, 2016.

[15] Y. A. Younis, K. Kifayat, and M. Merabti, "An access control model for cloud computing," Journal of Information Security and Applications, vol. 19, no. 1, pp. 45–60, Feb. 2014.

[16] A. Almutairi, M. Sarfraz, S. Basalamah, W. Aref, and A. Ghafoor, "A distributed access control architecture for cloud computing," IEEE software, vol. 29, no. 2, pp. 36–44, 2012.

[17] M. Auxilia and K. Raja, "Dynamic Access Control Model for Cloud Computing," in Advanced Computing (ICoAC), 2014 Sixth International Conference on, 2014, pp. 47–56.

[18] A. R. Khan, "Access control in cloud computing environment," ARPN Journal of Engineering and Applied Sciences, vol. 7, no. 5, pp. 613–615, 2012.

[19] C. Choi, J. Choi, and P. Kim, "Ontology-based access control model for security policy reasoning in cloud computing," The Journal of Supercomputing, vol. 67, no. 3, pp. 711–722, Mar. 2014.

[20] S. Ruj, A. Nayak, and I. Stojmenovic, "DACC: Distributed Access Control in Clouds," 2011, pp. 91–98.

[21] Y. Zhu, H. Hu, G.-J. Ahn, D. Huang, and S. Wang, "Towards temporal access control in cloud computing," in INFOCOM, 2012 Proceedings IEEE, 2012, pp. 2576–2580.

[22] C. Belbergui, N. Elkamoun, and R. Hilal, " A dynamic Access Control Model for Cloud Computing environments," In Proceedings of the 6th International Conference on Communication and Network Security, 2016, pp. 21-29. ACM.

AUTHORS PROFILE

Belbergui Chaimaa - was born in Morocco in 1991. She obtained her Masters in networks and systems from "Sciences and technologies Faculty of Settat" in 2013. She is currently studying for a doctorate at Chouaib Doukkali University in Morocco. Her field of interest is Modeling and assessment of the security of a companies' information systems.

Najib Elkamoun – received his Ph.D. degree in Optical and Microwave Communication from the National Polytechnic Institute of Grenoble, France, in 1990. He is currently Professor Researcher at Faculty of Science, University Chouaib Doukkali, El Jadida, Morocco. With over 20 years of expertise in information technology and communication, he has conducted several thesis and overseas missions in e-learning and telecommunication networks. His research interests include High Speed Network Architectures (MPLS), Mobility Management, security and QoS in Emerging Networks (MANET, VANET and WSN), Wireless Communications and Traffic Engineering for Computer and Telecommunication Networks.

Rachid Hilal – was born in Rabat, Morocco, in 1965. He received the PhD degree in telecommunications from the University of Limoges, France, in 1996. Since 2003, he was the General Secretary of the Cadi Ayyad University Marrakech. Since 2012, he is a vice president of the Chouaib Doukkali University El Jadida. He is member of the STIC Laboratory and ability professor. His research interests include distributed power amplifiers, microwave nonlinear circuit design, and inset-fed patch antenna.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

CLUSTERING ALGORITHMS: A REVIEW

NIKITHA JOHNSIRANI VENKATESAN, CHOONSUNG NAM, DONG RYEOL SHIN*

School of Electrical and Computer Engineering,

Sungkyunkwan UniversitySuwon, South Korea

E-mail: nikithajv@skku.edu, namgun99@gmail.com, drshin@skku.edu

ABSTRACT

In the last few years, large-scale data analysis has become a hot topic in both research and industry side. The traditional database system could no more handle the storage and processing of big data in an individual system. Clustering plays a major role in data analysis. Since 90 % of the total data we deal today are unstructured, classification and recommendation cannot be of great use. Clustering comes into play to deal with a large amount of unstructured data. Apache Mahout, an open source platform which is focused in of clustering and classification is used on top of Apache Hadoop. In this paper, we study various clustering algorithms on text and XML dataset. We also analyzed the best clustering algorithm by varying important parameters such as distance measure, convergence threshold, and the like. This paper not only gives a detailed survey of apache hadoop and mahout but also presents a case study evaluation of cluster quality. The working of mahout clustering algorithms and Hadoop distributed file system (hdfs) is reviewed in an exhausted manner. We used text and XML dataset of size 7.27 GB (7,811,654,084 bytes) for clustering. Comparison of results in accordance with a distance measure, a clustering algorithm, convergence parameter is provided. This paper helps researchers to analyze the merits and demerits of various clustering algorithms on different kinds of datasets.

KEYWORDS: big data, Apache mahout, clustering algorithms, Hadoop

1. INTRODUCTION

In today's digital era, Big Data is everywhere from social media to sensor data. The rapid evolution of technology is sweeping the entire world [1] with digital information. Humans are connected with each other through a social platform where neither time nor space act as barriers to communication. Statistics show that it will take 98 years for a person to watch a one day's YouTube videos back-to-back continuously. Clustering [2] is a hot topic of research in different sectors such as management, biology, social media, statistics, and the like. Any kind of objects can be clustered if we can distinguish the objects into similar and dissimilar. For example, Photos and pictures can be clustered depending upon the colors, date of the picture taken, animals or humans, Friends, and families etc.

Today data is being collected almost from everything and everywhere. Our entire life is dependent on data and information. The world has seen exponential growth in data because of social media, health organizations, sensors and so on. Big Data has earned a great significance among the researchers in the past years. "Big Data", as the name depicts, refers to really large data which cannot be handled by the traditional storage system. Hadoop comes into play for storing and processing the Big Data [3]. The data which is stored without any further processing or analyzing is of no use. The main purpose of storing a large amount of data is to find some useful pattern out of it or to predict any fault detection in advance. For example, in health organizations, the data is collected from the patients to predict the various kinds of disease that can occur in future [4]. A famous use case is predicting the cancerous tumor based on its characteristics and features [5]. Similarly, sensor data which are collected from vehicles are of great use to insurance companies [6]. The data analysts have information like driving speed, braking, mileage etc. which are useful to know about the driver. A business organization mainly collects the customer data to understand their churn [7]. After the collection of data, there are many problems in analyzing the data like, finding patterns from the data and identifying new patterns from the incoming datasets. Hadoop provides an open source tool called Mahout which can be very much

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

useful in finding out the patterns from the datasets. In Machine Learning [8], supervised and unsupervised learning is two broad categories. Supervised learning is further categorized into regression and classification. In the former, the input variables are mapped to some continuous function as output while in the latter, the input is predicted based on its characteristics and mapped into the predefined discrete category as output. Unsupervised learning has little or no idea about what the output should be. Based on the similarities of the input data, the output is predicted and clustered according to the relationships among the data.

In this paper, we will focus on clustering, which is unsupervised learning. We took different types of datasets say text and XML (Extensible Markup Language). The important algorithms are applied to the datasets and the resulting clusters are observed. Similarly, by varying the distance measure and the size of the datasets, the quality of the clusters and the time taken to compute the clusters are noted. Finally, we provide the justification for which algorithm is best suited for various kind of datasets. The organization of the paper is as follows: Section 2 talks about literature review and real-time applications of Mahout Clustering [9]. The working of Apache Mahout and Apache Hadoop is discussed in Section 3. Section 4 exhaustively describes the pseudo code of clustering algorithms step by step. Section 5 and 6 interpret our experiments and results. Finally, Section 7 concludes the paper with possible future work

2. BACKGROUND & LITERATURE REVIEW

The basic idea of clustering is to group the unordered items based on some manner of homogeneity. In technical words, clustering helps to find a proper55structure in a collection of unlabeled data. Clustering involves three important things:

•An Algorithm: A step by step procedure used to group items together

•An idea of similarity and dissimilarity: Evaluating which item belongs to which cluster

•A stopping condition: A condition in which the items cannot be grouped further.

The above is the general idea which is being used in various kinds of clustering algorithms based on the size of the data and the results we want. The following section briefs the various research articles that use Apache Mahout clustering algorithms for the respective datasets.

In [10], topic modeling in Mahout has been used for clustering the large amount of data in business organizations. To find out the hidden topics and to eliminate the unwanted noises in Hadoop framework, Latent Dirichlet Allocation is used. A reuters-21578 text corpus is feed into LDA [11] by extending the features of collapsed variational Bayesian (CVB). The results indicate better accuracy in finding the topics among multiple documents when compared to standard CVB. An interesting case study on tourism agency in Sri Lanka has used Apache Mahout collaborative filtering for data analysis [12]. The authors analyzed the data from social media, visitor's blog and the feedback from the tourists. They designed an architecture using Hadoop open source components, where Mahout plays the major role in analyzing the data. Similarly, Arantxa et al. Proposed an open source solution for analyzing the technical call centers data [13]. The architecture includes Mahout as the main components it does the clustering and analysis part. Though other open source components such as Hive and HBase are useful for storage and querying, analyzing the data is the main task to find out patterns from the data. [14] Did a comparison between a variety of data mining frameworks such as weka, rapid miner, Mahout and R. Log files are collected from the virtual campus of Open University of Catalonia and used as a dataset. The dataset is really large and Mahout stood out in performance as well as in clustering quality. Rui Mximo Esteves et al. Clusters all the Wikipedia articles in XML format using K-means and Fuzzy k-means algorithm in Mahout [15]. Since their research was conducted in an early stage, Mahout could not remove the noise in the dataset. The results concluded that the data should undergo lots of preprocessing for a better quality of clustering. Mahout provides the facility for creating our own distance measure as well clustering algorithm [16]. The authors have improved the cosine distance measure and applied the k-means clustering for Wikipedia and Reuters data set. The improvised version of the cosine distance measure shows better accuracy than the standard cosine distance measure. In [17], cloud-based mahout's k-means algorithm is utilized for analyzing data. As healthcare organizations deal with a large amount of data, the authors used healthcare data set for discussing the Mahout clustering. They concluded that Apache Mahout scales well for the enormous amount of data and as the number of nodes gets increased, the algorithm takes less amount of time to cluster. [18] and [19] has focused on social media analyzing. Twitter apparently uses lots of hashtags and they mean the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

emotions and moods of the users. Connections between different hashtags have been clustered based upon some sort of similarity. The experiment was done by distributed Hadoop environment using Mahout clustering algorithms. The results revealed great connections between hashtags and topmost words of the cluster. [20] also deals with twitter data clustering based on weighting mechanism and stop words. ChaoLung et al. did real-time crop data analysis in Taiwan using Mahout clustering [21]. The crop dataset contains the species of the crops from the agriculture department. Analysing this data using Mahout and Hadoop increases the seasonal effect in terms of price and volume. Venkateswara Reddy et al. did a comparative study on K-Means, fuzzy and Canopy clustering algorithms at [22]. Two different text data sets taken from UCI repository and Reuters data set are used for the experiments and comparison. The comparison was made between in-memory implementation and MapReduce phase. The results concluded that K-means is not the best suit for Big Data as it suffers in identifying the number of clusters initially. Most of the researchers in their articles use real-world datasets for analysis.5

Big organizations such as BuzzLogic, Breast cancer center, stack overflow which is an online discussion website are all using Apache Mahout for analyzing a large amount of data. In medical image processing, Mahout clustering is utilized for image segmentation for CAD (Computer Aided Disease) diagnosis. The mammogram in the breast is clustered into several parts for the future identification of cancer affected region. The real-time applications of Mahout clustering have been exhaustively discussed. From the literature review, it is clear that Mahout is a successful and satisfactory platform for data analysis. The next section gives denotation about the in-depth working of Apache Hadoop and Mahout.

3. TOOLS & TECHNOLOGIES

From the literature review, it is well understood that most of the organizations exploit Mahout technology and Hadoop for data analysis. When implementing Big Data projects, selecting appropriate technologies is a part of processing the data sets. Almost everywhere we go online nowadays, Apache Hadoop plays a vital part in statistical analysis, ETL processing, and business intelligence. Popular website or web services like Facebook, eBay, Yelp, and Etsyall uses Hadoop and Mahout to analyses their huge datasets. The organizations use Mahout to either generate data about the user's behavior or even for their own operations. This section details the working of Hadoop and Mahout in detail.

3.1 Hadoop

Google faced a problem of handling billions of searches and indexing millions of web pages. But, they could not find any distributed, large and scalable computing platforms for analyzing their own data. To address this problem, the Google team decided to create their own algorithm that allowed the large data calculations to be chopped up into smaller chunks and map into many computers. Google released their own Google File system paper [23] and MapReduce paper [24] in 2003 and 2004 respectively. Inspired by these papers, Doug Cutting created an open source project called, Hadoop. Doug Cutting is also the creator of Apache Lucene. Apache Lucene is a popularly used text search library [25]. Hadoop helps to store and process Big Data in a distributed cluster of computers using simple programming models. Hadoop is a part of Apache project which is sponsored by Apache Software Foundation. It is written in Java language for distributed storage and processing of mountains of data on computer clusters built from commodity hardware. The core concepts of Hadoop are Hadoop Distributed File System(HDFS) and Map Reduce algorithms. We will take a closer look at these two components.

3.1.1 HDFS HDFS helps to store petabytes of data among multiple nodes and servers. HDFS [26] is a specially designed file system that allows multiple machines to store and process data from a single source. It consists of a Name Node and a Data Node which operates as a master-slave architecture. Name Node serves as a master component whereas Data Node serves as a slave component. Name Node comprises of metadata information of HDFS [26]. Metadata includes data like a number of blocks present in DataNode; a total number of times the data files being replicated; starting a cycle of the NameNode; the number of DataNodes which are under a particular NameNode; capacity of the NameNode; and information about space utilization. Data Node comprises information of all the processing data

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

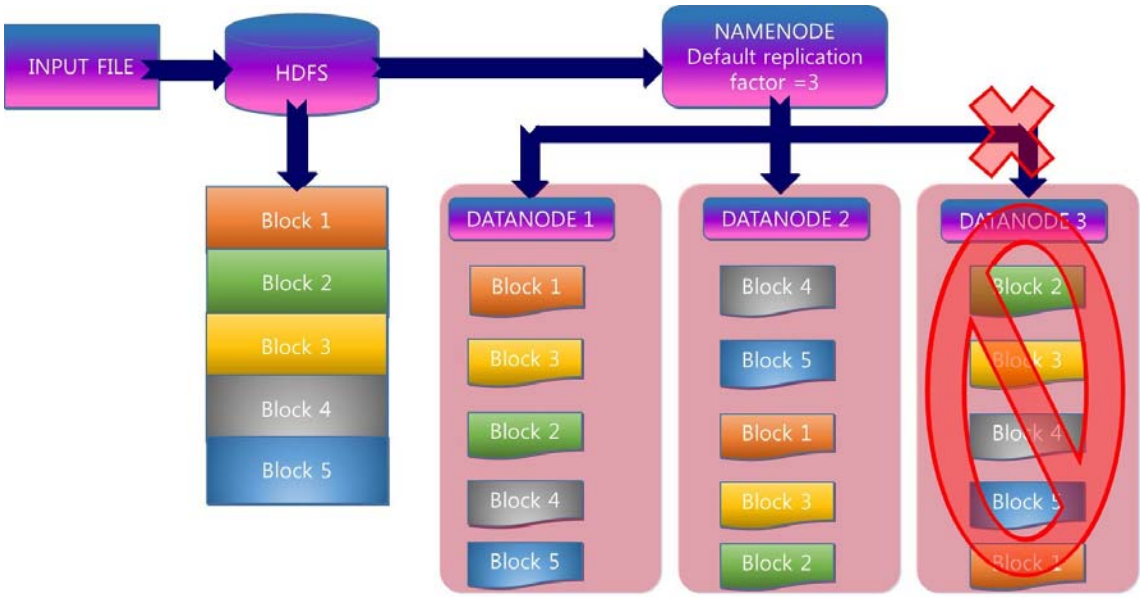


Figure 1: Replication process

which are stored in the Data Node and the machine it is being deployed. DataNode also deals with the actual storage of the file being processed and it serves READ and WRITE request for the client [27]. In the earlier versions of Hadoop, there was only one NameNode attached to multiple DataNodes which will result in a single point of failure. Later, the new versions of Hadoop provide multiple NameNodes where secondary NameNode can take over the entire process in case the primary NameNode fails. Secondary NameNode is responsible for conducting periodic checkpoints. HDFS also has self-healing architecture, where it replicates the same data across different nodes. So, it can process the data in a high availability environment. Let's say we have one NameNode and three DataNodes, and the data is transferred to the DataNodes from the client environment. HDFS has a replication factor which defines how many times a data block should be replicated in the clustered environment. For example, we have a file that is split into five data blocks across three DataNodes. The replication factor is set to three. If one of the nodes failed, the data from the failed node will get redistributed among the remaining active nodes. Hence, the other nodes complete the processing successfully. Figure 1 depicts the replication process in detail. In Hadoop 1.0 the NameNode use Jobtracker and data node uses Tasktrackerto manage resources.

3.1.2 Map Reduce Map Reduce is a parallel programming model which helps to retrieve the data from Hadoop cluster and process all the data across commodity computer clusters. Processing data could mean anything like searching, counting the keywords, aggregating or enriching the data. Map Reduce is composed of several components. Some of them are **Job Tracker**: It is the master node which manages all the resources and jobs in the cluster. It also schedules the jobs and keeps track of the map reduce tasks across all the nodes. In case of failure, JobTracker reassigns the job to some other node. To sum up, JobTracker makes sure that the Big Data process successfully and the client get backs the data in a reliable condition. **Task Trackers**: They are agents that are deployed in each cluster to run the map reduce tasks which are assigned by the JobTracker. It sends a periodic heartbeat message informing that the node is running without any interruption. **Job History Server**: It is a component which tracks all the completed jobs and is deployed with JobTracker. The implementation of the program is two parts: Map and Reduce. The data is initially fed into the Mapper function to form intermediate key and value pairs. Once the mapping function is done, Reducer gets all the intermediate results to produce a final output.

3.2 Apache Mahout

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Mahout is a machine learning library which is an open source from Apache. Apache Mahout aims to find out useful patterns from a large set of data, later predicting patterns from the new datasets as well. In this section, we shall discuss the fundamental concepts of Mahout clustering.

3.2.1 Distance Measure Distance measure technique is important implementation because the results of the same algorithm might differ depends upon the distance measuring technique. So, it is important to choose the distance measure based on the dataset. Distance measure affects the final quality of clustering. The data set to be clustered should be considered before selecting the distance measure. For clustering points which are well distributed in space, Euclidean distance measure suits well. If we need to cluster news snippets, cosine distance measure can be applied. For varying lengths of a document such as news articles, Tanimoto suits well. For complex clustering, like identifying friend root in the social network graph, weighted distance measure will apply. Different datasets have a different effect on the distance. If we knew the data set well enough, implementing our own distance measure will result in quality improvement. We used Euclidean, squared Euclidean, Tanimoto, Manhattan and cosine distance measure for clustering our datasets.

3.2.2 Representing Data Representation of input data as vectors help the algorithms to understand the data sets better and to calculate the similarity between the datasets. Datasets can be described as an ordered list of values, which is a vector. In 2-dimensions, vectors are represented as (5,8) first value for x dimension and the second one for y. Mahout can contain any number of dimensions when we deal with various kinds of input data. All the datasets contains features. Vectorization starts with assigning each feature to each dimension.

3.2.3 Vectorization Understanding the procedure of vectorization is essential to achieve accurate clusters. Apache takes vectors as input to produce good results. Vectors can be implemented as two classes which are explained below,

•Dense Vector: It implements the vector as an array of doubles because dense vector assumes that the size equals the number of features in the input datasets. It can handle zero values as well.
•Sparse Vector: Sparse vectors consists of two types named, Random and Sequential. The former is implemented between a double and an integer. Random Sparse vector allocates non-zero features. The latter vector which is sequential is implemented in parallel arrays where one for integers and other for doubles.

However, the vector selection does not affect the final results but, will increase the calculation speed and decrease memory usage. The vectors are stored in the sequence file format, which can be read by Mahout algorithms. Sequence files are files which consist of key value pairs. Hadoop Map Reduce use input and output data sets in sequence file format. As Mahout runs on top of Hadoop, it is necessary to give the input in sequence file format. Among two kinds of dataset, one of the dataset we used is text format. We shall discuss the process of converting text into vector format. In real time scenarios, the text documents that are to be clustered will come in petabytes of sizes. And also all the documents will not be of uniform length. In the above subsection, we gave example for 2-dimension vectors. But, in the real case, the dimension of the text document vectors will be really large. Since the count of words in a document of size say terabyte and petabyte will be incredibly big, the vectors will have infinite dimensions. One important factor to be considered is the frequency of words in a given file. A number of occurrences of a word are referred as a value of the vector. This is defined as Term Frequency (TF) weighting. In Mahout, the value of the vector is technically referred as weight. In text document clustering, the similarity between any two documents is calculated based on distance measure. TF does not know about the common words like “the”, “is”, “a” and it calculates the weight by including all the words in the document. So the result will be disagreeable, as the distance value will be influenced by the weight of the frequent words. Such frequently occurring words are referred to as stop words. This stop-words problem can be overcome by Term Frequency- Inverse Document Frequency (TF-IDF) which is discussed in the following section.

3.2.4. TF-IDF Term frequency-inverse document frequency (TF-IDF) weight is a statistical measure used to detect the importance of a particular term in the document. Intuitively, a word is given high score if it occurs frequently in a document. But, the word is given low score if it occurs frequently in many documents. Thus, we can conclude that the word is not unique identifier. Words like “the”, “for”, “a”, “what” should be

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

scaled down since they are most common words which occur across multiple documents, whereas words which occurs frequently in single document should be scaled up. In TF-IDF, Inverse document frequency (IDF) is used to improve the weighting. DF is the number of documents a word occurs in and N is the total number of documents present. The log value is determined to reduce the problem of final term weight as the value wont be ideal. The formula for IDF for a particular word, w_i is mentioned below.

$$IDF_i=\log(N/DF_i)$$

Now, to calculate the TF-IDF weight for a word w_i is,

$$W_i=T\ F_i*\log(N/DF_i)$$

This above calculated weight for a word I will be assigned as a dimension. By this, stop words will have small weight whereas the unique words will get larger weights, resulting in best cluster formation. Vector space model assumes that all the words in the document are independent to each other. Still, word dependencies are another threat for weight calculation. Some words like "coca cola" has a high probability of occurring together which is called word dependencies. In the next subsection, we shall discuss how to overcome the word dependencies.

3.2.5 Normalization n-gram in general is defined as group of words that occur in a sequence. Combination of single word is defined as "unigram" and similarly combination of two words is called "bigram". More than two words combinations are referred as "trigram", "4-gram" and so on. Mahout has a technique to find out the highest probability of combined words like coco cola, Martin Luther. Vectors can be created to point at bigrams instead of single words. From a single sentence, bigrams can be created by picking all the possible combination of two words. Some may give meaningful units, while most of the bigrams represent worst meaning [28]. If bigrams are created for all the documents, eventually we will end up with lots of meaningless bigrams with big TF-IDF weight. To outdo this problem, Mahout conduct log-likelihood test. This test can conclude whether the bigrams forms notable meaning or not. It selects the bigrams with most appropriate meaning and eliminates the remaining. With the help of TF-ID weight and n-gram collocation, text documents are converted into input vectors. Thus resulting in high quality of clusters.

The text document is converted into sequence format for the sake of HDFS by using the Mahout launcher script. Later, the sequence format of input data is converted into vector for better cluster formation. Once the vector is formed from the input data, the following output are written in HDFS as a result of vectorization process.

•**Tokenized documents:** The documents are segregated as individual words, using Standard Analyzer and stored in this file.

•**Word Count**: The n-gram is generated by iterating through the tokenized documents and the important meaningful words are stored in this folder.

•**TF vectors:** Using term frequency , TF vectors are created from the tokenized documents.

•**DF count**: The frequency of the words across multiple documents is calculated for TF-IDF weight.

•**TF-IDF vectors**: The vectorizer uses TF-IDF weight, so after the generation of TF vectors, the weight for TD-IDF is calculated.

•**Dictionary file**: The final folder "dictionary.file-0" consists of mapping between a word and the integer. This file is used as input for the various algorithms in Mahout. This file contains all the meaningful bigrams, trigrams as well.

Normalization is a method for cleaning up the edge cases. Data with ab-normal characteristics change the direction of the resulting clusters disproportionately. In general, it is common for any two documents which are not all similar but they pop up claiming they are similar. This is due to the size of the document. When a larger document is being compared with a smaller document, it will results as they are similar because of the large number of non-zero dimensions. The error should be negated as the incoming datasets are not of same length and size. Normalization decreases the magnitude of the large vectors and increases the magnitude of the small vectors. The quality of the final clusters are increased to an extent by normalization. Further increase in clusters quality depends upon the selection of distance measure and suitable algorithms for the input datasets. Distance measures were already discussed in the previous section. The working of the clustering algorithms are exhaustively explained in the following section.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

4. CLUSTERING ALGORITHMS

Clustering in Mahout refers to organizing the unknown elements based on their similarity. In this section, we present a brief overview of some important clustering algorithms in Mahout. The general working of each algorithm and the way it works in Mahout as well are discussed below.

4.1 K-means Clustering

K-means is the traditional simple algorithm which ages 50 years old. As the name illustrates, k is the important parameter which is used to set the number of output clusters [29]. The quality of the cluster is dependent on the value of k. In our paper, one of the dataset we clustered is news dataset. The expected out-put should be grouped into various news categories say politics, health, sports and the like The first step is to assign the centroids among the data points. Mahout chooses the initial centroid randomly by random seed generator. The second step is to assign the data points to the nearest centroids to form the initial clusters and calculate the mean. The above steps are repeated again and again until the data points are assigned to the same cluster in successive rounds. Figure 2 shows the working of the K-means algorithm as a flow chart. Mahout computes the K-means algorithms as Expectation Maximization (EM) algorithm [30]. Based on the initial centroids, the expected points are assigned to the cluster. Later the cluster center is improved by varying distance measure until it reaches the convergence point. Convergence point is that point where the centroids don't move further after the computing. Then no further iterations are required and the final clusters are displayed. In Mahout, convergence threshold, maximum iterations, distance measure are used as necessary parameters to execute the K-means algorithm. K-means will read from the dictionary file, which is created as a result of vectorization and form the final clusters.

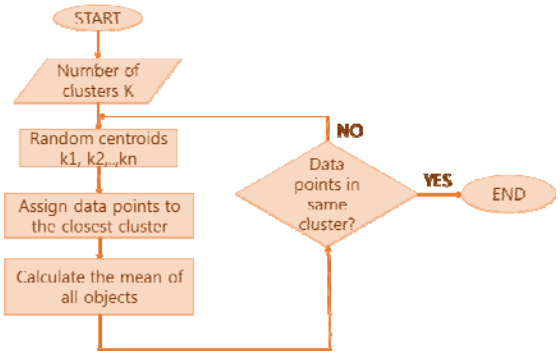


Figure 2: Work flow of K-means algorithm

K-means algorithm can be executed both by java code and MapReduce. Java implementation is feasible only up to certain limit of dataset. When dealing with large amount of data, MapReduce gives best result in clustering. As Hadoop and MapReduce splits the input into several chunks for parallel processing, Mahout framework takes care of the spilt chunks of the vectors. After the formation of vectors, the algorithm command is executed with all the necessary input parameters. After the implementation of K-means, the output directory contains all the clusters which are formed form first iteration until the final iteration. Clustered Points directory consists of the mapping information for documents and cluster. Figure 3 shows the final output directory of cluster formation in HDFS. The final clusters formed in the directory is in vector format which is not understandable by humans.

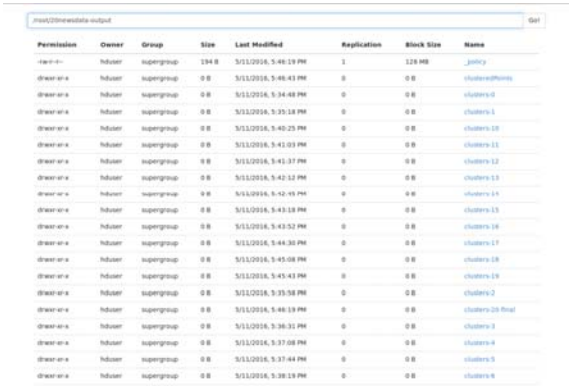


Figure 3: Clusters output directory

Mahout provides a utility called ClusterDumper which can convert vector format into words and dump them in the local directory.

4.2. Canopy Clustering

Canopy is not only fast clustering process but also accurate in grouping390the data points based on some sort of similarity. It mainly uses two distance measures named T1 and T2 where T1>T2. The first step is to select any point randomly from the input datasets to form the canopy center. From that random point the distance to all other data points is calculated. By keeping all the points which falls into the distance valueT1it forms an initial cluster. Remove all the data points that fall with in the distance value T2. These points can neither be a new centroid nor be capable of forming new canopies. The above steps are

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

repeated again and again till there are no more points to form a new cluster.

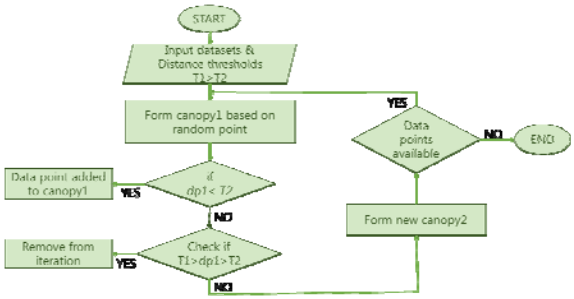


Figure 4: Flow chart of Canopy algorithm

The advantage of canopy clustering is its quick formation of clustering, but the quality of the final clusters cannot be guaranteed. The one disadvantage in K-means is the random initial centroid formation. If the initial centroid selection is not good, then again the quality of the final clusters might be bad. In order to overcome both the disadvantages, Mahout provides a way to combine canopy and k-means algorithm. Canopy algorithm is used to form the centroids instead of random seed generator.

The dataset about news articles contain may overlapping topics which cannot be clustered under a single topic. Let's say, a news article which talks about celebrity's death might be clustered in both health and movies. In such case, the algorithm should be able to replicate the news article in both the categories. In such cases, the fuzzy K-means clustering algorithm serves better which is discussed in the following section.

4.3. Fuzzy K-means Clustering

Fuzzy k-means otherwise known as fuzzy c-means clustering algorithm. This algorithm tries to create overlapping clusters between the datasets. Fuzzy k-means working is based on soft cluster where one data point can belong to more than one cluster.

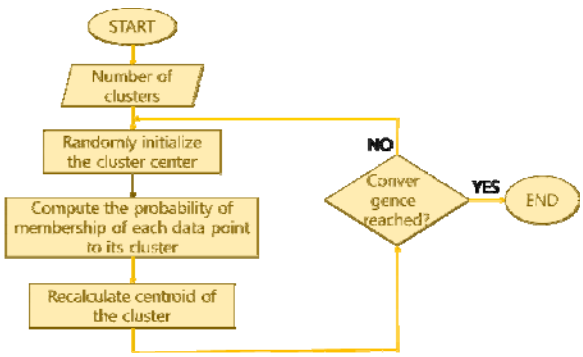


Figure 5: Working flow of fuzzy k-means algorithm

In Mahout, the distance measure, convergence threshold, number of iterations and fuzziness factor are given as input parameters. The value of the fuzziness factor should be greater than 1. As the fuzziness factor becomes greater than 1 the datasets might have more overlap. Figure 5 gives overall working of fuzzy k-means algorithm. The algorithm runs in MapReduce mode involving the following three steps.

•Mapper: Reads the input dataset and calculate the membership probability of the data point to its cluster.

•Combiner: Receives the output from mapper and calculates the partial sums of probability of each vector.

•Reducer: Calculates the total sum of the membership probability and creates a new centroid for the cluster.

If the data is uncertain, fuzzy k-means gives better cluster quality than K-means. There is one disadvantage that either of the above three algorithms cannot handle. Standard k-means is well suitable for exclusive clustering and fuzzy k-means is for overlapping clustering. But if the datasets are not in the normal distribution, then applying the above algorithms might not generate a best fit for the data points. For an instance, to form a cluster based on specific topic which is debatable. A word can be included in two topics because it has same meaning, but not both. To prevail over this issue, topic modeling algorithms are implemented which are discussed further in the upcoming sections.

4.4. LDA

Latent Dirichlet Allocation (LDA) clustering algorithm learns the pattern from the input data based on the similarity of the context, it forms the final clusters. In the dense context, LDA assumes that the datasets has t number of topics and the entire input datasets are revolving around the topic. Single input document is considered to have mixed probabilities of each t number of topics. In topic modeling, the algorithm takes word vector as input instead of document vector as it deals with each word of the document to find out the probability of the topics. In the word vector, IDs are features and the total number of documents they occur in will be considered as weights. Using word vector, the cluster of the words are figured out by executing the algorithm and the topics are discovered. After the discovery of the topics, the probabilities of each word occurring in that topic is calculated. This is the general working of topic modeling.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

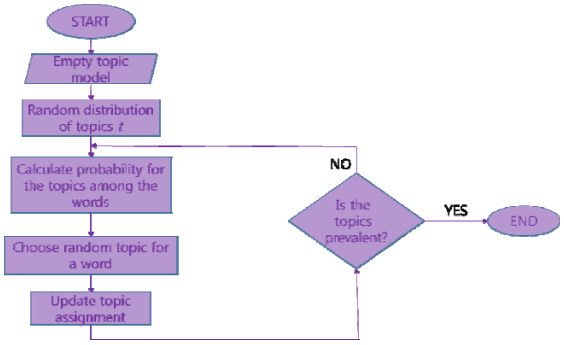


Figure 6: Flow model of LDA algorithm

LDA works one step ahead of topic modeling. Even if any words which have the same meaning but do not occur together, LDA scrutinize across the documents, to figure out the similar words and the content of the document. After examining the words which have the same concept, the words are considered as a single topic. The following figure 7 shows the required input to be given to run LDA algorithm. The matrix results represent the most frequently occurring words in the documents. This is later used for building an LDA model. The model based algorithm, LDA creates good quality of clusters when the datasets are static. The prediction of pattern in the data and the creation of cluster quality is high when we deal with a large amount of stored data set. But, when dealing with real-time datasets, finding a pattern from the dataset might be difficult if the data keeps on accumulating in the directory.

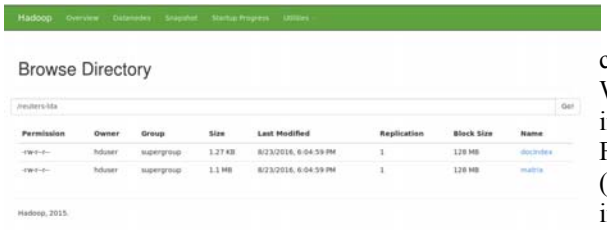


Figure 7: Output of LDA

To overcome this problem, streaming k-means is used which is detailed in the following section.

4.5. Streaming K-means Clustering

Streaming K-means clustering algorithm is most suited for dynamic kind of data [31]. It is mainly used for online clustering, where the input data points are received every second. In our paper, we do not deal with real time data.

5. Experiments

5.1. Datasets

In order to compare the results of the algorithms, two different types of datasets are used, text and xml. We downloaded bag of words dataset for the former and Multivariate for the latter. The size of the text and xml dataset are 7.27 GB combining both test and training data uncompressed.

5.2. Setup

We had 8 node Hadoop clusters which is based on commodity hardware. Each node has the following same configurations as follows: 8GB RAM; 4 integrated Gigabyte Ethernet 1000BASE-T ports(RJ-45); 0.5 TB hard drive; 1CPU Inter Core i4; 30 MB L3 cache with Ubuntu version 14.04, Hadoop v2.7and Mahout v0.9. The evaluation is detailed corresponding to distance measure, convergence and clustering speed.

5.3. Use cases

The following are the different scenarios we considered for doing the Mahout case study. (i) Text dataset 7 GB tested with different distance measures. (ii)XML dataset 6 GB tested with different distance measure. (iii) Text dataset with K-means; Single node; experimented by varying the convergence threshold.(iv) XML dataset with Fuzzy K-means; Single node; experimented by varying the fuzziness factor. (v) Text dataset clustering with LDA; pseudo distributed with 1 server and 8 nodes.

6. RESULTS & INTERPRETATION

In this section, the analysis results of the clustering algorithms are discussed exhaustively. We considered four clustering algorithms for the interpretation: K-means, K-means with Canopy, Fuzzy K-means and Latent Dirichlet Allocation (LDA). The above mentioned datasets were loaded into HDFS. The block size of HDFS was 128 MB. The replication factor of the dataset is set to default value which is 3. The input parameters such as initial cluster, total number of clusters, distance measure, distance thresholds make a huge difference in the resulting quality of the clusters and the accuracy.

Table 1: Clustering parameters

PARAMETERS	Text Values	XML values	Fixed clusters
K-means	20	20	Yes
Canopy Clustering T1	2050	2000	No
Canopy C;ustering T2	1050	1500	No

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Fuzziness factor m	-m 1.05, 2, 2.5, 3.0	2	Yes

In K-means, selecting the number of clusters (k) depends upon the input data and the expected output. In our text dataset, the value of the parameter k should be 10 to 20 if we want to cluster only general topics say sports, health, politics etc. On the other hand, if the resulting category needed to be finely clustered, then the input number of clusters should be increased to 35 to 40. If the volume of data is too large to be stored in the main memory available, the K-Means algorithm not suitable, as it's batch processing mechanism iterates over all the data points. Also, the K-Means algorithm is sensitive to the noise and outliers in data. Moreover, its random initialization step causes problems when it comes to computation time and accuracy. Canopy clustering has been applied for both text and XML datasets. We used canopy clustering as a pre-clustering algorithm. The result of the canopy clustering was used to estimate the values of centroids and number of clusters. Later, the initial set of clusters has been used as centroids in K-means algorithm. Canopy integrated with K-means gives quick and accurate results. From the results shown in the graph8, it is clear that canopy with k-means clustering gives quick results for XML dataset than text dataset. For large dataset which has high dimensions, canopy clustering does not work well. This is because high dimensional data points can be involved in many canopies, thus resulting in poor quality of cluster formation.

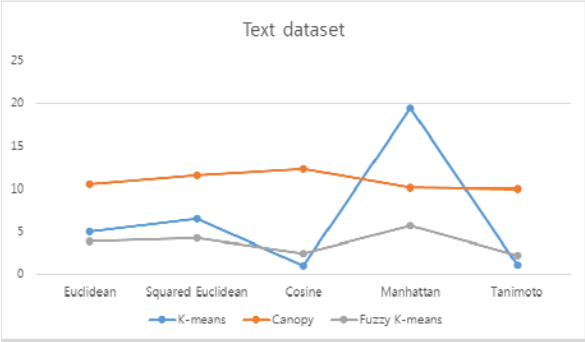


Figure 8: Clustering algorithms results of text dataset

Figure 8 depicts the processing speed of clustering algorithm by varying the distance measure. The X-axis shows the distance measure while the Y-axis indicates the time taken for the processing of the clusters in minutes. From the graph it is clear that Manhattan distance measure stands out extraordinarily huge from all other distance measures apart from the convergence value. The maximum iterations for all the algorithms have been kept constant as 20. Fora small size dataset like 2GB, Manhattan distance measure seems to be taking very long time. Surprisingly, for high dimensionality vectors Manhattan with K-means algorithm gives better quality cluster. It is clear that fuzzy K-means algorithm suits well for text dataset than XML dataset. Text dataset has to be soft clustered rather than hard clustering. The nature of the news datasets that they cannot be categorized in one single category. The news which talks about health issues of a sportsman should be clustered in both health category and sport category. The Euclidean distance measure is not fit for the text documents of varying size. This is because, the distance measure fails to consider the length of the documents. Hence, Fuzzy K-means is best suited for our text dataset. Except Manhattan distance measure, all others forms the final clusters more or less in a same time period. The maximum iteration parameter for all the distance measures and the clustering algorithms is set constant values 20.

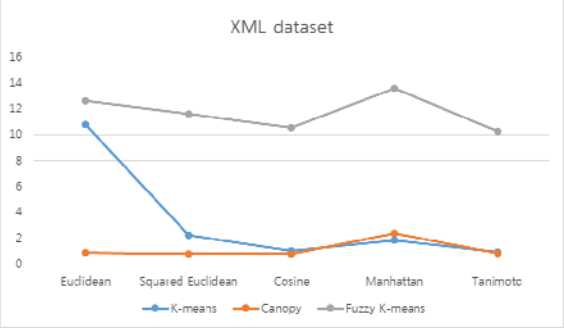


Figure 9: Clustering results of XML dataset

The figure 9 shows the results of the various clustering algorithm for XML dataset executed using different distance measure. The X-axis indicates various distance measure while the Y-axis shows the time taken in minutes for the formation of the clusters. Overall from the graph it is shown that canopy clustering algorithm works well for XML dataset. As Fuzzy K-means algorithm has overlapping nature, it showed better and quick results for text dataset. We expected the same kind of results for XML dataset. But we observed that the cluster formation takes much longer time even though we minimized the dataset to 2 GB. The cluster quality of fuzzy K-means was not good as most of the data points have been assigned to almost all the clusters. Figure 10 shows

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

the clustering results of the text dataset by varying the convergence threshold.

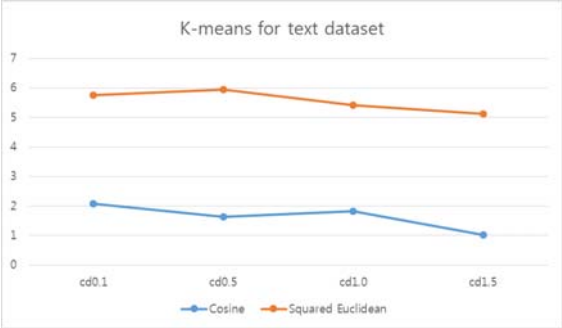


Figure 10: Clustering results of text dataset by varying the convergence threshold.

Hence, the centroids were not properly representing the clusters. In this case, even if the centroids are randomized, K-means clustering algorithm produced better clusters for XML dataset. With canopy as centers, K-means converge even faster irrespective to distance measure.

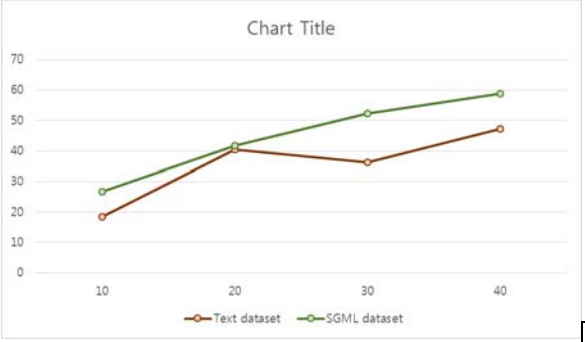


Figure 11: LDA model results of both text and sgml dataset

The results of the figure 11 indicates the k-means clustering results by varying the convergence threshold for text dataset. The X-axis indicates the values of convergence threshold while the Y-axis shows the execution time in minutes. It is clear that using cosine distance measure for clustering makes the dataset converge much faster when compared to squared Euclidean distance measure. When the distance measure we used is cosine, we observe that as the convergence value increases, the clustering time gets decreased. But, In case of squared Euclidean distance measure, none of the clusters reached convergence. All the clusters formed 10 clusters because it reached maximum iterations. It is known from the results that the squared Euclidean distance measure need more convergence for the resulting clusters. In the case of cosine distance, the number of resulting clusters also varies based on the convergence factor. Anyways, K-means clustering initialize its centroid using random seed generator. So, the results may vary when we use canopy centroids. From the table 2, it is visible that as the fuzziness factor gets increasing, there is slight decrement in the execution time. Though in our XML dataset, none of the clusters reached convergence. All the clusters reached maximum iterations and got stopped. From this results it can be decided that fuzzy k-means algorithm is not best suited for XML dataset.

Table 2: Clustering results for XML by varying the fuzziness value

Parameters	Fuzzy factor	Time taken (mins)
Cd 1.0, dm Squared Euclidean	-m 1.5	5.8829
	-m 2.0	5.6485
	-m 2.5	5.5325
	-m 3.0	5.3348

The results of the figure 11 shows that LDA gives quick results in case of text analysis. The X-axis indicates the maximum iteration whereas Y-axis indicates time in minutes. The input for the LDA analysis is raw text, and LDA models topics out of the raw input data. The table 3 lists top three topics among all the results and the corpus with highest probability values is displayed here. From the group of words, the topics they might be allocated is quickly understandable.

Table 3. Results of Topic model in LDA using 20 newsgroup dataset

Topic 1	Probability	Topic 2	Probability	Topic 3	Probability
Food	0.018	Space	0.023	Leaders	0.021
Treatment	0.012	Earth	0.012	Government	0.011
Cancer	0.0009	Launch	0.009	People	0.009
Medicine	0.006	Nasa	0.005	Organization	0.005
Health	0.006	Shuttle	0.004	Law	0.004
Disease	0.005	Research	0.003	Policy	0.002

Topic 1, 2 and 3 talks about health news, space exploration and politics respectively. LDA is successful because of its ability to produce interpretable topics among large amount of documents. Still, parallelization in LDA is not straightforward. From the graph 11, it is evident that none of the lda model attains convergence. The topics keep on modeling till the maximum iteration which is 40 in our case. Using EM

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

algorithm the convergence can be achieved in short number of iterations.

7. CONCLUSION

Mahout and Hadoop are promising platform for large data analysis due to open source and scalable characteristics. In-built libraries of Mahout and parallel processing feature of Hadoop provide better quality of resulting clusters. From the paper, it is clear that not all the algorithms are suited for all kind of datasets. There are many factors such as distance measure, convergence threshold, iterations, number of nodes to be considered before feeding the input datasets. We have experimented with text dataset and XML dataset. Although the selection of algorithm entirely depends upon the types of dataset, we still can choose the best suited distance measure from the above results.

Future work includes applying classification and recommendation algorithms for different kinds of datasets like CSV (Comma Separated Values).

8. ACKNOWLEDGEMENT

This research was supported by Basic Science Re-search Program through the National Research Foun-dation of Korea(NRF) funded by the Ministry ofEducation(NRF-2016R1A6A3A11932892)

REFERENCES:

[1] X. Zhu, B. Song, Y. Ni, Y. Ren, R. Li, Business Trends in the Digital Era:Evolution of Theories and Applications, Springer, 2016.

[2] J. A. Hartigan, J. Hartigan, Clustering algorithms, Vol. 209, Wiley NewYork, 1975.

[3] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. H.Byers, Big data: The next frontier for innovation, competition, and pro-ductivity.

[4] V. M. Rohokale, N. R. Prasad, R. Prasad, A cooperative internet of things(iot) for rural healthcare monitoring and control, in: Wireless Communi-cation, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology (Wireless VITAE), 2011 2nd International Con-ference on, IEEE, 2011, pp. 1–6.

[5] S. Loi, N. Sirtaine, F. Piette, R. Salgado, G. Viale, F. Van Eenoo, G. Rouas,P. Francis, J. P. Crown, E. Hitre, et al., Prognostic and predictive value oftumor-infiltrating lymphocytes in a phase iii randomized adjuvant breast cancer trial in node-positive breast cancer comparing the addition of do-cetaxel to doxorubicin with doxorubicin-based chemotherapy: Big 02-98,Journal of clinical oncology 31 (7) (2013) 860–867.30

[6] I. S. Kweon, T. Kanade, High resolution terrain map from multiple sen-sor data, in: Intelligent Robots and Systems' 90.'Towards a New Frontier of Applications', Proceedings. IROS'90. IEEE International Workshop on,IEEE, 1990, pp. 127–134.

[7] K. Coussement, D. Van den Poel, Churn prediction in subscription services:An application of support vector machines while comparing two parameter-selection techniques, Expert systems with applications 34 (1) (2008) 313–327.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel,M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Ma-chine learning in python, Journal of Machine Learning Research 12 (Oct)(2011) 2825–2830.

[9] S. Owen, S. Owen, Mahout in action.

[10] W. Romsaiyud, Big data topic modeling with mahout for managing busi-ness analysis services, in: System Integration (SII), 2014 IEEE/SICE In-ternational Symposium on, IEEE, 2014, pp. 514–519.

[11] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (Jan) (2003) 993–1022.

[12] R. Irudeen, S. Samaraweera, Big data solution for sri lankan development:A case study from travel and tourism, in: Advances in ICT for EmergingRegions (ICTer), 2013 International Conference on, IEEE, 2013, pp. 207–216.

[13] A. D. Barrachina, A. ODriscoll, A big data methodology for categorisingtechnical support requests using hadoop and mahout, Journal of Big Data1 (1) (2014) 1.

[14] F. Xhafa, D. Ramirez, D. Garcia, S. Caball´e, Performance evaluation ofdata mining frameworks in hadoop cluster using virtual campus log files, 31 in: Intelligent Networking and Collaborative Systems (INCOS), 2015 In-ternational Conference on, IEEE, 2015, pp. 217–222.

[15] C. Rong, et al., Using mahout for clustering wikipedia's latest articles: acomparison between k-means and fuzzy c-means in the cloud, in: CloudComputing Technology and Science (CloudCom), 2011 IEEE Third International Conference on, IEEE, 2011, pp. 565–569.

[16] L. Sahu, B. R. Mohan, An improved k-means algorithm using modified co-sine distance measure for document clustering using mahout with hadoop,in: 2014 9th International Conference on

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Industrial and Information Sys-tems (ICIIS), IEEE, 2014, pp. 1–5.
[17] S. Rallapalli, R. Gondkar, G. V. M. Rao, Cloud based k-means cluster-ing running as a mapreduce job for big data healthcare analytics usingapache mahout, in: Information Systems Design and Intelligent Applica-tions, Springer, 2016, pp. 127–135.
[18] C. I. Muntean, G. A. Morar, D. Moldovan, Exploring the meaning behind twitter hashtags through clustering, in: Business Information SystemsWorkshops, Springer, 2012, pp. 231–242
[19] E. Jain, S. Jain, Categorizing twitter users on the basis of their interestsusing hadoop/mahout platform, in: 2014 9th International Conference onIndustrial and Information Systems (ICIIS), IEEE, 2014, pp. 1–5.
[20] T.-S. Moh, S. Bhagvat, Clustering of technology tweets and the impactof stop words on clusters, in: Proceedings of the 50th Annual SoutheastRegional Conference, ACM, 2012, pp. 226–231.
[21] C.-L. Yang, M. R. Nurtam, Data clustering on taiwan crop sales underhadoop platform, in: Proceedings of the Institute of Industrial Engineers Asian Conference 2013, Springer, 2013, pp. 827–835.
[22] V. R. Eluri, M. Ramesh, A. S. M. Al-Jabri, M. Jane, A comparative studyof various clustering techniques on big data sets using apache mahout, in: 2016 3rd MEC International Conference on Big Data and Smart City(ICBDSC), IEEE, 2016, pp. 1–4.
[23] S. Ghemawat, H. Gobioff, S.-T. Leung, The google file system, in: ACMSIGOPS operating systems review, Vol. 37, ACM, 2003, pp. 29–43.
[24] J. Dean, S. Ghemawat, Mapreduce: simplified data processing on largeclusters, Communications of the ACM 51 (1) (2008) 107–113.
[25] T. White, Hadoop: The definitive guide, ” O’Reilly Media, Inc.”, 2012.
[26] D. Borthakur, et al., Hdfs architecture guide, Hadoop Apache Project 53.
[27] P. Zikopoulos, C. Eaton, et al., Understanding big data: Analytics forenterprise class hadoop and streaming data, McGraw-Hill Osborne Media,2011.
[28] N. K. Visalakshi, K. Thangavel, Impact of normalization in distributed k-means clustering, International Journal of Soft Computing 4 (4) (2009)168–172.
[29] K. Wagstaff, C. Cardie, S. Rogers, S. Schr¨odl, et al., Constrained k-meansclustering with background knowledge, in: ICML, Vol. 1, 2001, pp. 577–584.
[30] F. Dellaert, The expectation maximization algorithm. 2002, URLhttp://www. cc. gatech. edu/dellaert/em-paper. pdf.
[31] K. Venkatalakshmi, S. M. Shalinie, Classification of multispectral imagesusing support vector machines based on pso and k-means clustering, in:Proceedings of 2005 International Conference on Intelligent Sensing and Information Processing, 2005., IEEE, 2005, pp. 127–133

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Implementation challenges in Campus Network security

Zaka Ullah
Department of Computer Science
Lahore Garrison University, Lahore, Pakistan
Email: zakaullah@lgu.edu.pk

Muhammad Zulkifl Hasan
Department of Computer Science
Lahore Garrison University, Lahore, Pakistan
Email: zulkifl.hasan@lgu.edu.pk

***Abstract*—Recently, due to economical escalation & with the need of advancements in national strategy based on scientific technologies, maximum institutions have been adopting their own campus network. That is why setting up campus network is being realized as important part of school education information. The utilization of campus network shares opportunity for teaching, scientific research and management to work together by using resources and exchange maximum information in minimum time. Hence, campus network security positively influences school teaching activities.**

***Index Terms*—SSL, PKI, IDS, IPS, VPN, URL, DNS, DMZ, LAN, IT.**

I. Introduction

With the expansion of society, organize data is utilized by individuals to an ever-increasing extent and its significance is getting to be prominent, the system has turned into the most basic thing of a nation's political, financial, and military assets. Nowadays security is compromised due to various applications and can cause data spillage, illegal data entrance, and unlawful utilization of system assets, data altering and counterfeit. The security may include identification, authentication, authorization, and surveillance camera to protect integrity, availability, accountability, and authenticity of computer hardware or network equipment [Ali et al.(2015)Ali, Hossain, and Parvez].
As the pervasiveness of multi-PC arrange security dangers, avert "programmers" is feeble, corporate and government sites have been "assaulted" increasingly often, these have caused tremendous monetary misfortunes. Subsequently, the system data framework security and avoidance and its Secrecy appears to be progressively essential. With the quick extension of the grounds arrange availability, the system applications have expanded quickly, in the meantime, the grounds organize data security has caused more consideration today. A hierarchical architecture of campus network is configured with different types of traffic loads and security issues for ensuring the quality of service. [Ali et al.(2015)Ali, Hossain, and Parvez] In recent organize and application frameworks, there are no safety efforts which have been taken care off. Meanwhile security vulnerabilities in the host working framework and application framework are additionally with no handling, there are numerous issues inside framework administration, all of these shaped a genuine security issue, therefore truly debilitating the security of the grounds arrange. Currently the framework and the host was observed to endeavour to be attacked by others, a substantial number of security vulnerabilities exists in the framework, and there are numerous security vulnerabilities which are hard to keep away from and destroy; Also, virus transmitted through the system seriously influenced the ordinarily running of the grounds organize. Therefore, network architecture and its security are vital issues for any university. In this work, a network infrastructure is proposed on the basis of the practical and experimental requirements. The proposed network infrastructure is realizable with adaptable infrastructure [Das and Behera(2017)].

II. Challenges in Designing Campus Network

School and University are generally furnished with a PC room; there are a few PCs which have open access to the grounds PC organize in this room, understudies also, personnel are normally accessible to utilize these registers to access to web to get data and learn on the web. The goal is to minimize cost based on these elements while delivering service that does not compromise established availability requirements. There may be two primary concerns: availability and cost. These issues are essentially at odds. Any increase in availability must generally be reflected as an increase in cost. [Das and Behera(2017)] However, the absence of bound together administration programming and framework for observing and logging, these PC rooms can't be basic in the administration state. Most rooms have genuine imperfections in enlistment and administration framework, so the web client's character cannot be perceived. It also must scale, offer operational simplicity and flexibility, accommodate new computing trends without much alterations in the original structure and design.
[K C Ramakrishnegowda(2009)] Its very convenient for us to use functions provided by the campus network, but it also has become a quick way to transfer the virus. The virus attack can invade the confidentiality of user and can results in information leakage. Network virus can deteriorate network resources that can decrease network performance rapidly. Hence network security is most vital component in information security, because it is responsible for securing all information passed through networked computers.[MADJE(2017)]
In the grounds organize, assaults, interruption the machines, robbery of another record, the unlawful utilization of the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

system, illicit access to unapproved documents, badgering via mail and different episodes frequently happen, et cetera, demonstrating the clients' security awareness in the grounds organize are exceptionally unremarkable. Presently the internet is flooded with the hacking tools; hackers use network protocols, server and operating system security vulnerabilities and management oversight to illegally access to network resources, deletion of data, damage the system, these attacks caused to the adverse effects of the campus network and the damage to the reputation of the school.[Gursimrat Singh(2013)] A Most difficult issue is the deficiency of assets for the organize development in school or college, restricted assets are fundamentally put resources into arrange gear, efficient contribution for development and administration of system security has not been considered genuinely. In light of absence of the mindfulness in significant colleges, administration organizations are not perfect, administration framework is flawed, administration innovation is not propelled; these elements influence the college's system administration to focus can't take any measures and preventive measures for data security. In the interim, Countries don't have all around created and thorough system security framework, there is no strict execution of gauges for the grounds organize security administration; this is an imperative purpose behind the multiplication of system security 1 I is the normal sorts of system security dangers.

By the investigation over, the grounds arrange security issues is for the most part in the accompanying areas: Password exposure can bring about information spillage. An assortment of database frameworks is running on-line in the grounds organize, for example, showing administration framework, understudy accomplishment administration framework, grounds card administration framework, test bank et cetera. The overall topology is based on tree topology which is conveniently serving the network. The network is based on client-server architecture.[Begum and Nandyal()] The client individual offense or carelessness of wellbeing measures can prompt these database secret key be lost, the information might be illicitly evacuated or repeated, bringing about data divulgence, in genuine condition, may bring about genuine unlawful removal of information. Subsequently, setting secret key is likewise imperative.

Grounds systems can associate with the web with switches; obviously, web clients can appreciate the accommodation of quick and boundless assets of this stage, yet in addition need to face to the danger of an assault. The overall topology is based on tree topology which is conveniently serving the network. The network is based on client-server architecture. There is significant security chance inside the grounds organize, interior clients are generally see more about the arrange structures and utilizations of models than the outside clients, subsequently the interior security dangers are the principle dangers. As discussed earlier, hacking devices overwhelmed in the Internet, programmers utilize arrange conventions, server and working framework security vulnerabilities and administration

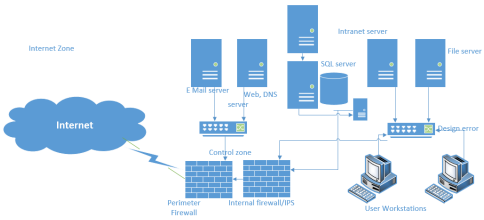


Fig. 1. Simulation results for the network.

oversight to illicitly access to organize assets, erasure of information, harm the framework, these assaults caused to the unfavourable impacts of the grounds system and the harm to the notoriety of the school.

There is a great deal of security vulnerabilities inside authentic working framework, these security vulnerabilities posture numerous genuine dangers, for example, the data security, utilizing of the framework, arrange operations et cetera. Security policy weaknesses can create unforeseen security threats. [Alabady(2009)] Attributable to the absence of attention to copyright, robbery programming, movies and TV assets is utilized as a part of general in the grounds arrange, while the spread of the product takes up a considerable measure of system data transmission; on the other hand it brings a specific system security dangers. Test case, Microsoft's 8.1 working framework organization has few extent of PC applications which are step by step extended in the grounds, get to focus to the grounds arrange have expanded to a great extent, however a large portion of these hubs is the new and a portion of the defensive measures have not been received, this may cause the virus flooding whenever, data misfortune, information debasement, arrange assaults, System crumple and other genuine results. Constraint in refreshing the product; if the clients introduce Microsoft's 8.1 working framework with the pilfered one, the PC framework will leave countless vulnerabilities. Then again, downloading from the Web programming which may has shrouded Trojans virus, indirect accesses and different malevolent code, in this way, numerous frameworks are intrusive and utilized by the aggressor.

A portion of the aggressors and merchants send spam and spread hurtful data utilizing unmanaged grounds server as a travel station, this has conveyed a genuine impact to the organize and expedited incredible impact the school's notoriety. In rundown, the security circumstance of the grounds organize is intense, keeping in mind the end goal to understand these security dangers and vulnerabilities, as indicated by the basic highlights of the grounds system and security issues which the grounds arrange confronted, security arrangements ought to be taken to the grounds system and it ought to be executed as snappy as could reasonably be expected.

III. 3 Campus Network Security Solutions

The campus network safety provides shield for IT system resources against external threats like intruders and malicious code. The network security system includes firewalls, intrusion detection and prevention system (IPS/IDS), VPN protections

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

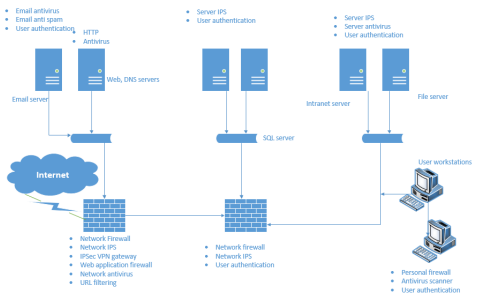


Fig. 2. Simulation results for the network.

and content inspection systems such as anti-virus, anti-spam, anti-malware and URL filtering. The security mechanisms in terms of operating systems database and applications are strengthen by hardware and software solutions. The strong security system built for large networks needs to be designed carefully after organization's risk analysis that fulfils security standards.

A. Security Rules

While designing network safety system, the main IT systems security principles that should be implemented are characterized below

1) Defence-in-depth: IT systems resources safety comprises of several security layers as illustrated in 1 Defence-in-depth is followed by rules given below: Layered protections-number of security layers to amplify each other, what one misses the other catches.

Defence in multiple places network security system is placed at various locations of IT system. The security system of IT network resources depends on variety of safety layers that performs function of safeguard in different ways. Moreover, layers like two network firewalls having similarities must be originating from different vendors.

Meanwhile due to increased complexity of security system along with expensive management the implementation of this rule must be done carefully.

2) Compartmentalization of information: As shown in figure2 given below for different sensitivity levels of IT network, there must be different security zones. In addition to this information hiding is also part of this rule that says IT system will provide only that data which is obligatory for carrying out IT system tasks (as only those servers will be registered in public DNS that are responsible for providing services to the internet).

3) Principle of least privilege: Another principle for IT system subjects is that they must be given with least benefits only required for proper functioning of organization. Same rule will be applied for external users using data and services. "Need-to-know" is extension of this rule which illustrates only that information will be provide to administrators of IT system that is necessary for them to perform their functions and duties.

4) Security zones: Firewall provides control on network traffic flow and it will flow according to command. Firewall includes firewall devices, firewall functions in IPS devices and list of access control for system routers and switches. Firewalls plays important role in securing architecture by categorizing the IT system infrastructure into security zones having effective communication between them. This division rule is comprised of following points:

1) IT system resources of different sensitivity levels should be located in different security zones:
   - Devices and computer systems providing services for external networks (e.g., the Internet) should be located in different zones (De-Militarized Zone - DMZ) than internal network devices and computer systems
   - Strategic IT system resources should be located in dedicated security zones.
   - Devices and computer systems of low trust level such as remote access servers and wireless network access points should be located in dedicated security zones.
2) IT system resources of different types should be located in separate security zones
   - User workstations should be located in different security zones than servers
   - Network and security management systems should be located in dedicated security zones
   - Systems in development stage should be located in different zones than production systems

5) Intrusion prevention: IPS devices are responsible for detecting and blocking penetrations and attacks conducted by intruders and malicious malware applications. They should be installed in the network path between potential threat sources and sensitive IT system resources. When designing IPS systems, attacks through encrypted sessions (e.g., SSL) should also be taken into account. Since the IPS would not able to inspect these encrypted sessions, an effective method would be to decrypt the sessions prior to IPS devices in order to inspect unencrypted packets. An important requirement for intrusion prevention tightness is the proper design of network protections and control rules. For one, internal networks should not have direct access to the Internet so a Trojan sent to a user's workstation through a phishing attack would not allow the intruder to connect to the external network.

IV. Security Implementation

It is compulsory to develop plan based to determine security risk for campus network. For this purpose, number of technologies like firewall technology, PKI technology, virtual private network (VPN) technology, virtual exchange network can be utilized to develop centralized configuration, and to strengthen monitoring management system. In short, formulating system must be made strong and specifications must be introduced to keep security confidential and implement it.

The whole network system must be made secure in order to increase physical security of several equipment's of computer information system. It usually covers three features

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

that are lines security, environmental safety and equipment safety. Moreover, equipment safety covers several domains like anti- theft of devices, anti-electromagnetic information radiation leaks, anti-crash to secure power supply and anti-electromagnetic interference. Whereas environmental safety facilitates the security of system environment like disaster protection along with regional safety. Campus can be categorized into two parts that is internal network and external network. Internal network covers LAN, LAN office automation, library LAN and teaching LAN, whereas external network covers various public servers. According to network security viewpoint it is important to install internet firewall in the border along with implementation of security policy control. The internal network must be made secure by installation of probe for switcher on the main segment. This will work as intrusion detection system which keeps network access secure. As all computer terminals, servers and workstations are based on operating system to continue optimal function, so it is important to make secure operating system. Hence, genuine operating system with server version is appropriate for use in the critical servers and workstations such as database server, Email server, www server, backup server and network management stations, etc.

In the meantime, operating systems are at greater risk of insecurities because of inappropriate design and version, another reason behind security vulnerabilities is lack of knowledge about use of security settings. Without other higher security level commercial operating system for decision, working framework security administration and improvement of upgraded safety efforts is the key variables, furthermore, issues of operating system security may rise from virus threats and hackers penetrated the network to destroy the information. Therefore, anti-virus tools must have ability to secure network from every kind of virus attack through any internet source. In addition to, anti-virus program installation must be capable of removing virus program detected.

Hackers can hack the data for misuse and can make changes in system illegally. It is beneficial for system to install risk assessment programs so that administrators can easily recognize the threats and accept or reject user privileges command accordingly. Moreover, its fruitful to install real-time intrusion detection system IDS because it can trace out user's activity, and continuous check can secure internal staff from intranet destruction. Soon after detection of any harm, the system will immediately inform the administrator, then administrator is authorized to record the relevant test results and can use information to track hacker or intruder, and further can identify security risks to system and develop strategies to secure system's useful information and data from misuse or damage. Mostly student's computer room and campus network is connected to each other, and there is chance that virus can spread as students will definitely use the usb storage device on PC and students may frequently download some product from the Internet on the machine. Hence, to keep campus network safe from virus its important resolve the problem by using server antivirus, stand-alone anti-virus and gateway anti-virus. Administrator can make network more secure by setting password and permissions in the network for every user, it not only saves user password and name but can also facilitate to keep user complete record with confidentiality. Network administrator must keep database of network users and system log in organized manner. To evaluate and audit the security circumstance frequently on the campus network, give careful consideration to the dynamic system security concerns and alter the significant security settings for interruption aversion, and recovery systems after emergency. Various security plans like network access control, the directory security control plan, access control policy, network monitoring, lock control and operation access control can be developed to make campus network security more strong. Another strategy i.e. information encryption strategy is beneficial to prevent unauthorized person from misuse or stealing of valuable data by applying encryption algorithm. There is various encryption programming which can encode messages; files et al. Utilize encryption programming can viably ensure information, documents, secret key, and control the data to transmit securely through the system. Back up and imaging technology can also be used to strengthen campus network. Data integrity can be increased by using backup technology, as data is stored in another place to make a backup, so data can be recovered easily once deleted from anywhere. When the development of campus network is being done, network security needs of a unified security system particular that ought to be combined with actual situation, and afterward the unified security system details ought to be actualized in the entire network. The administration ought to be executed first before the Implementation of network methods wellbeing, the development of the security association system is basic. Experts must be involved in designing and implementation of security plan and experts must coordinate with other departments for the implementation of security process and they should facilitate other departments with guidance to enhance the level of system's security.

The campus network data system can be made secure by taking technical measures, but besides this, management system should be improved and constructed in effective way. The campus network security is always at greater risk, so to expand usefulness for the security administrations, enhance network safety efforts, we should attempt awesome endeavours to fortify the system security administration, build up a secure administration framework and execute well secured administration standards, data arrange security is an exceptionally orderly work, just when the system security administration and the system security innovation are utilized at the same time, at that point they can develop a complete data organize security framework.

V. Conclusion

Recently number of colleges and schools develop their own particular IT systems, and the use of campus network turns out to be increasing generally, for this reason data security is an inescapable factor to guarantee the systems running

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

easily with maximum efficacy. Facilitating campus network with educational environment not only provides strong back up model but also promotes the information technique education. Maintenance of high efficiency of campus network is main problem that need to be resolved. This paper has highlighted various factors which can compromise campus security and give suggestions to sort out security related issues.

REFERENCES

[Ali et al.(2015)Ali, Hossain, and Parvez] Mohammed Nadir Bin Ali, Mohamed Emran Hossain, and Md Masud Parvez. Design and implementation of a secure campus network. *International Journal of Emerging Technology and Advanced Engineering*, 5:370–374, 2015.

[Das and Behera(2017)] Kajaree Das and Rabi Narayan Behera. A survey on machine learning: Concept, algorithms and applications. 2017.

[K C Ramakrishnegowda(2009)] B Shankarappa K C Ramakrishnegowda, U Kannappanavar. Campus network management: Best practice by kuvempu university. 2009.

[MADJE(2017)] UMA MADJE. Design and analysis of network security model. pages 48–50, 2017.

[Gursimrat Singh(2013)] Dr. Amardeep Singh Gursimrat Singh. Campus network security policies: Problems and its solutions. 2013.

[Begum and Nandyal()] Faqarunnisa Begum and Suvarna Nandyal. Implementing different set of network security policies for a well infrastructure campus network.

[Alabady(2009)] Salah Alabady. Design and implementation of a network security model for cooperative network. *Int. Arab J. e-Technol.*, 1(2): 26–36, 2009.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Improving the Data Quality in the Research Information Systems

Otmane Azeroual[abc]

[a] German Center for Higher Education Research
and Science Studies (DZHW), Schützenstraße 6a, Berlin, 10117, Germany
[b] Otto-von-Guericke-University Magdeburg, Department of Computer Science, Institute for Technical and Business Information Systems Database Research Group, P.O. Box 4120; 39106 Magdeburg, Germany
[c] University of Applied Sciences HTW Berlin, Study Program Computational Science and Engineering, Wilhelminenhofstraße 75A, 12459 Berlin, Germany
Azeroual@dzhw.eu

Mohammad Abuosba[c]

[c] University of Applied Sciences HTW Berlin, Study Program Computational Science and Engineering, Wilhelminenhofstraße 75A, 12459 Berlin, Germany
Mohammad.Abuosba@HTW-Berlin.de

Abstract — **In order to introduce an integrated research information system, this will provide scientific institutions with the necessary information on research activities and research results in assured quality. Since data collection, duplication, missing values, incorrect formatting, inconsistencies, etc. can arise in the collection of research data in different research information systems, which can have a wide range of negative effects on data quality, the subject of data quality should be treated with better results. This paper examines the data quality problems in research information systems and presents the new techniques that enable organizations to improve their quality of research data.**

Keywords — **Research Information Systems, CRIS, RIS, Data Quality, Research Data, Data Cleaning, Data Profiling, Science System, Standardization**

I. INTRODUCTION

With the introduction of a research information system, universities and non-university research institutions can provide an up-to-date overview of their research activities, record, process and manage information on their scientific activities, projects and publications, as well as integrate them into their web presence. For scientists, they offer opportunities to collect, categorize and use research information, be it for publication lists, for the preparation of projects, for the reduction of the effort required to produce reports, or for the external presentation of their research and scientific expertise.
The data quality is of great importance. Only correct data can provide resilient, useful results and allow for a profound understanding of the research data of research establishment that are always up-to-date. The completeness, correctness and timeliness of the data are thus essential for successful operational processes. The data errors extend across different areas and weaken the entire research activities of an establishment. Therefore, the aim of this paper is to define and classify the problems of the quality of data that can occur in the research information systems, and then present new techniques that are used to recognize, quantify, resolve data quality problems in research information systems, to improve their Data quality.

II. Research Information System (RIS) - Architecture

A RIS is a central database that can be used to collect, manage and provide information on research activities and research results. The information considered here provides metadata about research activities such as projects, third-party funds, patents, cooperation partners, prices and publications. The RIS architecture usually consists of (see figure 1):

- Data Access Layer
- Application Layer
- Presentation layer

The following figure provides an overview of the individual processes and shows which components belong to which process step:

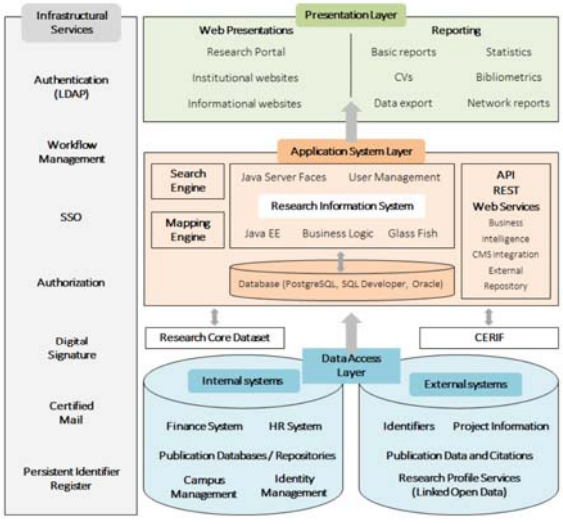


Figure 1: Architecture of RIS

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The **data access layer** contains the internal and external data sources. This level contains, for example, databases from the administration or publication repositories of libraries, identifiers such as ORCID or bibliographic data from the Web of Science or Scopus etc. The **application layer** contains the research information system and its applications, which merge, manage and analyze the data held at the underlying level.
In the **presentation layer**, the target group-specific preparations and representations of the analysis results are depicted for the user. In addition to various possibilities of reporting, you can also fill portals and websites of the establishment here.
Orthogonal to the described layers, there are the Infrastructural Services, the overlapping services for the entire information systems, such as authentication, authorization, single sign on, etc.
Offers for the standardized collection, provision and exchange of research information in RIS are Research Core Dataset (RCD) and the Common European Research Information Format (CERIF) data model maintained by the non-profit organization euroCRIS Version CERIF 2008 1.0 is available. This data model describes the entities as well as their relationship to each other.

III. Problems of data quality

Collecting data in a database system is a standard process. At each facility, personal data, information about their scientific activities, projects and publications are entered and recorded. The processing and management of this data usually needs to be in a good quality, so that users can get better results.
The quality of data is often defined as the suitability of the data for use in certain required usage objectives, which must be error free, complete, correct, up to date and consistent. Requirements can be set by different stakeholders, in the RIS context e.g. especially by users of a RIS, but also by the RIS administrator. Poor quality data includes data errors, typographical errors, missing values, incorrect formatting, contradictions and further more. Such quality issues in RIS need to be analyzed and then remedied by data transformation and data cleansing. The following are the typical quality issues of data in the context of a RIS (see figure 2):

- Missing values (features completeness)
- Incorrect information caused by input, measurement or processing errors (characteristic correctness)
- Duplicates in the dataset (feature redundancy-free)
- Unevenly represented data (feature uniformity)
- Logically contradictory values (consistency characteristic)

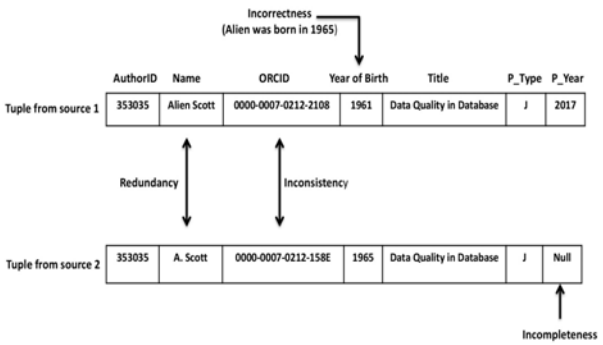


Figure 2: Examples of Data Quality Problems in RIS

IV. Improvement of data quality

Due to the integration of different internal data sources of the establishment and of external sources in research information systems, problems as stated in Chapter 3 have to be overcomes. Now it is important to oppose the causes in this step and to improve the data quality.
The process of identifying and correcting errors and inconsistencies with the aim of increasing the quality of given data sources in RIS, is referred to as data cleansing (or "Data Cleaning") [8].
Data Cleansing includes all necessary activities to clean dirty data (incomplete, incorrect, not up to date, inconsistency, redundant, etc.). The data cleaning process can be roughly structured as follows [3]:

1. Defining and determining the actual problem
2. Find and identify faulty instances
3. Correction of the found errors

Data cleansing uses a variety of specialized methods and technologies within the data cleansing process. [9] subdivides them into the following phases (see figure 3):

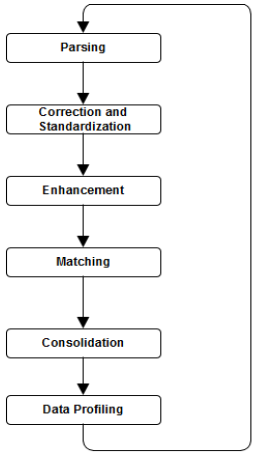


Figure 3: Data Cleaning Process

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Parsing is the first critical component of data cleansing, helping the user to understand and transform the attributes more accurately. Individual data elements are referenced according to the metadata. This process locates, identifies and isolates individual data elements. For this process, e.g. for names, addresses, zip code and city. The parser Profiler analyzes the attributes and generates a list of tokens from them, and with these tokens the input data can be analyzed to create application-specific rules. The biggest problems here are different field formats that have to be detected.

Correction & Standardization is further necessary to check the parsed data for correctness, and then correct it afterwards. Standardization is the prerequisite for successful matching and there is no way around using a second reliable data source. For address data, a postal validation is recommended.

Enhancement or data enrichment is the process that expands existing data with data from other sources. Here additional data is added to close existing information gaps. Typical enrichment values are demographic, geographic or address information.

Matching: There are different types of matching: for reduplicating, matching to different datasets, consolidating or grouping. The adaptation enables the recognition of the same data. For example, redundancies can be detected and condensed for further information.

Consolidation (merge): Matching data items with contexts are recognized by bringing them together.

Data profiling is what the data is to be understood, i. E. capture technical structures, analyze them for the purpose of detecting data quality issues, and identify business inconsistencies. Profilers do not describe business rules and do not make any changes. They are only for analyzing the data. Profilers are often used at the beginning of a data analysis, but can also help to better illustrate the results of the analysis.

All of these steps are essential for achieving and maintaining maximum data quality in research information systems. Errors in the integration of multiple data sources in a RIS are eliminated by the clearing up.

The following Table 1 illustrates an example of identifying records with faulty names in a publication list to show how data cleansing processes (clearing up, standardization, enrichment, matching, and merging) can improve the quality of data sources.

The clearing up process adds missing entries, and completed fields are automatically adjusted to a specific format according to set rules.

Original Data before clearing up

Data Source				
Author ID	**Name**	**ORCID**	**Birth Date**	**Address**
353035	Alien Scott	0000-0007-0212-2108	10/25/1965	145 F. Concord Street, Orlando, 32801
353035	Dr. Alien Scott	0000-0000-0000-0000	25.10.1965	Concord Street, 32801 145F
353035	Alien William Scott	0000-0007-0212-2108	652510	25 Concord 32801 Street
353036	A. Scott		11/25/56	12 Ford Ave 32801
353036	Scott Alien	0000-0007-0212-2108	25.11.1965	
	Alien Scoth	0000000702122108	25.10.1956	Street C., 32801 145F. US
410003	Olivia Svenson	0450-1254-3598-F156	6-2-1983	745-7801 P.B. Las Vegas 29502
410003	Svenson Olivia	045012543598F156	1983	7801 P.B. Las Vegas 29502

Data after clearing up

In this example, the missing zip code is determined based on the addresses and added as a separate field. Enrichment rounds off the content by comparing the information against external content, such as demographic and geographic factors, and dynamically expanding and optimizing it with attributes.

Cleansed Data					
Author ID	**First**	**Last**	**ORCID**	**Birth Date**	**Address**
353035	Alien	Scott	0000-0007-0212-2108	1965-10-25	32801; FL; Orlando; 145 F. Concord Street
353035	Alien	Scott		1965-10-25	32801; FL; Orlando; 145 F. Concord Street
353035	Alien	Scott	0000-0007-0212-2108		32801; FL; Orlando; 145 F. Concord Street
353036	Alien	Scott		1965-11-25	32801; FL; Orlando; 12 Ford Ave
353036	Alien	Scott	0000-0007-0212-2108	1956-11-25	
	Alien	Scott		1965-10-25	32801; FL; Orlando; 145 F. Concord Street
410003	Olivia	Svenson	0450-1254-3598-F156	1971-02-06	29502; NV; Las Vegas; 745-7801 PO Box
410003	Olivia	Svenson	0450-1254-3598-F156		29502; NV; Las Vegas; 745-7801 PO Box

Data before enrichment

Cleansed Data					
Author ID	**First**	**Last**	**ORCID**	**Birth Date**	**Address**
353035	Alien	Scott	0000-0007-0212-2108	1965-10-25	FL; Orlando; 145 F. Concord Street
353035	Alien	Scott		1965-10-25	FL; Orlando; 145 F. Concord Street
353035	Alien	Scott	0000-0007-0212-2108		FL; Orlando; 145 F. Concord Street
353036	Alien	Scott		1965-11-25	FL; Orlando; 12 Ford Ave
353036	Alien	Scott	0000-0007-0212-2108	1956-11-25	
	Alien	Scott		1965-10-25	FL; Orlando; 145 F. Concord Street
410003	Olivia	Svenson	0450-1254-3598-F156	1971-02-06	NV; Las Vegas; 745-7801 PO Box
410003	Olivia	Svenson	0450-1254-3598-F156		NV; Las Vegas; 745-7801 PO Box

Data after enrichment

The example shows how the reconciliation and merge process runs. Merging and matching promote consistency because related entries within a system or across systems can be automatically recognized and then linked, tuned, or merged.

Enriched Data						
Author ID	**First**	**Last**	**ORCID**	**Birth Date**	**Address**	**Zip**
353035	Alien	Scott	0000-0007-0212-2108	1965-10-25	FL; Orlando; 145 F. Concord Street	32801
353035	Alien	Scott		1965-10-25	FL; Orlando; 145 F. Concord Street	32801
353035	Alien	Scott	0000-0007-0212-2108		FL; Orlando; 145 F. Concord Street	32801
353036	Alien	Scott		1965-11-25	FL; Orlando; 12 Ford Ave	32801
353036	Alien	Scott	0000-0007-0212-2108	1956-11-25		
	Alien	Scott		1965-10-25	FL; Orlando; 145 F. Concord Street	32801
410003	Olivia	Svenson	0450-1254-3598-F156	1971-02-06	NV; Las Vegas; 745-7801 PO Box	29502
410003	Olivia	Svenson	0450-1254-3598-F156		NV; Las Vegas; 745-7801 PO Box	29502

Matching

This example finds related entries for Alien Scott and Olivia Svenson. Despite the similarities between the datasets, not all information is redundant. The adjustment functions evaluate the data in the individual records in detail and determine which ones are redundant and which ones are independent.

Cleansed Data					
Author ID	**First**	**Last**	**ORCID**	**Birth Date**	**Address**
353035	Alien	Scott	0000-0007-0212-2108	1965-10-25	32801; FL; Orlando; 145 F. Concord Street
353035	Alien	Scott		1965-10-25	32801; FL; Orlando; 145 F. Concord Street
353035	Alien	Scott	0000-0007-0212-2108		32801; FL; Orlando; 145 F. Concord Street
353036	Alien	Scott		1965-11-25	32801; FL; Orlando; 12 Ford Ave
353036	Alien	Scott	0000-0007-0212-2108	1956-11-25	
	Alien	Scott		1965-10-25	32801; FL; Orlando; 145 F. Concord Street
410003	Olivia	Svenson	0450-1254-3598-F156	1971-02-06	29502; NV; Las Vegas; 745-7801 PO Box
410003	Olivia	Svenson	0450-1254-3598-F156		29502; NV; Las Vegas; 745-7801 PO Box

Consolidation

The merge makes the reconciled data a comprehensive data set. In this example, the duplicate entries for Alien Scott are merged into a complete record containing all the information.

Cleansed Data					
Author ID	**First**	**Last**	**ORCID**	**Birth Date**	**Address**
353035	Alien	Scott	0000-0007-0212-2108	1965-10-25	32801; FL; Orlando; 145 F. Concord Street
353035	Alien	Scott		1965-10-25	32801; FL; Orlando; 145 F. Concord Street
353035	Alien	Scott	0000-0007-0212-2108		32801; FL; Orlando; 145 F. Concord Street
410003	Olivia	Svenson	0450-1254-3598-F156	1971-02-06	29502; NV; Las Vegas; 745-7801 PO Box
410003	Olivia	Svenson	0450-1254-3598-F156		29502; NV; Las Vegas; 745-7801 PO Box

Golden Record					
Author ID	**First**	**Last**	**ORCID**	**Birth Date**	**Address**
353035	Alien	Scott	0000-0007-0212-2108	1965-10-25	32801; FL; Orlando; 145 F. Concord Street
410003	Olivia	Svenson	0450-1254-3598-F156	1971-02-06	29502; NV; Las Vegas; 745-7801 PO Box

For the frontend of the RIS could be checked with the profiling process, when it comes to the production of statistics,

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

reports on research data. Profiling makes it easy to assess the overall condition of the data, to identify, prioritize and correct errors, and to remedy the cause of quality issues. Once a profile is created, a facility can respond to quality issues by continually monitoring profile-related parameters.

V. Discussion

The clearing up of data errors in RIS is one of the possible ways to improve existing data quality. Following the defined data cleansing processes, the following developed use cases could be identified in the target system and should serve as a model to show how to detect, quantify, correct and improve them in the case of data errors in research information systems in the establishment.

The following figure introduces the just-mentioned use cases for improving data quality in RIS.

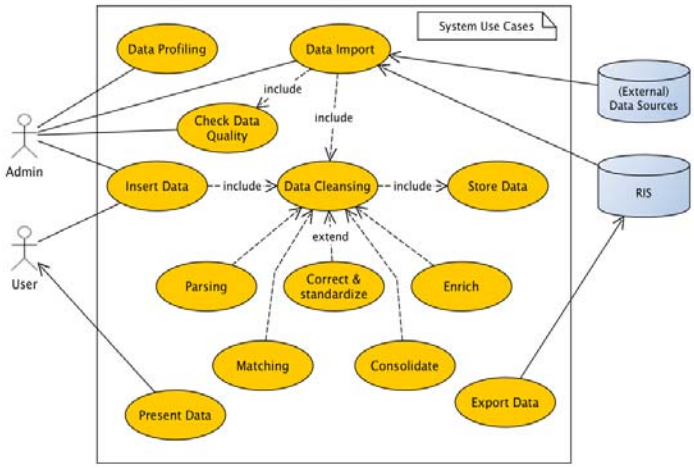


Figure 4: Use Case for Improving Data Quality in the RIS

The meta process flow can be viewed as shown in the following figure:

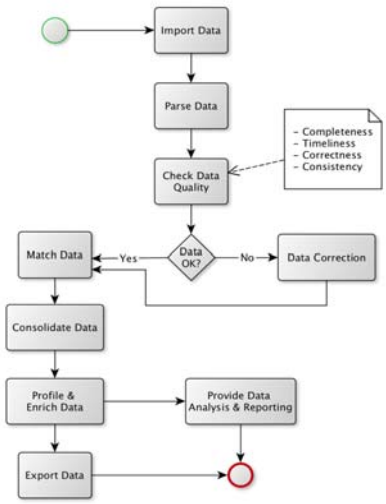


Figure 5: Main Workflow of the Process

The techniques presented by Data Cleaning help establishments that overcome problems. With these steps every establishment can successfully enforce their data quality.

VI. Conclusion and Outlook

In this paper, the question was addressed, which quality problems can occur in research information systems and how to fix and improve them with new techniques or methods, such as Data Cleansing.
As a result, it were been shown that the improvement of the data quality can be performed at different stages of the data cleansing process in any research information system and that high data quality can be obtained from universities and research institutions to operate e.g. research information systems successfully. In this respect, the review and improvement of data quality are always targeted. The illustrated concept can be used as a basis for the using facilities. It offers a suitable procedure and a use case, on the one hand to be able to better evaluate the data quality in RIS, to be able to prioritize problems better and to prevent them in the future as soon as they occur. On the other hand, these data errors must be corrected and improved with Data Cleansing. It says: "The sooner quality defects are detected and remedied, the better." Already in the acquisition phase, the author himself or a downstream control authority can correct software errors, such as typing errors, missing values, incorrect formatting, contradictions, etc. To support the universities and all research institutions in the implementation of the data cleansing process, there are numerous tools. With these tools, all facilities can significantly increase the completeness, correctness, timeliness, and consistency of their key data, and they can successfully implement and enforce formal data quality policies.
Data-Cleansing tools are primarily commercial and available for both small application contexts and large data integration application suites. In recent years, a market for data cleansing is developing as a service.

In future work, possible expert interviews and / or quantitative surveys are planned with the research facilities and universities in order to find out how high the data quality is in their research information system and what measures to improve the quality of data are applied to research information.

REFERENCES

[1] Batini, C. and Scannapieco, M. Data Quality - Concepts, Methodologies and Techniques; Springer-Verlag, Heidelberg, 2006.

[2] Gebauer, M. and Windheuser, U. Structured Data Analysis, Profiling and business rules, Springer Fachmedien Wiesbaden 2015.

[3] Helmis, S. and Hollmann, R. Web-based data integration, approaches to measuring and securing the quality of information in heterogeneous data sets using a fully web-based tool, 1. edition © Vieweg+Teubner | GWV Fachverlage GmbH, Wiesbaden 2009.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[4] Hildebrand, K. Gebauer, M. Hinrichs, H. and Mielke, M. Data and information quality. On the way to the information excellence, 3rd, extended edition, Springer Fachmedien Wiesbaden 2015.

[5] Martin, M. Measuring and Improving Data Quality. Part II: Measuring Data Quality. In: NAHSS Outlook. Ausgabe 5, April 2005.

[6] Müller, H. and Freytag, J.-C. Problems, Methods and Challenges in Comprehensive Data Cleansing, Technical Report HUB-IB- 164; Humboldt Universität, Berlin, 2003.

[7] Naumann, F. and Rolker, C. Assessment Methods for information Quality Criteria. In: Proceedings of the MIT Conference on Information Quality, pp. 148-162.

[8] Naumann, F. and Leser, U. Informationsintegration: Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen. Dpunkt Verlag, 1. Edition, Oktober 2007.

[9] Rahm, E. and Do, H.H. Data cleaning: Problems and current approaches. IEEE Bulletin of the Technical Committee on Data Engineering, 23(4), 2000.

AUTHORS PROFILE

Otmane Azeroual, is working as a researcher in the project "Helpdesk for the Introduction of the Research Core Dataset" at the Department 2 Research System and Science Dynamics, German Center for Higher Education Research and Science Studies (DZHW). After studying Business Information Systems at the University of Applied Sciences Berlin (HTW), he began his PhD in Computer Science at the Institute for Technical and Business Information Systems (ITI), Database Research Group of the Otto-von-Guericke-University Magdeburg and at the University of Applied Sciences (HTW) Berlin in the Department of Study Program Computational Science and Engineering. His research interest is in the area of Research Information Systems, Quality Management, Database Systems, IT-Security, Cloud Data Management, Data Science, Project Management and Industry 4.0.

Prof Dr.-Ing. Mohammad Abuosba, is a Professor at the Department of Study Program Computational Science and Engineering, University of Applied Sciences (HTW) Berlin. His Research areas are Engineering, IT Systems Engineering (focus on Database Systems, Product Data Management), Modeling, Quality Management and Project Management.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

One hop forwarding technique for QOS routing with an Improved Fault tolerant Clustering Mechanism for Multimedia Sensor Networks

[1] R. Guru, [2] Dr. Siddaraju, [3] Ananda babu
[1]Research scholar, [2]Professor, [3]Assistant Professor
[1]Sri Jayachamarajendra College of Engineering, Mysuru, India.
[2]Dr. Ambedkar Institute of Technology, Bangalore, India
[3]Kalpataru Institute of Engineering, Tiptur, India.
[1]guruirg@sjce.ac.in, [2] siddaraju_b@yahoo.co.in, [3] babu.tiptur@gmail.com

Abstract: *Wireless Sensor Network (WSN) is a wireless network consists of large number of sensor nodes. Routing protocols are in charge of discovering and maintaining the routes in the network. Finding the shortest path is one of challenging factor which greatly enhances the performance of the network. The first section of this paper focuses on finding best possible next hop to route the packets to reach the sink node using TPGF protocol. Next section of this paper concentrate on multimedia applications where they have taken an essential role in our daily lives, and their usage is growing day by day. The multimedia sensors have the ability to capture video, image, audio, and scalar sensor data and deliver the multimedia content through sensors network. However, due to the reason of their location and size of data they transmit for multimedia may cause the system failure in many cases. To overcome this paper presents an Improved Fault tolerant distributed clustering mechanism (IFCMM). This mechanism use run time for the recovery of sensor nodes due to the failure of head of cluster nodes. Experimental results show an improvement in the performance of the algorithm with respect to various metrics of QOS.*

Keywords: Geographic routing, Transmission of multipath, Realistic conditions of sensor.

I. INTRODUCTION

Effectively transmitting sight and sound streams in remote interactive media sensor systems (WMSNs) is a noteworthy testing issue, because of the constrained transmission transfer speed and control asset of sensor hubs. Three late reviews [1–3] on interactive media correspondence in wireless multimedia sensor networks which demonstrates that present conventions in both interactive media and sensor systems fields that are not appropriate for both sight and sound correspondence in WMSNs, in light of the fact that they don't have enough thought on the attributes of sight and sound gushing information also, characteristic obliges of sensor systems in the meantime.

These three overviews additionally expounded that there is no arrangement concentrating on tending to the steering issue of sight and sound spilling in geographic WMSNs.

1.1 Transmission of Multipath: Packets of sight and sound spilling information for the most part are vast in size and the transmission prerequisites can be a few times higher than the most extreme transmission limit (data transfer capacity) of sensor hubs. This requires that multipath transmission ought to be utilized to increment transmission execution in WSNs [4].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

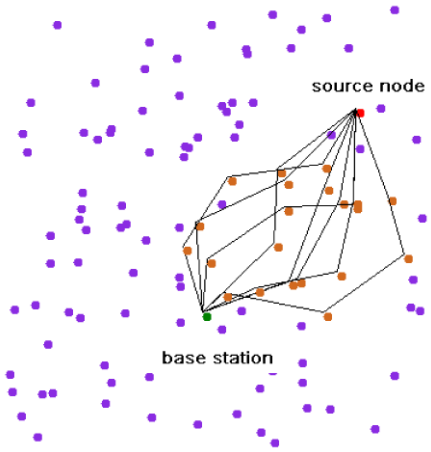


Figure 1: Dynamic Hole: Constructed by a group of sensor nodes present in the present routing paths

1.2 Bypassing the Hole: This includes how the openings of dynamic may happen due to sensor hubs which are little in range over-burden because of the sight and sound transmission, Figure 1 describes how proficiently bypassing these element openings is an important part of transmission in WSNs.

1.3 Shortest way transmission: Multimedia applications by and large have a postpone imperative, which requires that the sight and sound spilling in WSNs ought to dependably utilize the most limited steering way, which has the base end-to-end transmission delay.

Media transmission in WMSNs requires another directing calculation that can bolster all these three necessities in the meantime. This paper proposes another Two-Phase geographic Greedy Forwarding (TPGF) directing calculation for investigating one or numerous most limited (close briefest) hole bypassing transmission ways in WMSNs. The principal period of TPGF is in charge of investigating the conceivable directing way. The second period of TPGF is in charge of improving the discovered steering way with minimal number of bounces. TPGF can be executed over and again to discover various on-request node disjoint steering ways. TPGF has the accompanying essential elements that make it be not quite the same as existing geographic steering calculations [5–8].

TPGF is an immaculate geographic voracious sending directing calculation. It does exclude the face steering idea, e.g., right/left hand guidelines and tally/clockwise edges, which is not the same as many existing geographic sending directing calculations, e.g., GPSR [5].

TPGF does not require the calculation and protection of the planar diagram in WMSNs. This point permits more connections to be accessible for TPGF to investigate more node disjoint steering ways, since utilizing the planarization calculations really restrains the useable connections for investigating conceivable steering ways.

TPGF does not have the outstanding Local Minimum Issue [5], which is characterized as "a sensor hub finds no next-bounce hub that is nearer to the base station than itself". Explore work in this paper has made both hypothetical also, viable commitments to comprehend the geographic directing in WMSNs. The hypothetical commitments are: It is demonstrated that: there exists a geographic covetous sending directing calculation (TPGF) that can ensure bundle conveyance (bypassing openings) in any 2D/3D sensor systems without utilizing the face directing strategy, when sensor hubs just think about their 1-bounce neighbor hubs.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

It is demonstrated that: there exists a geographic avaricious sending directing calculation (TPGF) that can locate the most limited directing way (or close most limited steering way when openings exist) for limiting the end-to-end transmission delay, at the point when the gaps data is not recognized ahead of time. The handy commitments in this paper are as taking after four viewpoints:

- **Key curiosity:** To the best of our insight, TPGF is the to start with unadulterated geographic avaricious sending steering calculation that spotlights on supporting interactive media spilling in WMSNs, which bolsters the accompanying three elements at the same time.
- **Supporting multipath transmission:** TPGF can discover one directing way per execution and can be executed over and again to discover more on-request hub disjoint directing ways.
- **Supporting gap bypassing:** TPGF gives a superior arrangement for gap bypassing in both 2D and 3D sensor systems than other related research work.
- **Supporting most brief way transmission:** TPGF can discover the most brief steering way (or close most limited directing way when openings exist) for limiting the end-to-end transmission delay.

TPGF controlling computation can make a tremendous effect on both adaptable sight and sound and remote sensor frameworks (WSNs) research gatherings.

Remote sensor systems have increased tremendous consideration for their extensive variety of utilizations, for example, natural checking, military reconnaissance, human services, and debacle administration [1]. One of the real imperatives of WSNs is the constrained and by and large indispensable power wellsprings of the sensor hubs. Undoubtedly, in various applications, it is impossible to supplant the sensor center points as they work under savage condition. In this way, diminishing imperativeness usage of the sensor centers is considered as the most essential test for long run operation of WSNs. Expansive looks at have been finished in arranging essentialness saving traditions which consolidate low-control radio correspondence hardware, imperativeness careful MAC traditions, et cetera.

In any case, vitality productive grouping and steering calculations [2, 3] are the most two promising territories that have been contemplated widely for WSNs. In a bunch based WSN (allude Figure . 1), the sensor hubs are composed into unmistakable gatherings, called groups. Each gathering has a pioneer, called group head (CH) and every sensor hub has a place with one and just a single bunch. Bunching WSN has following points of interest. (1) It empowers information conglomeration at group make a beeline for dispose of the repetitive and uncorrelated information, subsequently diminishing vitality utilization of the sensor hubs. (2) Routing can be all the more effectively oversaw on the grounds that lone CHs need to keep up the neighborhood course setup of different CHs and hence requiring little directing data. This thusly enhances the versatility of the system essentially. (3) It likewise moderates correspondence transmission capacity as the sensor hubs speak with their CHs just and in this way stay away from the trading of repetitive messages among themselves. Be that as it may, in grouping approach, a CH bears some additional work stack, i.e., getting detected information sent by part sensor hubs, information total and information scattering to the BS.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

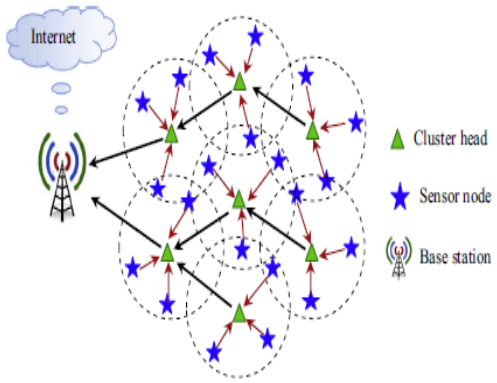


Figure 2: Model of wireless sensor network

Additionally, in numerous WSNs, the CHs are typically chosen among the ordinary sensor hubs, which can bite the dust rapidly as they devour more vitality because of such additional work stack. In this specific situation, numerous analysts [4– 10] have proposed the utilization of some exceptional hub called entryways or transfer hubs that are provisioned with additional vitality. These passages are dealt with as the bunch heads (CHs) which are in charge of a similar usefulness of the CHs. Sadly the entryways are likewise battery worked and thus control compelled.

In other grouping calculations, the bunching is framed by picking chose number of group heads and changing the CHs on each cycle. Be that as it may, the calculation does not consider if the chose CH comes up short amid its cycle. This may prompt disappointment of parcels because of system soften up the framework. Subsequently, the proposed framework thinks about disappointment case.

Intellectual nodes are most promising because of innovation in the WSNs [1]. It bolsters extremely introverted correspondence worldview suits remote sensor organizes that convey well proportioned movement because of their sudden behavior. These intellectual sensor nodes can help defeat the issues of failure of sensor nodes and these nodes will exorbitant dispute in WSNs that emerge because of the sending of a few sensors associated through radio connections. In correlation, ordinary WSNs utilize settled range allotment approach and their execution is restricted because of low preparing and correspondence energy of sensor hubs which are ordinarily asset compelled. In any case, WSNs work over unlicensed groups which are getting to be plainly immersed because of a huge development in the quantity of remote applications utilizing similar groups. Subsequently, the test is to proficiently use the range and this test can be tended to by utilizing psychological radio innovation for WSNs.

This paper also considers video perception based on different condition for blended sensor nodes in the network. Scholarly the framework this sensor nodes include the circumstance, which intuitive media sensor that centers outfitted with cameras is related with remote associations in exceptionally designated way. No less than one of these center points is dynamic meanwhile. They endlessly screen the earth and transmit the got accounts to the sink center point. These center points take a shot at batteries and accordingly imperativeness capability is a fundamental issue for this structure. Likewise, for remote transmission of video the range should be utilized successfully as nowadays run in urban regions is an uncommon resource. The approved range gatherings (3G, 4G, et cetera.) are exceedingly utilized and the unlicensed range gatherings, for instance, ISM bunches are winding up perceptibly exceptionally drenched. That is the reason we propose an imperativeness profitable approach in light of subjective radio for a successful utilization of the range.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The key issues and challenges of MWSN are high transmission limit ask for, high essentialness use, nature of organization (QoS) provisioning, data taking care of, and compacting frameworks, and cross-layer design:

- MWSN requires high transmission limit with a particular true objective to pass on media substance, for instance, a video stream, sound stream or pictures.

Multimedia transmission requires certain QoS guarantees. Regardless, QoS provisioning is to a great degree testing in WSNs as radio associations can have variable breaking point and deferral.

Energy capability is fundamental for MWSN as sensor centers frequently have low essentialness resources.

CR innovation can answer the above prerequisites of MWSNs. CR can give additional data transmission and enhance the nature of administration. Be that as it may, such intellectual radio methodologies, particular to MWSNs, should be outlined which concentrate on vitality productive correspondence to abuse transmission power and range qualities versus execution and unwavering quality exchange off. Moreover, minimal effort calculations should be intended for range detecting and dynamic range use. A large portion of the works in the writing concentrating on CRSNs are identified with just range detecting [20]-[22]. A couple of works concentrate on sight and sound transmission over intellectual radio systems [23]-[26], however they don't consider the WSN condition and the related limitations. This paper presents CR technology for MWSNs and the key contributions of this paper are:

- The spectrum aware based clustering mechanism for CRWSN: The clustering mechanism is aware of spectrum and takes it considers the accepted state of channels.
- This paper also includes head election mechanism for energy aware cluster: The proposed mechanism also considers the residual energy of the candidate cluster head
  And energy consumed by the elected node.

II. RELATED WORKS

Some present works related to CR [1] [2], focus on go identifying, dynamic range get to, MAC traditions [3], controlling traditions and QoS [4]. Above is just a reviewing of the composition and various diverse works exists for CR for improvised frameworks and cell frameworks since it is particularly packed in composing. In any case, the examination on applying CR to WSNs is still in its underlying stage. The examination game plans proposed for all around helpful CR frameworks can't be particularly associated with CRSNs by virtue of the remarkable components of WSNs, for instance, the confined resources and imperativeness restrictions. Thusly, more a la mode CR plans are required that should be changed for WSNs to speak to resource and essentialness goals. If we especially focus on MWSNs by then there are various essentialness beneficial techniques, as assembled in a diagram on MWSN [5] [6], however mental radio isn't considered is those procedures. In particular, using CR with MWSN incorporates a tradeoff between go availability, QoS, essentialness usage and resource capability.

A couple of works related to blended media spilling in CRSNs are delineated in the going with content. The going with works conveys issues related to different layers, for instance, transport, framework, MAC and PHY layers. Bicen et al. [7] discuss the primary arrangement challenges for sight and sound and deferral delicate data transport in CRSN. They research particular transport layer traditions from the point of view of sight and sound transport in

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

CRSNs. In any case, unique issues like packing and guiding are not taken care of. Remembering the ultimate objective to address the issue of high information exchange limit solicitations and QoS necessities, Rehmani et al. [8] proposed channel holding for CRSNs. The thinking is to bond the flanking channels when appeared differently in relation to channel add up to which can add up to non-neighboring channels, yet is more personality boggling for WSNs. They proposed the channel holding thought and called consideration an excessive number of issues and troubles. In any case, the computations and traditions to perform perfect channel holding were not viewed as and were left for future work.

Liang et al. [9] thought about the QoS execution in CRSNs. They propose a need based framework in MAC layer. The steady development is arranged in guaranteed plan openings and the best effort action is served in strife get to period. They mull over the execution using investigative models and exhibit that appealing execution can be expert for consistent development. An expansion of this examination is presented in [10]. In any case, they don't consider the imperativeness use and simply the case related to one gathering of sensor centers are inspected. Multi-gatherings, batching instruments and bury cluster correspondences are not considered.

Gathering is considered by Fard et al. [11], who propose a QoS MAC tradition for remote sight and sound sensor frameworks using multi-channel approach. The remote development is requested into different classes and assorted channels are adaptively named to different action. This approach upgrades the throughput and delays 4 displays while being imperativeness capable. Regardless, PU development isn't contemplated and each cluster head is acknowledged to have various radio interfaces, which isn't useful. Wen et al. [12] propose a QoS coordinating tradition for extraordinarily named CRNs by considering parameters, for instance, open time, repeat gatherings, transmission run, botch rate, basic customer (PU) impedance rate and transmission run. Nevertheless, they simply focus on end-to-end delay.

Zhou et al. [13] propose a scattered booking figuring for video spouting over multi-channel multi-radio and multi-bounce remote frameworks. This approach makes appropriated booking designs with the objective of constraining the video mutilation, considering conventionality. The booking is characterized as a bended change issue, and an answer is proposed by together considering channel errand, rate circulation, and controlling. The work considers the tradeoff between constraining video twisting of each stream as opposed to upgrading overall execution by restricting framework blockage. Nevertheless, they don't consider subjective radio. They acknowledge all channels are open and dynamic range get to considering PU development isn't considered. What's more, observe that imperativeness organization isn't considered in the above works.

Low et al. [24] have proposed a figuring which uses a bipartite chart of the sensor center points and the entryways for finding a most outrageous planning of naming a sensor center to a CH. The figuring has space plan savvy multifaceted nature of O (mn2) for n sensor center points and m CHs. This is high for an immense scale WSN. It in like manner requires building a BFS tree for an individual sensor center point which takes a liberal measure of memory space. In [26] creators have proposed a figuring with O (nlogn) time, which is a change over [6]. Regardless, both the figuring's having not considered the essentialness use issue. In [10], an Energy Efficient Load-Balanced Clustering Algorithm (EELBCA) has been proposed with O (nlogn) time. EELBCA watches out for imperativeness adequacy and likewise stack changing. Low-vitality versatile bunching chain of importance

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

(LEACH) [29] is a well known procedure that structures groups by utilizing a conveyed calculation. It progressively pivots the work heap of the CH among the sensor hubs, which is valuable for stack adjusting. Be that as it may, the primary burden of this approach is that a hub with low vitality might be chosen as a CH which may pass on rapidly. In MHRM (least bounce directing model) [30], each hand-off hub finds a way to the BS such that the jump check is limited. Hence, much measure of vitality is spent to transmit information as it chooses farthest transfer hub in its correspondence go. In [26], creators have built up a disseminated blame tolerant system called CMATO (Cluster-Member-based blame Tolerant component) for WSNs. The creators use the catching methods to recognize the blame. At whatever point a CH falls flat, group individuals identify it and re-select another CH to shape another bunch.

III. PROPOSED ALGORITHM

This paper presents the Two-Phase Geographic Greedy Forwarding (TPGF) routing algorithm for which explores one or more multiple shortest hole by transmitting the paths in wireless multimedia sensor networks, the paper introduces clustering algorithm with energy efficient fault tolerant mechanism. Protocol uses Greedy forwarding technique. A forwarding node finds its next hop node which is closer to the base station using shortest distance vector path. This resolves finding shortest route. Shortest path always leads to depletion of energy of nodes falling in the path. This accounts for creation of dynamic hole. To overcome this, step back rule is introduced. Alternative path is searched through reverse path to ensure the packet will not be dropped due to dynamic hole formation.

- TPGF mechanism is used to convert the WSN to Distance based Search Tree (DST) using the routing algorithms presented in GPSR [5].
- While converting the nodes into TPGF_DST, all neighbor nodes will be added to the search tree.
- Path circle may be formed if two or more nodes within the path are neighbor nodes of the sensor node fall in the path.
- TPGF eliminates the path circle problem by implementing label based path optimization. Path optimization eliminates unnecessary nodes which are part of the path formed.
- Source node establishes a path to destination node at beginning. The response message is received all through from destination node with label number of each node that participated. If any node finds two of its neighbors involved in path formation, then the farthest node from base station will be eliminated from path. This improves the performance of the network.

This paper also presents an **I**mproved **F**ault **T**olerant **C**lustering **M**echanism for **M**ultimedia **T**ransmission of Wireless Sensor Networks **(IFCMM** WSNs). This mechanism described in two different ways:

3.1 Clustering system

- Network is divided into three types of nodes: CH, non-CH and gateways
- Gateways are selectively chosen based on their location
- Gateways play major role while transmitting packets

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

- In case of CH failover, nonCH nodes within the network will apply fault tolerance mechanism. They initiate a broadcast request to know neighbor CH information. Nearby gateway upon receiving the request will initiate them as CH and respond back to the requested nonCH node.
- Gateways broadcasts its change of role to all nearby nodes and nodes will update in their memory

3.2 Multimedia

- All nodes are operated in multi channels. And nodes are divided into Primary User (PU) and Secondary User(SU)
- PUs will start their transmission by default
- SUs will look for channel availability by sensing the channel
- Intra-cluster communication will be achieved by TDMA algorithm
- Inter-cluster communication will be achieved by CDMA algorithm by SUs

3.3 Distributed clustering algorithm

There are many clustering protocols developed so far. This paper extends the clustering to a next level by introducing gateways. Gateways are special nodes deployed selectively within the network. These play a substitute role when CH node dies. When no response is received by nonCH node, issues broadcast message. Gateways nearby hearing the message will initiate the role of CH. Figure 3.1 explains distributed fault tolerant algorithm.

Algorithm: Clustering Fault Tolerance

Step 1: $ComCH(S_i) = \{\}$;

Step 2: send data to CH, wait till timeout

If (S_i doesn't receive a HELLO message from CH) then

- Si becomes member of *UnCOset*

End if

Step 3: Si broadcast message

Step 4: Nearby G_k receives, and responds to Si

Step 5: Si receives pkt

- $ComCH(S_i) = ComCH(S_i) + G_k$
- Si becomes member of *COset*

Figure 3.1: Distributed Fault Tolerant Clustering Algorithm

VI. Performance Analysis

In all the experiments, we have randomly deployed the nodes within the network. In the first phase, the protocol is implemented without optimizing the transmission path. Fault tolerance modal is implemented to evaluate the performance of the protocol. Delivery of the packets is calculated against distinct number of nodes taken into consideration. Results are compared with GPSR protocol. Proposed protocol shows a good result with respect to GPSR in terms of packet delivery ratio as shown in the figure 4.1 and 4.2 with fault tolerance value of 10% and 20% respectively.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

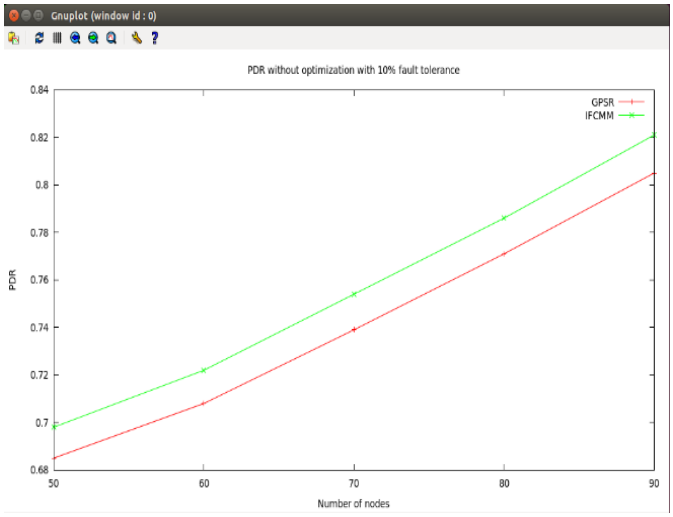


Figure 4.1: PDR without path optimization F=10%

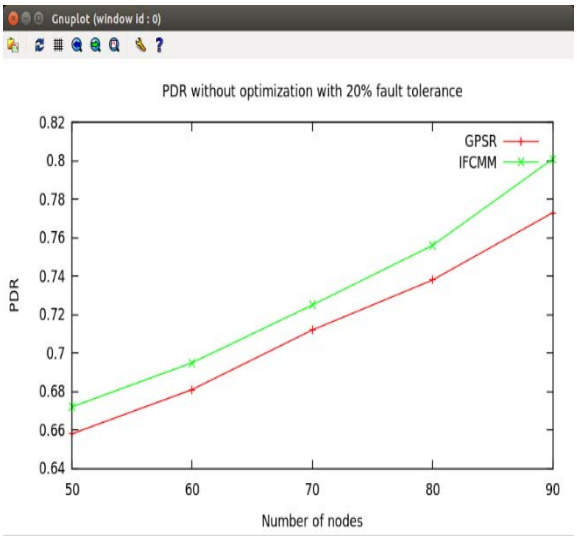


Figure 4.2: PDR without path optimization F=20%

Also, the results are calculated with path optimization. Figure 4.3 and 4.4 shows packet delivery ratio results with fault tolerance of 10% and 20% respectively. In GPSR, the routing is just to the next hop of the node from which the packet is being sent. If the receiving node doesn't have next forwarding node, then the packet will be lost. And this causes the degraded performance. Whereas IFCMM uses greedy forwarding with step-back technique to find a different route and ensure a good result in delivering the packets.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

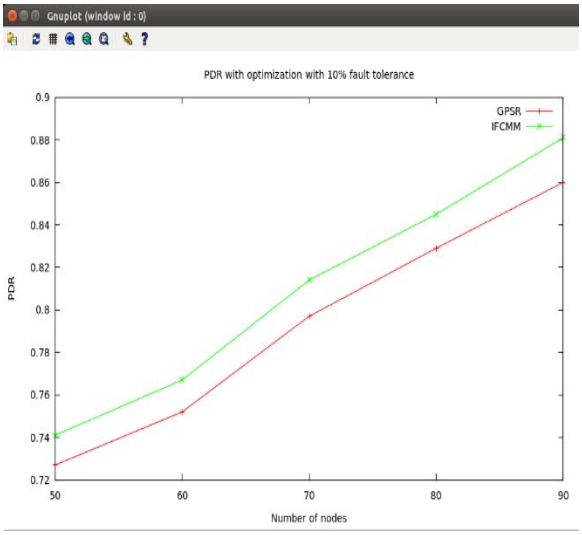


Figure 4.3: PDR with path optimization F=10%

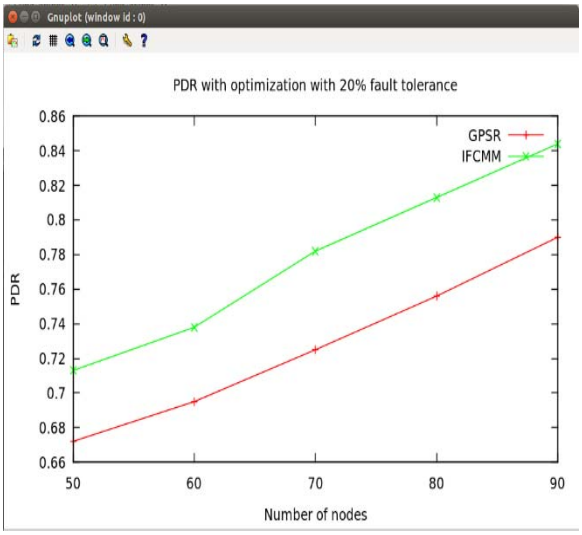


Figure 4.4: PDR with path optimization F=20%

The results are calculated for the energy consumption against the number of data packets delivered. Figure 4.5 shows the energy consumption results plotted against Direct Diffusion protocol. The results show the energy consumed by the system is better that DD, since IFCMM uses the path optimization with fault detection. In DD, the location identification of nodes is through receiving signal strength and this may cause inaccurate identification and may result in disjoint path.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

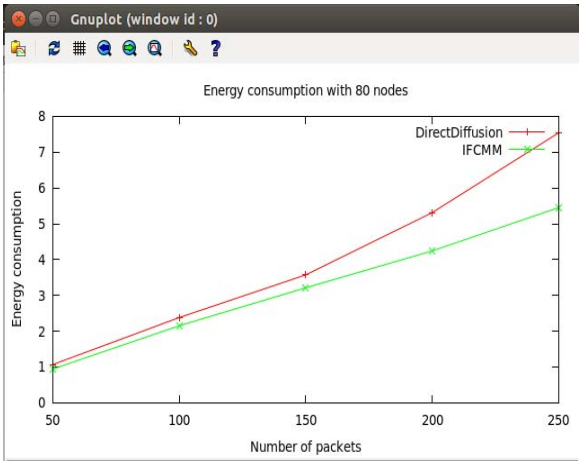


Figure 4.5: Energy consumption with N=80 nodes

4.2 Parameter Value

The proposed protocol is simulated using NS2 simulator with the following parameters as shown in below table 1.

Table 1: simulation parameters Proposed algorithm is evaluated by considering the below parameters:

Area	200m * 500m
Number of nodes	[50-100]
Number of SU	150
Number of channels	[2, 18]
Sensing time per channel	5ms
Channel switching time	80μs

4.3 Dead CH

The results in below Figure 4.6 show the comparison of dead CHs round by round. Dead CH refers to the nodes which are dead due to complete energy depletion or device failure. The results are compared with DEBR [12].

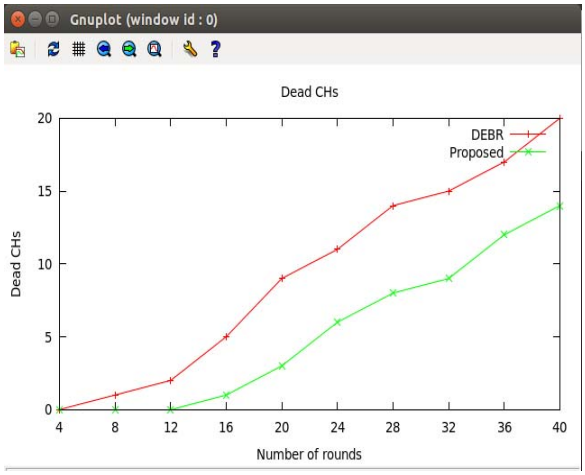


Figure 4.6: Dead Cluster Head vs. Number of nodes compared proposed with DEBR

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

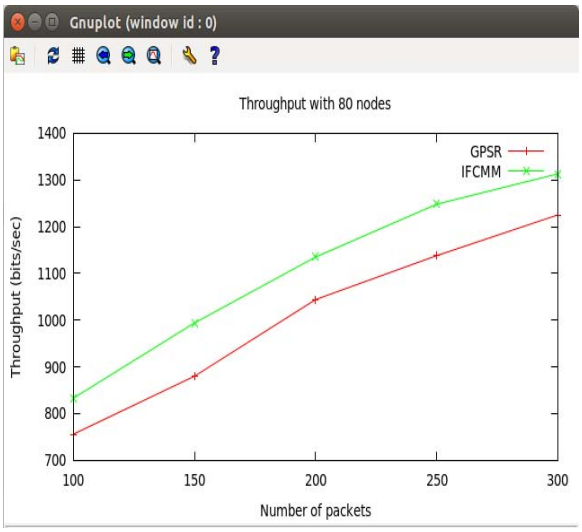


Figure 4.7: Throughput with 80 nodes

Figure 4.7 shows the throughput for total number of packets. In this graph we considered for 80 nodes. When throughput is compared with existing algorithm the proposed algorithm results in high throughput.

Figure 4.8 describes the delay for total 80 numbers of nodes. When compared to existing algorithm the delay is less in the proposed algorithm

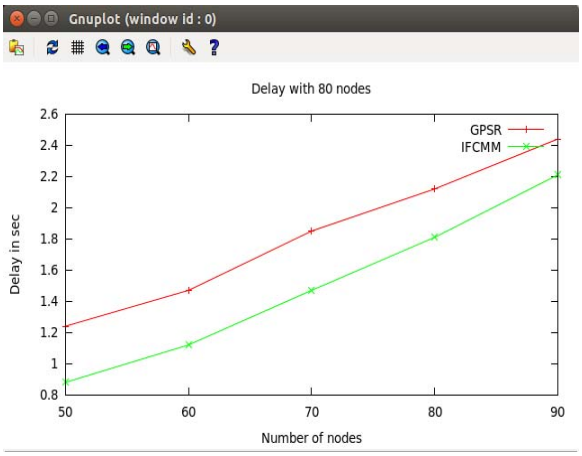


Figure 4.8: Delay with 80 nodes

V. CONCLUSION

This paper presents fault tolerant mechanism for multimedia. It considers the partition of the network into clusters. Energy conservation of the cluster head node is the main aim of this design. Cluster formation is based on residual energy of the CH. Proper selection of gateway nodes which act as replacement of CH node in case of failure will help in retaining the continuity of the network. Also, video streaming technique which operates in various spectrum condition is proposed, which takes higher bandwidth and low delay. The simulation result shows that the performance of the developed scheme is better than the earlier work.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

REFERENCES

[1] WANG, Beibei et LIU, KJ Ray. Advances in cognitive radio networks: A survey. Selected Topics in Signal Processing,IEEE Journal of, 2011, vol. 5, no 1, p. 5-23.

[2] Akyildiz IF, Lee W-Y, and Chowdhury KR: CRAHNs: Cognitive radio ad hoc networks. Ad Hoc Networks 2009,7(5):810–836.

[3] Jha SC, Rashid MM, Bhargava VK, Despins C: Medium access control in distributed cognitive radio networks. Wireless Communications 2011,18(4):41–51.

[4] Jha SC, Phuyal U, Rashid MM, and Bhargava VK: Design of OMC-MAC: an opportunistic multi-channel MAC with QoS provisioning for distributed cognitive radio networks. IEEE Trans. Wireless Commun 2011,10(10):3414–3425.

[5] Ehsan, Samina, and Bechir Hamdaoui. "A survey on energy-efficient routing techniques with QoS assurances for wireless multimedia sensor networks." Communications Surveys & Tutorials, IEEE 14.2 (2012): 265-278.

[6] Priyanka Rawat, Kamal Deep Singh, Hakima Chaouchi, and Jean Marie Bonnin. "Wireless sensor networks: a survey on recent developments and potential synergies." The Journal of Supercomputing (2013), Springer

[7] Bicen, A. Ozan, V. Cagri Gungor, and Ozgur B. Akan. "Delay-sensitive and multimedia communication in cognitive radio sensor networks." Ad Hoc Networks 10.5 (2012): 816-830.

[8] Rehmani, Mubashir Husain, Stéphane Lohier, and Abderrezak Rachedi. "Channel bonding in cognitive radio wirelesssensor networks." Mobile and Wireless Networking (iCOST), 2012 International Conference on Selected Topics in. IEEE, 2012.

[9] Liang, Zhongliang, and Dongmei Zhao. "Quality of service performance of a cognitive radio sensor network."Communications (ICC), 2010 IEEE international conference on. IEEE, 2010.

[10] Liang, Zhongliang, et al. "Delay performance analysis for supporting real-time traffic in a cognitive radio sensor network." Wireless Communications, IEEE Transactions on 10.1 (2011): 325-335.

[11] G. H. E. Fard, M. H. Yaghmaee, R. Monsefi, “An Adaptive Cross-Layer Multichannel QoS-MAC Protocol for Cluster Based Wireless Multimedia Sensor networks,", Proc. of Ultra Modern Telecommunications & Workshops (ICUMT’09),pp.1-6, 12-14 Oct. 2009.

[12] Y.-F. Wen, W. Liao, “On QoS Routing in Wireless Ad-Hoc Cognitive Radio Networks," Proc. of 71st IEEE Vehicular Technology Conference (VTC’10), pp.1-5, 16-19 May 2010.

[13] L. Zhou, X. Wang, W. To, G. Mutean, and B. Geller, “Distributed Scheduling Scheme for Video Streaming over Multi-Channel MultiRadio Multi-Hop Wireless Networks," IEEE Journal on Selected Areas in Communications, vol. 28, no.3, pp. 409-419, Apr. 2010.

[14] K.R. Chowdhury, M.D. Felice, “SEARCH: A routing protocol for mobile cognitive radio ad-hoc networks,", Computer Communications, vol. 32 pp. 1983-1997, 2009.

[15] Digham, F.F.; Alouini, M.S.; Simon, M.K. On the energy detection of unknown signals over fading channels. IEEE Trans. Commun. 2007, 55, 21–24.

[16] KUMAR, Dilip, ASERI, Trilok C., et PATEL, R. B. EECDA: energy efficient clustering and data aggregation protocol for heterogeneous wireless sensor networks. International Journal of Computers Communications & Control, 2011, vol.6, no 1, p. 113-124.

[17] T. Wiegand, and G. J. Sullivan, “The H.264/AVC Video Coding Standard," IEEE Signal Proc. Mag., pp. 148-153 Mar2007.

[18] SHAH, G., ALAGOZ, Fatih, FADEL, Etimad, et al. A Spectrum-Aware Clustering for Efficient Multimedia Routing in Cognitive Radio Sensor Networks. 2014.

[19] Akyildiz IF, Su W, Sankara subramaniam Y. Wireless sensor networks: a survey. Computer Network 2002;38:393–422.

[20] Abbasi AH, Younis M. A survey on clustering algorithms for wireless sensor networks. Computer Commun 2007;30:2826–41.

[21] Akkaya K, Younis M. A survey on routing protocols for wireless sensor networks. Ad Hoc Network 2005;3:325–49.

[22] Kuila P, Gupta SK, Jana PK. A novel evolutionary approach for load balanced clustering problem for wireless sensor networks. Swarm Evol Comput2013;12:48–56.

[23] Bari A et al. Design of fault tolerant wireless sensor networks satisfying survivability and lifetime requirements. Computer Commun 2012;35(3):320–33.

[24] Low CP et al. Efficient load-balanced clustering algorithms for wireless sensor networks. Computer Commun 2008;31(4):750–9.

[25] Kuila P, Jana PK. Improved load balanced clustering algorithm for wireless sensor networks. In: Proc. of int. conf. ADCONS 2011, LNCS 7135; 2012. p.399–404.

[26] Kuila P, Jana PK. Approximation schemes for load balanced clustering in wireless sensor networks. J Supercomput 2014;68(1):87–105.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[27] Gupta G, Younis M. Fault-tolerant clustering of wireless sensor networks. Proc Int Conf IEEE WCNC 2003;3:1579–84.

[28] Kuila P, Jana PK. Energy efficient load-balanced clustering algorithm for wireless Sensor Network. Proc Techno 2012;

[29] Chiang SS et al. A minimum hop routing protocol for home security systems using wireless sensor networks. IEEE Trans Consume Electron2007;53(4):1483–9

[30] Ok Chang-So et al. Distributed energy balanced routing for wireless sensor networks. Computer Ind Eng 2009;57:125–35.

[31] Heinzelman W, Chandrakasan A, Balakrishnan H. Application specific protocol architecture for wireless micro sensor networks. IEEE Trans Wireless Communication 2002;1(4):660–70.

[32] Tyagi S, Kumar N. A systematic review on clustering and routing techniques based upon LEACH protocol for wireless sensor networks. J Network Computer Appllications 2013;36:623–45.

[33] Kuila P, Jana PK. An energy balanced distributed clustering and routing algorithm for wireless sensor networks. In: Proc. OF INT. CONF. PDGC 2012,IEEE Xplore; 2012. p. 220–225.

[34]. Gurses, E., & Akan, O. B. (2005). Multimedia communication wireless sensor networks. *Annals of Telecommunications*,*60*(7–8), 799–827.

[35]. Akyildiz, I. F., Melodia, T., & Chowdhury, K. R. (2007). A survey on wireless multimedia sensor networks. *Computer Networks*,*51*(4), 921–960.

[36]. Mira, S., Reisslein, M., & Xue, G. (2008). A survey of multimedia streaming in wireless sensor networks. *IEEE Communications Surveys and Tutorials*, *10*(4), 18–39. doi:10.1109/SURV.2008.080404.

[37]. He, Z., & Wu, D. (2006). Resource allocation and performance analysis of wireless video sensors. *IEEE Transactions on Circuits and Systems for Video Technology*, *16*(5), 590–599.

[38]. Karp, B., & Kung, H. T. (2000). GPSR: greedy perimeter stateless routing for wireless networks. In *Proceedings of the annual international conference on mobile computing and networking* (Mobi-Com 2000), Boston, USA, August.

[39]. Kuhn, F., Wattenhofer, R., & Zollinger, A. (2003). Worst-case optimal and average-case efficient geometric ad-hoc routing. In *Proceedings of the 4th ACM international symposium on mobile ado networking and computing (MobiHoc 2003)*, Annapolis, MD,USA, June, 2003.

[40] Kuhn, F.,Wattenhofer, R., Zhang, Y., & Zollinger, A. (2003). Geometricad -hoc routing: of theory and practice. In *Proceedings of the 22nd ACM international symposium on the principles of distributed computing (PODC 2003)*, Boston, Massachusetts, USA, July 13–16, 2003.

[41] Leong, B., Mitra, S., & Liskov, B. (2005). Path vector face routing: geographic routing with local face information. In *Proceedings of the 13th IEEE international conference on network protocols(ICNP 2005)*, Boston, Massachusetts, USA, November 6–9,2005.

[42] Gabriel, K., & Social, R. (1969). A new statistical approach to geographic variation analysis. *Systematic Zoology*, *18*, 259–278.

[43] Toussaint, G. T. (1980). The relative neighborhood graph of a finite planar set. *Pattern Recognition*, *12*, 261–268.

[44] Frey, H., & Stojmenovic, I. (2006). On delivery guarantees of face and combined greedy-face routing in ad hoc and sensor networks. In *Proceedings of the international conference on mobile computing and networking (Mobi Com 2006)*, Los Angeles, USA, September,2006.

[45] Seada, K., Helmy, A., & Govindan, R. (2007). Modeling and analyzing the correctness of geographic face routing under realistic conditions. *Ad Hoc Networks*, 855–871. doi:10.1016/j.adhoc.2007.02.008.

[46] Fang, Q., Gao, J., & Guibas, L. J. (2004). Locating and bypassing routing holes in sensor networks. In *Proceedings of the 23rd conference of the IEEE communications society (INFOCOM 2004)*,Hong Kong, China, March, 2004.

[47] Jia, W., Wang, T., Wang, G., & Guo, M. (2007). Hole avoiding in advance routing in wireless sensor networks. In *Proceedings of theIEEE wireless communication & networking conference (WCNC2007)*, USA, March, 2007.

[48] Yu, F., Lee, E., Choi, Y., Park, S., Lee, D., Tian, Y., & Kim, S.(2007). A modeling for hole problem in wireless sensor networks. In *Proceedings of the international wireless communications and mobile computing conference (IWCMC 2007)*, Honolulu, Hawaii, USA, August, 2007.

[49] Tsai, J., & Moors, T. (2006). A review of multipath routing protocols: from wireless ad hoc to mesh networks. In *Proceedings of A CoRN early career researcher workshop on wireless Multihop networking*, Sydney, July 17–18, 2006.

[50] Johnson, D. B., & Maltz, D. A. (1996). Dynamic source routing in ad hoc wireless networks. In K. Imielinski (Ed.), *Mobile computing*. Dordrecht: Lower Academic.

[51] Perkins, C. (2003). *Ad hoc on-demand distance vector (AODV)routing*. RFC 3561.

Constructions of Design Elements from Software Requirement Specifications (SRS)

Syed Naimatullah Hussain, CIT, Taif University, Taif, KSA
Email:s.naimatullah@tu.edu.sa

ABSTRACT

The software project development normally starts with customer's requirements. The customers will collect the requirements from the end users of the organization, and prepares a document called the SRS. Unified modeling language (UML) designed to provide a standard way to visualize the design of an information system. Apart from other internal lacunae of (UML) itself, the main reasons for low success rate are due to the non-availability of a correct and complete methodology for abstraction of the view elements for the design of an information system. This paper abstracts the view elements from Software Requirement Specification (SRS) for any information System in the form of Classes, Object of the Classes, Actors its interfaces their characterizing attributes, and these abstractions are further refined using good database design principles and the use of data flow diagrams (DFDs).

Keywords: Classes, Attributes, Interfaces

1. INTRODUCTION:

The software development project normally starts with SRS of the customers [1, 2]. The customers are in general, strategic management people of the organization who work in classical ambience. So the requirements of the expected system reflect their processing mindset. These requirements are influenced by either the data oriented approach or the procedure oriented approach as with the available information of the organization. Presently, since these are not natural ways of processing the information system, these will not serve the development process effectively. Now a day, people feel the object-oriented paradigm is more towards naturalness and will survive for long time. So it is required to transform the requirements into object-oriented paradigm(Classes, Object of Classes, Actors and its interfaces) and then proceed for the development. The main objective of this paper is to develop a system of software tools, that takes SRS and then transform it into Actors, Classes and Objects and its attributes.. There may also need to limit our ambition, as some of the sub processes may not be automated due to English as Natural Language. In such case, there is a need to provide guidelines for each of these sub processes to minimize the human dependency. the aim to develop a sequence of semi-automated software tools with embedded guidelines for inevitable subtasks at some stages then, the set of guidelines may give the scope to develop software agents to take up the role of semi automated processes.

Few researchers[3] have suggested some techniques for identification of object class hierarchies in procedural object oriented, but in our methodology we are using a perfect balance of procedural as well data oriented paradigm i.e nothing but object oriented paradigm which more towards naturalness for certain stages of the design of object classes. [4, 9] in this authors are abstracting only object class and its attributes based on parsed use case description not all the design components from the SRS. Author's[8,10] this paper helps in the designing only the class and subclass structures but fails to identify the object class structure i.e. object class name, attributes, actors and its interfaces. Although, these give good guidelines to the design, the authors could not derive any concrete procedure and/or guidelines to the design in its totality. To develop a methodology that identifies Classes and Objects of the Classes, Actors and its attributes, , their characterizing attributes. This semi automatic methodology comprises of a sequence of steps like feasibility analysis, for object structure identification, resolution of synonyms & homonyms issues, regrouping of attributes of entities & functionalities through the design of data flow diagrams and elimination of imbalance data & procedure selection along with authentication of correctness & completeness of the abstractions at each stage. There is manual intervention at few stages is necessitated because of the need for human intelligence in these steps. Even for these manual intervention steps, attempt is made to provide clear-cut guidelines to streamline the design process.

2. Methodology

The algorithm presumes that, based on SRS from customer's this SRS is already available with the developer. Who can seek needy information from client's team of users (CTU). This paper addresses a sequence of semiautomatic methodology.

2.1 Requirement gathering.

As per the SRS, the detailed requirement is gathered from CTU. The input of any information system is SRS, output is Classes, Objects of the Classes, Actors and its attributes, If CTU is very large, then similar information is collected through the response to the questionnaire from the client.

2.2 Authentication of correctness and completeness of the process

Now we have two sets of entities, attributes, interrelationships, business rules, work processes and business process. The developer need to establish the one-to-one and on-to correspondence separately between each pair of items of two sets.

2.3 Resolve synonyms/homonyms issue.

As English is a Natural Language, in a multi-user system, though each user assigns meaningful

names to attributes, the semantic flexibility in the use of English words leads synonymous words for the same attribute. The set of synonymous words of the same meaning forms a synonymy and each such synonymy is replaced by a generic name. Similarly, the use of context specific word leads to homonyms. The presence of each such homonym in attributes/entities/actors should be replaced by different names.

2.4 Model the business process through to measure the paradigms imbalance.

- Entities, which are modifiable within the system
- Entities which are non modifiable within the systems form actors.
- The identified functionalities form processes
- The intersection of input/output attributes of a functionality with each entity attributes forms input/output dataflow from/to the concerned entity /actor to/form the process

This logical DFD (Fig-2.2) reveals the degree of imbalance in the paradigms opted while choosing structural and behavioral components. Figures 2.51 and 2.52 depict the possible reasons tilt respectively towards Procedure oriented and data oriented paradigms

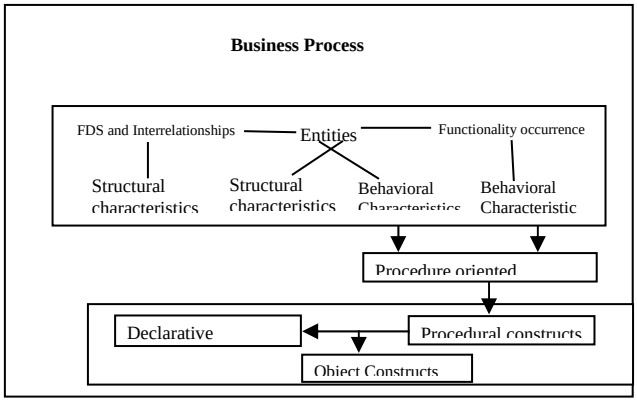


Fig. 1.1(Procedure Oriented)

Real World

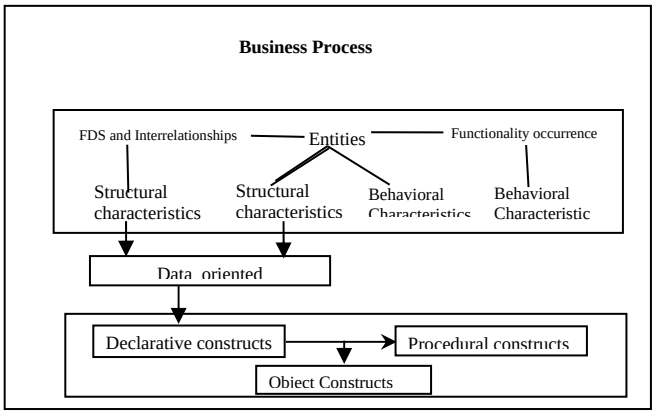


Fig.1.2 (Data Oriented)

2.5 Eliminate superfluity and redundancy in attributes/entities presence.

Study each attributes of each entity/actor in isolation with other entities/attributes for absolute necessity of their presence in the information system. This can be identified by the participation of the attribute in any of the functionalities. Discard the attributes that are not participating in the functionalities. If an attribute or group of attributes is present in two or more entities, form a separate entity with each such group. This participation can be tested through the establishment of one-to-one and on-to correspondence between attributes referenced in data flows of logical DFD (Fig 2.0)

2.6 Identify keys and design extended ER-diagram

Abstract the functional dependencies amongst attributes of each entity from the business rules of the information system. Identify the primary key and foreign keys for each entity/actor. If an attribute or group of attributes of an entity of data store is independent of the primary key, take it out and form separate entity. The entities and attributes are abstracted from the data dictionary and the interrelationships are abstracted from the business rules of the system.

2.7 Minimize the imbalance between the procedure and data oriented paradigms.

Figure-1.3 (Object Oriented)

[Fig-1.1 &1.2] Regroup the attributes of input data flows, based on their characterization of Person, place, thing, event or concept. This can be achieved through the grouping the attributes such that each attribute of a group establishes a functional dependence with one primary key. The input data flows may contain subset of attributes of a group so that each group of attributes may be in input data flows of one or more processes. Each such group forms a first cut object structure. [Fig-1.3] shows the perfect balance between data and procedure oriented i.e object oriented which is more towards naturalness.

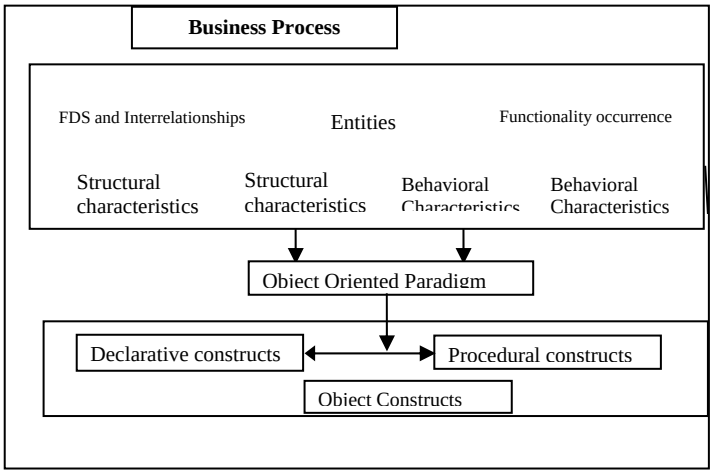


Figure-1.3 Object Oriented

2.8 Refine the abstracts by brining the perfect balance between the paradigms.

[5.6.7] Apply good database design principles to each first cut object structure [section 3.2].

Since the normalization is a way of good database design, the refined group should be normalized at least up to Boyce coded normal form. Figure-2.2 indicates the constructs to bring perfect balance between procedure oriented and data oriented paradigms.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

2.9 Design the logical DFD with object structure

Fig-2.2 Refine the logical data flow diagram with each object structure group as data store, maintaining attributes of total input data flow to each process and redesigning input flows such that each flow emanates from first cut object structure.

2.10 Decompose functionality

If two or more data flows directed towards single process, study the possibility of decomposing the process, so that each decomposed process either receives data flow from single entity or additionally through parameter passing from other entities. Refine the logical data flow diagram into physical data flow diagram.

2.11 Software Requirement Specification[SRS]

.

The input to the any information system is the requirements that contain the business rules of the information system along with various applications. For example, SRS for Banking Application information system may contain some of the business rules as follows.

3.0 Theory

Automate the customer transaction in the banking system
There are two types of customer.
Deposit customer

- Each customer is provided with a unique account number
- Periodically credits/debits from his/her account.
- The customer earns interest on minimum balance of each month.
- There may be overdraft facility for some customer who has developed some kind of rapport with the bank.
- The customer can also transfer amount from his/her account to any account.

Loan Customer

- Loan customer initially applies for loan for specific purpose.
- The bank grants loan to customers after authenticating their repayments in installments.
- The bank charges interest either on floating-point rate or fixed-point rate.
- Interest rate for bank loan is generally much higher than interest rate for bank deposit
- Thus the bank earns profit by liaising between savings customer and loan customer
- Each branch manager needs to submit periodical summary report indicating the total number of transactions, the total amount deposited, the total interest paid to the customer, the total amount sanctioned, the interest earn by customer and thus the profit earn by the bank in the specific

3.1 Identification of Classes, attributes from SRS

Depositor (Acctno,D-name,D-address,Contact-no)
Transaction (Acctno,Pre-balance,Transtype,Transamt,New-balance,OD-facility)
Daccount (acctno, period, minibalance, int-rate, intAmt)
Transfer (acctno, minbalance,ch-amt,payee-acct,newbalance,payeebalance)
Borrower (acctno, loan-type, Amt-sanctioned, Amtavailed, newbalance, int-rate)
EMI (acctno, Balance, Rp-period, Principal, Int, EMI)
Baccount (acctno, period, maxbal, int-rate, intamt)
Report (period, no of depositor, amt deposited, int-paid, amt-sanctioned, no of borrowers, amt-disbursed, int-earned, profit)
Loan (Bacctno, Bname, Baddress, Bcontact no)

3.2. Synonyms and homonyms

These are Homonyms found in various entities like, Acctno, period, Int-rate, Int-amt, NewBalance.
Depositor (Acctno,D-name,D-address,Contact-no)
Transaction (Acctno,Pre-balance,Transtype,Transamt,New-balance,OD-facility)
Daccount (Dacctno, Dperiod, minibalance, Dint-rate, DintAmt)
Transfer (Dacctno, minbalance,ch-amt,payee-acct,Dnewbalance,payeebalance)
Borrower (Bacctno, loan-type, Amt-sanctioned, Amtavailed, Bnewbalance, Bint-rate)
EMI (Bacctno, Balance, Rp-period, Principal, Int, EMI)
Baccount (Bacctno, Bperiod, maxbal, Bint-rate, Bintamt)
Report (period, no of depositor, amt deposited, int-paid, amt-sanctioned, no of borrowers, amt-disbursed, int-earned, profit)
Loan (Bacctno, Bname, Baddress, Bcontact no)
I could not fine any of synonyms in this application

4. DATA FLOW DIAGRAM (DFD) (Context Level Diagram and Logical DFD refer Page no.4, 5)

Results:

Actors: Customer, Depositor

Attributes of Actor: report, transaction, EMI, Bank database

Classes: Sanction Loan, Transaction Amount, Transfer Amount,Compute EMI,Prepare Report, Assign Account No.

Objects of Classes: Acctno,D-name,D-address,Contact-no,Acctno,Pre-balance,Transtype,Transamt,New-balance,OD-facility) Dacctno, minbalance,ch-amt,payee-acct,Dnewbalance,payeebalance
(Bacctno, Balance, Rp-period, Principal, Int, EMI)
period, no of depositor, amt deposited, int-paid, amt-sanctioned, no of borrowers, amt-disbursed, int-earned, profit, Bacctno, Bname, Baddress, Bcontact no.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Discussion:

. Unified Modeling Language [UML][11,12] has become a de facto standard for the design and implementation of object oriented information system. Though the UML [13] is a boom for the design of business processes in an object oriented paradigm, the success rate of projects involving the design using UML is very slow.

The SRS is the abstraction of number of end user's requirements. Because of the environment in which the individual is working, the end user may use different words for the same meaning. Thus, the SRS contains synonymous words. In addition, the use of contexts specific words compels the end user to use the same word for different meanings. Thus, SRS also contains homonymous. The non-resolution of synonymous issue results in the redundancy model elements whereas; non-resolution homonym issue results in violation of uniqueness of the model elements. Thus any information system design needs to resolve the synonym and homonym issues. Thus, the success rate of the information system developed using such error prone approaches swings. There is an urgent need to isolate the dependency of software development success rate on human skills and to elevate it to the height of the automatic process.

5. CONCLUSION:

UML is a graphical language for visualizing, specifying, constructing, and documenting the artifacts of a software-information system. But suffer from the correctness and completeness aspect. Developed methodology helps in designing, design components (Classes, Objects and its Attributes) in a semiautomatic way incorporating correctness and completeness in each of its stages. This methodology minimizes human interventions that too giving clear-cut guidelines. In future these, intervention also can be automated.

6. REFERENCES:

[1] Pankaj Jalote (2005). An Integrated approach to Software engineering, 3rd edition springer's publications, India.
[2] Roger S Pressman (2005). Software Engineering, 6th edition McGraw-Hill companies international edition, India
[3] Mosa Emhamed Elbendak(2005) "Framework for using a Natural Language Approach to Object Identification Framework for using a Natural Language" http://www.aclweb.org/anthology/R09-2005.
[4] Muhammed Usman, Stepahane Ducane and Marianne Huchard (IEEE-2008) 15th working conference on reverse engineering "Reconsidering classes in procedural object oriented code" page 257- 266.
[5] Jonathan & charles J Hannon (IEEE-2009) Eighth Mexican International Conference on Current Trends in Computer Science. "An algorithm for identifyingAuthors using synonyms" page 99 – 104.
[6] Alen Lovrencie and Tonirnir Kisasondi (IEEE-2007) 11th International Conference on Intelligent Engineering System "Modelling Functional Dependencies in Databases using mathematical logic" page 307 – 312.
[7] Steffen Rudle and LarsSchmidt-thieme (IEEE-2006) Sixth International Conference in Data Mining "Object Identification with constraints" page 1 -7.
[8] Sukhamay Kundu & Migel (IEEE-2005) 29th International Conference Software and Application Conference "A Formal approach to Designing a class subclass structure using a partial order on the functions" page 1 -8.
[9] Mosa Emhamed Elbendak "Framework for using a Natural Language Approach to Object Identification Framework for using a Natural Language" Student Research Workshop, RANLP 2009 - Borovets, Bulgaria, pages 23–28
[10] .S. Anandha Mala and Dr.G.V.Uma "Object Oriented Visualization of Natural Language Requirement Specification and NFR Preference Elicitation. IJCSNS International Journal of Computer Science and Network Security, VOL.6 No.8, August 2006.
[11] Lilac A. E. Al-Safadi "Natural Language Processing for Conceptual Modeling" A semi-automatic approach for designing databases. International Journal of Digital Content Technology and its Applications Volume 3, Number 3, September 2009
[12] Sunguk Lee "UML for Database System and computer application. Interantional Journal of Database System and Computer Application. Volume 5, number 3, 2012
[13] K.P Jayant, Garg,Kumar, Rana. "An approach of Software Design Testing Based on UML diagram. International Journal of Advanced Research in Computer Science and Software Engineering. Volume-4, Issue-2, February 2014
[14] Tiwari,Alpika,Dubey "Merging of Dataflow Diagrams with Unified Modeling Language" Internation Journal of Scientific and Research Publication Volume-2,issue-8, August-2012
[15] Rosziati,Yen "A Formaol Model for Data Flow Diagrams Rules" APRN Jounal of System and Software, volume-1 No.2, May-2011

Context Level Diagram [14]:
Fig-2.1 Depicts that the interaction between the banking information systems with the external agents which acts as data sources and data sinks.

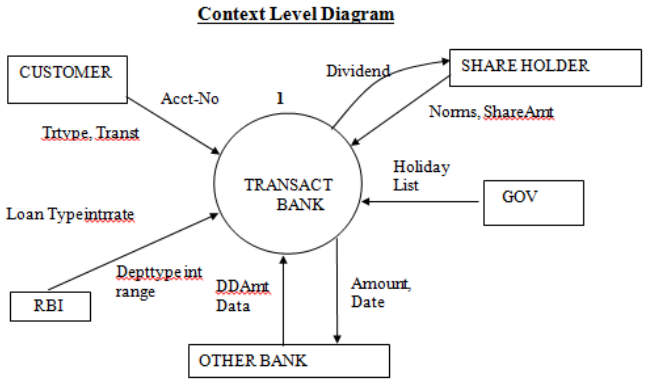


Figure-2.1 Context Level Diagram

Figure-2.2 logical DFD [15] captures the data flows that are necessary for a banking information system to operate. It describes the processes that are undertaken, the data required and produced by each process.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

#

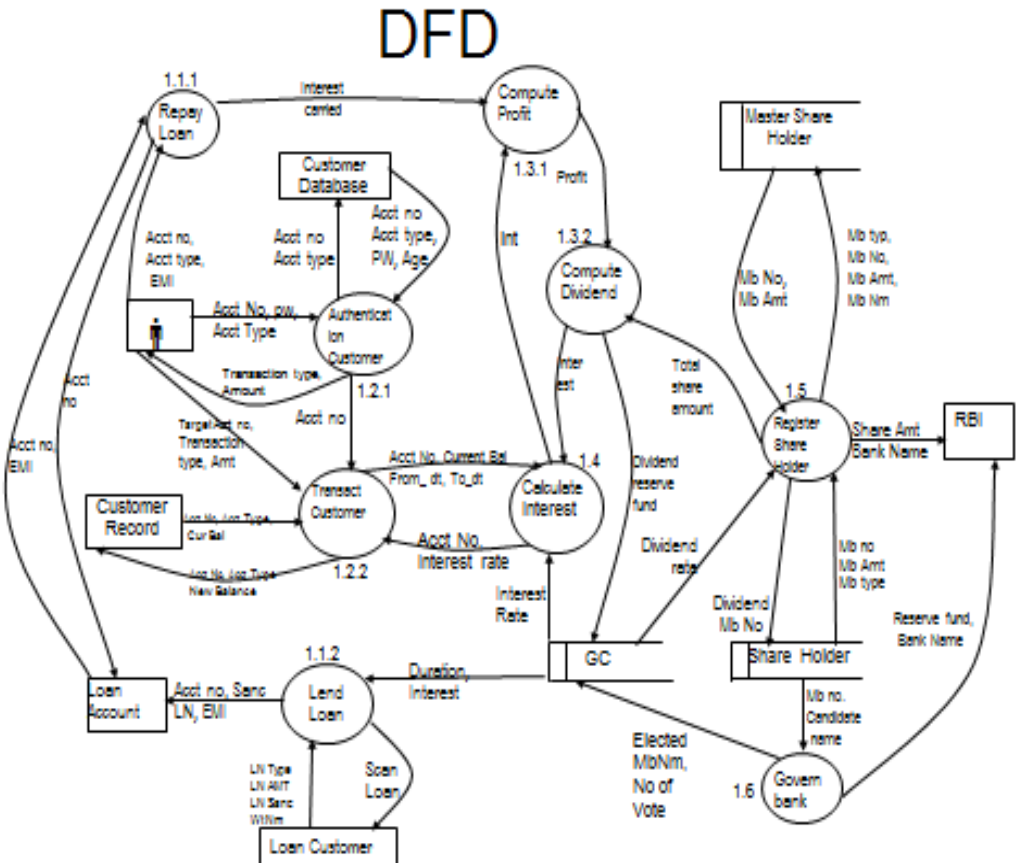


Figure-2.2 Logical DFD

An Efficient Way of Medical Image Encryption using Cat Map and Chaotic Logistic Function

Ranu Gupta [1*], Rahul Pachauri [2], Ashutosh K Singh [3]
[1, 2]Jaypee University of Engineering and Technology, Raghogarh, Guna (M.P.) India, 473226
[3]Thapar Institute of Engineering and Technology University, Patiala, 147004, India
*Corresponding author, E-mail: ranugupta02@yahoo.com
pachauri.123@gmail.com (2), aksingh@thapar.edu(3)

***Abstract*: Encryption is used to securely transmit data in open networks. The proposed image encryption method uses Arnold cat map and one dimensional chaotic map. The pixel shuffling is done using cat map. Then the encryption is done by using 128 bit long external secret key. The initial condition is calculated using secret key. Two matrices of size of the image are formed using chaotic logistic function. Finally the image is encrypted by performing XOR operation with the two matrices formed by the chaotic function and the shuffled image.**

I INTRODUCTION

In today's Hi-Tech world of innovations in the field of internet, medical imaging systems, military message communication, it is needed to transmit the images through network confidentially. For reliable and secure image transmission, we heavily rely on image cryptography which is the base of security in digital communication. Many image encryption methods are available to protect the content of digital images but, few of them are at par with the expected goals i.e. speed, reliability and security. Since the images play a vital role in the field of medical treatment of the patient. These medical images need to be transmitted through public network in order to consult the doctors. Therefore security plays a vital role while transmitting the images. While the general images are important in everyday life. The traditional encryption schemes .like simple-Data Encryption Standard (DES), triple- DES, Rivest Shamir Adleman (RSA), International Data Encryption Algorithm (IDEA) and Advanced Encryption Standard (AES) do not fit for modern image transmission requirement. Many researchers have tried to innovate better solutions for image encryptions. In particular, application of chaos theory in multimedia encryption is one of the important research directions. The field of chaotic cryptography has undergone tremendous growth over the past few decades. The primary motivation of employing chaotic systems is its simplicity in form and complexity in dynamic are not appropriate to make cryptosystems for large digital data. For encrypting the digital images, plenty of encryption schemes have been proposed [1-15]. Scanning procedure for the encryption and simultaneously applying compression to the image [2] was proposed. In [6], a symmetric encryption scheme based on 2D chaotic map is proposed. A two or higher dimensional discretized chaotic maps is adopted for pixel permutation together with 1D map for diffusion. The superiorities of such kind of chaos-based approaches are mainly relatively large block size and a high encryption rate. The analysis of nonlinear chaotic algorithm (NCA) map and a no. of attacks were proposed in [7]. Image encryption scheme based on 3D baker with dynamical compound chaotic sequence cipher generator was proposed in [8]. Divide dynamic block of 3D baker map by using compound chaotic map, and compare with 2D baker map. The 3D baker scheme is 2-3 time faster of 2D baker map. The proposed scheme used in real-time secured image transmission.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

This work combines the confusion/diffusion in single unit for image encryption. Permutation and diffusion are two separate and iterative stages, and they both require scanning the image in order to obtain the pixel values in fastest way. The paper is organized as follows:

Section-2 and 3 gives a short overview of chaotic logistic function and the proposed method respectively. Section 4 evaluates the proposed method with the performance parameters. Section-5 gives the conclusion.

II CHAOTIC LOGISTIC FUNCTION

The chaotic logistic map is random and simple to implement and therefore cryptographers are diverted to this approach for image encryption. The one dimensional chaotic function [1] is expressed as:

$$X_{i+1} = aX_i(1 - X_i) \quad (1)$$

where X_i takes values in the range from [0-1]. It is the simplest method in chaos. Assuming initial condition to be X_0 = 0.4, for different value of *a*, the logistic equation is evaluated through simulation and shown in Fig. 1. It can be concluded that when *a* is in the interval [3.5-4], it is highly chaotic.

The characteristics of chaotic systems are [2]:

i It is deterministic and follows some mathematical equation.

ii But it appears to be highly random.

iii It is unpredictable and follows non linear relation.

iv It is very sensitive to initial condition i.e with a slight change in the initial condition there is a vast change in the succeeding values.

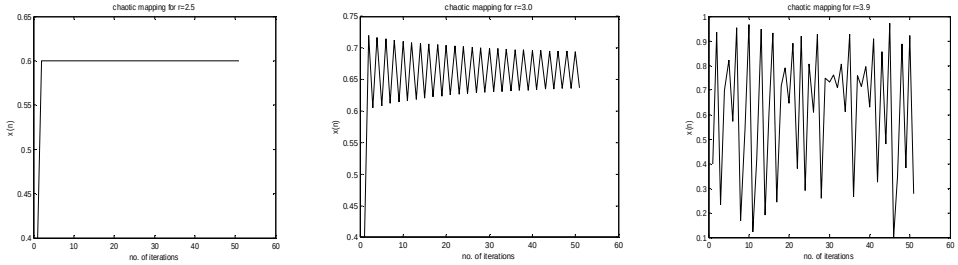


Figure 1. Iteration property

III PROPOSED METHOD

In the proposed image encryption method, the image is permutated by Arnold cat map. Thereafter, the external secret key of 16 ASCII characters is entered. The flow diagram of the proposed method is shown in Fig. 2. Steps involved in the proposed method for the encryption of the image are as follows:

(a) The image is shuffled by using Arnold cat map.

$$x' = (x + ay) \quad mod(N)$$
$$y' = (bx + (ab + 1)y) \quad mod(N) \quad (2)$$

Where a, b are controlling parameters which are positive integers and x', y' is the new pixel position of the original image pixel of position x, y respectively. It is applied once in the image.

(b) A secret key of 16 ASCII characters (128-bit long) is used in the proposed method. The secret key can be represented as:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

$K_i = K_1\ K_2\ K_3$ .....A_{16} (in ASCII) (i= 0-16) (3)

(c) The initial condition (X_0) for the chaotic map is calculated as follows:

$$T = (\sum_{i=1}^{16} S_i * (K_i)) \bmod 256 \quad (4)$$

$$X_0 = T - \lfloor T \rfloor \quad (5)$$

where, $S_i = 0.123 + S_{i-1}$ (6)

and Ki, $\lfloor\ \rfloor$, are the decimal equivalent of the keys, the floor function respectively. The initial value of S_i is taken to be zero.

(d) A chaotic logistic map as in equation (1) is used in the proposed method to generate two matrices of the same as image for image encryption. The initial values calculated are ranging from 0-1, therefore these values are converted into whole numbers < 256.

(e) Then one by one the consecutive byte is read from the shuffled image. The encryption is done by XORing the corresponding values of the three matrices. Two matricesX_i, Y_i generated from logistic function and third is the image matrix P_i.

$$C_i = P_i \oplus X_i \oplus Y_i \quad (7)$$

(f). In case of decryption process all the steps from (a) to (e) are same but decryption is done in the reverse order.

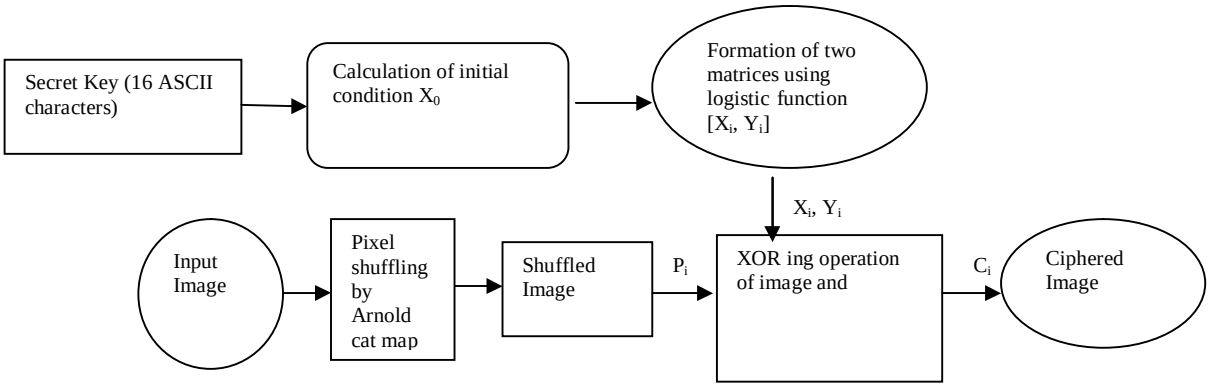
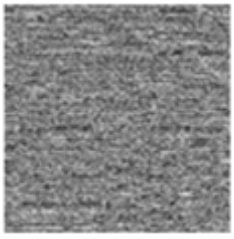


Figure 2. Block Diagram of Proposed Method

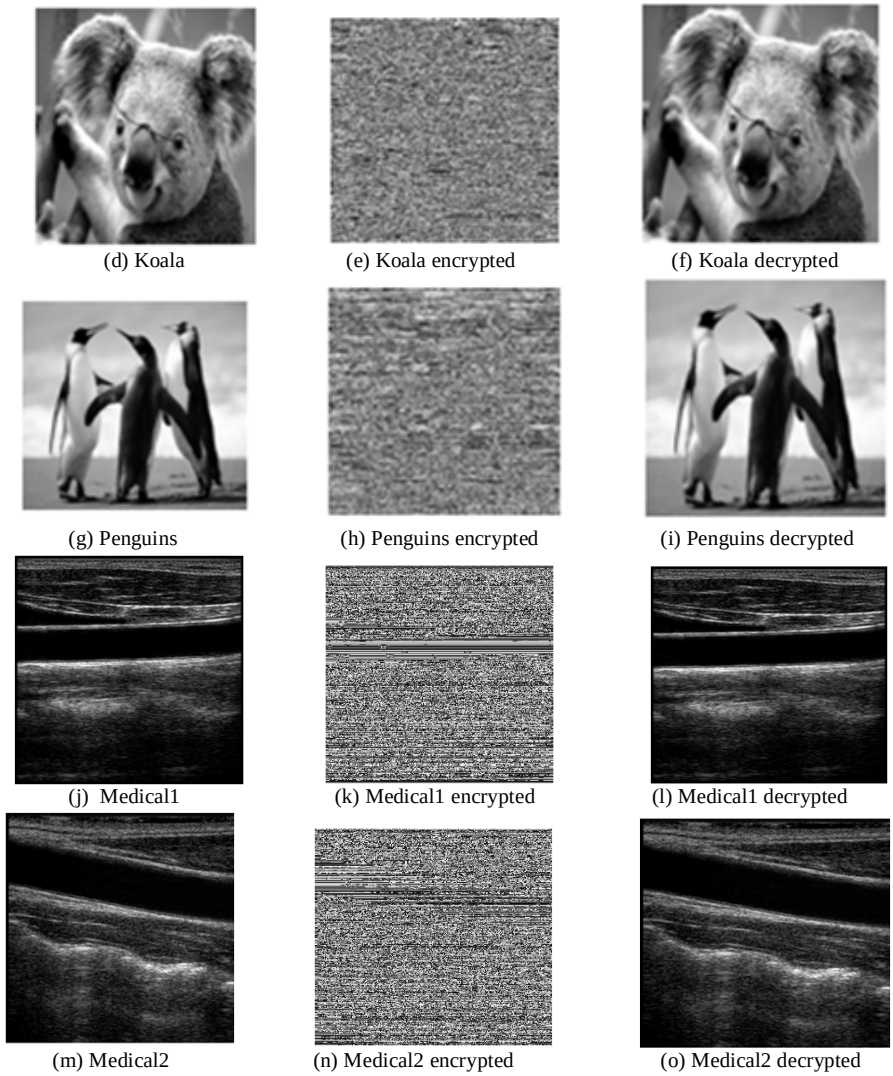
(a) Lena



(b) Lena encrypted

(c) Lena decrypted

https://sites.google.com/site/ijcsis/
ISSN 1947-5500



(d) Koala (e) Koala encrypted (f) Koala decrypted

(g) Penguins (h) Penguins encrypted (i) Penguins decrypted

(j) Medical1 (k) Medical1 encrypted (l) Medical1 decrypted

(m) Medical2 (n) Medical2 encrypted (o) Medical2 decrypted

Figure 3. Original and encrypted images with the proposed method

As shown in Fig. 3 five images (a), (d), (g), (j), and (m) are encrypted by the proposed method whereas Fig. 3 (b), (e), (h), (k), and (n) shows the encrypted images. The encrypted images show that it is quite difficult to trace the original information, making the proposed method more efficient. While Fig. 3 (c), (f), (i), (l), and (o) shows its corresponding decrypted images.

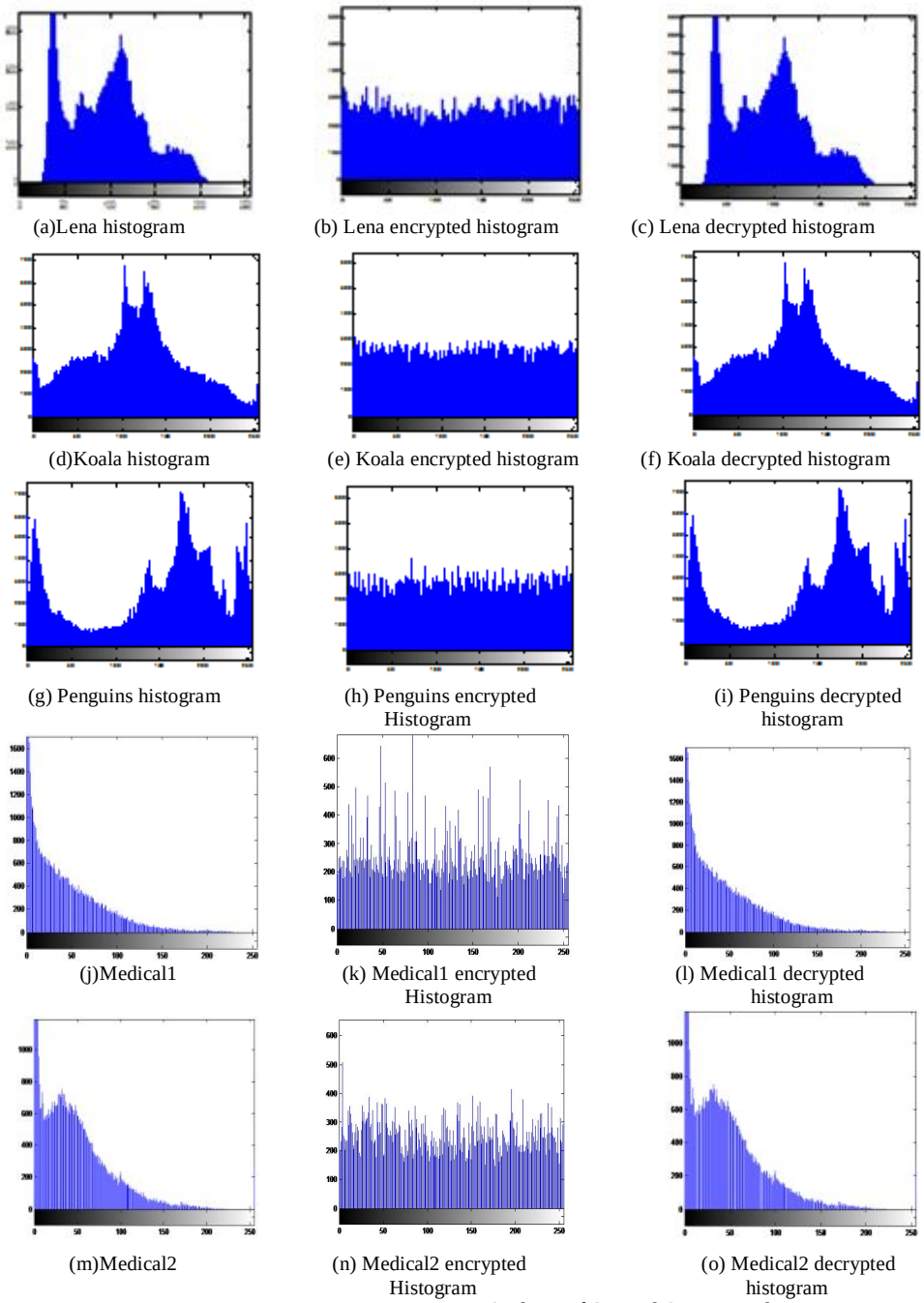
IV` PERFORMANCE ANALYSES

The various performance analyses are done of the proposed method in the following sections:-

A Histogram Analysis

The histograms of various encrypted and original images is shown in Fig. 4 (b), (e), (h), (k), and (n) and Fig. 4 (a), (d), (g), (j), and (m) respectively.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500



(a)Lena histogram (b) Lena encrypted histogram (c) Lena decrypted histogram

(d)Koala histogram (e) Koala encrypted histogram (f) Koala decrypted histogram

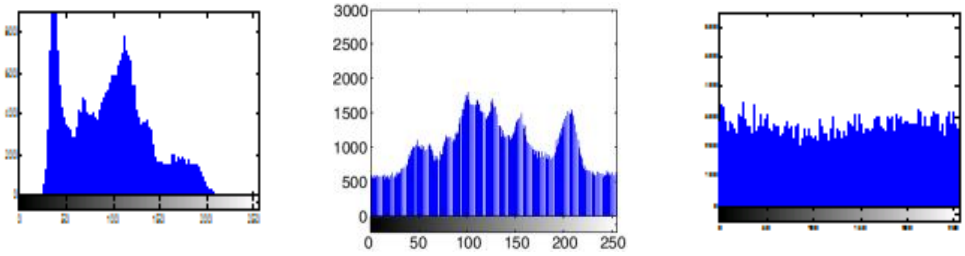
(g) Penguins histogram (h) Penguins encrypted Histogram (i) Penguins decrypted histogram

(j)Medical1 (k) Medical1 encrypted Histogram (l) Medical1 decrypted histogram

(m)Medical2 (n) Medical2 encrypted Histogram (o) Medical2 decrypted histogram

Figure 4. Histogram Analyses of Original & Encrypted Images

https://sites.google.com/site/ijcsis/
ISSN 1947-5500



Original Lena Histogram | Encrypted Lena Histogram [17] | Encrypted Histogram [proposed method]

Figure 5. Comparison of histogram with the proposed method

B Correlation Coefficient Analysis

In addition to the histogram analysis, we have also analyzed the correlation between two vertically adjacent pixels, two horizontally adjacent pixels and two diagonally adjacent pixels in plain-image/cipher-image respectively. Following formula is used for calculation [4].

$$C_r = \frac{N\sum_{j=1}^{N}(x_j \times y_j) - \sum_{j=1}^{N} x_j \times \sum_{j=1}^{N} y_j}{\sqrt{\left(N\sum_{j=1}^{N} x_j^2 - \left(\sum_{j=1}^{N} x_j\right)^2\right) \times \left(N\sum_{j=1}^{N} y_j^2 - \left(\sum_{j=1}^{N} y_j\right)^2\right)}} \quad (8)$$

where x and y are the two neighboring pixels and *N* is number of pixels in the image. Table I and II shows the correlation coefficients of the original as well as encrypted images respectively. The images are encrypted using the secret key **"**12ghUO3456HJKLjx**".** Table II shows that encrypted image are weakly correlated. The pictorial representation of coefficient of correlation of Lena image and the encrypted image with the proposed method is shown in Fig. 6.The comparison of coefficient of correlation of the proposed method with other method is shown in Table III.

Table I Correlation coefficients for the neighboring pixels in the original images

S.No.	Images	Size	Vertical	Horizontal	Diagonal
1.	Lena	256x256	0.9414	0.9204	0.9412
2.	Koala	683x512	0.9553	0.9470	0.9553
3.	Penguins	1024x768	0.9785	0.9576	0.9785
4.	Medical1	348x412	0.8801	0.9836	0.8801
5.	Medical2	389x367	0.8278	0.9194	0.8278
6.	Medical3	400x307	0.9275	0.9718	0.9275

Table II Correlation coefficients for neighboring pixels in the encrypted images

S.No.	Images	Size	vertical	horizontal	Diagonal
1.	Lena	256x256	-0.00035	0.00074	-0.00013
2.	Koala	683x512	0.00064	0.00066	-0.0006
3.	Penguins	1024x768	0.00062	0.00124	0.0035
4.	Medical1	348x412	0.00151	0.00121	0.0031
5.	Medical2	389x367	0.00085	0.00104	0.0041
6.	Medical3	400x307	0.00271	0.00144	0.0075

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table III Comparison of correlation coefficient of the proposed method (Lena)

Correlation Coefficient	**[18]**	**[19]**	**[20]**	**[21]**	**[22]**	**[23]**	**Proposed method**
Vertical	0.028	0.040	0.065	0.027	0.0024	0.0021	-0.0035
Horizontal	0.045	0.082	0.016	0.012	−0.0086	0.0046	0.0074
Diagonal	0.021	0.005	0.032	0.007	0.0402	0.0033	-0.0013

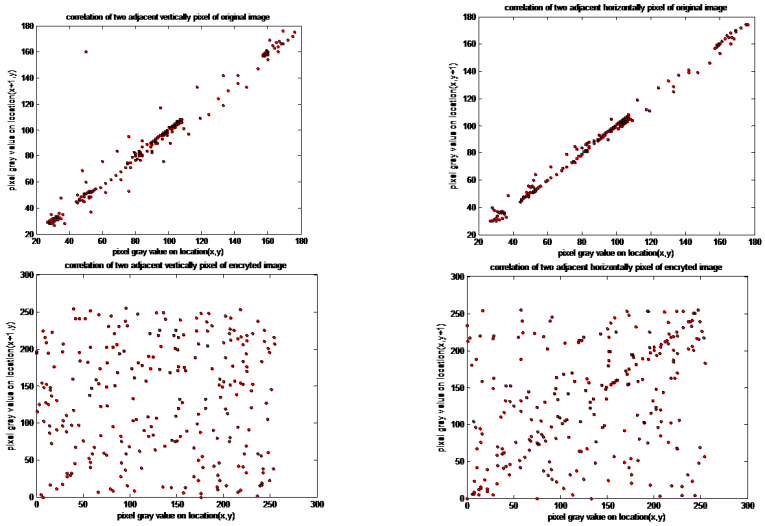


Figure 6. Horizontal and vertical Correlation coefficient of Lena original and encrypted image

C Sensitivity Analysis

Sensitivity analysis can be done on the secret key as well as on the image. To test the sensitivity of the proposed method the original image is encrypted by three different session keys, first by changing the MSB and second by LSB. The three session keys are '12ghUO3456HJKLjx', 'M2ghUO3456HJKLjx', and '12ghUO3456HJKLj4' respectively. The encrypted images and its correlation are shown in Fig. 7 and Table IV.

(i) Original lena image (ii) Encrypted image

(iii) Encrypted image (iv) Encrypted image

Figure 7. Sensitivity test I: Frame (i) Original image, (ii), (iii), (iv) Encrypted images

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table IV Correlation coefficients of the three encrypted images

S.No.	Images	Images	Correlation coefficient
1.	(ii)	(iii)	-0.00084
2.	(iii)	(iv)	-0.00075
3.	(iv)	(ii)	-0.00043

Table IV shows the correlation coefficients between the corresponding pixels of the three encrypted images (ii), (iii) and (iv) and indicate that the proposed method is very sensitive to even a slight change in the secret key.

D Number of Pixels Change Rate (NPCR)

The NPCR [10] is defined by the following equation:

$$NPCR = \sum_{i=1,j=1}^{x,y} \frac{D(i,j)}{x \times y} \times 100\% \quad (8)$$

If $C_1(i,j) = C(i,j)$ then $D(i,j) = 0$ else $D(i,j) = 1$

where *x* and *y* are the width and height of encrypted image. The *NPCR* for various images is calculated by the proposed encryption method and found to be above ***99%***. Table V shows the comparison of the NPCR with the proposed method.

E *Unified Average Change Intensity (UACI)*

UACI can be defined as

$$UCAI = \sum_{i=j=1}^{w,\ h} \frac{|E1(i,j) - E2(i,j)|}{255wh} \times 100 \quad (9)$$

where w and h are the width and height of the image. It shows the average change in the intensities of the pixel.The comparison of UACI with other methods is shown in Table V.

Table V Comparison of NPCR and UACI criteria of proposed method and the other methods for Lena image

Methods	NPCR%	UACI %
[20]	NA	NA
[21] 2nd round	25	8.50
[18] 1st round	37	9
[26]	97.62	32.90
[19]	98.669	33.362
[25]	99.52	33.14
[23]	99.63	33.71
[27]	99.61	33.35
[24]	99.70	29.3
Proposed Method	**99.9975**	**29.6477**

G Time Analysis

The time is also calculated for the encryption and decryption of Lena image by the proposed method. The time analysis is done on Intel(R) Core (™) 2 Duo CPU T5870 @2.00GHz with 3GB RAM computer. The coding is done on MATLAB 7.9.0(R2009b). Table VI shows the encryption/ decryption time taken by the proposed method.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table VI Time taken in Encryption and Decryption by the proposed method

Image	Size	Encryption Time (secs)	Decryption Time (secs)
Lena	256x256	0.00222775	0.00223049

V CONCLUSION

In this work a technique is proposed which replaces the traditional preprocessing complex system and utilizes the basic operations like confusion, diffusion which provide better encryption using cat map and chaotic function. The proposed method is successfully and efficiently implemented to various images. The performance of various parameters shows that the proposed method is robust efficient, secured and fast. It can also be used in real time applications. Theoretical analyses and computer simulations on the basis of histogram analysis, correlation analysis, NPCR and UACI confirm that the new algorithm minimizes the possibility of brute force attack for decryption and fast for practical image encryption. Its future scope is that the method can be varied by increasing the key length and modifying the mathematical calculation for calculating the initial condition.

DISCLOSURE

There was no funding from any agencies either government or non-government.

ACKNOWLEDGMENTS

The general images were taken from the site http://sipi.usc.edu/database/ which is freely available whereas medical images were taken from Government medical college, India.

REFERENCES

[1] G. Chen, X.Y. Zhao, “A self-adaptive algorithm on image encryption”, International Journal Software, 16, 1987, pp. 1975–1982.
[2] N. Bourbakis, C. Alexopoulos, “Picture data encryption using SCAN patterns”, Pattern Recognition, 25(6), 1992, pp. 567–581.
[3] K.L. Chung, L.C. Chang, “Large encrypting binary images with higher security”, Pattern Recognition Letters, 19(5-6), 1998, pp. 461– 468.
[4] J.C. Yen and J.I. Guo, “A new image encryption algorithm and its VLSI architecture”, IEEE Workshop on Signal Processing System, 3 1999, pp. 430–437.
[5] H. Cheng, X.B. Li, “Partial encryption of compressed images and videos”, IEEE Transaction in Signal Processing, 48(8), 2000, pp. 2439–2451.
[6] J. Fridrich, “Symmetric ciphers based on two-dimensional chaotic maps”, International Journal of Bifurcation and Chaos, 8(6), 1998, pp. 1259–1284.
[7] G. Alvarez, Shujun Li, “Crypt analyzing a nonlinear chaotic algorithm (NCA) for image encryption”, Commun Nonlinear Sci. Numer. Simulation, 14, 2009, pp. 3743– 3749.
[8] M.S. Baptista, “Cryptography with chaos”, Physics Letter A, 240(1-2), 1999, pp. 50–54.
[9] X. Liao, X. Li, J. Pen and G. Chen, “A digital secure image communication scheme based on the chaotic chebyshev map”, International Journal of Communication Systems, 17(5), 2004, pp. 437–445.
[10] F. Han, X. Yu and S. Han, “Improved baker map for image encryption”, International Symposium on Systems & Control in Aerospace & Astronautics, 2, 2006, pp. 1273–1276.
[11] S. Lian, J. Sun, Z. Wang, “A block cipher based on a suitable use of chaotic standard map”, Chaos Solitons & Fractals, 26(1), 2005, pp. 117–129.
[12] B. He, F. Zhang, L. Luo, M. Du, Y. Wang, “An image encryption algorithm based on spatiotemporal chaos”, International Congress on Image and Signal Processing, 2009, pp. 1–5.
[13] G. Chen, Y. Mao, C.K. Chui, “A symmetric image encryption schemes based on 3D chaotic cat maps”, Chaos Solitons and Fractals, 21(3) 2004, pp. 749–761.

[14] Z.H. Guan, F. Huang, W. Guan, “Chaos-based image encryption algorithm”, Physics Letters A, 346(1-3), 2005, pp. 153–157.

[15] T. Gao, Z. Chen, “Image encryption based on a new total shuffling algorithm”, Chaos Solitons & Fractals, 38(1), 2008, pp. 213–220.

[16] Xiaojun Tong, MinggenCui, “Image encryption scheme based on 3D baker with dynamical compound chaotic sequence cipher generator”, Signal Processing, 89, 2009, pp. 480–491.

[17] G. Zhang, Q. Liu, “A novel image encryption method based on total shuffling scheme”, Optics Communications, 284 (12), 2011, pp. 2775–2780.

[18] Y. Mao, G. Chen, S. Lian, “A novel fast image encryption scheme based on 3D chaotic baker maps”, International Journal of Bifurcation and Chaos, 14(10), 2004, pp. 3613–3624.

[19] L. Zhang, X. Liao, X. Wang, “An image encryption approach based on chaotic maps”, Chaos Solitons & Fractals, 24 (3), 2005, pp. 759–765.

[20] H. Gao, Y. Zhang, S. Liang, D. Li, “A new chaotic algorithm for image encryption”, Chaos Solitons & Fractals, 29(2), 2006, pp. 393–399.

[21] Q. Zhou, K.W. Wong, X. Liao, T. Xiang, Y. Hu, “Parallel image encryption algorithm based on discretized chaotic map”, Chaos Solitons & Fractals, 38(4), 2008, pp. 1081–1092.

[22] Z. Hua, Y. Zhou, C.M. Pun, C.L.P. Chen, “2D Sine logistic modulation map for image encryption”, Information Sciences, 297, 2015, pp. 80–94.

[23] L. Liu, S. Miao, “A new image encryption algorithm based on logistic chaotic map with varying parameter”, Springer plus, 2016, doi: 10.1186/s40064-016-1959-1.

[24] S.E. Borujeni, M. Eshghi, “Chaotic image encryption design using Tompkins-paige algorithm”, Hindawi Publishing Corporation Mathematical Problem in Engineering, 2009, 2009, 22 pages.

[25] H. Khanzadi, M.A. Omam, F. Lotfifar, M. Eshghi, “Image encryption based on gyrator transforms using chaotic maps”, IEEE 10th International Conference in Signal Processing (ICSP), 2009, pp. 2608–2612.

[26] Y. Wang, K.W. Wong, X. Liao, G. Chen, “A new chaos-based fast image encryption algorithm”, Applied Soft Computing, 11(1), 2011, pp. 514–522.

[27] H. Khanzadi, M. Eshghi, S.E. Borujeni, “Image Encryption Using Random Bit Sequence Based on Chaotic Maps”, Arabian Journal of Science and Engineering, 39(2), 2014, pp. 1039–1047.

[28] F. Zhou, G. Cao, B. Li, “Design of digital image encryption algorithm based on compound chaotic system”, Journal of Harbin Institute of Technology, 14 (Supplement 2), 2007, pp. 30–33.

THE METHODOLOGY OF DATA COLLECTING AND THE REAL TIME SYNCHRONIZATION IN ETL

Zijadin Krasniqi [1], Enver Ahmeti [2], Adriana Gjonaj [3]

[1] Ph.D. in Information Systems, Information Technology, Pristina, Kosovo.
[2] M.sc in Information Technology, Pristina, Kosovo.
[3] Professor emeritus, European University of Tirana, Albania.

[1] zijadinkrasniqi@hotmail.com, [2] eenver_ahmetii@hotmail.com,[3] adriana.gjonaj@uet.edu.al

***Abstract*-Taking into account the problems and challenges which the Emergency Management Agency (EMA) is facing today, and the way operational data scattered across various institutions is being used, we have set an hypothesis to build a DataMart, which has the capability for deep and detailed data analysis. Here we shall present the step by step modeling of this DataMart, following the Kimball model, which will be useful for detailed analysis by medical, hydro-meteorological and seismic institutes. In collecting, transforming and loading the data we have applied the ETL methods, which are considered as the most critical tasks in building a DWH system. Then the development of a professional modeling in Talend Open Studio for Big Data, which enabled the integration of data from many data sources. In addition to this, we will have the possibility to synchronize and build the data in a way that DWH will be able to communicate in real time, something which has not been possible until now in the field of emergency management**

Keywords: ETL, DataMart, TOS and Big Data, DWH.

I. INTRODUCTION

Upon building the data mart model, we have to develop all the processes which will enable to gather the information from the original sources, then to transform and load it in its destination [7]. These three processes are considered to be the most complex processes during the phase of building the ETL systems. [5] In accordance with the specific architecture, the destination of the data will be a data mart built for the process of managing the emergency and natural hazard situations.

In a nutshell, real time data warehouses aim for decreasing the time it takes to make decisions and try to attain zero latency between cause and effect for that decision, closing the gap between intelligent reactive systems and system processes. Our aim is transforming a standard DW using batch loading during update windows (during which analytical access is not allowed) into near zero latency analytical environment providing current data, in order to enable (near) real time dissemination of new information across an organization.[9].

II. DATA EXTRACTING

The first part of an ETL process involves extracting the data from the source system(s). In many cases this represents the most important aspect of ETL, since extracting data correctly sets the stage for the success of subsequent processes. In general, the extraction phase aims to convert the data into a single format appropriate for transformation processing. [4]

To keep the data up to date in data warehouse, data has to be extracted several times in a periodic manner. There are two logical methods for extraction: Full extraction and incremental extraction.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

1. Full extraction – In this type of extraction, data from the source systems is completely extracted. As full extraction extracts all the data from source systems there is no need to keep track of changes made in the source system with respect to previous extraction.

2. Progressive extraction – In this type of extraction only the changes made to the source systems will be extracted with respect to previous extraction. This comparison results in large performance impact on data warehouse ETL processes.

There are two physical methods of extraction: Online extraction and Offline extraction.

-Online extraction – The extraction process of ETL connects to source system to extract the source tables or store them in a pre-configured format in intermediary systems, e.g. log tables.

-Offline extraction – The data extracted is staged outside the source systems (ETL Data Extraction Methods – Part Two) [6].

A. *Transformation*

In the data transformation stage, a series of rules or functions are applied to the extracted data in order to prepare it for loading into the end target. Some data does not require any transformation at all [4].

B. *Data loading*

The load phase loads the data into the end target that may be a simple delimited flat file or a data warehouse Depending on the requirements of the organization, this process varies widely. Some data warehouses may overwrite existing information with cumulative information; updating extracted data is frequently done on a daily, weekly, or monthly basis. The timing and scope to replace or append are strategic design choices dependent on the time available and the EMA needs [6].

III. DATA SYNCHRONIZATION

Data synchronization, usually, is needed to keep two or more systems synchronized and up to date. A common reason to synchronize data is system migration (as a result of consolidating or reducing the business etc.). In this situation, it is required to employ two parallel systems for a period of time, sometimes even for several years
This makes real time data warehouse (RTDW) support a critical issue in this kind of applications. The demand for updated data in data warehouse has always been very strong. Data warehouse update (the integration of the latest data) traditionally is done using the off-line method. This means that while undergoing updating processes, the users and the OLAP applications cannot use these data. Still, the users are in search of higher updating levels, taking into consideration the fact that the agency of natural hazards operates at 24x7 time shifts. Active data storage refers to a new trend, where the data warehouse is refreshed as many times as possible, because of great user demand to have access to the latest refreshed data [2].

CASE STUDY-MEDICAL INSTITUTE OF KOSOVO (MIK)

A. *The tool used for ETL–Talend Open Studio for Big Data*

The aim of this chapter is to design and implement ETL processes through the data integration of heterogonous databases using Talend Open Studio (TOS).
The usage of Talend Open Studio and Big Data **(**TOS) for designing and semi-autonomous implementing ETL tasks in Java enables a fast way of adopting to new data sources. By doing this we are able to integrate various sources of heterogeneous academic data sample and populate them in one central repository using Talend Open Studio and Big Data (TOS). It is important to ensure the capabilities of open source ETL are equal to any commercial products available, because it will help in implementing DW project with lower costs [1].

Figure 1: The components of Talend Open Studio for Big Data.

Based on the feature of data warehouse system separation, the ETL processes have to be done on a different database, which is independent from the source and the data warehouse. The application of the ETL techniques, by means of the Talend Open Studio

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

for Big Data component, will be done according to the following steps:

Step 1: Explaining the possibilities of extracting the data from the sources

The source identification is evident, given that the files generated from the platform will serve as the source of the information. If our focus is on the files generated by the Medical Institute of Kosovo (MIK), then these are unstructured files. So, we create a fact table: Tabela_Fact_MIK which contains the data.Five dimension tables connect to Tabela_Fact_MIK and they are the following:Dim_Address, Dim_Blood_group,Dim_Data,Dim_Time and Dim_Health_institu.Dim_Address consist of data regarding dwelling, such as address,residence, contact, phone number, e-mail.
The other table Dim_Blood_group,contains data regarding the blood group, the blood group type and the allergies.Dim_Data contains all types of data, including the following data: the date when the treatment starts, the date the treatment ends. These relate to the medical treatment. While, the other data, such as the Date of Birth, the Day of the Month, the Day of the Year, the Day of the Week, the Number of the Month and the Year show various dates as needed. The table Dim_Time contains time measured by: Second, Minute, Hour, Departure Time, and Arrival Time. The table Dim_Health_Institu contains specific data on: Name and Place of Institution, Department, Unit.

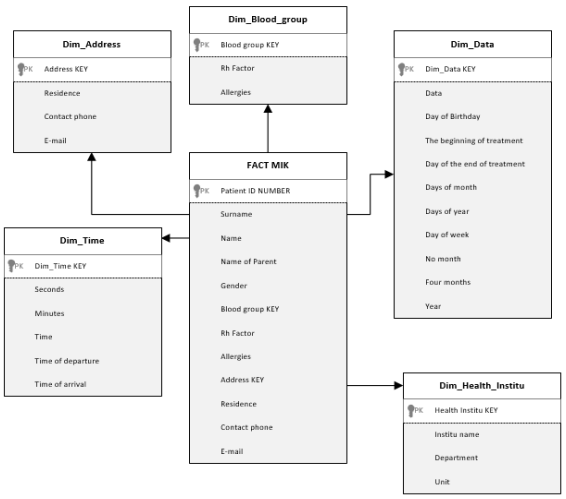


Figure 2: It shows Entity Relationship Diagram of target systems.

In the following chart we can see the sources of all the dimensions,such as Dim_Address, Dim_Blood_group,Dim_Data,Dim_Time and Dim_Health_institu, which are Excelfile. We can see also the other source, the table Fact_MIK, built in ORACLE

Table I.: The DBMS source

#	**Source DBMS**	**Table Name**
1	Oracle	Fact_KHI
2	Excelfile	Dim_Time
3	Excelfile	Dim_Blood_Group
4	Excelfile	Dim_Data
5	Excelfile	Dim_Address
6	Excelfile	Dim_Health_Institu

This means that each record type in the source file has an assigned numbers of fields different from other types. We need to build a process based on the records of the Medical Institute of Kosovo (MIK). This process will then implement the data structure.

Meanwhile, another source of information, such as the Civil Registration Agency (CRA), generates structured information, where each record of the generated file has the same content. For our data mart, we are very interested on the information generated by MIK. The data extraction process from the source involves:
- The initial process of extracting information;
- The process of incremental extraction.

The initial process takes place when there is no information in the destination database, while the process of incremental extraction provides the extraction of that part of information which has been changed or added since the last version.

Step 2: Information structuring steps

Developing the structuring of information involves the following steps:
a) In the first step, we will divide the information in separate models, according to the rules of each type of information. The division will be done based on the rules of information differentiation.
b) The second step is the proper structuring and storing of data, in order to enable their further manipulation. In our case, the data will be stored in Stage tables, according to the chosen architecture. Based on the logic of transformation, we will set the rules to

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

structure the unstructured information at our disposal. [7].

c) The third step involves using tJoin and tMap operations in Talend Open Studio.It is most convenient to load the input data schemas into the *Repository*. The Repository maintains a list of input and output record structures (schemas) that can be referenced and maintained globally in the system.

B. tJoin component of Talend Open System

Joining two files consists of combining the fields (columns) of one data source with the fields of another data source whenever a key field matches. Joining is accomplished by the tJoin component in Talend Open Studio.

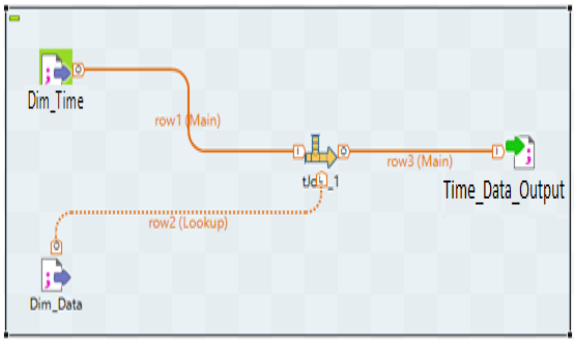


Figure 3: Shows a sample join job designed to combine two data sources

Each data source must be configured to access the appropriate data set according to the requested information in the Basic Settings pane at the bottom. To see the join conditions, double-click on the tJoin component to open the dialog box shown in Figure 4

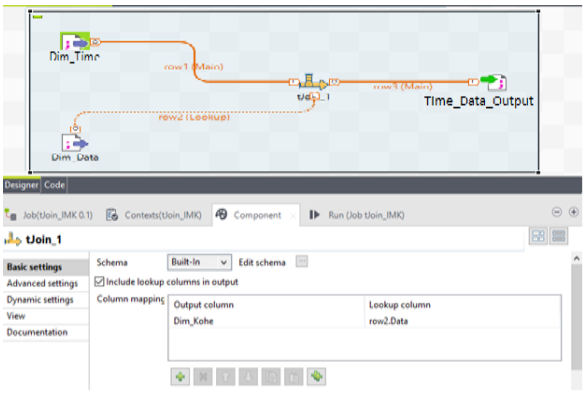


Figure 4: The tJoin component- dialog box

The tJoin dialog box, showing options to specify an inner join and to edit the schema of the join.

If you want to perform a join for areas B and C in Figure 5 (which constitutes a logical right join), just connect the Dim_Data. The common field in each data source (the key) is specified in the definition of the schema in the Repository.

To change the default mapping of input columns, click on the Edit schema box in the tJoin dialog box to display the mapping schema shown in Figure 5 (select View Schema at the prompt, and click OK).

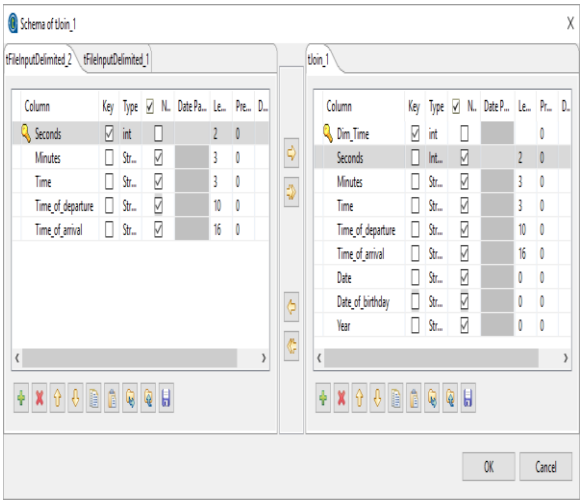


Figure 5: The tJoin field mapping schema for the Dim_Data data source.

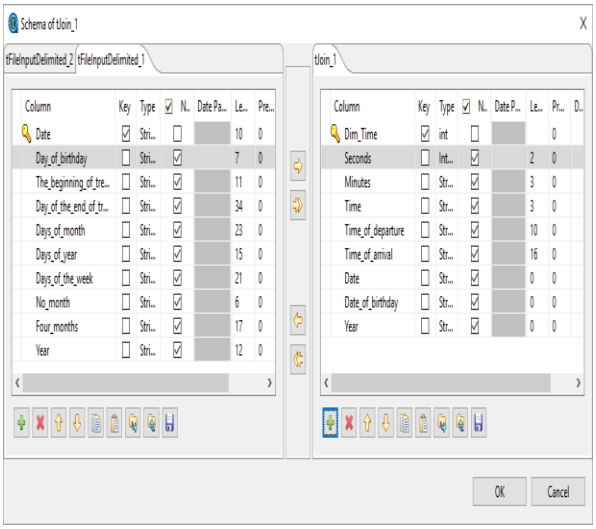


Figure 6: The tJoin field mapping schema for data source Dim_Time.

The goal of the join operation is to add the fields from data source Dim_Time to those of data source

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Dim_Data, resulting in a record for a given customer (Time_Data_Output) with all the fields included.

C. tMap component of Talend Open System

Figure 7 shows a job that uses the tMap component to map which input fields relate to each field in the output data structure. Fields can be repositioned in the data records with this component.

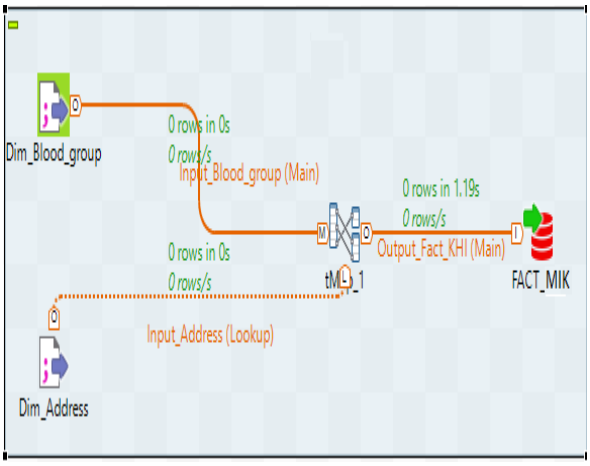


Figure 7. tMap component of Talend Open Studio.

The Figure 7 shows the Talend tMap component for mapping fields to an output data structure.
Double-clicking on the tMap_1 component displays the tMap configuration screen in the Figure 8 as below.

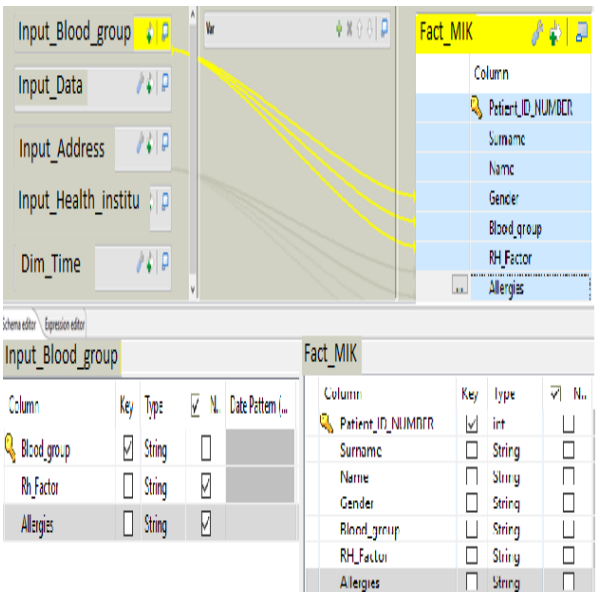


Figure 8. tMap configuration.

Step 3: Specifying the transformation that will happen

After extracting them from the sources, the data undergo the processes of transformation. The transform actions in our case relate to the data extracted from the sources or temporary objects from the Stage zone. The transformations itself can create temporary result that will be stored in the Stage zone.
The main transforming processes that will happen are:
- Conversion and normalization that works on both sides of the information to convert the uniform data.
- Joining process, which enables the logical join of equivalent fields from different sources;
- Selection, which reduces the number of the source fields and records

Step 4: Loading the data

During the process of loading the Fact_MIK table, we use the created references. All the references tables needed have to be stored in the memory, so they can be captured again any time a record of this Fact table is processed with the information of the natural key of dimension. Therefore this reference tables are different from the original tables of data mart's Dimension. The Figure 9 below shows the processes that the information undergoes before being loaded in a Fact table [7].

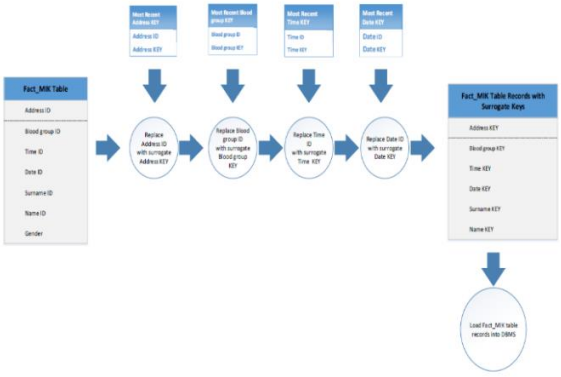


Figure 9: The diagram of loading the table Fact_MIK [3].

The purpose is to show that Talend Open Studio for Big Data (TOS) is capable to perform jobs as any commercial ETL tools.
To prove the proposed framework, we will use dummy data from two different DBMS, which are Oracle and Excelfile to test ETL process in academic environment. Other than that, some record files as Dim_Address,Dim_Blood_group,Dim_Date,Dim_

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Time and Dim_Health_institu as text files in Microsoft Excel format have been added as a source data to integrate with DBMS.

Oracle has been chosen as target database because most of enterprises companies are using Oracle in their enterprise applications; in order to load data in the table Fact_MIK we have built the following job in Talend Open Studio for Big Data, as shown in the Figure 10

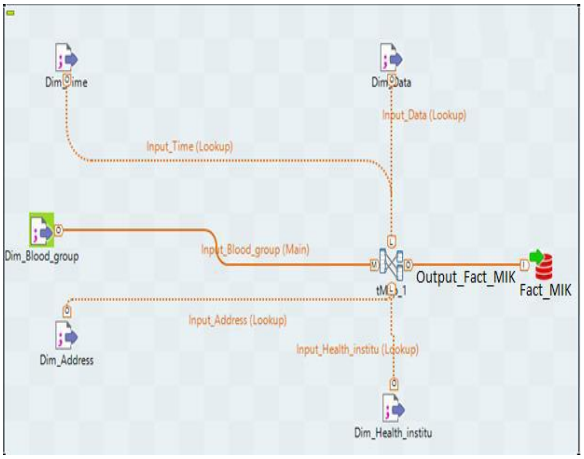


Figure 10. Talend's Job for loading the table Fact_MIK

Meanwhile, in Oracle, a table known as Fact_MIK is created to store the records of Medical Institute of Kosovo, including the patients' number, surname, name, maid's surname, gender, blood group, Rh factor, allergies, location, address, contact number, email, the prescribed medicaments.
Whereas Microsoft Excel files contain data about time, address, date, blood group and medical institution. [1].
During the loading process, we can add many more elements of information, such as the case of the fields after the planning, like surname, name, gender, needed to keep in this table the record's loading time in the database, as shown below.
This added element are not present in the source, but are generated by the system via standard functions of Talend Open Studio for Big Data (TOS).

The abovementioned fields serve to manage the loading process of information or to make examinations in case there might be any problem. [7].

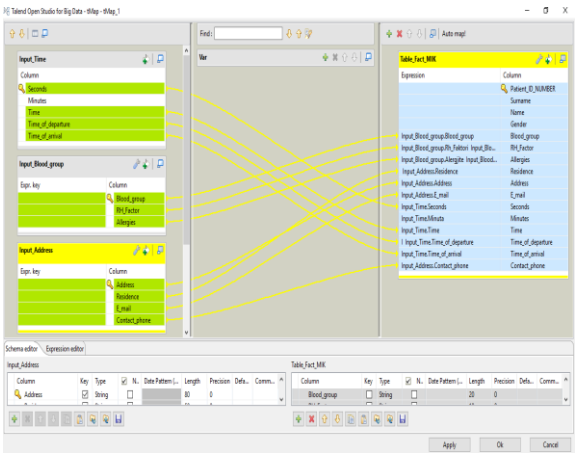


Figure 11 TOS Job to add information.

This means that any record in the original source will have an assigned number of fields different from other types. We need to create a precedence process based on the records of MIK. This process will enable the structuring and adding of the data, as shown in the Figure below.

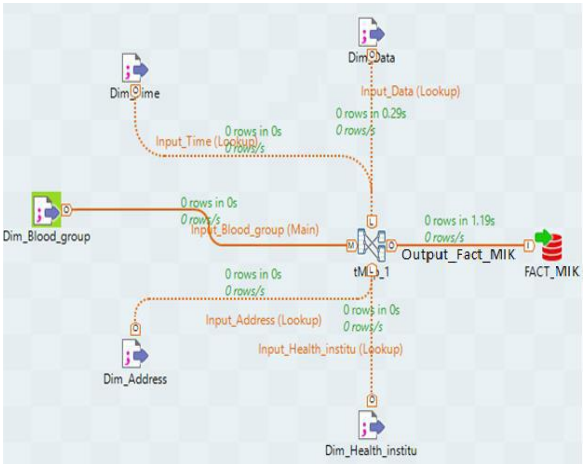


Figure 12. The TOS Job for adding and loading information.

IV. THE RESULTS DERIVED FROM THE ANALYSIS IN THE CASE OF MIK DATA MART

The analysis for the MIK data mart case involves the MIK data mart, which was built to enable exchanging and integrating the data from many dimensions (Dim_time,Dim_blood_group,Dim_address,Dim_date,Dim_health_institu)in a single fact table (FACT_MIK).
Thus, the data mart model above offers high and fast performance in generating the desired results and information, which enables performing analysis within the required deadline. Similar to this data mart model, we can build other data mart models which will

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

serve to generate information in internet and in small mobile devices.

Taking into consideration the MIK data mart, we can do various analyses regarding the communication form in the Medical Institute of Kosovo. If we take as a cause the requirements of the medical personnel regarding the information relating to the patients, we can easily have the answers for these requirements, simply based on the dimension tables.

Here we can include the time required by the procedures which enable extraction, loading and transforming data. The following table shows the result we got from these procedures:

Table II: Comparing the jobs execution.

#	Job Name	Frequency	Duration
1	Dim_Data	Rows/s=600;ByHour=0;ByMinute=0;BySecond=0.1	00:00:01
2	Dim_Time	Rows/s=189.19;ByHour=0;ByMinute=0;BySecond=0.4	00:00:04
3	Dim_Blood_Group	Rows/s=160;ByHour=0;ByMinute=0;BySecond=0.3	00:00:03
4	Dim_Address	Rows/s=3000;ByHour=0;ByMinute=0;BySecond=0	00:00:00
5	Dim_Health_Institu	Rows/s=6;ByHour=0;ByMinute=0;BySecond=0	00:00:00
6	Fact_MIK	Rows/s=1379.41;ByHour=0;ByMinute=1;BySecond=0.6	00:01:06

Looking at the above results, it is evident that Dim_Data, from 600 rows/s are loaded in 0.1 second, Dim_Time, from 189 rows/s are loaded in 0.4 seconds, Dim_Blood_Group,from 160 rows/s are loaded in 0.3 second, Dim_Health_Institu, from 6 rows/s are loaded in zero second. Thus, all the data from all the dimensions is loaded into Fact_MIK table with 1397 rows/s at a total time of 1.6 min. The MIK data mart case depends on the update of this data warehouse (DWH) within a specific time interval (ore in continuity). Failure to fulfill that responsibility puts the reliability and dependability of the data warehouse MIK in question. Therefore, the data warehouse cannot succeed without efficient and persistent data feeds. [8]

CONCLUSION

The fact is that the Emergency Management Agency possesses a large amount of data, which can easily go up to many megabyte or gigabyte per day. In fact, it is very difficult to run an effective update of this data, so through theoretical explanation based on literature and scientific facts, we have been able to apply real time data synchronization. The ETL process extracts data from the source systems, transforms the data according to the rules of the institution and assigns the results to a target DWH. Talend Open Studio and Big data software, which enables the integration of data from many different sources, providing a positive example of integration and data modeling.

REFERENCES

[1] A. Azwa, W.Abdul, H. Nazi rah, "integration of heterogeneous databases in academic environment using open source ETL tools", February 4, 2015

[2] A.Hoffer, V. Ramesh, H. Topi "Modern database management-Tenth Edition", Pearson, New Jersey, 2011.

[3] C. Johnson, "Surrogate Key in Data Warehouse, What, When and Why", 2013.

[4] ETL "Extract transform load "May 12, 2015.

[5] E. Johnson, J. Jones, **"**Building ETL Processes for Business Intelligence Solutions Built on Microsoft SQL Server", March 12, 2015

[6] K., Kakish Th. Kraft , "ETL Evolution for Real-Time Data Warehousing", New Orleans Louisiana, Vol. 5,2012, pp 1-12.

[7] L. Juliana," Studying data warehouse systems and building a reporting model based on business intelligence technology", Tirana, 2014

[8] R. Kimball, J. Caserta, "The Data Warehouse ETL Toolkit-Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data", Wiley Publishing, USA, 2004.

[9] R. Santos, J. Bernardino, "Real-Time Data Warehouse Loading Methodology, ACM, Portugal, 2008.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Conceptual Model of Intelligent Collaborative Educational System: Possible Solutions

Sabina Katalnikova
Faculty of Computer Science and Information Technology
Riga Technical University
Riga, Latvia
sabina.katalnikova@rtu.lv

Leonids Novickis
Faculty of Computer Science and Information Technology
Riga Technical University
Riga, Latvia

***Abstract* - In modern society, an increasingly important role is played by knowledge workers whose professional competence is one of the basic resources of the company and organisation evolution. In order to be competitive, knowledge workers have to keep up with new knowledge, which makes the process of continuous education of employees as one of the priorities in the staff management. One type of lifelong education - collaborative education - provides benefits for cognitive achievements. The Article deals with the characteristics, basic tasks and the operational process of collaborative education. The conceptual model of intelligent collaborative educational system is presented; the functions of system basic models and the scheme for setting up the training scenario are described. Besides, the focus is on the group formation issues.**

Keywords - collaborative education; competence approach; intelligent collaborative educational system; knowledge worker

I. INTRODUCTION

In today's society, no one can call into question the effect of new technologies on the educational system. In this new reality, at any level of education, the study process should be organised in such a way that graduates become the so-called knowledge workers in the full sense of this term [7]. This means that the training becomes more learner-centred, providing training opportunities for individual and professional development purposes, and educational institutions seek for flexible and dynamic response to changes on the labour market and in daily life.

One of the characteristics of the 21st century is the transition to the knowledge economics and the transition of information to the source of economic growth and profit. The staff and their knowledge as well as their growing professional competence have become the basic resources of organization development. In order to be competitive, it is necessary for knowledge workers to continuously acquire innovative knowledge. This makes the process of staff continuous training as one of the priorities in the functions of staff management in modern organisations. Thus lifelong education is becoming increasingly important.

According to [14], lifelong education is the process of personal, social and professional development of the individual throughout their lifetime to improve the quality of life. It is a comprehensive and uniting idea which includes formal, beyond-formal and non-formal education.

Lifelong education is a continuous process in which a person is involved from youth to old age, it requires flexibility of mind, the desire to complement and broaden the knowledge as well as the desire for continuous development as a full and versatile person, and it enables an individual to adapt to the changes on the modern labour market. The rationale for adults to engage in lifelong education can be varied, both internal and external. Internal motives refer to the wishes of the person to acquire new competences; the external motives are the various demands of person's workplace.

On the other hand, information and communication technology opportunities and capacity are constantly increasing, leading to a continuous increase in the volume and variety of service activity in any area, including education. Therefore, the concept of collaborative educational system has appeared in the modern society. On the modern labour market, there is an increasing demand for creative, responsible, dynamic and well-prepared employees who can quickly adapt to changes, who know how to cooperate with other employees, who are open to different cultures, with flexible thinking, etc. [1]. In this context, collaborative education is becoming increasingly important.

II. COLLABORATIVE EDUCATION WITHIN THE FRAMEWORK OF COMPETENCE APPROACH

Collaborative education can be defined as a training environment in which students participate with a purpose to tackle the challenges together [16]. Collaboration activities enable students to provide explanations based on their understanding, which assist them in developing and reorganising their knowledge [3]. Collaborative education provides the following advantages [13]:

- the development of critical thinking skills,
- joint creation of knowledge and meaning,
- deliberation,
- transformational leaning.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

There are four basic elements of collaborative education [5]:

- positive interdependence – students must be fully involved in the work and strive to achieve the common objective of the group; in this case, each team member has an assignment, and the student understands that the execution of an individual task affects the operational effectiveness of the whole group;
- "eye to eye" interaction – each member contributes to the success of other participants. Students explain to each other what they have learned and help each other to understand and complete the task;
- social skills – collaboration in the field of education to be successful, everybody needs to develop effective communication types;
- group evaluation – in the process of collaborative education, the group often has to assess operational efficiency and to decide how to increase it; in order to achieve a significant improvement, students have to possess two features: each of them is working for the group aim, and success depends on everybody's results; the contribution and responsibility of each student are recognized in the assessment.

Another topical issue at present: competence approach in education. For a long time, the qualifications of employees were defined by their knowledge, based on what the person was studying and how long. However, at present a knowledge-based approach is not topical because nowadays, the specialist's abilities after studying matter as well as their ability to respond adequately to the changes in the situation and to act creatively.

Another cause of the decline in the importance of knowledge is linked to global Web development because at the moment, information is available to anyone without any expert mediation.

The idea of knowledge, skills and abilities makes it possible to act by analogy with the model, but the competence is based on independent activities at the basis of universal knowledge. As a conceptual definition, it is possible to adopt the term "competence" as defined in the European Union project TUNING: The term "competence" includes knowledge and understanding (theoretical knowledge in an academic field, ability to know and understand), the knowledge of how to take action (practical application of knowledge in particular situations) and the knowledge of how to exist (values as an integral part of life in a social context).

This European study in the field of the assessment of qualifications and competence is a part of international project Tuning Educational Structures. The project is based on the introduction of the suggestions of Bologna Process, the harmonisation of the parameters of education programmes at all levels as well as an improvement in the interaction of higher education and entrepreneurship being at the basis of competence approach [18].

The basic principles of competence approach in education are as follows:

- constant updating of the study programme in accordance with the requirements of the labour market;
- development a study programme according to the individual student's needs;
- necessity of serious motivation for the acquisition of theoretical knowledge and practical skills in the chosen field;
- evaluation of the will to study (strengths and weaknesses of the student, individual characteristics, etc.).

How should competences be created? In general, the development of one competence must correspond to a number of different disciplines, grouped in a specific way (of course, there are situations where one competence is generated within the course of one study course, or even within a part of the course). A set of study courses constituting a single competence or a group related competence is referred to as a module. The formation of one competence can include studying different disciplines, participation in student conferences, various kinds of traineeship and unassisted work.

As it has been said, collaborative education is a topical issue in today's society, the main economic resource of which is knowledge and where knowledge workers are of great importance. The term "knowledge workers" was introduced by Peter Ferdinand Drucker – a scientist, economist, educator (USA) - in 1959 [12] to describe an ever-increasing number of employees working mainly with information or using knowledge. The key aspects of the operations of knowledge workers are innovation, responsibilities and collaboration that have led to the great importance of collaborative educational systems for knowledge workers. The above approach to the implementation of the concept of lifelong learning is therefore complemented with a need to take into account the characteristics of knowledge workers.

In general, the following approach is proposed to implement the concept of lifelong education (see figure 1):

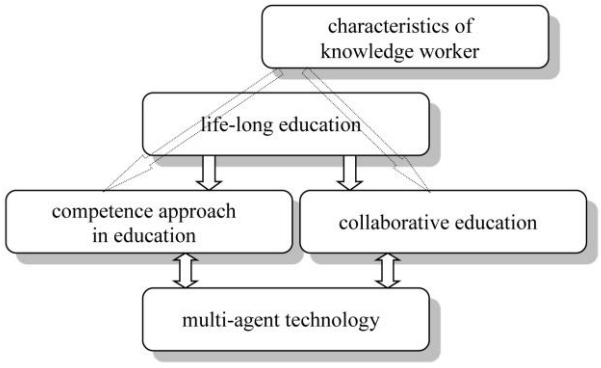


Figure 1. Proposed approach to the implementation of an idea for the lifelong education of knowledge workers

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

III. Systems of Collaborative Education

A distinctive feature of collaborative educational systems is the development of domain knowledge, the use of individualised education and training strategies based on the model of the user. The benefit of collaborative educational systems is an advantage of independence from a student placement, but the difference between the classes of users is an important problem, as the system should be used for a very diverse set of students.

Systems of collaborative education possess the following characteristics:

- the aim of the system is to provide students with adequate support in the process of problem solving as an instructor-human would have done;
- any system of collaborative education is based on a user (student) model, taking into account the set of characteristics of the adults as well as the current level of professional competence and knowledge;
- at the basis of studies, there is the underlying study programme based on the student characteristics and needs, the purpose of education and the required range of competencies;
- education is geared to the needs of employers;
- during studies, there are developed skills for the user to apply knowledge for specific practical tasks;
- self-motivation is of great importance in such systems;
- education takes place in collaboration, promoting equal partnership.

In general, the key objectives of collaborative educational system are considered to be:

- the development of a model of competence based on existing standards and labour market requirements;
- student registration and development of his or her personal space;
- the development of the current model of each participant based on a deep and thorough diagnosis and a comparison with the competence model with a view to determining the goal of the education process of each participant;
- development of a training strategy (programme) at the basis of the goal of student education process;
- formation of collaborative education groups, taking place separately for each study course module based on the current status and personal characteristics of students;
- developing the content (scenario) based on the existing training strategy and the groups formed;
- formation of teaching staff and auxiliary staff team, based on employee characteristics and study programmes;
- determination of labour-intensity and labour input of the training scenario;
- implementation of training, support for maintaining the training process, formation of teaching-methodology group;
- implementation of control, updating of the student model, determination and evaluation of study results.

The operational process of the collaborative educational system is summarized below (figure 2).

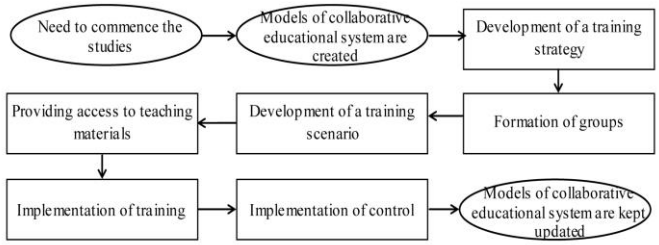


Figure 2. Process of the collaborative educational system

As part of this process, an inspection is carried out as to whether each unit established corresponds to the model of the student (specialty, training objectives and the level of knowledge, skills and competences) and the model of competence. If there is conformity, the unit is considered to be created, otherwise the process will recur.

IV. Conceptual Model for the Intelligent Collaborative Educational System

The term "intelligent system" is used in the field of artificial intelligence and identifies a system using artificial intelligence with a view to providing better support for system users. There is no uniform definition of intelligent system at present, for example [20] gives the following explanation: "The technical or software system, the operational resolution of which is traditionally regarded as creative and belonging to a particular item of understanding, the knowledge of which is kept in the system memory; an intelligent system consists of three main blocks: the knowledge base, solver (output mechanism) and an intelligent interface".

Mostly intelligent systems are knowledge based systems and focus on processing knowledge (not on processing data or information). Therefore, the intelligent system is a system based on knowledge, i.e. a domain model, described in the language of knowledge interpretation (ultra-high level programming language, close to natural language) [21].

Intelligent systems are widely used in different fields to meet versatile challenges. One of such tasks is planning – organisation of a system with a view to achieving a defined objective taking into account certain restrictions. Intelligent educational systems also belong to these systems.

A possible conceptual model for the intelligent collaborative educational system, namely, the main sub-systems and data flow between them, is presented in figure 3:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

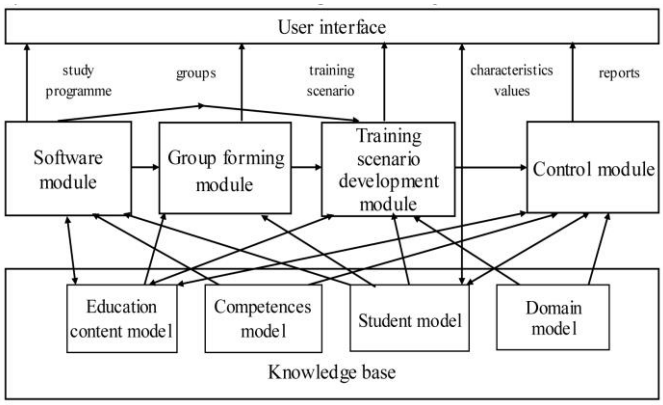


Figure 3. Conceptual model
of the intelligent collaborative educational system

The education content model includes the strategies (regulations on the behaviour of the system) of study programme and study content (scenario) development and management. The student model contains information about the student, the set of characteristics (personal data, level of knowledge, etc.). Competence model: a set of required competences based on industrial requirements comparing current and acquired competences of students. Domain model includes study objects (study courses, tasks, tests, control questions, etc.).

The models are in accordance with the base modules, the rest of modules provide ancillary functions. This could be: module design and analysis module, module for the implementation of training, module for knowledge base validation (correctness check) module, etc.

We will review the functions of the base modules.

User interface:

- registration of users and development of personal space;
- access to the knowledge base;
- interaction with other modules;
- input of current values for the assessment of student competences, values for student characteristics, etc.

Module for programme development:

- comparison of the value of student's competence rating values with the value required for the assessment of competence (i.e., competence model);
- determination of the study programme (training strategy), taking into consideration the student's characteristics as well as the difference between the value of student competence assessments and the values required according to the assessment of competence, i.e. the definition of a set of study modules for each student.

Group formation module:

- based on the established programme for each student, his/her personal characteristics and current situation, collaborative groups are created for module studies of each study course;
- checking the adjustments of groups created for each study course module. If a group is not formed, the students outside groups are united in groups, not taking into account the characteristics and present situation.

Training scenario development module:

- sequence planning for study course modules;
- implementation of training, support for the maintenance of the study process.

Control module:

- implementation of the control of acquired knowledge;
- upgrading of the student module;
- generation of management solutions appropriate for study process to achieve study objectives;
- determination and evaluation of study results.

All the activities of the system are carried out, taking into account the limitations of the resources distributed, and directly linked to the student. Development of model training scenario could therefore be presented as shown in figure 4.

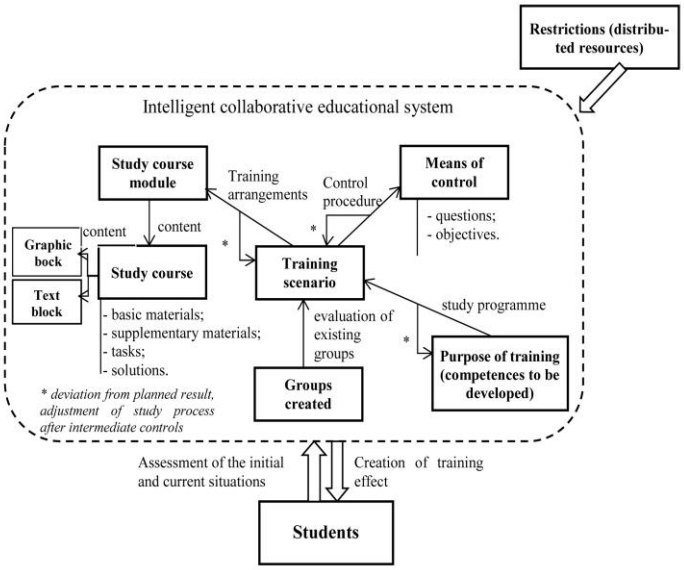


Figure 4. Diagram for development a training scenario

The following concepts are highlighted in figure 4:

- the text and graphical block is the minimum unit of presenting information in the text and in the graphic form respectively (e.g. paragraph, sentence of the text block, etc.);
- the study course allows the combination of different information in the form of text and graphical block in various structures;

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

- a study module is a set of study courses (it may also be traineeship) which has a logical completeness and that is intended to a certain competence, i.e. study of interrelated courses.

V. Questions Related to User Group Formation of Intelligent Collaborative Educational System

Collaborative education provides benefits for cognitive achievements, but their effectiveness depends on such factors as the previous knowledge of group participants, team composition and quality of explanations [8]. Without adequate prior knowledge, students fail to provide high-quality explanations or properly understand the perspectives provided by other members of the group [10].

One of the problems related to collaborative educational system: how to divide students into study groups and according to which criteria. One approach: dividing into groups based on the level of knowledge and current competencies of each member, taking into account also his/her cognitive and social capacity.

There are also a number of approaches in determining the optimum number of participants in the group. It is most often recommended to divide the participants into as small groups as possible, 2-5 people in a group [6, 9].

In the research [4], the author offers a review of literature sources related to research of collaborative education. It has been noted that, within the framework of collaboration, both interaction models and the success of collaboration vary across groups with different levels of participants' capacity. In the research, studies were considered in which groups with both a high spectrum and a narrow spectrum participated (where there were students with high – mean capabilities and mean – low capabilities). In the groups with a wide spectrum of abilities, high and low capacity participants represented the teacher–pupil relationship, while the medium level capacity participants stayed away. In the groups with narrow capabilities, medium capacity participants worked much better. In these groups, all the students had a tendency to participate actively in the work, and medium and low capacity participants showed a higher achievement. The research has shown that collaboration has a strong impact on student learning outcomes, especially as it relates to participants with low skills, and the groups with a narrow difference in capacity levels are the most successful ones.

Development of an efficient collaborative educational system includes sub-tasks: how to choose and quantify appropriate characteristics to develop a system user model to divide users into optimal groups for study purposes [2].

The research [11] provides a methodology for determining the level of competence of knowledge workers: the level of competence of staff members is calculated by means of an audit procedure based on different methods and techniques of competence analysis. From an audit point of view, each competence includes the title and attributes set to define it. Each attribute of a particular worker is assessed in some way (for example, through a survey or an interview). The collected attributes make it possible to calculate the staff member's level of competence.

Each employee *i* possesses a set of competences characterized by competence vector Com_i $[com_{1,i}, com_{2,i}, ..., com_{n,i}]$ where *n* is the number of competences and the value $com_{n,i} >= 0$ is the initial assessment of competences carried out with an audit procedure In addition, value $com_{n,i}$ can be changed after the loss or transfer of knowledge, training or other knowledge related processes.

Another question which has to be addressed arises: is competence a binary property or not? In the field of education, competence may be assessed by scale, with degree or "size". In the research [15], it is noted that competence may consist of several subdomains and may include the use of a number of related elements, so that a lower level competence can be considered binary, whereas a higher level competence - able to be partially completed. For example, the level of competence may be normalized and linked to the audit procedure of a specific employee: beginner (0-0,2), initiator (0,2-0,4), trainee (0,4-0,6), skilled employee (0,6-0,8), expert (0,8-1) and master (1) [17].

VI. Conclusion

There is no doubt that major changes in higher education are currently taking place. Educational institutions offer ever more lifelong education opportunities for lifelong education and adult education. The main aim of this activity is to ensure the availability of quality and efficient education for all. In the circumstances where demand for such opportunities is constantly rising, the importance of the intelligent collaborative educational system and development issues associated with it should not be exaggerated, since the development of education should take place directly from the "closed" education model to an open one based on the use of a virtual education environment as the main means of communication between students and educators.

The article deals with the concept and core functions of the concept of the intelligent collaborative educational system, as well as the conceptual model of the intelligent collaborative educational system has been promoted, which might be helpful to the developers of this type of intelligent systems.

References

[1] A.C. Damian, M.Georgescu. Collaborative and Self-directed Learning in a Virtual Campus Environment: a Potential Solution for Our Years? EcoForum, Vol. 3, No. 1 (2014). [Online]. Available at: http://ecoforumjournal.ro/index.php/eco/article/download/52/52 [Accessed: October 25, 2017.

[2] C. Long, Y. Qing-hong. A group Division Method Based on Collaborative Learning Elements // Proceedings of the 26th Chinese Control and Decision Conference, 2014, pp. 1701-1705.

[3] C. Van Boxtel, J. Van der Linden,G. Kanselaar,. Collaborative Learning Tasks and the Elaboration of Conceptual Knowledge. Learning and Instruction, 10(4), 2000, pp. 311–330.

[4] E.R. Lai. Collaboration: A Literature Review. Research Report, 2011, 49 p. [Online]. Available: images.pearsonassessments.com/images/tmrs/collaboration-review.pdf [Accessed: August 10, 2017].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[5] H. Brown, D.C. Ciuffetelli. Foundational methods: Understanding teaching and learning. Toronto: Pearson Education, 2009. – 584 p. ISBN: 978-0558371968.

[6] J. Bernacki, A. Kozierkiewicz-Hetmańska. The Comparison of Creating Homogeneus and Heterogeneus Collaborative Learning Groups in Intelligent Tutoring Systems// Proceedings of 7th Asian Conference of Intelligent Information and Database Systems, 2015. Part I, pp. 46-56.

[7] J. Grundspenkis. The Conceptual Framework for Integration of Multiagent Based Intelligent Tutoring and Personal Knowledge Management Systems in Educational Settings // Workshops on Business Informatics Research: International Workshops and Doctoral Consortium. – 2011. - pp. 143-157.

[8] J. Janssen, F. Kirschner, G. Erkens, P.A. Kirschner, and F. Paas. Making the black box of collaborative learning transparent: Combining process-oriented and cognitive load approaches. Educational Psychology Review, 22, 2010, pp. 139–154.

[9] J.K. Olsen, D.M. Belenky, V. Aleven, M. Ringenberg, N. Rummel, J. Sewall. Authoring Collaborative Intelligent Tutoring Systems // Proceedings of the Workshops at the 16th International Conference on Artificial Intelligence in Education, 2013, p.10.

[10] N.M. Webb, K.M. Nemer, A.V. Chizhik, and B. Sugrue. Equity Issues in Collaborative Group Assessment: Group composition and Performance. American Educational Research Journal, 35(4), 1998, pp. 607–651.

[11] P. Różewski, J. Jankowskia, P. Bródka, R. Michalski. Knowledge Workers' Collaborative Learning Behavior Modeling in an Organizational Social Network. Computers in Human Behavior, Volume 51, Part B, October 2015, p. 1248–1260.

[12] P.F. Drucker. Landmarks of tomorrow. Harper Business, USA, 1st edition. -1959. - p. 270.

[13] R.M. Palloff, K. Pratt. Collaborating Online: Learning Together in Community. San Francisco, CA: Jossey-Bass. 2004, p. 128. ISBN: 978-0-7879-7614-9.

[14] R.N. Dave. Foundation of Lifelong Education: Some Methodological Aspects. Hamburg, Pergamon. 1976. – 388 p. ISBN: 978-0-08-021192-3.

[15] S. Grant, R. Young. Concepts and Standardization in Areas Relating to Competence // International Journal of IT Standards and Standardization Research, 8(2), 2010, pp. 29-44, DOI: 10.4018/jitsr.2010070103.

[16] S.D. Teasley, J. Roschelle. Constructing a Joint Problem Space: The computer as a tool for sharing knowledge. In Lajoie, SP and Derry, SJ (eds), The Computer as a Cognitive Tool. Hillsdale, NJ, USA. Erlbaum, 1993, pp. 229–258.

[17] T. Farrington-Darby, J.R. Wilson. The nature of expertise: A review // Applied Ergonomics, 37(1), 2006, pp. 17-32.

[18] Tuning Educational Structures in Europe. [Online] Available at: http://www.unideusto.org/tuningeu/ [Accessed: August 2, 2017].

[19] J. Griesbaum. Mehrwerte des kollaborativen Wissensmanagements in der Hochschullehre: Integration asynchroner netzwerkbasierter Szenarien des CSCL in der Ausbildung. Hülsbusch, W, 2009. 480 p. [In German].

[20] A.N. Averkin, M.G. Gaaze-Rapoport, D.A. Pospelov. Explanatory Dictionary on Artificial Intelligence. M .: Radio and communication, 1992. – p. 256 [In Russian].

[21] G.V. Ribina. Basics of Development of Intelligent Systems. M.: Finance and Statistics, 2010, p. 432 [In Russian].

Develop a distributed Intrusion detection system based on Cloud Computing

Khoi Nguyen
Datic Lab, DaNang University
DaNang, Viet Nam
ntkhoi@dut.udn.vn

Trang Dang Thi
IT Faculty, PhamVanDong University
QuangNgai, VietNam
thuytrang1401@gmail.com

***Abstract*— An intrusion detection system (IDS) is a hardware device or software application that monitors network and/or system or host activities for malicious activities policy violations. IDS has been used widely in network systems to detect malicious behaviors which can harm system or computers. Snort is an open source network intrusion detection and prevention system. It can analyze real-time traffic analysis and data flow in a network. It is able to detect many different type of attack. In this paper, we introduce a Cloud IDS model as a solution to implementing a network Snort based on cloud computing. The Cloud IDS provides network IDS as a service over the internet which can be simple in deployment, maintenance, scalability without investing in the new infrastructure.**

Keywords: IDS; Snort; OpenStack; Cloud Computing

I. INTRODUCTION *(HEADING 1)*

In recent years, Cloud computing technology has developed and become popular with many providers like Amazon, Microsoft, Google, etc. The abilities to share resources among several users, dynamic scalability, and pay-per-use basis etc. bring the convenience to users. The network security is one of the most important parts of a network system, but the traditional technologies like firewall, encryption… is not good enough to protect the network against many new attacking mechanisms. The IDS has been proposed and widely used as an efficient security system to identify internal and external malicious behaviors. Although considered as an efficient tool to protect a network system, IDS has to analyze all packets in network to detect malicious behaviors. Therefore, the processing time will be high and may lead to congestions, and packets may be delayed or dropped. Besides, because of the complexity of deployment and operation, companies must invest in infrastructure and personnel training.

Intrusion detection system plays an important role in the security and perseverance of active defense system against intruder hostile attacks for any business and IT organization. IDS implementation in cloud computing requires an efficient, scalable and virtualization based approach. In cloud computing, user data and application is hosted on cloud service providers remote servers and cloud user has a limited control over its data and resources. In such case, the administration of IDS in cloud becomes the responsibility of cloud provider. Although the administrator of cloud IDS should be the user and not the provider of cloud services. In the paper [1][6] authors have proposed an integration solution for central IDS management that can combine and integrate various renowned IDS sensors output reports on a single interface. The intrusion detection message exchange format (IDMEF) standard has been used for communication between different IDS agent.

In this paper, we introduce Cloud IDS which is built as a SaaS (Software as a Service) to provide IDS as a service to cloud based user through the internet. This model helps users reduce costs and becomes simple in deployment and operation.

II. PROPOSED AN CLOUD IDS MODEL

A. *Intrusion Detection System (IDS)*

Intrusion Detection System (IDS) is a system that analyzes packets in a network to detect malicious activities or policy violations and produces reports to the administrator. The IDS can identity and detect malicious behaviors from inside (agents, customers...) or outside (hacker...)

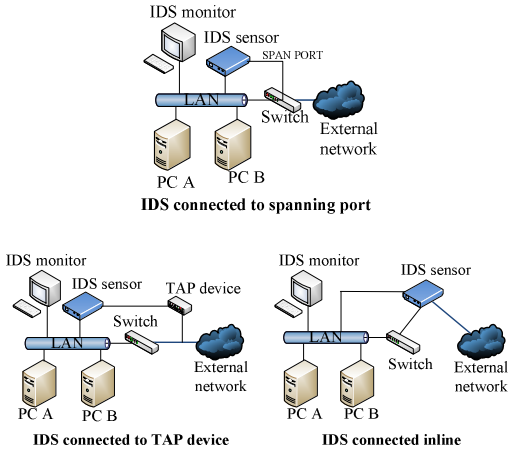


Figure 1. Network IDS deployment strategies

Beside many advantages, there are some challenges that face traditional IDS. For example, the user must invest in the new infrastructure, personnel training, etc. in order to deploy a network IDS. The IDS must capture and analyze all packets in the network to identify malicious activities. Therefore, in a large network it can lead to congestions and packets may be delayed or dropped.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

B. *Cloud computing*

Cloud computing is an innovative computing model in which resources are provided as a service over the internet based on the user demand. Users of cloud computing services can request, scale up, and use services without intervention of the provider. User doesn't need to invest in infrastructure, reduce maintenance cost and just pay only for the resources that they use. Cloud computing providers offer their services according to several fundamental models: IaaS, PaaS, SaaS [1][5].

C. *Proposed Model*

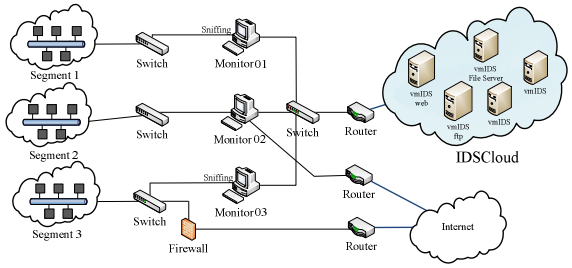


Figure 2. Proposed IDSCloud Model

In this section, we propose a model with the goal of providing IDS as a service to users in order to solve the problems of the traditional IDS and inherit advantage features of cloud computing. We named it CloudIDS.

In this proposed model, the user integrates a light weight sensor into the protected network and installs a central management application to monitor reports. The Detection Unit is rented from IDSCloud service. Based on specific demand, users can rent as many Detection Units as they need, customize the signature database. The User is the only one that completely possesses the analysis results. This model promises to provide users with an IDS with easy deployment, no costly investment in infrastructure, less maintenance cost, scalability...

Cloud IDS consists of four components:

- Capture component: the software or hardware is integrated *into the user network; the purpose of this* component is to capture packets in network. According to where this component is placed, it can capture packets in the whole network or network segment.
- Transportation component: this component consists of two elements, the transmitter and the receiver. The transmitter is placed at users' side and send collected packets to receiver, the elements which are placed at IDSCloud service. A Secure connection is established between two elements in order to protect the transmission of data from the external intrusion

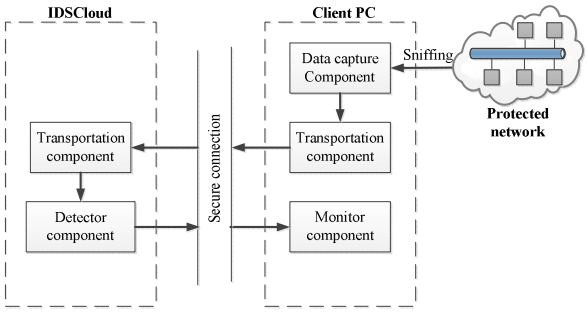


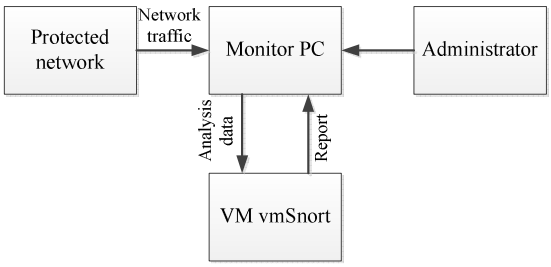
Figure 3. IDSCloud Architecture

- Detector component: The most important part of IDSCloud, a detector component analyses packets which the Transportation component received in order to detect malicious activities. Usually, Analysis process consists of four stages: packets decoder, preprocessor, engine detection, logging and alert system.
- Monitor component: This component is a real-time network monitoring tool. In order to guarantee that users possess analysis results and data does not lose when the detector unit IDS Cloud is terminated by user, the monitor component is placed at user's side and receive report and alert from detector component via a secure connection path.

IDS Cloud Requirements are as follows:

- Capture component: In order to protect user's network completely, IDS Cloud needs to monitor all the traffic on protected network; therefore, the capture component must be placed where it can capture all packets in protected network such as Test Access Port (TAP), Switched Port Analyzer(SPAN). It has the ability to work with high-speed networks and user can apply filter to it to reduce analyzed packets.
- Detector component: because the purposes of each user is different, this component should have ability to customize the detector to optimize the user's network data analysis. Besides, like other IDSs signatures database can be updated to improve detection ability.
- Secure connection: IDSCloud's user needs a secure connection to send and receive data with IDSCloud service. The secure connection can protect user's data from external intrusion.

In the next part, we are going to implement and evaluate our proposed IDSCloud model in local test.



https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Figure 4. IDSCloud activity diagram

1) Infrastructure as a service(IaaS): What IDSCloud provides is a computing resource that has the ability to analyse network packets and identify malicious activities. The computing resource is a virtual machine(VM) with the preinstalled IDS software, it will analyse data received from user and send reports to the central management. Openstack is used to deploy as an IaaS solution and aimed at users such as cloud service providers, goverment, company etc that want to implement large-scale public cloud computing, private cloud computing ot hydrid cloud computing.
2) IDS software: We chose Snort as the IDS software in the proposed model. Snort is one of the most popular free and open source IDS developed by Sourcefire. Snort is a signature-based NIDS but it can be expanded to use anomaly based detection plugins as well. The signature database is updated frequently and Snort can analyse protocols, header packets to identify many attack methods like buffer overflow, steath ports scan, CGI attacks... In this experimental scenario, Snort is configured to use DAQ_pcap_spooler as Data Acquisition library in order to monitor received folder and start analysing data file as soon as it arrives.
3) Secure Connection: The data transmission between clients and IDSCloud service is sent through Virtual Private Network(VPN) in order to ensure data is kept safe. Besides, users can use VPN to manage, config virtual machines at IDSCloud service. User's computer is connected to IDS VM through a VM named vmGateway.

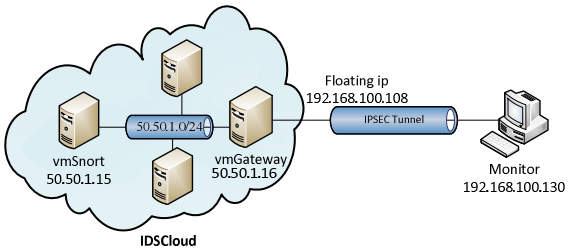


Figure 5. VPN Configuration with StrongSwan

III. EXPERIMENTAL EVALUATION

In order to evaluate the performance of the IDS Cloud model we have built at section 5, we propose a attack detection scenario and compare the performance result between IDSCloud and the traditional IDS. In figure 6 illustrate the experimental scenario.

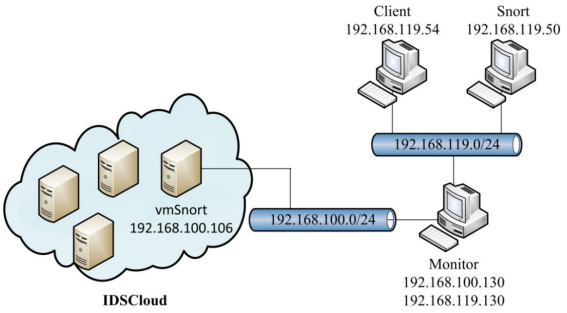
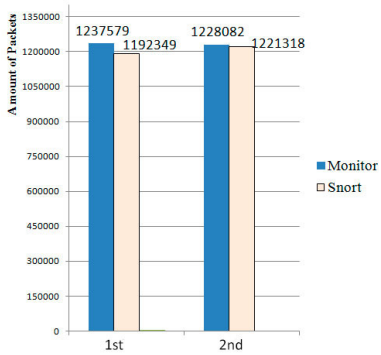


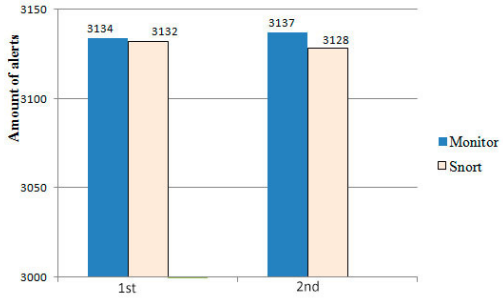
Figure 6. Experimental scenario

The evaluation environment consists of 4 PCs, the PC Monitor using IDSCloud service and PC Snort using IDS snort to protect network segment 192.168.119.0/24. The PC Client (192.168.119.54) will attack the protected network. The VM vmSnort is a virtual machine on IDSCloud. The Monitor PC,Snort PC and Client PC have same specification(1 CPU, 768MB RAM, OS Ubuntu 12.10 64 bit). The Snort PC and VM vmSnort use the same Snort version 2.9.6.0 with rule database version 2955 downloaded from Snort website.

The testing pcap data file contains 1.240.251 network packets which trigger snort and generate 3139 alerts with preinstalled rule database. The Client PC send packets to protect network by using Bittwist tool. We run evaluation case twice and evaluate the number of packets captured and alerts generated by two IDS. The results have been shown in Figure 7.



a) Number of captured packets comparison



b) Number of Alerts generated comparison

Figure 7. Compare the results between the traditional Snort system and the IDSCloud

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The results in Fig 7 show that the IDSCloud can capture more packets than the traditional Snort in the same condition (it also means the less dropped packets) and the result is that the IDSCloud detects more malicious activities.

IV. CONCLUSION

In this paper, we propose an Cloud IDS model as a solution to providing IDS as a service to users. This model has advantages over traditional IDS; it helps users reduce maintenance cost, dynamic scalability and pay-as-you-use feature. In the implementation and evaluation, we deploy IDSCloud model in virtualization environment and evaluate it with pre-captured network traffic file. The results show the effectiveness of proposed model but it is just the first step. There are many factors that can add an extra cost, so for the future works, we will research in the direction of simplifying deployment of Capture Component, optimizing the packet filter to reduce amount of data transferred between client and Cloud IDS, and applying more powerful detection mechanism to Snort as well as evaluating model on real internet environment.

REFERENCES

[1] Gartner (2013), Forecast Overview: Public Cloud Services, Worldwide, 2011-2016, 4Q12

[2] Hassen Sallay, Khalid A.Alshalfan, Ouissem Ben Fred (2009), A scalable distributed IDS Architecture for High speed Networks, International Journal of Computer Science and Network Security, VOL.9, No.8.

[3] Kevin Jackson (2012), OpenStack Cloud Computing Cookbook, Packt Publishing.

[4] Ms. Parag K. Shelke, Ms. Sneha Sontakke, Dr. A. D. Gawande (2012), Intrusion Detection System for Cloud Computing, International Journal of Scientific & Technology Research Volume 1.

[5] Richard Hill, Laurie Hirsch, Peter Lake, Siavash Moshiri (2013), Guide to Cloud Computing, Springer.

[6] S. Roschke, F. Cheng, and C. Meinel, "Intrusion detection in the cloud," in 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing. IEEE,2009, pp. 729–734.

AUTHORS PROFILE

Khoi Nguyen received the B.S. degree in Information Technology from the Danang University of Technology, Vietnam, in 1997, the M.Sc. degree in Computer Science in 2003, and the Ph.D. degree in Computer Science from the Aix-Marseille University in 2010. He now is lecturer at University of Science and Technology, Danang, Vietnam. His current research interests include Distributed System, Network Security.

Trang Dang Thi received the the M.Sc. degree in Computer Science in 2014. He now is lecturer at PhamVanDong University, QuangNgai province, Vietnam. His current research interests include Computer network, Network Security.

IJCSIS REVIEWERS' LIST

Assist Prof (Dr.) M. Emre Celebi, Louisiana State University in Shreveport, USA
Dr. Lam Hong Lee, Universiti Tunku Abdul Rahman, Malaysia
Dr. Shimon K. Modi, Director of Research BSPA Labs, Purdue University, USA
Dr. Jianguo Ding, Norwegian University of Science and Technology (NTNU), Norway
Assoc. Prof. N. Jaisankar, VIT University, Vellore,Tamilnadu, India
Dr. Amogh Kavimandan, The Mathworks Inc., USA
Dr. Ramasamy Mariappan, Vinayaka Missions University, India
Dr. Yong Li, School of Electronic and Information Engineering, Beijing Jiaotong University, P.R. China
Assist. Prof. Sugam Sharma, NIET, India / Iowa State University, USA
Dr. Jorge A. Ruiz-Vanoye, Universidad Autónoma del Estado de Morelos, Mexico
Dr. Neeraj Kumar, SMVD University, Katra (J&K), India
Dr Genge Bela, "Petru Maior" University of Targu Mures, Romania
Dr. Junjie Peng, Shanghai University, P. R. China
Dr. Ilhem LENGLIZ, HANA Group - CRISTAL Laboratory, Tunisia
Prof. Dr. Durgesh Kumar Mishra, Acropolis Institute of Technology and Research, Indore, MP, India
Dr. Jorge L. Hernández-Ardieta, University Carlos III of Madrid, Spain
Prof. Dr.C.Suresh Gnana Dhas, Anna University, India
Dr Li Fang, Nanyang Technological University, Singapore
Prof. Pijush Biswas, RCC Institute of Information Technology, India
Dr. Siddhivinayak Kulkarni, University of Ballarat, Ballarat, Victoria, Australia
Dr. A. Arul Lawrence, Royal College of Engineering & Technology, India
Dr. Wongyos Keardsri, Chulalongkorn University, Bangkok, Thailand
Dr. Somesh Kumar Dewangan, CSVTU Bhilai (C.G.)/ Dimat Raipur, India
Dr. Hayder N. Jasem, University Putra Malaysia, Malaysia
Dr. A.V.Senthil Kumar, C. M. S. College of Science and Commerce, India
Dr. R. S. Karthik, C. M. S. College of Science and Commerce, India
Dr. P. Vasant, University Technology Petronas, Malaysia
Dr. Wong Kok Seng, Soongsil University, Seoul, South Korea
Dr. Praveen Ranjan Srivastava, BITS PILANI, India
Dr. Kong Sang Kelvin, Leong, The Hong Kong Polytechnic University, Hong Kong
Dr. Mohd Nazri Ismail, Universiti Kuala Lumpur, Malaysia
Dr. Rami J. Matarneh, Al-isra Private University, Amman, Jordan
Dr Ojesanmi Olusegun Ayodeji, Ajayi Crowther University, Oyo, Nigeria
Dr. Riktesh Srivastava, Skyline University, UAE
Dr. Oras F. Baker, UCSI University - Kuala Lumpur, Malaysia
Dr. Ahmed S. Ghiduk, Faculty of Science, Beni-Suef University, Egypt
and Department of Computer science, Taif University, Saudi Arabia
Dr. Tirthankar Gayen, IIT Kharagpur, India
Dr. Huei-Ru Tseng, National Chiao Tung University, Taiwan
Prof. Ning Xu, Wuhan University of Technology, China
Dr Mohammed Salem Binwahlan, Hadhramout University of Science and Technology, Yemen
& Universiti Teknologi Malaysia, Malaysia.
Dr. Aruna Ranganath, Bhoj Reddy Engineering College for Women, India
Dr. Hafeezullah Amin, Institute of Information Technology, KUST, Kohat, Pakistan

CALL FOR PAPERS
International Journal of Computer Science and Information Security

IJCSIS 2017-2018
ISSN: 1947-5500
http://sites.google.com/site/ijcsis/

International Journal Computer Science and Information Security, IJCSIS, is the premier scholarly venue in the areas of computer science and security issues. IJCSIS 2011 will provide a high profile, leading edge platform for researchers and engineers alike to publish state-of-the-art research in the respective fields of information technology and communication security. The journal will feature a diverse mixture of publication articles including core and applied computer science related topics.

Authors are solicited to contribute to the special issue by submitting articles that illustrate research results, projects, surveying works and industrial experiences that describe significant advances in the following areas, but are not limited to. Submissions may span a broad range of topics, e.g.:

Track A: Security

Access control, Anonymity, Audit and audit reduction & Authentication and authorization, Applied cryptography, Cryptanalysis, Digital Signatures, Biometric security, Boundary control devices, Certification and accreditation, Cross-layer design for security, Security & Network Management, Data and system integrity, Database security, Defensive information warfare, Denial of service protection, Intrusion Detection, Anti-malware, Distributed systems security, Electronic commerce, E-mail security, Spam, Phishing, E-mail fraud, Virus, worms, Trojan Protection, Grid security, Information hiding and watermarking & Information survivability, Insider threat protection, Integrity
Intellectual property protection, Internet/Intranet Security, Key management and key recovery, Language-based security, Mobile and wireless security, Mobile, Ad Hoc and Sensor Network Security, Monitoring and surveillance, Multimedia security ,Operating system security, Peer-to-peer security, Performance Evaluations of Protocols & Security Application, Privacy and data protection, Product evaluation criteria and compliance, Risk evaluation and security certification, Risk/vulnerability assessment, Security & Network Management, Security Models & protocols, Security threats & countermeasures (DDoS, MiM, Session Hijacking, Replay attack etc,), Trusted computing, Ubiquitous Computing Security, Virtualization security, VoIP security, Web 2.0 security, Submission Procedures, Active Defense Systems, Adaptive Defense Systems, Benchmark, Analysis and Evaluation of Security Systems, Distributed Access Control and Trust Management, Distributed Attack Systems and Mechanisms, Distributed Intrusion Detection/Prevention Systems, Denial-of-Service Attacks and Countermeasures, High Performance Security Systems, Identity Management and Authentication, Implementation, Deployment and Management of Security Systems, Intelligent Defense Systems, Internet and Network Forensics, Large-scale Attacks and Defense, RFID Security and Privacy, Security Architectures in Distributed Network Systems, Security for Critical Infrastructures, Security for P2P systems and Grid Systems, Security in E-Commerce, Security and Privacy in Wireless Networks, Secure Mobile Agents and Mobile Code, Security Protocols, Security Simulation and Tools, Security Theory and Tools, Standards and Assurance Methods, Trusted Computing, Viruses, Worms, and Other Malicious Code, World Wide Web Security, Novel and emerging secure architecture, Study of attack strategies, attack modeling, Case studies and analysis of actual attacks, Continuity of Operations during an attack, Key management, Trust management, Intrusion detection techniques, Intrusion response, alarm management, and correlation analysis, Study of tradeoffs between security and system performance, Intrusion tolerance systems, Secure protocols, Security in wireless networks (e.g. mesh networks, sensor networks, etc.), Cryptography and Secure Communications, Computer Forensics, Recovery and Healing, Security Visualization, Formal Methods in Security, Principles for Designing a Secure Computing System, Autonomic Security, Internet Security, Security in Health Care Systems, Security Solutions Using Reconfigurable Computing, Adaptive and Intelligent Defense Systems, Authentication and Access control, Denial of service attacks and countermeasures, Identity, Route and

Location Anonymity schemes, Intrusion detection and prevention techniques, Cryptography, encryption algorithms and Key management schemes, Secure routing schemes, Secure neighbor discovery and localization, Trust establishment and maintenance, Confidentiality and data integrity, Security architectures, deployments and solutions, Emerging threats to cloud-based services, Security model for new services, Cloud-aware web service security, Information hiding in Cloud Computing, Securing distributed data storage in cloud, Security, privacy and trust in mobile computing systems and applications, **Middleware security & Security features:** middleware software is an asset on
its own and has to be protected, interaction between security-specific and other middleware features, e.g., context-awareness, **Middleware-level security monitoring and measurement:** metrics and mechanisms for quantification and evaluation of security enforced by the middleware, **Security co-design:** trade-off and co-design between application-based and middleware-based security, **Policy-based management:** innovative support for policy-based definition and enforcement of security concerns, **Identification and authentication mechanisms:** Means to capture application specific constraints in defining and enforcing access control rules, **Middleware-oriented security patterns:** identification of patterns for sound, reusable security, **Security in aspect-based middleware:** mechanisms for isolating and enforcing security aspects, **Security in agent-based platforms:** protection for mobile code and platforms, Smart Devices: Biometrics, National ID cards, Embedded Systems Security and TPMs, RFID Systems Security, Smart Card Security, Pervasive Systems: Digital Rights Management (DRM) in pervasive environments, Intrusion Detection and Information Filtering, Localization Systems Security (Tracking of People and Goods), Mobile Commerce Security, Privacy Enhancing Technologies, Security Protocols (for Identification and Authentication, Confidentiality and Privacy, and Integrity), Ubiquitous Networks: Ad Hoc Networks Security, Delay-Tolerant Network Security, Domestic Network Security, Peer-to-Peer Networks Security, Security Issues in Mobile and Ubiquitous Networks, Security of GSM/GPRS/UMTS Systems, Sensor Networks Security, Vehicular Network Security, Wireless Communication Security: Bluetooth, NFC, WiFi, WiMAX, WiMedia, others

This Track will emphasize the design, implementation, management and applications of computer communications, networks and services. Topics of mostly theoretical nature are also welcome, provided there is clear practical potential in applying the results of such work.

Track B: Computer Science

Broadband wireless technologies: LTE, WiMAX, WiRAN, HSDPA, HSUPA, Resource allocation and interference management, Quality of service and scheduling methods, Capacity planning and dimensioning, Cross-layer design and Physical layer based issue, Interworking architecture and interoperability, Relay assisted and cooperative communications, Location and provisioning and mobility management, Call admission and flow/congestion control, Performance optimization, Channel capacity modeling and analysis, Middleware Issues: Event-based, publish/subscribe, and message-oriented middleware, Reconfigurable, adaptable, and reflective middleware approaches, Middleware solutions for reliability, fault tolerance, and quality-of-service, Scalability of middleware, Context-aware middleware, Autonomic and self-managing middleware, Evaluation techniques for middleware solutions, Formal methods and tools for designing, verifying, and evaluating, middleware, Software engineering techniques for middleware, Service oriented middleware, Agent-based middleware, Security middleware, Network Applications: Network-based automation, Cloud applications, Ubiquitous and pervasive applications, Collaborative applications, RFID and sensor network applications, Mobile applications, Smart home applications, Infrastructure monitoring and control applications, Remote health monitoring, GPS and location-based applications, Networked vehicles applications, Alert applications, Embeded Computer System, Advanced Control Systems, and Intelligent Control : Advanced control and measurement, computer and microprocessor-based control, signal processing, estimation and identification techniques, application specific IC's, nonlinear and adaptive control, optimal and robot control, intelligent control, evolutionary computing, and intelligent systems, instrumentation subject to critical conditions, automotive, marine and aero-space control and all other control applications, Intelligent Control System, Wiring/Wireless Sensor, Signal Control System. Sensors, Actuators and Systems Integration : Intelligent sensors and actuators, multisensor fusion, sensor array and multi-channel processing, micro/nano technology, microsensors and microactuators, instrumentation electronics, MEMS and system integration, wireless sensor, Network Sensor, Hybrid

Sensor, Distributed Sensor Networks. Signal and Image Processing : Digital signal processing theory, methods, DSP implementation, speech processing, image and multidimensional signal processing, Image analysis and processing, Image and Multimedia applications, Real-time multimedia signal processing, Computer vision, Emerging signal processing areas, Remote Sensing, Signal processing in education. Industrial Informatics: Industrial applications of neural networks, fuzzy algorithms, Neuro-Fuzzy application, bioInformatics, real-time computer control, real-time information systems, human-machine interfaces, CAD/CAM/CAT/CIM, virtual reality, industrial communications, flexible manufacturing systems, industrial automated process, Data Storage Management, Harddisk control, Supply Chain Management, Logistics applications, Power plant automation, Drives automation. Information Technology, Management of Information System : Management information systems, Information Management, Nursing information management, Information System, Information Technology and their application, Data retrieval, Data Base Management, Decision analysis methods, Information processing, Operations research, E-Business, E-Commerce, E-Government, Computer Business, Security and risk management, Medical imaging, Biotechnology, Bio-Medicine, Computer-based information systems in health care, Changing Access to Patient Information, Healthcare Management Information Technology. Communication/Computer Network, Transportation Application : On-board diagnostics, Active safety systems, Communication systems, Wireless technology, Communication application, Navigation and Guidance, Vision-based applications, Speech interface, Sensor fusion, Networking theory and technologies, Transportation information, Autonomous vehicle, Vehicle application of affective computing, Advance Computing technology and their application : Broadband and intelligent networks, Data Mining, Data fusion, Computational intelligence, Information and data security, Information indexing and retrieval, Information processing, Information systems and applications, Internet applications and performances, Knowledge based systems, Knowledge management, Software Engineering, Decision making, Mobile networks and services, Network management and services, Neural Network, Fuzzy logics, Neuro-Fuzzy, Expert approaches, Innovation Technology and Management : Innovation and product development, Emerging advances in business and its applications, Creativity in Internet management and retailing, B2B and B2C management, Electronic transceiver device for Retail Marketing Industries, Facilities planning and management, Innovative pervasive computing applications, Programming paradigms for pervasive systems, Software evolution and maintenance in pervasive systems, Middleware services and agent technologies, Adaptive, autonomic and context-aware computing, Mobile/Wireless computing systems and services in pervasive computing, Energy-efficient and green pervasive computing, Communication architectures for pervasive computing, Ad hoc networks for pervasive communications, Pervasive opportunistic communications and applications, Enabling technologies for pervasive systems (e.g., wireless BAN, PAN), Positioning and tracking technologies, Sensors and RFID in pervasive systems, Multimodal sensing and context for pervasive applications, Pervasive sensing, perception and semantic interpretation, Smart devices and intelligent environments, Trust, security and privacy issues in pervasive systems, User interfaces and interaction models, Virtual immersive communications, Wearable computers, Standards and interfaces for pervasive computing environments, Social and economic models for pervasive systems, Active and Programmable Networks, Ad Hoc & Sensor Network, Congestion and/or Flow Control, Content Distribution, Grid Networking, High-speed Network Architectures, Internet Services and Applications, Optical Networks, Mobile and Wireless Networks, Network Modeling and Simulation, Multicast, Multimedia Communications, Network Control and Management, Network Protocols, Network Performance, Network Measurement, Peer to Peer and Overlay Networks, Quality of Service and Quality of Experience, Ubiquitous Networks, Crosscutting Themes – Internet Technologies, Infrastructure, Services and Applications; Open Source Tools, Open Models and Architectures; Security, Privacy and Trust; Navigation Systems, Location Based Services; Social Networks and Online Communities; ICT Convergence, Digital Economy and Digital Divide, Neural Networks, Pattern Recognition, Computer Vision, Advanced Computing Architectures and New Programming Models, Visualization and Virtual Reality as Applied to Computational Science, Computer Architecture and Embedded Systems, Technology in Education, Theoretical Computer Science, Computing Ethics, Computing Practices & Applications

Authors are invited to submit papers through e-mail ijcsiseditor@gmail.com. Submissions must be original and should not have been published previously or be under consideration for publication while being evaluated by IJCSIS. Before submission authors should carefully read over the journal's Author Guidelines, which are located at http://sites.google.com/site/ijcsis/authors-notes .

ISSN 1947 5500
http://sites.google.com/site/ijcsis/