

LNAI 6276

Rossitza Setchi  
Ivan Jordanov  
Robert J. Howlett  
Lakhmi C. Jain (Eds.)

# Knowledge-Based and Intelligent Information and Engineering Systems

14th International Conference, KES 2010  
Cardiff, UK, September 2010  
Proceedings, Part I

1  
Part I



Springer

Lecture Notes in Artificial Intelligence 6276

Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

Rossitza Setchi Ivan Jordanov  
Robert J. Howlett Lakhmi C. Jain (Eds.)

# Knowledge-Based and Intelligent Information and Engineering Systems

14th International Conference, KES 2010  
Cardiff, UK, September 8-10, 2010  
Proceedings, Part I

**Series Editors**

Randy Goebel, University of Alberta, Edmonton, Canada

Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Wolfgang Wahlster, DFKI and University of Saarland, Saarbrücken, Germany

**Volume Editors**

Rossitza Setchi

Cardiff University, School of Engineering

The Parade, Cardiff CF24 3AA, UK

E-mail: Setchi@cf.ac.uk

Ivan Jordanov

University of Portsmouth, Dept. of Computer Science and Software Engineering

Buckingham Building, Lion Terrace, Portsmouth, PO1 3HE, UK

E-mail: Ivan.Jordanov@port.ac.uk

Robert J. Howlett

KES International

145-157 St. John Street, London EC1V 4PY, UK

E-mail: rjhowlett@kesinternational.org

Lakhmi C. Jain

University of South Australia, School of Electrical and Information Engineering

Adelaide, Mawson Lakes Campus, SA 5095, Australia

E-mail: Lakhmi.Jain@unisa.edu.au

Library of Congress Control Number: 2010932879

CR Subject Classification (1998): I.2, H.4, H.3, I.4, H.5, I.5

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743

ISBN-10 3-642-15386-0 Springer Berlin Heidelberg New York

ISBN-13 978-3-642-15386-0 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

[springer.com](http://springer.com)

© Springer-Verlag Berlin Heidelberg 2010

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India  
Printed on acid-free paper 06/3180

## Preface

The 14<sup>th</sup> International Conference on Knowledge-Based and Intelligent Information and Engineering Systems was held during September 8–10, 2010 in Cardiff, UK. The conference was organized by the School of Engineering at Cardiff University, UK and KES International.

KES2010 provided an international scientific forum for the presentation of the results of high-quality research on a broad range of intelligent systems topics. The conference attracted over 360 submissions from 42 countries and 6 continents: Argentina, Australia, Belgium, Brazil, Bulgaria, Canada, Chile, China, Croatia, Czech Republic, Denmark, Finland, France, Germany, Greece, Hong Kong ROC, Hungary, India, Iran, Ireland, Israel, Italy, Japan, Korea, Malaysia, Mexico, The Netherlands, New Zealand, Pakistan, Poland, Romania, Singapore, Slovenia, Spain, Sweden, Syria, Taiwan, Tunisia, Turkey, UK, USA and Vietnam.

The conference consisted of 6 keynote talks, 11 general tracks and 29 invited sessions and workshops, on the applications and theory of intelligent systems and related areas. The distinguished keynote speakers were Christopher Bishop, UK, Nikola Kasabov, New Zealand, Saeid Nahavandi, Australia, Tetsuo Sawaragi, Japan, Yuzuru Tanaka, Japan and Roger Whitaker, UK.

Over 240 oral and poster presentations provided excellent opportunities for the presentation of interesting new research results and discussion about them, leading to knowledge transfer and generation of new ideas.

Extended versions of selected papers were considered for publication in the *International Journal of Knowledge-Based and Intelligent Engineering Systems*, *Engineering Applications of Artificial Intelligence*, *Journal of Intelligent Manufacturing*, and *Neural Computing and Applications*.

We would like to acknowledge the contribution of the Track Chairs, Invited Sessions Chairs, all members of the Program Committee and external reviewers for co-ordinating and monitoring the review process. We are grateful to the editorial team of Springer led by Alfred Hofmann. Our sincere gratitude goes to all participants and the authors of the submitted papers.

September 2010

Rossitza Setchi  
Ivan Jordanov  
Robert J. Howlett  
Lakhmi C. Jain

# **Organization**

KES 2010 was hosted and organized by the School of Engineering at Cardiff University, UK and KES International. The conference was held at the Mercure Holland House Hotel, September 8–10, 2010.

## **Conference Committee**

General Chair	Rossi Setchi, Cardiff University, UK
Conference Co-chair	Lakhmi C. Jain, University of South Australia, Australia
Executive Chair	Robert J. Howlett, University of Brighton, UK
Chair of the Organizing Committee	Y. Hicks, Cardiff University, UK
Program Chair	I. Jordanov, University of Portsmouth, UK

## **Organizing Committee**

KES Operations Manager	Peter Cushion, KES International
Publicity Chairs	D. Todorov, Cardiff University, UK
KES Systems Support	Yu-kun Lai, Cardiff University, UK
Members	Shaun Lee, KES International
	Engku Fadzli, Cardiff University, UK
	Lei Shi, Cardiff University, UK
	Nedyalko Petrov, Portsmouth University, UK
	Panagiotis Loukakos, Cardiff University, UK

## **Track Chairs**

Bruno Apolloni	University of Milan, Italy
Bojana Dalbelo Basic	University of Zagreb, Croatia
Floriana Esposito	University of Bari, Italy
Anne Hakansson	Stockholm University, Sweden
Ron Hartung	Franklyn University, USA
Honghai Liu	University of Portsmouth, UK
Ngoc Thanh Nguyen	Wroclaw University of Technology, Poland
Andreas Nuernberger	University of Magdeburg, Germany
Bernd Reusch	University of Dortmund, Germany
Tuan Pham	University of New South Wales, Australia
Toyohide Watanabe	Nagoya University, Japan

## Invited Sessions Chairs

3D Visualization of Natural Language	Minhua Eunice Ma, University of Derby, UK Nikolaos Antonopoulos, University of Derby, UK Bob Coyne, Columbia University, USA Kazuhiro Takeda, Shizuoka University, Japan Takashi Hamaguchi, Nagoya Institute of Technology, Japan
Intelligent Data Processing in Process Systems and Plants	
A Meta-Heuristic Approach to Management Engineering	Junzo Watada, Waseda University, Japan Taki Kanda, Bunri University of Hospitality, Japan Huey-Ming Lee, Chinese Culture University, Taiwan Lily Lin, China University of Technology, Taiwan Cesar Sanin, University of Newcastle, Australia
Knowledge Engineering and Smart Systems	
Skill Acquisition and Ubiquitous Human-Computer Interaction	Hirokazu Taki, Wakayama University, Japan Masato Soga, Wakayama University, Japan
Application of Knowledge Models in Healthcare	István Vassányi, University of Pannonia, Hungary György Surján, National Institute for Strategic Health Research, Hungary
Knowledge Environment for Supporting Creative Learning	Toyohide Watanabe, Nagoya University, Japan Tomoko Kojiri, Nagoya University, Japan
ICT in Innovation and Creativity	Toyohide Watanabe, Nagoya University, Japan Taketoshi Ushijima, Kyushu University, Japan
Intelligent Support for Designing Social Information Infrastructure	Toyohide Watanabe, Nagoya University, Japan Naoto Mukai, Tokyo University of Science, Japan
Intelligent Systems in Ambient-Assisted Living Environments	Antonio F. Gómez-Skarmeta, Universidad de Murcia, Spain Juan A. Botía, Universidad de Murcia, Spain
Knowledge-Based Systems for e-Business	Kazuhiko Tsuda, University of Tsukuba, Japan Nobuo Suzuki, KDDI Corporation, Japan
Quality Assurance and Intelligent Web-Based Information Technology	Anastasia N. Kastania, Athens University of Economics and Business, Greece Stelios Zimeras, University of the Aegean, Greece
Knowledge-Based Interface Systems	Yuji Iwahori, Chubu University, Japan Naohiro Ishii, Aichi Institute of Technology, Japan Yoshinori Adachi, Chubu University, Japan Nobuhiro Inuzuka, Nagoya Institute of Technology, Japan
Reasoning-Based Intelligent Systems	Kazumi Nakamatsu, University of Hyogo, Japan Jair Minoro Abe, University of Sao Paulo, Brazil
Data Mining and Service Science for Innovation	Katsutoshi Yada, Kansai University, Japan Takahira Yamaguchi, Keio University, Japan Maria Alessandra Torsello, University of Bari, Italy

Web 2.0: Opportunities and Challenges for Social Recommender Systems	Jose J. Pazos-Arias, University of Vigo, Spain Ana Fernandez-Vilas, University of Vigo, Spain
Innovations in Chance Discovery	Akinori Abe, University of Tokyo, Japan
Personalization of Web Contents and Services	In-Young Ko, Korea Advanced Institute of Science and Technology (KAIST), Korea Juan D. Velásquez, University of Chile, Chile
Advanced Knowledge-Based Systems	Alfredo Cuzzocrea, ICAR-CNR and University of Calabria, Italy Susumu Kunifugi, Jaist, Japan Kazuo Misue, University of Tsukuba, Japan Hidehiko Hayashi, Naruto University of Education, Japan
Knowledge-Based Creativity Support Systems	Motoki Miura, Kyushu Institute of Technology, Japan Toyohisa Nakada, Niigata University of International and Information Studies, Japan Tessai Hayama, JAIST, Japan Jun Munemori, Wakayama University, Japan Takaya Yuizono, JAIST, Japan
Intelligent Network and Service	Rashid Mehmood, Swansea School of Engineering, UK Omer F. Rana, Cardiff University, UK Ziad Salem, Aleppo University, Syria
Real-World Data Mining and Digital Intelligence	Sorin Hintea, Technical University of Cluj-Napoca, Romania Hernando Fernández-Canque, Glasgow Caledonian University, UK Gabriel Oltean, Technical University of Cluj-Napoca, Romania
Advanced Design Techniques for Adaptive Systems	Norio Baba, Osaka Kyoiku University, Japan
Soft Computing Techniques and Their Intelligent Utilizations Toward Gaming, Robotics, Stock Markets etc.	Argyris Dentsoras, University of Patras, Greece Nikos Aspragathos, University of Patras, Greece Vassilis Moulianitis, University of the Aegean, Greece
Methods and Techniques of Artificial and Computational Intelligence in Engineering Design	Vesa A. Niskanen, University of Helsinki, Finland
Philosophical and Methodological Aspects of Reasoning and Decision Making	

Semantic Technologies for Knowledge Workers	Andreas Dengel, German Research Center for Artificial Intelligence (DFKI), Germany Ansgar Bernardi, German Research Center for Artificial Intelligence (DFKI), Germany
Tools and Techniques for Effective Creation and Exploitation of Biodiversity Knowledge	Andrew C. Jones, Cardiff University, UK Richard J. White, Cardiff University, UK Gabriele Gianini, Università degli Studi di Milan, Italy Antonia Azzini, Università degli Studi di Milan, Italy Stefania Marrara, Università degli Studi di Milan, Italy
Immunity-Based Systems	Yoshiteru Ishida, Toyohashi University of Technology, Japan Takeshi Okamoto, Kanagawa Institute of Technology, Japan Yuji Watanabe, Nagoya City University, Japan Koji Harada, Toyohashi University of Technology, Japan

## Program Committee

Abe, Akinori	IREIIMS University, Japan
Alexandre de Matos Araujo	Rui, University of Coimbra, Portugal
Angelov, Plamen	Lancaster University, UK
Anwer, Nabil	LURPA - ENS CACHAN, France
Aoki, Shingo	Osaka Prefecture University, Japan
Apolloni, Bruno	University of Milan, Italy
Aspragathos, Nikos A.	University of Patras, Greece
Bannore, Vivek	University of South Australia, Australia
Barb, Adrian S.	Penn State University, USA
Becker-Asano	Christian, Intelligent Robotics and Communication Labs, Japan
Bianchini, Monica	University of Siena, Italy
Bichindaritz, Isabelle	University of Washington, USA
Boeva, Veselka	Technical University of Sofia, Bulgaria
Boutalis, Yiannis	Democritus University of Thrace, Greece
Brna, Paul	University of Glasgow, UK
Buckingham, Christopher	Aston University, UK
Camastra, Francesco	University of Naples Parthenope, Italy
Cao, Cungen	Chinese Academy of Sciences, China
Ceccarelli, Michele	University of Sannio, Italy
Chalup, Stephan	The University of Newcastle, Australia
Chang, Bao Rong	National University of Kaohsiung, Taiwan
Chen, Lihui	Nanyang Technological University, Singapore
Chen, Toly	Feng Chia University, Taiwan
Cheng, Kai	Brunel University, UK
Cheung, Benny	Honk Kong Polytechnic University, Hong Kong
Cobos Pérez, Ruth	Universidad Autónoma de Madrid, Spain
Crippa, Paolo	Università Politecnica delle Marche, Italy

Cuzzocrea, Alfredo	University of Calabria, Italy
Damiana, Maria Luisa,	University of Milan, Italy
Dasiopoulou, Stamatia	Informatics and Telematics Institute, Greece
De Cock, Martine	University of Washington, USA
De Wilde, Philippe	Heriot-Watt University, UK
Dengel, Andreas	German Research Center for Artificial Intelligence (DFKI), Germany
Duro, Richard J.	Universidade da Coruña, Spain
Dustdar, Schahram	Vienna University of Technology, Austria
Elomaa, Tapiو	Tampere University of Technology, Finland
Fernandez-Canque, Hernando	Glasgow Caledonian University, UK
Georgieva, Petia	University of Aveiro, Portugal
Godoy, Daniela	UNICEN University, Argentina
Grabot, Bernard	LGP-ENIT, France
Graña Romay, Manuel	Universidad del País Vasco, Spain
Grecos, Christos	University of West Scotland, UK
Hara, Takahiro	Osaka University, Japan
Hintea, Sorin	Cluj-Napoca University, Romania
Honda, Katsuhiro	Osaka Prefecture University, Japan
Hong, Tzung-Pei	National University of Kaohsiung, Taiwan
Hu, Chenyi	University of Central Arkansas, USA
Hurtado Larrain, Carlos	University of Chile, Chile
Ichalkarane, Nikhil	University of South Australia, Australia
Ishibuchi, Hisao	Osaka Prefecture University, Japan
Ishida, Yoshiteru	Toyohashi University of Technology, Japan
Ito, Takayuki	Massachusetts Institute of Technology, USA
Ivancevic, Tijana	University of South Australia, Australia
Janicki, Ryszard	McMaster University, Canada
Jastroch, Norbert	MET Communications GmbH, Germany
Jensen, Richard	Aberystwyth University, UK
Jones, Andrew	Cardiff University, UK
Jordanov, Ivan	University of Portsmouth, UK
Jung, Jason J.	Yeungnam University, Korea
Juric, Matjaz B.	University of Maribor, Slovenia
Katagiri, Hideki	Hiroshima University, Japan
Ko, In-Young	KAIST, Korea
Kodogiannis, Vassilis S.	University of Westminster, UK
Koenig, Andreas	Technische Universität Kaiserslautern, Germany
Kojadinovic, Ivan	University of Auckland, New Zealand
Kompatsiaris, Yiannis	Informatics and Telematics Institute, Greece
Konar, Amit	Jadavpur University, India
Koshizuka, Takamasa	Honda R&D Co., Ltd., Japan
Koychev, Ivan	University of Sofia, Bulgaria
Kwong, C.K.	The Hong Kong Polytechnic University, Hong Kong
Lee, Dah-Jye	Brigham Young University, USA
Lee, W.B.	Hong Kong Polytechnic University, Hong Kong
Likas, Aristidis	University of Ioannina, Greece

Lim, C.P.	Universiti Sains Malaysia, Malaysia
Liu, Lei	Beijing University of Technology, China
Maglogiannis, Ilias	University of Central Greece, Greece
Maier, Patrick	The University of Edinburgh, UK
Marinov, Milko T.	University of Ruse, Bulgaria
McCauley Bush, Pamela	University of Central Florida, USA
Montani, Stefania	Università del Piemonte Orientale, Italy
Moreno Jimenez, Ramón	Universidad del País Vasco, Spain
Nguyen, Ngoc Thanh	Wroclaw University of Technology, Poland
Nishida, Toyoaki	Kyoto University, Japan
Niskanen, Vesa A.	University of Helsinki, Finland
Ohkura, Kazuhiro	Hiroshima University, Japan
Palade, Vasile	Oxford University, UK
Pallares, Alvaro	Plastiasite S.A., Spain
Paranjape, Raman	University of Regina, Canada
Pasek, Zbigniew J.	University of Windsor, Canada
Pasi, Gabriella	University of Milan, Italy
Passerini, Andrea	Università degli Studi di Trento, Italy
Pazos-Arias, Jose	University of Vigo, Spain
Petrosino, Alfredo	Università di Napoli Parthenope, Italy
Prada, Rui	IST-UTL and INESC-ID, Portugal
Pratihar, Dilip Kumar	Osaka Prefecture University, Japan
Putnik, Goran D.	University of Minho, Portugal
Reidsema, Carl	University of New South Wales, Australia
Resconi, Germano	Catholic University in Brescia, Italy
Rovetta, Stefano	University of Genoa, Italy
Sansone, Carlo	Università di Napoli Federico II, Italy
Sarangapani, Jagannathan	Missouri University of Science and Technology, USA
Sato-Ilic, Mika	University of Tsukuba, Japan
Schockaert, Steven	Ghent University, Belgium
Seiffert, Udo	Fraunhofer-Institute IFF Magdeburg, Germany
Simperl, Elena	University of Innsbruck, Austria
Smrz, Pavel	Brno University of Technology, Czech Republic
Soroka, Anthony	Cardiff University, UK
Szczerbicki, Edward	The University of Newcastle, Australia
Tanaka, Takushi	Fukuoka Institute of Technology, Japan
Teng, Wei-Chung	National Taiwan University of Science and Technology, Taiwan
Tichy, Pavel	Rockwell Automation Research Centre, Czech Republic
Tino, Peter	The University of Birmingham, UK
Tolk, Andreas	Old Dominion University, USA
Toro, Carlos	VICOMTech, Spain
Torra, Vicenc	IIIA-CSIC, Spain
Tsihrintzis, George	University of Piraeus, Greece
Tsiporkova, Elena	Sirris, Belgium
Turchetti, Claudio	Università Politecnica delle Marche, Italy

Uchino, Eiji	Yamaguchi University, Japan
Urlings, Pierre	DSTO, Department of Defence, Australia
Vadera, Sunil	University of Salford, UK
Vald��z Vela, Mercedes	Universidad de Murcia, Spain
Vellido, Alfredo	Universitat Polit��nica de Catalunya, Spain
Virvou, Maria	University of Piraeus, Greece
Wang, Zidong	Brunel University, UK
Watts, Mike	TBA, New Zealand
White, Richard J.	Cardiff University, UK
Williams, M. Howard	Heriot-Watt University, UK
Yang, Zijiang	York University, Canada
Yoshida, Hiroyuki	Harvard Medical School, USA
Zanni-Merk, Cecilia	LGeCo - INSA de Strasbourg, France
Zheng, Li-Rong	Royal Institute of Technology (KTH), Sweden

## Reviewers

Adam Nowak	Bao Rong Chang	David Vallejo
Adam Slowik	Benjamin Adrian	Davor Skrlec
Adrian S. Barb	Bernard Grabot	Dickson Lukose
Akinori Abe	Bernd Reusch	Dilip Pratihar
Akira Hattori	Bettina Waldvogel	Doctor Jair Abe
Alan Paton	Bj��rn Forcher	Don Jeng
Alessandra Micheletti	Bob Coyne	Donggang Yu
Alfredo Cuzzocrea	Bojan Basrak	Doris Csipkes
Ammar Aljer	Bojana Dalbelo Basic	Eduardo Cerqueira
Amparo Vila	Bozidar Ivankovic	Eduardo Merlo
Ana Fernandez-Vilas	Branko Zitko	Edward Szczerbicki
Anastasia Kastania	Bruno Apolloni	Eiji Uchino
Anastasius Moumtzoglou	Calin Ciufudean	Elena Pagani
Andrea Visconti	Carlo Sansone	Elena Simperl
Andreas Abecker	Carlos Ocampo	Esmail Bonakdarian
Andreas Dengel	Carlos Pedrinaci	Esmiralda Moradian
Andreas Oikonomou	Carlos Toro	Francesco Camastrra
Andrew Jones	Cecilia Zanni-Merk	Frane Saric
Annalisa Appice	Cesar Sanin	Fujiki Morii
Anne H��kansson	Chang-Tien Lu	Fumihiko Anma
Ansgar Bernardi	Christian Becker-Asano	Fumitaka Uchio
Anthony Soroka	Christine Mumford	Gabbar Hossam
Antonio Gomez-Skarmeta	Chunbo Chu	Gabor Csipkes
Antonio Zippo	Costantino Lucisano	Gabriel Oltean
Aristidis Likas	C.P. Lim	Gabriella Pasi
Armando Buzzanca	Cristos Orovas	George Mitchell
Artur Silic	Daniela Godoy	George Tsihrintzis
Athina Lazakidou	Danijel Radosevic	Gergely H��ja
Azizul Azhar Ramli	Danilo Dell'Agnello	Gianluca Sforza
Bal��zs Ga��l	David Martens	Giovanna Castellano

Giovanni Gomez Zuluaga	Kazuhiro Ohkura	Narayanan
Gunnar Grimnes	Kazuhiro Takeda	Kulathuramaiyer
Gyorgy Surjan	Kazuhsisa Seta	Nikica Hlupi
Haoxi Dorje Zhang	Kazumi Nakamatsu	Nikola Ljubesic
Haruhiko H. Nishimura	Kazunori Nishino	Nikos Tsourveloudis
Haruhiko Haruhiko Nishimura	Kazuo Misue	Nobuhiro Inuzuka
Haruki Kawanaka	Keiichiro Mitani	Nobuo Suzuki
Hector Alvarez	Kenji Matsuura	Norbert Jastroch
Hernando Fernandez-Canque	Koji Harada	Norio Baba
Hideaki Ito	Kouji Yoshida	Noriyuki Matsuda
Hidehiko Hayashi	Lars Hildebrand	Omar Rana
Hideo Funaoi	Laura Caponetti	Orleo Marinaro
Hideyuki Matsumoto	Lei Liu	Paolo Crippa
Hirokazu Miura	Lelia Festila	Pasquale Di Meo
Hisayoshi Kunimune	Leonardo Mancilla Amaya	Pavel Tichy
Hrvoje Markovic	Lily Lin	Philippe Wilde
Huey-Ming Lee	Ljiljana Stojanovic	Rafael Batres
Ilias Maglogiannis	Lorenzo Magnani	Raffaele Cannone
Ing. Angelo Ciccazzo	Lorenzo Valerio	Ramón Jimenez
Ivan Koychev	Ludger van Elst	Rashid Mehmoed
Ivan Stajduhar	Manuel Grana	Richard Pyle
J. Mattila	Marek Malski	Richard White
Jair Abe	Maria Torsello	Robert Howlett
Jari Kortelainen	Mario Koeppen	Roberto Cordone
Jayanthi Ranjan	Marko Banek	Ronald Hartung
Jerome Darmont	Martin Lopez-Nores	Roumen Kountchev
Jessie Kennedy	Martine Cock	Rozália Lakner
Jesualdo Tomás	Masakazu Takahashi	Ruediger Oehlmann
Fernández-Breis	Masaru Noda	Ruth Cobos
Jiangtao Cao	Masato Soga	Ryohei Nakano
Jim Sheng	Masayoshi Aritsugi	Ryuuki Sakamoto
Johnson Fader	Mayumi Ueda	Sachio Hirokawa
Jose Manuel Molina	Melita Hajdinjak	Satoru Fujii
Juan Botía	Michelangelo Ceci	Sebastian Rios
Juan Manuel Corchado	Michele Missikoff	Sebastian Weber
Juan Pavón	Miguel Delgado	Seiji Isotani
Julia Hirschberg	Milko Marinov	Seiki Akama
Jun Munemori	Minhua Ma	Setsuya Kurahashi
Jun Sawamoto	Minoru Minoru Fukumi	Shamshul Bahar Yaakob
Junzo Watada	Monica Bianchini	Shinji Fukui
Jure Mijic	Motoi Iwashita	Shusaku Tsumoto
Katalina Grigorova	Motoki Miura	Shyue-Liang Wang
Katsuhiro Honda	Nahla Barakat	Simone Bassis
Katsumi Yamashita	Naohiro Ishii	Sophia Kossida
Kazuhiko Tsuda	Naoto Mukai	Stamatia Dasiopoulou
	Naoyuki Naoyuki Kubota	Stefan Zinsmeister
		Stefania Marrara

Stefania Montani	Toru Fukumoto	Yiannis Boutalis
Stephan Chalup	Toshihiro Hayashi	Yoshifumi Tsuge
Steven Schockaert	Toshio Mochizuki	Yoshihiro Okada
Sunil Vadera	Toumoto	Yoshihiro Takuya
susumu hashizume	Toyohide Watanabe	Yoshinori Adachi
Susumu Kunifugi	Toyohis Nakada	Yoshiyuki Yamashita
Takanobu Umetsu	Tsuyoshi Nakamura	Youji Ochi
Takashi Hamaguchi	Tuan Pham	Young Ko
Takashi Mitsuishi	Valerio Arnaboldi	Yuichiro Tateiwa
Takashi Yukawa	Vassilis Kodogiannis	Yuji Iwahori
Takaya Yuizono	Vassilis Moulianitis	Yuji Wada
Takeshi Okamoto	Vesa Niskanen	Yuji Watanabe
Taketoshi Kurooka	Veselka Boeva	Yuki Hayashi
Taketoshi Ushijima	Vivek Bamnore	Yukio Ohsawa
Takushi Tanaka	Wataru Sunayama	Yumiko Nara
Tapio Elomaa	Wei-Chung Teng	Yurie Iribe
Tatiana Tambouratzis	William Hochstettler	Zdenek Zdrahal
Tessai Hayama	Winston Jain	Ziad Salem
Thomas Roth-Berghofer	Wolfgang Stock	Zijiang Yang
Tomislav Hrkac	Xiaofei Ji	Zlatko Drmac
Tomoko Kojiri	Yi Xiao	Zuwairie Ibrahim

# Table of Contents – Part I

## Keynote Talks

Evolving Integrative Brain-, Gene-, and Quantum Inspired Systems for Computational Intelligence and Knowledge Engineering (Abstract) .... <i>Nikola Kasabov</i>	1
A Semiotic View of Social Intelligence for Realizing Human-Machine Symbiotic Systems (Abstract) .... <i>Tetsuo Sawaragi</i>	2
Embracing Uncertainty: The New Machine Intelligence (Abstract) .... <i>Christopher Bishop</i>	3
Exploiting Social Structures and Social Networks (Abstract) .... <i>Roger Whitaker</i>	4
Knowledge Visualization for Engineered Systems .... <i>Saeid Nahavandi, Dawei Jia, and Asim Bhatti</i>	5
Proximity-Based Federation of Smart Objects: Liberating Ubiquitous Computing from Stereotyped Application Scenarios .... <i>Yuzuru Tanaka</i>	14

## Artificial Neural Networks, Connectionists Systems and Evolutionary Computation

A Neural Network Model to Develop Urban Acupuncture .... <i>Leandro Tortosa, José F. Vicent, Antonio Zamora, and José L. Oliver</i>	31
Discovering Process Models with Genetic Algorithms Using Sampling ... <i>Carmen Bratosin, Natalia Sidorova, and Wil van der Aalst</i>	41
A Multi-Objective Evolutionary Approach for the Antenna Positioning Problem .... <i>Carlos Segura, Yanira González, Gara Miranda, and Coromoto León</i>	51
CLONAL-GP Framework for Artificial Immune System Inspired Genetic Programming for Classification .... <i>Hajira Jabeen and Abdul Rauf Baig</i>	61
Solving Industrial Based Job-Shop Scheduling Problem by Distributed Micro-Genetic Algorithm with Local Search .... <i>Rubiyah Yusof, Marzuki Khalid, and Tay Cheng San</i>	69

## XVIII Table of Contents – Part I

Data Mining via Rules Extracted from GMDH: An Application to Predict Churn in Bank Credit Cards .....	80
<i>Nekuri Naveen, V. Ravi, and C. Raghavendra Rao</i>	

Sensitivity Analysis and Automatic Calibration of a Rainfall-Runoff Model Using Multi-objectives .....	90
<i>Fan Sun and Yang Liu</i>	

University Course Timetabling Using ACO: A Case Study on Laboratory Exercises .....	100
<i>Vatroslav Dino Matijaš, Goran Molnar, Marko Čupić, Domagoj Jakobović, and Bojana Dalbelo Bašić</i>	

## Machine Learning and Classical AI

Nonparametric Statistical Analysis of Machine Learning Algorithms for Regression Problems .....	111
<i>Magdalena Graczyk, Tadeusz Lasota, Zbigniew Telec, and Bogdan Trawiński</i>	

Adaptive Learning of Nominal Concepts for Supervised Classification ...	121
<i>Nida Meddouri and Mondher Maddouri</i>	

A Novel Approach of Process Mining with Event Graph .....	131
<i>Hui Zhang, Ying Liu, Chunping Li, and Roger Jiao</i>	

A Classification Algorithm for Process Sequences Based on Markov Chains and Bayesian Networks .....	141
<i>Katharina Tschumitschew, Detlef Nauck, and Frank Klawonn</i>	

Coaching to Enhance the Online Behavior Learning of a Robotic Agent .....	148
<i>Masakazu Hirokawa and Kenji Suzuki</i>	

## Agent and Multi-agent Systems

Cooperation of AGVs' Head-on Collision Avoidance by Knowledge Exchange in Autonomous Decentralized FMS .....	158
<i>Hidehiko Yamamoto and Takayoshi Yamada</i>	

A Log Analyzer Agent for Intrusion Detection in a Multi-Agent System .....	168
<i>Iago Porto-Díaz, Óscar Fontenla-Romero, and Amparo Alonso-Betanzos</i>	

A Proof System for Time-Dependent Multi-agents .....	178
<i>Norihiro Kamide</i>	

Toward Emotional E-Commerce: Formalizing Agents for a Simple Negotiation Protocol .....	188
<i>Veronica Jascanu, Nicolae Jascanu, and Severin Bumbaru</i>	
Distributed Ant Colony Clustering Using Mobile Agents and Its Effects .....	198
<i>Ryotaro Oikawa, Masashi Mizutani, Munehiro Takimoto, and Yasushi Kambayashi</i>	
Monitoring a Multi-Agent System Evolution through Iterative Development .....	209
<i>Yves Wautelet and Manuel Kolp</i>	
An Agent for Ecological Deliberation .....	220
<i>John Debenham and Carles Sierra</i>	
A Framework to Compute Inference Rules Valid in Agents' Temporal Logics .....	230
<i>Sergey Babenyshev and Vladimir Rybakov</i>	
Statecharts-Based JADE Agents and Tools for Engineering Multi-Agent Systems .....	240
<i>Giancarlo Fortino, Francesco Rango, and Wilma Russo</i>	
Telco Agent: Enabler of Paradigm Shift towards Customer-Managed Relationship .....	251
<i>Vedran Podobnik and Ignac Lovrek</i>	
Multi-attribute Auction Model for Agent-Based Content Trading in Telecom Markets .....	261
<i>Ana Petric and Gordan Jezic</i>	
Applying Possibility and Belief Operators to Conditional Statements ...	271
<i>Grzegorz Skorupa and Radostaw Katarzyniak</i>	
A Computer Adaptive Testing Method for Intelligent Tutoring Systems .....	281
<i>Adrianna Kozierkiewicz-Hetmańska and Ngoc Thanh Nguyen</i>	
<b>Intelligent Vision, Image Processing and Signal Processing</b>	
Combining Patient Metadata Extraction and Automatic Image Parsing for the Generation of an Anatomic Atlas .....	290
<i>Manuel Möller, Patrick Ernst, Michael Sintek, Sascha Seifert, Gunnar Grimnes, Alexander Cavallaro, and Andreas Dengel</i>	
Parallel Processing with CUDA in Ceramic Tiles Classification .....	300
<i>Tomislav Matić and Željko Hocenski</i>	

Signal Receiving and Processing Platform of the Experimental Passive Radar for Intelligent Surveillance System Using Software Defined Radio Approach .....	311
<i>Boguslaw Szlachetko and Andrzej Lewandowski</i>	
Automated Anticounterfeiting Inspection Methods for Rigid Films Based on Infrared and Ultraviolet Pigments and Supervised Image Segmentation and Classification .....	321
<i>Michael Kohlert, Christian Kohlert, and Andreas König</i>	
Vowel Recognition by Using the Combination of Haar Wavelet and Neural Network .....	331
<i>Mohammad Mehdi Hosseini, Abdorreza Alavi Gharahbagh, and Sedigheh Ghofrani</i>	
Bayesian Classification Using DCT Features for Brain Tumor Detection .....	340
<i>Qurat-ul Ain, Irfan Mehmood, Syed M. Naqi, and M. Arfan Jaffar</i>	
A New Strategy of Adaptive Nonlinear Echo Cancelling Volterra-Wiener Filter Structure Selection .....	350
<i>Pawel Biernacki</i>	
Intelligent System for Commercial Block Recognition Using Audio Signal Only .....	360
<i>Pawel Biernacki</i>	
Viewpoint Insensitive Actions Recognition Using Hidden Conditional Random Fields .....	369
<i>Xiaofei Ji, Honghai Liu, and Yibo Li</i>	
Fuzzy Hyper-Prototype Clustering .....	379
<i>Jin Liu and Tuan D. Pham</i>	
<b>Knowledge Management, Ontologies and Data Mining</b>	
Clustering Using Difference Criterion of Distortion Ratios .....	390
<i>Fujiki Morii</i>	
Computer-Generated Conversation Based on Newspaper Headline Interpretation .....	400
<i>Eriko Yoshimura, Seiji Tsuchiya, and Hirokazu Watabe</i>	
Using Regression Analysis to Identify Patterns of Non-Technical Losses on Power Utilities .....	410
<i>Iñigo Monedero, Félix Biscarri, Carlos León, Juan I. Guerrero, Jesús Biscarri, and Rocío Millán</i>	

Enhancing the Symbolic Aggregate Approximation Method Using Updated Lookup Tables .....	420
<i>Muhammad Marwan, Muhammad Fuad and Pierre-François Marteau</i>	
Which XML Storage for Knowledge and Ontology Systems? .....	432
<i>Martin Bukatovič, Aleš Horák, and Adam Rambořík</i>	
Finding Temporal Patterns Using Constraints on (Partial) Absence, Presence and Duration .....	442
<i>S. Peter and F. Höppner</i>	
Clustering Based on Kolmogorov Information .....	452
<i>Fouchal Said, Ahat Murat, Lavallée Ivan, Bui Marc, and Benamor Sofiane</i>	
Rule Extraction from Support Vector Machine Using Modified Active Learning Based Approach: An Application to CRM .....	461
<i>M.A.H. Farquad, V. Ravi, and S. Bapi Raju</i>	
Factorizing Three-Way Binary Data with Triadic Formal Concepts .....	471
<i>Radim Belohlávek and Vilem Vychodil</i>	
Application of Ontological Engineering in Customs Domain .....	481
<i>Panagiotis Loukakos and Rossitza Setchi</i>	
Classification and Prediction of Academic Talent Using Data Mining Techniques .....	491
<i>Hamidah Jantan, Abdul Razak Hamdan, and Zulaiha Ali Othman</i>	
Test-Cost Sensitive Classification on Data with Missing Values in the Limited Time .....	501
<i>Chang Wan</i>	
Modified K-Means Clustering for Travel Time Prediction Based on Historical Traffic Data .....	511
<i>Rudra Pratap Deb Nath, Hyun-Jo Lee, Nihad Karim Chowdhury, and Jae-Woo Chang</i>	
An Ontology-Based Approach for Autonomous Systems' Description and Engineering: The OA Sys Framework .....	522
<i>Julita Bermejo-Alonso, Ricardo Sanz, Manuel Rodríguez, and Carlos Hernández</i>	
Search Space Reduction for an Efficient Handling of Empty Answers in Database Flexible Querying .....	532
<i>Mohamed Ali Ben Hassine, Chaker Abidi Nasri, and Habib Ounelli</i>	
Using Association Rules to Discover Color-Emotion Relationships Based on Social Tagging .....	544
<i>Haifeng Feng, Marie-Jeanne Lesot, and Marcin Detyniecki</i>	

A Conceptual Framework for Role-Based Knowledge Profiling Using Semiotics Approach .....	554
<i>Nazmona Mat Ali and Kecheng Liu</i>	

Using Biased Discriminant Analysis for Email Filtering .....	566
<i>Juan Carlos Gomez and Marie-Francine Moens</i>	

Use of Geospatial Analyses for Semantic Reasoning.....	576
<i>Ashish Karmacharya, Christophe Cruz, Frank Boochs, and Franck Marzani</i>	

## Application of Knowledge Models in Healthcare

Computer-Based Dietary Menu Planning: How to Support It by Complex Knowledge? .....	587
<i>Barbara Koroušić Seljak</i>	

Flexible Semantic Querying of Clinical Archetypes .....	597
<i>Catalina Martínez-Costa, José Antonio Miñarro-Giménez, Marcos Menárguez-Tortosa, Rafael Valencia-García, and Jesualdo Tomás Fernández-Breis</i>	

A Formal Domain Model for Dietary and Physical Activity Counseling .....	607
<i>Erzsébet Mák, Balázs Pintér, Balázs Gaál, István Vassányi, György Kozmann, and Istvánné Németh</i>	

## Semantic Technologies for Knowledge Workers

An Ontology Based Approach to Measuring the Semantic Similarity between Information Objects in Personal Information Collections .....	617
<i>Lei Shi and Rossitza Setchi</i>	

Ontology Based Graphical Query Language Supporting Recursion .....	627
<i>Arun Anand Sadanandan, Kow Weng Onn, and Dickson Lukose</i>	

Using Concept Maps to Improve Proactive Information Delivery in TaskNavigator .....	639
<i>Oleg Rostanin, Heiko Maus, Takeshi Suzuki, and Kaoru Maeda</i>	

A Vocabulary Building Mechanism Based on Lexical Semantics for Querying the Semantic Web .....	649
<i>Yu Asano and Yuzuru Tanaka</i>	

Designing a Knowledge Mapping Tool for Knowledge Workers .....	660
<i>Heiko Haller and Andreas Abecker</i>	

<b>Author Index .....</b>	<b>671</b>
---------------------------	------------

## Table of Contents – Part II

### Web Intelligence, Text and Multimedia Mining and Retrieval

Semantics-Based Representation Model for Multi-layer Text Classification . . . . .	1
<i>Jiali Yun, Liping Jing, Jian Yu, and Houkuan Huang</i>	
Frequent Itemset Based Hierarchical Document Clustering Using Wikipedia as External Knowledge . . . . .	11
<i>G.V.R. Kiran, Ravi Shankar, and Vikram Pudi</i>	
Automatic Authorship Attribution for Texts in Croatian Language Using Combinations of Features . . . . .	21
<i>Tomislav Reicher, Ivan Krišto, Igor Belša, and Artur šilić</i>	
Visualization of Text Streams: A Survey . . . . .	31
<i>Artur šilić and Bojana Dalbelo Bašić</i>	
Combining Semantic and Content Based Image Retrieval in ORDBMS . . . . .	44
<i>Carlos E. Alvez and Aldo R. Vecchietti</i>	
A Historically-Based Task Composition Mechanism to Support Spontaneous Interactions among Users in Urban Computing Environments . . . . .	54
<i>Angel Jimenez-Molina and In-Young Ko</i>	
Multi-criteria Retrieval in Cultural Heritage Recommendation Systems . . . . .	64
<i>Pierpaolo Di Bitonto, Maria Laterza, Teresa Roselli, and Veronica Rossano</i>	
An Approach for the Automatic Recommendation of Ontologies Using Collaborative Knowledge . . . . .	74
<i>Marcos Martínez-Romero, José M. Vázquez-Naya, Cristian R. Munteanu, Javier Pereira, and Alejandro Pazos</i>	
Knowledge Mining with ELM System . . . . .	82
<i>Ilona Bluemke and Agnieszka Orlewicz</i>	
DOCODE-Lite: A Meta-Search Engine for Document Similarity Retrieval . . . . .	93
<i>Felipe Bravo-Marquez, Gaston L'Huillier, Sebastián A. Ríos, Juan D. Velásquez, and Luis A. Guerrero</i>	

## Intelligent Tutoring Systems and E-Learning Environments

Group Formation for Collaboration in Exploratory Learning Using Group Technology Techniques .....	103
<i>Mihaela Cocea and George D. Magoulas</i>	
Applying Pedagogical Analyses to Create an On-Line Course for e Learning .....	114
<i>D.-L. Le, V.-H. Tran, D.-T. Nguyen, A.-T. Nguyen, and A. Hunger</i>	
Adaptive Modelling of Users' Strategies in Exploratory Learning Using Case-Based Reasoning .....	124
<i>Mihaela Cocea, Sergio Gutierrez-Santos, and George D. Magoulas</i>	
An Implementation of Reprogramming Scheme for Wireless Sensor Networks .....	135
<i>Aoi Hashizume, Hiroshi Mineno, and Tadanori Mizuno</i>	
Predicting e-Learning Course Adaptability and Changes in Learning Preferences after Taking e-Learning Courses .....	143
<i>Kazunori Nishino, Toshifumi Shimoda, Yurie Iribé, Shinji Mizuno, Kumiko Aoki, and Yoshimi Fukumura</i>	

## Intelligent Systems

A Logic for Incomplete Sequential Information .....	153
<i>Norihiro Kamide</i>	
A Power-Enhanced Algorithm for Spatial Anomaly Detection in Binary Labelled Point Data Using the Spatial Scan Statistic .....	163
<i>Simon Read, Peter Bath, Peter Willett, and Ravi Maheswaran</i>	
Vertical Fragmentation Design of Distributed Databases Considering the Nonlinear Nature of Roundtrip Response Time .....	173
<i>Rodolfo A. Pazos R., Graciela Vázquez A., José A. Martínez F., and Joaquín Pérez O.</i>	
Improving Iterated Local Search Solution For The Linear Ordering Problem With Cumulative Costs (LOPCC) .....	183
<i>David Terán Villanueva, Héctor Joaquín Fraire Huacuja, Abraham Duarte, Rodolfo Pazos R., Juan Martín Carpio Valadez, and Héctor José Puga Soberanes</i>	
A Common-Sense Planning Strategy for Ambient Intelligence .....	193
<i>María J. Santofimia, Scott E. Fahlman, Francisco Moya, and Juan C. López</i>	

Dialogue Manager for a NLIDB for Solving the Semantic Ellipsis Problem in Query Formulation . . . . .	203
<i>Rodolfo A. Pazos R., Juan C. Rojas P., René Santaolaya S., José A. Martínez F., and Juan J. Gonzalez B.</i>	
Hand Gesture Recognition Based on Segmented Singular Value Decomposition . . . . .	214
<i>Jing Liu and Manolya Kavaklı</i>	
Reasoning and Inference Rules in Basic Linear Temporal Logic $\mathcal{BLTL}$ . . . . .	224
<i>S. Babenyshev and V. Rybakov</i>	
Direct Adaptive Control of an Anaerobic Depollution Bioprocess Using Radial Basis Neural Networks . . . . .	234
<i>Emil Petre, Dorin Sendrescu, and Dan Selișteanu</i>	
Visualisation of Test Coverage for Conformance Tests of Low Level Communication Protocols . . . . .	244
<i>Katharina Tschumitschew, Frank Klawonn, Nils Obermöller, and Wolfhard Lawrenz</i>	
Control Network Programming with SPIDER: Dynamic Search Control . . . . .	253
<i>Kostadin Kratchanov, Tzanko Golemanov, Emilia Golemanova, and Tuncay Ercan</i>	
Non-procedural Implementation of Local Heuristic Search in Control Network Programming . . . . .	263
<i>Kostadin Kratchanov, Emilia Golemanova, Tzanko Golemanov, and Tuncay Ercan</i>	
Meta Agents, Ontologies and Search, a Proposed Synthesis . . . . .	273
<i>Ronald L. Hartung and Anne Håkansson</i>	
Categorizing User Interests in Recommender Systems . . . . .	282
<i>Sourav Saha, Sandipan Majumder, Sanjog Ray, and Ambuj Mahanti</i>	
Architecture of Hascheck – An Intelligent Spellchecker for Croatian Language . . . . .	292
<i>Šandor Dembitz, Gordan Gledec, and Bruno Blašković</i>	
Light-Weight Access Control Scheme for XML Data . . . . .	302
<i>Dongchan An, Hakim Kim, and Seog Park</i>	
A New Distributed Particle Swarm Optimization Algorithm for Constraint Reasoning . . . . .	312
<i>Sadok Bouamama</i>	

Simulation of Fuzzy Control Applied to a Railway Pantograph-Catenary System . . . . .	322
<i>Simon Walters</i>	
Floor Circulation Index and Optimal Positioning of Elevator Hoistways . . . . .	331
<i>Panagiotis Markos and Argyris Dentsoras</i>	
Rapid Evaluation of Reconfigurable Robots Anatomies Using Computational Intelligence . . . . .	341
<i>Harry Valsamos, Vassilis Moulianitis, and Nikos Aspragathos</i>	
Incremental Construction of Alpha Lattices and Association Rules . . . . .	351
<i>Henry Soldano, Véronique Ventos, Marc Champesme, and David Forge</i>	
Intelligent Magnetic Sensing System for Low Power WSN Localization Immersed in Liquid-Filled Industrial Containers . . . . .	361
<i>Kuncup Iswandy, Stefano Carrella, and Andreas König</i>	
<b>Intelligent Data Processing in Process Systems and Plants</b>	
An Overview of a Microcontroller-Based Approach to Intelligent Machine Tool Monitoring . . . . .	371
<i>Raees Siddiqui, Roger Grosvenor, and Paul Prickett</i>	
Use of Two-Layer Cause-Effect Model to Select Source of Signal in Plant Alarm System . . . . .	381
<i>Kazuhiro Takeda, Takashi Hamaguchi, Masaru Noda, Naoki Kimura, and Toshiaki Itoh</i>	
Coloured Petri Net Diagnosers for Lumped Process Systems . . . . .	389
<i>Attila Tóth, Erzsébet Németh, and Katalin M. Hangos</i>	
Proactive Control of Manufacturing Processes Using Historical Data . . . . .	399
<i>Manfred Grauer, Sachin Karadgi, Ulf Müller, Daniel Metz, and Walter Schäfer</i>	
A Multiagent Approach for Sustainable Design of Heat Exchanger Networks . . . . .	409
<i>Naoki Kimura, Kizuki Yasue, Tekishi Kou, and Yoshifumi Tsuge</i>	
Consistency Checking Method of Inventory Control for Countermeasures Planning System . . . . .	417
<i>Takashi Hamaguchi, Kazuhiro Takeda, Hideyuki Matsumoto, and Yoshihiro Hashimoto</i>	

Fault Semantic Networks for Accident Forecasting of LNG Plants .....	427
<i>Hossam A. Gabbar</i>	

## A Meta Heuristic Approach to Management Engineering

Fuzzy Group Evaluating the Aggregative Risk Rate of Software Development .....	438
<i>Huey-Ming Lee and Lily Lin</i>	

Fuzzy Power System Reliability Model Based on Value-at-Risk .....	445
<i>Bo Wang, You Li, and Junzo Watada</i>	

Human Tracking: A State-of-Art Survey .....	454
<i>Junzo Watada, Zalili Musa, Lakhmi C. Jain, and John Fulcher</i>	

Ordinal Structure Fuzzy Logic Predictor for Consumer Behaviour .....	464
<i>Rubiyah Yusof, Marzuki Khalid, and Mohd. Ridzuan Yunus</i>	

Kansei for Colors Depending on Objects .....	477
<i>Taki Kanda</i>	

A Hybrid Intelligent Algorithm for Solving the Bilevel Programming Models .....	485
<i>Shamshul Bahar Yaakob and Junzo Watada</i>	

## Knowledge Engineering and Smart Systems

Using Semantics to Bridge the Information and Knowledge Sharing Gaps in Virtual Engineering .....	495
<i>Javier Vaquero, Carlos Toro, Carlos Palenzuela, and Eneko Azpeitia</i>	

Discovering and Usage of Customer Knowledge in QoS Mechanism for B2C Web Server Systems .....	505
<i>Leszek Borzemski and Grażyna Suchacka</i>	

Conceptual Fuzzy Model of the Polish Internet Mortgage Market .....	515
<i>Aleksander Orłowski and Edward Szczerbicki</i>	

Translations of Service Level Agreement in Systems Based on Service Oriented Architecture .....	523
<i>Adam Grzech and Piotr Rygielski</i>	

Ontology Engineering Aspects in the Intelligent Systems Development .....	533
<i>Adam Czarnecki and Cezary Orłowski</i>	

Supporting Software Project Management Processes Using the Agent System .....	543
<i>Cezary Orłowski and Artur Ziółkowski</i>	
Knowledge-Based Virtual Organizations for the E-Decisional Community .....	553
<i>Leonardo Mancilla-Amaya, Cesar Sanín, and Edward Szczerbicki</i>	
Decisional DNA Applied to Robotics .....	563
<i>Haoxi Zhang, Cesar Sanin, and Edward Szczerbicki</i>	
Supporting Management Decisions with Intelligent Mechanisms of Obtaining and Processing Knowledge .....	571
<i>Cezary Orłowski and Tomasz Sitek</i>	
Finding Inner Copy Communities Using Social Network Analysis.....	581
<i>Eduardo Merlo, Sebastián A. Ríos, Héctor Álvarez, Gaston L'Huillier, and Juan D. Velásquez</i>	
Enhancing Social Network Analysis with a Concept-Based Text Mining Approach to Discover Key Members on a Virtual Community of Practice .....	591
<i>Héctor Alvarez, Sebastián A. Ríos, Felipe Aguilera, Eduardo Merlo, and Luis A. Guerrero</i>	
Intelligence Infrastructure: Architecture Discussion: Performance, Availability and Management.....	601
<i>Giovanni Gómez Zuluaga, Cesar Sanín, and Edward Szczerbicki</i>	
<b>Skill Acquisition and Ubiquitous Human Computer Interaction</b>	
Geometric Considerations of Search Behavior .....	611
<i>Masaya Ashida and Hirokazu Taki</i>	
A Web-Community Supporting Self-management for Runners with Annotation.....	620
<i>Naka Gotoda, Kenji Matsuura, Shinji Otsuka, Toshio Tanaka, and Yoneo Yano</i>	
An Analysis of Background-Color Effects on the Scores of a Computer-Based English Test .....	630
<i>Atsuko K. Yamazaki</i>	
Message Ferry Route Design Based on Clustering for Sparse Ad Hoc Networks .....	637
<i>Hirokazu Miura, Daisuke Nishi, Noriyuki Matsuda, and Hirokazu Taki</i>	

Affordance in Dynamic Objects Based on Face Recognition.....	645
<i>Taizo Miyachi, Toshiki Maezawa, Takanao Nishihara, and     Takeshi Suzuki</i>	
<b>Author Index .....</b>	<b>653</b>

# Table of Contents – Part III

## Knowledge-Based Systems for e-Business

A Study on Traveling Purpose Classification Method to Extract Traveling Requests .....	1
<i>Nobuo Suzuki, Mariko Yamamura, and Kazuhiko Tsuda</i>	
Variable Selection by $C_p$ Statistic in Multiple Responses Regression with Fewer Sample Size Than the Dimension .....	7
<i>Mariko Yamamura, Hirokazu Yanagihara, and Muni S. Srivastava</i>	
Customer Path Controlling in the Retail Store with the Vertex Dominating Cycle Algorithms .....	15
<i>Takeshi Sugiyama</i>	

## Quality Assurance and Intelligent Web-Based Information Technology

A Framework for the Quality Assurance of Blended E-Learning Communities .....	23
<i>Iraklis Varlamis and Ioannis Apostolakis</i>	
Quality of Content in Web 2.0 Applications .....	33
<i>Iraklis Varlamis</i>	
Telepediatrics Education on the Semantic Web .....	43
<i>Sofia Sidirokastriti and Anastasia N. Kastania</i>	
Web Applications and Public Diplomacy .....	53
<i>Antigoni Koffa and Anastasia N. Kastania</i>	

## Knowledge-Based Interface Systems

A Hybrid Face Recognition System for Managing Time of Going to Work and Getting away from Office .....	63
<i>Yoshinori Adachi, Zeng Yunfei, Masahiro Ozaki, and Yuji Iwahori</i>	
Multi-Relational Pattern Mining System for General Database Systems .....	72
<i>Nobuhiko Inuzuka and Toshiyuki Makino</i>	
Recovering 3-D Shape Based on Light Fall-Off Stereo under Point Light Source Illumination and Perspective Projection .....	81
<i>Yuji Iwahori, Claire Rouveyrol, Robert J. Woodham, Yoshinori Adachi, and Kunio Kasugai</i>	

Shadow Detection Method Based on Dirichlet Process Mixture Model . . . . .	89
<i>Wataru Kurahashi, Shinji Fukui, Yuji Iwahori, and Robert J. Woodham</i>	

Vowel Sound Recognition Using a Spectrum Envelope Feature Detection Method and Neural Network . . . . .	97
<i>Masashi Kawaguchi, Naohiro Yonekura, Takashi Jimbo, and Naohiro Ishii</i>	

Information Extraction Using XPath . . . . .	104
<i>Masashi Okada, Naohiro Ishii, and Ippei Torii</i>	

Information Visualization System for Activation of Shopping Streets . . . . .	113
<i>Ippei Torii, Yousuke Okada, Takahito Niwa, Manabu Onogi, and Naohiro Ishii</i>	

## Reasoning Based Intelligent Systems

Introduction to Intelligent Network Routing Based on EVALPSN . . . . .	123
<i>Kazumi Nakamatsu, Jair Minoro Abe, and Takashi Watanabe</i>	

Introduction to Intelligent Elevator Control Based on EVALPSN . . . . .	133
<i>Kazumi Nakamatsu, Jair Minoro Abe, Seiki Akama, and Roumen Kountchev</i>	

Monadic Curry System $N_1^*$ . . . . .	143
<i>Jair Minoro Abe, Kazumi Nakamatsu, and Seiki Akama</i>	

A Sensing System for an Autonomous Mobile Robot Based on the Paraconsistent Artificial Neural Network . . . . .	154
<i>Claudio Rodrigo Torres, Jair Minoro Abe, Germano Lambert-Torres, João Inácio Da Silva Filho, and Helga Gonzaga Martins</i>	

Paraconsistent Artificial Neural Networks and EEG Analysis . . . . .	164
<i>Jair Minoro Abe, Helder F.S. Lopes, Kazumi Nakamatsu, and Seiki Akama</i>	

A Reasoning-Based Strategy for Exploring the Synergy among Alternative Crops . . . . .	174
<i>Hércules Antonio do Prado, Edilson Ferneda, and Ricardo Coelho de Faria</i>	

Reasoning Elements for a Vehicle Routing System . . . . .	182
<i>Edilson Ferneda, Bernardo A. Mello, Janaína A.S. Diniz, and Adelaide Figueiredo</i>	

A Mechanism for Converting Circuit Grammars to Definite Clauses . . . . .	190
<i>Takushi Tanaka</i>	

Constructive Discursive Reasoning . . . . .	200
<i>Seiki Akama, Kazumi Nakamatsu, and Jair Minoro Abe</i>	

Formal Concept Analysis of Medical Incident Reports . . . . .	207
<i>Takahiro Baba, Lucing Liu, and Sachio Hirokawa</i>	

Compression of Multispectral Images with Inverse Pyramid Decomposition . . . . .	215
<i>Roumen Kountchev and Kazumi Nakamatsu</i>	

## Data Mining and Service Science for Innovation

Econometric Approach for Broadband Market in Japan . . . . .	225
<i>Takeshi Kurosawa, Hiromichi Kawano, Motoi Iwashita, Shinsuke Shimogawa, Shouji Kouno, and Akiya Inoue</i>	

Opinion Exchange Support System by Visualizing Input History . . . . .	235
<i>Yukihiro Tamura, Yuuki Tomiyama, and Wataru Sunayama</i>	

Extracting Promising Sequential Patterns from RFID Data Using the LCM Sequence . . . . .	244
<i>Takanobu Nakahara, Takeaki Uno, and Katsutoshi Yada</i>	

Relation between Stay-Time and Purchase Probability Based on RFID Data in a Japanese Supermarket . . . . .	254
<i>Keiji Takai and Katsutoshi Yada</i>	

Implementing an Image Search System with Integrating Social Tags and DBpedia . . . . .	264
<i>Chie Iijima, Makito Kimura, and Takahira Yamaguchi</i>	

The Influence of Shopping Path Length on Purchase Behavior in Grocery Store . . . . .	273
<i>Marina Kholod, Takanobu Nakahara, Haruka Azuma, and Katsutoshi Yada</i>	

Existence of Single Input Rule Modules for Optimal Fuzzy Logic Control . . . . .	281
<i>Takashi Mitsuishi, Hidefumi Kawakatsu, and Yasunari Shidama</i>	

## Innovations in Chance Discovery

Temporality and Reference Place: Discovering Chances for Conflict Avoidance in Teamwork . . . . .	290
<i>Ruediger Oehlmann</i>	

Discovering Research Key Terms as Temporal Patterns of Importance Indices for Text Mining . . . . .	297
<i>Hidenao Abe and Shusaku Tsumoto</i>	

Categorized and Integrated Data Mining of Medical Data from the Viewpoint of Chance Discovery .....	307
<i>Akinori Abe, Norihiro Hagita, Michiko Furutani, Yoshiyuki Furutani, and Rumiko Matsuoka</i>	
Support System for Thinking New Criteria of Unclassified Diseases .....	315
<i>Yoko Nishihara, Yoshimune Hiratsuka, Akira Murakami, Yukio Ohsawa, and Toshiro Kumakawa</i>	
Interpretation of Chance Discovery in Temporal Logic, Admissible Inference Rules .....	323
<i>Vladimir Rybakov</i>	
Faking Chance Cognitive Niche Impoverishment .....	331
<i>Lorenzo Magnani and Emanuele Bardone</i>	
<b>Advanced Knowledge-Based Systems</b>	
Summarization for Geographically Distributed Data Streams .....	339
<i>Anna Ciampi, Annalisa Appice, and Donato Malerba</i>	
Gradual Data Aggregation in Multi-granular Fact Tables on Resource-Constrained Systems .....	349
<i>Nadeem Iftikhar and Torben Bach Pedersen</i>	
A Refinement Operator Based Method for Semantic Grouping of Conjunctive Query Results .....	359
<i>Agnieszka Lawrynowicz, Claudia d'Amato, and Nicola Fanizzi</i>	
Semantic Network of Ground Station-Satellite Communication System .....	369
<i>Katarzyna Dqbrowska-Kubik</i>	
W-kmeans: Clustering News Articles Using WordNet .....	379
<i>Christos Bouras and Vassilis Tsogkas</i>	
An Efficient Mechanism for Stemming and Tagging: The Case of Greek language .....	389
<i>Giorgos Adam, Konstantinos Asimakis, Christos Bouras, and Vassilis Poulopoulos</i>	
Co-clustering Analysis of Weblogs Using Bipartite Spectral Projection Approach .....	398
<i>Guandong Xu, Yu Zong, Peter Dolog, and Yanchun Zhang</i>	
Talking Biology in Logic, and Back .....	408
<i>Hasan Jamil</i>	

Analysis of Medical Pathways by Means of Frequent Closed Sequences . . . . .	418
<i>Elena Baralis, Giulia Bruno, Silvia Chiusano, Virna C. Domenici, Naeem A. Mahoto, and Caterina Petrigni</i>	
Inheriting Access Control Rules from Large Relational Databases to Materialized Views Automatically . . . . .	426
<i>Alfredo Cuzzocrea, Mohand-Said Hacid, and Nicola Grillo</i>	
MySQL Data Mining: Extending MySQL to Support Data Mining Primitives (Demo) . . . . .	438
<i>Alfredo Ferro, Rosalba Giugno, Piera Laura Puglisi, and Alfredo Pulvirenti</i>	
A Genetic Algorithm to Design Industrial Materials . . . . .	445
<i>E. Tenorio, J. Gómez-Ruiz, J.I. Peláez, and J.M. Doña</i>	
<b>Intelligent Network and Service</b>	
A Proposal of P2P Content Retrieval System Using Access-Based Grouping Technique . . . . .	455
<i>Takuya Sasaki, Jun Sawamoto, Takashi Katoh, Yuji Wada, Norihisa Segawa, and Eiji Sugino</i>	
The Effects of Individual Differences in Two Persons on the Distributed and Cooperative KJ Method in an Anonymous Environment . . . . .	464
<i>Takaya Yuizono and Zhe Jin</i>	
Pictograph Chat Communicator III: A Chat System That Embodies Cross-Cultural Communication . . . . .	473
<i>Jun Munemori, Taro Fukuda, Moonyati Binti Mohd Yatid, Tadashi Nishide, and Junko Itou</i>	
Distance Learning Support System for Game Programming with Java . . . . .	483
<i>Kouji Yoshida, Takumu Yaoi, Isao Miyaji, Kunihiro Yamada, and Satoru Fujii</i>	
Evidence Analysis Method Using Bloom Filter for MANET Forensics . . . . .	493
<i>Takashi Mishina, Yoh Shiraishi, and Osamu Takahashi</i>	
Diminished Reality for Landscape Video Sequences with Homographies . . . . .	501
<i>Kosuke Takeda and Ryuuki Sakamoto</i>	
Prediction of Combinatorial Protein-Protein Interaction Networks from Expression Data Using Statistics on Conditional Probability . . . . .	509
<i>Takatoshi Fujiki, Etsuko Inoue, Takuya Yoshihiro, and Masaru Nakagawa</i>	

Development and Evaluation of a Historical Tour Support System Using 3D Graphics and Mobile Terminal .....	519
<i>Satoru Fujii, Takahiro Shima, Megumi Takahashi, and Koji Yoshida</i>	
Repetition of Dialogue Atmosphere Using Characters Based on Face-to-Face Dialogue .....	527
<i>Junko Itou and Jun Munemori</i>	
<b>Soft Computing Techniques and Their Intelligent Utilizations Toward Gaming, Robotics, Stock Markets etc.</b>	
CMOS-Based Radiation Movie and Still Image Pickup System with a Phototimer Using Smart Pattern Recognition .....	535
<i>Osamu Yuuki, Hiroshi Mineno, Kunihiro Yamada, and Tadanori Mizuno</i>	
Optimal H <sub>2</sub> Integral Controller Design with Derivative State Constraints for Torsional Vibration Model .....	545
<i>Noriyuki Komine and Kunihiro Yamada</i>	
Utilization of Evolutionary Algorithms for Making COMMONS GAME Much More Exciting .....	555
<i>Norio Baba, Hisashi Handa, Mariko Kusaka, Masaki Takeda, Yuriko Yoshihara, and Keisuke Kogawa</i>	
Education of Embedded System by Using Electric Fan .....	562
<i>Osamu Yuuki, Junji Namiki, and Kunihiro Yamada</i>	
Development and Evaluation of a Routing Simulator for a Mutually Complementary Network Incorporating Wired and Wireless Components .....	572
<i>Hiroki Morita, Naoki Yusa, Noriyuki Komine, Kouji Yoshida, Masanori Kojima, Tadanori Mizuno, and Kunihiro Yamada</i>	
On the Impact of the Metrics Choice in SOM Learning: Some Empirical Results from Financial Data .....	583
<i>Marina Resta</i>	
Reinforcement Learning Scheme for Grouping and Characterization of Multi-agent Network .....	592
<i>Koichiro Morihiro, Nobuyuki Matsui, Teijiyo Isokawa, and Haruhiko Nishimura</i>	
Extracting Principal Components from Pseudo-random Data by Using Random Matrix Theory .....	602
<i>Mieko Tanaka-Yamawaki</i>	

Music Impression Detection Method for User Independent Music Retrieval System . . . . .	612
<i>Masato Miyoshi, Satoru Tsuge, Hillary Kipsang Choge, Tadahiro Oyama, Momoyo Ito, and Minoru Fukumi</i>	
Applying Fuzzy Sets to Composite Algorithm for Remote Sensing Data . . . . .	622
<i>Kenneth J. Mackin, Takashi Yamaguchi, Jong Geol Park, Eiji Nunohiro, Kotaro Matsushita, Yukio Yanagisawa, and Masao Igarashi</i>	
<b>Immunity-Based Systems</b>	
Evaluations of Immunity-Based Diagnosis for a Motherboard . . . . .	628
<i>Haruki Shida, Takeshi Okamoto, and Yoshiteru Ishida</i>	
A Note on Dynamical Behaviors of a Spatial Game Operated on Intercrossed Rules . . . . .	637
<i>Kouji Harada and Yoshiteru Ishida</i>	
Asymmetry in Repairing and Infection: The Case of a Self-repair Network . . . . .	645
<i>Yoshiteru Ishida and Kei-ichi Tanabe</i>	
A Note on Symmetry in Logic of Self-repair: The Case of a Self-repair Network . . . . .	652
<i>Yoshiteru Ishida</i>	
An Immunity-Based Scheme for Statistical En-route Filtering in Wireless Sensor Networks . . . . .	660
<i>Yuji Watanabe</i>	
<b>Author Index . . . . .</b>	<b>667</b>

# Table of Contents – Part IV

## Knowledge Based and Expert Systems

Emotion Judgment Method from an Utterance Sentence . . . . .	1
<i>Seiji Tsuchiya, Eriko Yoshimura, and Hirokazu Watabe</i>	
Local Model Update with an Application to Sliding Window Protocol . . . . .	11
<i>Michael Kelly and Yan Zhang</i>	
Development of RP <sup>3</sup> CA-EMP, a Knowledge-Based System for Applying Environmental Management Plan (EMP) in the Malaysian Construction Industry . . . . .	22
<i>Leila Ooshaksaraie, Noor Ezlin Ahmad Basri, Azuraliza Abu Bakar, and Khairul Nizam Abdul Maulud</i>	
Reasoning with Multiple Points of View: A Case Study . . . . .	32
<i>P. Bouché, C. Zanni-Merk, N. Gartiser, D. Renaud, and F. Rousselot</i>	
Syntax and Semantics for Business Rules . . . . .	41
<i>Xiaofan Liu, Natasha Alechina, and Brian Logan</i>	
Dialect Recognition Method Using Emotion Judgment . . . . .	51
<i>Noriyuki Okumura</i>	
Applying a Knowledge Based System for Metadata Integration for Data Warehouses . . . . .	60
<i>Dan Wu and Anne Håkansson</i>	
A Ubiquitous Intelligent Tutoring System for Aiding Electronic Learning . . . . .	70
<i>Sergio Ciruela, Miguel Delgado, and Nicolás Marín</i>	
Towards a Proposal for a Vessel Knowledge Representation Model . . . . .	80
<i>I. Macía, M. Graña, and C. Paloc</i>	
Developing a Probabilistic Graphical Structure from a Model of Mental-Health Clinical Risk Expertise . . . . .	88
<i>Olufunmilayo Obembe and Christopher D. Buckingham</i>	
Controlling Security of Software Development with Multi-agent System . . . . .	98
<i>Esmiralda Moradian and Anne Håkansson</i>	
On Lukasiewicz' Infinie-Valued Logic and FuzzyL . . . . .	108
<i>Jorma K. Mattila</i>	

A Meta-level Approach to Approximate Probability .....	116
<i>Vesa A. Niskanen</i>	
Comparing Ontologies Using Multi-agent System and Knowledge Base .....	124
<i>Anne Håkansson, Ronald Hartung, Esmiralda Moradian, and Dan Wu</i>	
ACTL Local Model Update with Constraints .....	135
<i>Michael Kelly, Fei Pu, Yan Zhang, and Yi Zhou</i>	
<b>Knowledge Environment for Supporting Creative Learning</b>	
Bridging Multiple Motor-Skills in a Community Site.....	145
<i>Kenji Matsuura, Naka Gotoda, Tetsushi Ueta, and Yoneo Yano</i>	
Topic Visualization for Understanding Research Paper in Collaborative Discussion .....	153
<i>Masato Aoki, Yuki Hayashi, Tomoko Kojiri, and Toyohide Watanabe</i>	
Building a Framework to Design and Evaluate Meta-learning Support Systems .....	163
<i>Kazuhisa Seta, Minoru Fujiwara, Daijiro Noguchi, Hiroshi Maeno, and Mitsuru Ikeda</i>	
Promoting Learning Attention with Gaze Tracking Integrated e-Learning Contents .....	173
<i>Kai Li and Yurie Iribe</i>	
System For Creative Distance Learning Environment Development Based On Competence Management .....	180
<i>Przemysław Różewski and Bartłomiej Małachowski</i>	
A Blended Project-Based Learning Program on Embedded Software Design with Collaboration Support Tools .....	190
<i>Takashi Yukawa, Tomonori Iwazaki, Keisuke Ishida, Hirotaka Takahashi, Yoshimi Fukumura, Makoto Yamazaki, Naoki Hasegawa, and Hajime Miura</i>	
Multilingual Discussion in Metaverse among Students from the USA, Korea and Japan .....	200
<i>Hideyuki Kanematsu, Yoshimi Fukumura, Dana M. Barry, So Young Sohn, and Ryosuke Taguchi</i>	
Improvement of an Annotation Sharing System on Web-Based Materials to Activate Discussions .....	210
<i>Hisayoshi Kunimune, Yuuji Gonda, Yuuki Tominaga, and Masaaki Niimura</i>	

## Information Communication Technology in Innovation and Creativity

Where to Crawl Next for Focused Crawlers.....	220
<i>Yuki Uemura, Tsuyoshi Itokawa, Teruaki Kitasuka, and Masayoshi Aritsugi</i>	
Extraction of Co-existent Sentences for Explaining Figures toward Effective Support for Scientific Papers Reading .....	230
<i>Ryo Takeshima and Toyohide Watanabe</i>	
Semantic Approach to Image Retrieval Using Statistical Models Based on a Lexical Ontology .....	240
<i>Syed Abdullah Fadzli and Rossitza Setchi</i>	
IEC-Based Motion Retrieval System Using Laban Movement Analysis .....	251
<i>Yuki Wakayama, Seiji Okajima, Shigeru Takano, and Yoshihiro Okada</i>	
Automatic Composition of Personalized Appreciation Route Based on Semantic Relationship between Exhibits in Museum .....	261
<i>Chihiro Maehara, Kotaro Yatsugi, Daewoong Kim, and Taketoshi Ushijima</i>	

## Intelligent Support for Designing Social Information Infrastructure

Design and Implementation of a Context-Aware Guide Application for Mobile Users Based on Machine Learning .....	271
<i>Yuichi Omori, Yuki Nonaka, and Mikio Hasegawa</i>	
Adaptive Traffic Signal Control Based on Vehicle Route Sharing by Wireless Communication .....	280
<i>Hiroyasu Ezawa and Naoto Mukai</i>	
A System to Share Arrangements for Daily Tasks and Life Events on the Web .....	290
<i>Hitomi Sato, Akira Hattori, and Haruo Hayami</i>	
Population Estimation of Internet Forum Community by Posted Article Distribution .....	298
<i>Masao Kubo, Keitaro Naruse, Hiroshi Sato, and Takashi Matsubara</i>	
Development of Delay Estimation Method Using Probe Data for Adaptive Signal Control Algorithm .....	308
<i>Hisatomo Hanabusa, Morihisa Iijima, and Ryota Horiguchi</i>	

## Intelligent Systems in Ambient Assisted Living Environments

OVACARE: A Multi-Agent System for Assistance and Health Care . . . . .	318
<i>Juan F. De Paz, Sara Rodríguez, Javier Bajo,     Juan M. Corchado, and Emilio S. Corchado</i>	
Talking Agents in Ambient-Assisted Living . . . . .	328
<i>José M. Fernández de Alba and Juan Pavón</i>	
A System for Recognizing Activities of Daily Living Using Everyday Objects . . . . .	337
<i>María Ros, Miguel Delgado, and Amparo Vila</i>	
A Normality Analysis-Based Approach to Monitor Behaviors in AAL Domains . . . . .	347
<i>D. Vallejo, J. Albusac, C. Glez-Morello, and L. Jimenez</i>	
Adaptation of an Evaluation System for e-Health Environments . . . . .	357
<i>Nayat Sánchez-Pi and José Manuel Molina</i>	

OutCare: Supporting Dementia Patients in Outdoor Scenarios . . . . .	365
<i>Jie Wan, Caroline Byrne, Gregory M.P. O'Hare, and Michael J. O'Grady</i>	

## 3D Visualisation of Natural Language

Frame Semantics in Text-to-Scene Generation . . . . .	375
<i>Bob Coyne, Owen Rambow, Julia Hirschberg, and Richard Sproat</i>	
SenticSpace: Visualizing Opinions and Sentiments in a Multi-dimensional Vector Space . . . . .	385
<i>Erik Cambria, Amir Hussain, Catherine Havasi, and Chris Eckl</i>	
Supporting Collaborative Transcription of Recorded Speech with a 3D Game Interface . . . . .	394
<i>Saturnino Luz, Masood Masoodian, and Bill Rogers</i>	
Text-to-Video: Story Illustration from Online Photo Collections . . . . .	402
<i>Katharina Schwarz, Pavel Rojtberg, Joachim Caspar,     Iryna Gurevych, Michael Goesele, and Hendrik P.A. Lensch</i>	
Improving Communication Using 3D Animation . . . . .	410
<i>Laurent Ruhlmann, Benoit Ozell, Michel Gagnon,     Steve Bourgoin, and Eric Charton</i>	

Visualization and Language Processing for Supporting Analysis across the Biomedical Literature .....	420
<i>Carsten Görg, Hannah Tipney, Karin Verspoor, William A. Baumgartner Jr., K. Bretonnel Cohen, John Stasko, and Lawrence E. Hunter</i>	

SceneMaker: Multimodal Visualisation of Natural Language Film Scripts .....	430
<i>Eva Hanser, Paul Mc Kevitt, Tom Lunney, Joan Condell, and Minhua Ma</i>	

## Knowledge-Based Creativity Support Systems

Interaction Technique Combining Gripping and Pen Pressures .....	440
<i>Yu Suzuki, Kazuo Misue, and Jiro Tanaka</i>	

Destination Board System Based on Photographs .....	449
<i>Toyohisa Nakada</i>	

Practicing on Stage: Increasing Transparency and Interaction in Class Activity with Digital Pen system .....	457
<i>Taro Sugihara, Motoki Miura, and Susumu Kunifugi</i>	

Extracting a Keyword Network of Flood Disaster Measures .....	465
<i>Motoki Miura, Mitsuhiro Tokuda, and Daiki Kuwahara</i>	

Study on Classification of the Tacit Knowledge Using an Eye-Tracking Interface: Experiment of Observation Pictures and Assessment of Reproduction Drawing .....	475
<i>Yuta Watanabe, Issei Kodama, Hidehiko Hayashi, and Akinori Minaduki</i>	

## Tools and Techniques for Effective Creation and Exploitation of Biodiversity Knowledge

Evolution of the Catalogue of Life Architecture .....	485
<i>Andrew C. Jones, Richard J. White, Jonathan Giddy, Alex Hardisty, and Hardik Raja</i>	

A Consensus Method for Checklist Integration .....	497
<i>Kevin Richards, Aaron Wilton, Christina Flann, and Jerry Cooper</i>	

Tools for Semantic Annotation of Taxonomic Descriptions .....	506
<i>Hong Cui, Partha Pratim Sanyal, and Chunshui Yu</i>	

Automated Pre-processing Strategies for Species Occurrence Data Used in Biodiversity Modelling .....	517
<i>Marshall J. Heap and Alastair Culham</i>	

## Real World Data Mining and Digital Intelligence

A Hybrid Approach for Indexing and Retrieval of Archaeological Textual Information .....	527
<i>Ammar Halabi, Ahmed-Derar Islim, and Mohamed-Zakaria Kurdi</i>	
Prognosis of Breast Cancer Using Genetic Programming .....	536
<i>Simone A. Ludwig and Stefanie Roos</i>	
Classification of Software Artifacts Based on Structural Information .....	546
<i>Yuhanis Yusof and Omer F. Rana</i>	
Clustering Client Honeypot Data to Support Malware Analysis .....	556
<i>Yaser Alosefer and Omer F. Rana</i>	
LocPriS: A Security and Privacy Preserving Location Based Services Development Framework .....	566
<i>Gareth Ayres and Rashid Mehmood</i>	
Controlling Real World Pervasive Environments with Knowledge Bases .....	576
<i>Atif Alvi, Zubair Nabi, David Greaves, and Rashid Mehmood</i>	
Automatically Finding Answers to “Why” and “How to” Questions for Arabic Language .....	586
<i>Ziad Salem, Jawad Sadek, Fairouz Chakkour, and Nadia Haskkour</i>	
From Information to Sense-Making: Fetching and Querying Semantic Repositories .....	594
<i>Tope Omitola, Ian C. Millard, Hugh Glaser, Nicholas Gibbins, and Nigel Shadbolt</i>	

## Advanced Design Techniques for Adaptive Systems

A variable Topology Analog Filter Suitable for Multi-mode Wireless Applications .....	603
<i>Sorin Hintea, Doris Csipkes, Gabor Csipkes, and Hernando Fernandez-Canque</i>	
A Reconfigurable Voltage Reference without Resistors .....	613
<i>Lelia Festilă, Lorant Andras Szolga, Mihaela Cîrlugea, and Sorin Hintea</i>	
A Double-Layer Genetic Algorithm for Gm-C Filter Design .....	623
<i>Paul Farago, Sorin Hintea, Gabriel Oltean, and Lelia Festila</i>	

Optimised Dielectric Totally Internally Reflecting Concentrator for the Solar Photonic Optoelectronic Transformer System: Maximum Concentration Method .....	633
<i>Firdaus Muhammad-Sukki, Roberto Ramirez-Iniguez, Scott G. McMeekin, Brian G. Stewart, and Barry Clive</i>	
<b>Author Index .....</b>	<b>643</b>

# Evolving Integrative Brain-, Gene-, and Quantum Inspired Systems for Computational Intelligence and Knowledge Engineering

Nikola Kasabov

Knowledge Engineering and Discovery Research Institute, KEDRI

Auckland University of Technology, Auckland, New Zealand

[nkasabov@aut.ac.nz](mailto:nkasabov@aut.ac.nz)

<http://www.kedri.info>

**Abstract.** The talk presents theoretical foundations and practical applications of evolving intelligent information processing systems inspired by information principles in Nature in their interaction and integration. That includes neuronal-, genetic-, and quantum information principles, all manifesting the feature of *evolvability*. First, the talk reviews the main principles of information processing at neuronal-, genetic-, and quantum information levels. Each of these levels has already inspired the creation of efficient computational models that incrementally evolve their structure and functionality from incoming data and through interaction with the environment. The talk also extends these paradigms with novel methods and systems that integrate these principles. Examples of such models are: evolving spiking neural networks; computational neurogenetic models (where interaction between genes, either artificial or real, is part of the neuronal information processing); quantum inspired evolutionary algorithms; probabilistic spiking neural networks utilizing quantum computation as a probability theory. The new models are significantly faster in feature selection and learning and can be applied to solving efficiently complex biological and engineering problems for adaptive, incremental learning and knowledge discovery in large dimensional spaces and in a new environment. Examples include: incremental learning systems; on-line multimodal audiovisual information processing; evolving neuro-genetic systems; bio-informatics; biomedical decision support systems; cyber-security. Open questions, challenges and directions for further research are presented.

## References

- [1] Kasabov, N.: Evolving Connectionist Systems: The Knowledge Engineering Approach. Springer, London (2007), <http://www.springer.de>
- [2] Kasabov, N.: Evolving Intelligence in Humans and Machines: Integrative Connectionist Systems Approach. Feature article, IEEE CIS Magazine 3(3), 23–37 (2008), <http://www.ieee.cis.org>
- [3] Kasabov, N.: Integrative Connectionist Learning Systems Inspired by Nature: Current Models, Future Trends and Challenges. Natural Computing, Int. Journal 8(2), 199–210 (2009)
- [4] Kasabov, N.: To spike or not to spike: A probabilistic spiking neural model. Neural Networks 23(1), 16–19 (2010)

# A Semiotic View of Social Intelligence for Realizing Human-Machine Symbiotic Systems

Tetsuo Sawaragi

Kyoto University, Japan

[sawaragi@me.kyoto-u.ac.jp](mailto:sawaragi@me.kyoto-u.ac.jp)

**Abstract.** In a coming ubiquitous society, the collaboration between the human and the semi-automated machine is inevitable. The core difficulty herein is that cognitive agents (i.e., human and robot) are characterized as *creative* and *adaptable* to and within their environments. To tackle this problem, we have to clarify how the cognitive agent recognizes the external environment as well as other agents' behavioural performances, and how the context determines their cognition and makes the agent extract a particular *meaning* out of the physical and/or social environment. In this talk, we focus on the design issues of the mutual and inseparable relationships between the external environment including others and the internal constructs of the agent that is an actor, an observer, a cognizer, and an interpreter. For this purpose, we introduce the subject of "semiosis", which is any form of activity, conduct, or process that involves *signs*, including the production of *meanings*. After reviewing the original idea of Peirce's semiosis, our extended definition of the semiosis will be provided. That is, we define semiosis as "a process of constructing internal models within the cognitive agent" coherent to a target system in the environment to use, to monitor, to control, to collaborate with, etc. Based upon this, the research can be divided into the following three kinds of topics. The first one is on the semiotic analysis of the complex behaviours of the existing artefact systems and of the complex tasks that the user is forced to perform and to be instructed by some others. The second one is on the semiotic design of artefacts so that they should be coherent to cognitive agents' recognition, wherein the targets to be designed are focused on "signs" that are visible and eligible to the cognitive agents either directly or via interface systems. Finally, the third one is on the design of collaborative activities by a pair of cognitive agents (i.e., teaching tasks and/or learning tasks) via a variety of signs. With respect to this, the design and communication issues of human-centered automation and of ambient intelligence will be discussed in terms of the semiotic frame. This work is fully supported by a Grant-in-Aid Creative Scientific Research 2007-2011 (19GS0208) funded by the Ministry of Education, Culture, Sports, Science and Technology, Japan.

# Embracing Uncertainty: The New Machine Intelligence

Christopher Bishop

Microsoft Research Cambridge, UK

**Abstract.** Many of the early applications of machine intelligence were based on expert systems constructed using rules elicited from human experts. Limitations in the applicability of this approach helped to drive black-box statistical methods, such as neural networks, based on learning from data. These black-box methods too are hitting limitations, due to the challenges of incorporating background knowledge. In this talk I will describe a new paradigm for machine intelligence which has emerged over the last five years, and which allows prior knowledge from domain experts to be integrated with machine learning techniques to enable a new generation of large-scale applications. The talk will be illustrated with tutorial examples as well as real-world case studies.

# Exploiting Social Structures and Social Networks

Roger Whitaker

Cardiff University, UK

**Abstract.** In recent years there has been much recent interest in the use of “online” social networks for maintaining and building relationships with others. In this talk we explore some of the key characteristics of social networks and how they can potentially be exploited to provide intelligent content sharing in the pervasive and mobile computing domain.

Wireless and mobile devices such as phones, MP3 players, sensors, phones and PDAs are becoming increasingly capable of creating and sharing content. Harnessing this across devices that are only intermittently connected requires new adaptive approaches to networking and may facilitate new future applications. The basis for intermittent or temporary connectivity directly between devices, known as opportunistic networking, allows wireless devices to store, carry and forward information between themselves. In this talk we show how devices can build up, detect and potentially exploit social structures to fulfill functions such as community detection, cooperation, trust and content sharing between peers that may repeatedly interact with each other on a local basis. We also show how this technology may emerge in future applications. Coordinated by the School of Informatics at Cardiff University, this work is supported by the EC FP7 SOCIALNETS project, funded under the “Future emerging technologies” programme in collaboration with University of Cambridge (UK), University of Oxford (UK), CNR (Italy), Eurecom (France), University of Athens (Greece) and University of Aveiro (Portugal). See: <http://www.social-nets.eu/>

# Knowledge Visualization for Engineered Systems

Saeid Nahavandi, Dawei Jia, and Asim Bhatti

Centre for Intelligent Systems Research, Deakin University, Victoria 3217, Australia  
saeid.nahavandi@deakin.edu.au

**Abstract.** In this information age computer modelling and simulation is proving an indispensable tool for system design, operation, analysis, decision-making, optimisation, education and training. World leading operations are increasingly relying on modelling and simulation to develop more efficient systems and to produce higher quality products and services. Modelling and simulation allows scientists and engineers a better understanding of three-dimensional and time-dependent phenomena, as well as providing a platform for predicting future behaviour. This paper covers aspects of a keynote speech delivered by Saeid Nahavandi which focus on the challenges associated with the modelling and simulation of engineered systems and discusses how knowledge visualisation can provide effective communication to various levels of organisational management. Through examining the concepts of knowledge visualization, performance and spatial cognition and its relationship with user performance, perceptions and feedback on a series of assembly operations, tangible benefits of knowledge creation and representation of a 3 dimensional engineered system for training of complex motor and technical skills are shown.

**Keywords:** Knowledge visualization, Training, Virtual environments, Performance, Spatial Cognition.

## 1 Introduction

The use of visual representations to improve the creation and transfer of knowledge between individuals has been widely studied in the past [1, 2]. It has been argued that all graphic means can be used to construct and convey complex information and enable the correct reconstruction, remembrance and appliance of such information [3]. Earlier research has shown benefits of information visualization to amplify cognition using computer-supported, interactive, visual representations of abstract data [4]. In 3 dimensional (3D) virtual environments (VEs), interaction and display techniques between computer and human can be handled with greater bandwidth of communication through advanced visualization technologies from users' perspective. By adapting an interactive approach, VEs not only capable of presenting multimodal information and system feedbacks to the user in a timely manner, but also is effective to engage user's perceptual, aural and tactile sensory channels in knowledge acquisition and training.

The goal of the present research is to show tangible benefits of knowledge creation and information representation in such an immersive, haptic-audio visual environment from user's perspective. Through investigating the ways of human exposure to a

computer generated replica of machine assembly environment; we hypothesize that effective knowledge visualization of a VE should facilitate the creation of knowledge as well as transfer of such knowledge from computer-generated virtual world to the user. Specifically, we hypothesis that: H1: Effective knowledge visualization of a VE will enable high level of task performance and spatial knowledge development using visual, auditory and haptic representations to amplify cognition and learning of assembly tasks. H2: Combined effects of visual, auditory and haptic display technologies of a 3D VE will result in high level of user satisfaction, immersion and presence.

## 2 Related Work

Haptic-Audio-Visual (haptic + audio + visual) interfaces involves integration of visual, audio and haptic displays which present information to the user through the human visual, auditory and haptic systems [5]. These multimodal interactions between the user and such computer simulated virtual world is perceived to enhance user's spatial awareness, immersion and presence, as well as sense of control in task performance. By extending the power of information visualization methods [6], VEs can improve spatial cognition, and learning, presented in the virtual world [7, 8].

### 2.1 Knowledge Visualization

Knowledge visualization in VEs is achieved through interaction techniques and display devices. For instance, HMD enabling 3D viewing of the virtual world, and head tracking of user's head position and orientation which is used rendering the world from the user's point of view in space [9, 10]. Haptic-rendering engine and hand-tracking devices (e.g. glove-attached tracker) are also effective in enabling communication between computer and user and allowing users hand's to feel kinematics and force feedback through direct manual interaction approach during [6, 11]. Thus, the user can move within the synthetic environment by interacting with the hand-based haptic device for simulated 'realistic' force/touch feedback. This is enabled by a haptic user interface, which allows computer simulations of various tasks to relay realistic, tangible sensations to a user [12]. Auditory (aural) information can be included to provide another critical source of sensory information in VEs [13]. Furthermore, graphical representation of the user's virtual hand and force sensations the user experiences during object manipulation can enhance both psychological and physical fidelity that unique to 3D VEs [14]. Such fidelity accommodated by visualization technologies shown promising results for effective training and transfer of training to real environment.

Researchers [13] claim three interrelated fidelity factors influencing the effectiveness of a training system design i.e. 1) functional fidelity: the ability of any system to support the appropriate stimulus response set; 2) psychological fidelity: the degree to which the system affords the appropriate performance cue; and 3) physical fidelity: the extent to which the system provides multi-modal sensory stimulation. It is generally assumed that if all other factors are held constant, VEs provides more realistic features as human interact with real environment, and high fidelity VEs lead to

high/better performance outcomes, compare with the ones with low fidelity. Furthermore, since generic 3D VEs often include several interface devices together to form an immersive, interactive and intuitive interaction and user experience, and that variety of fidelity factors associated with user's psychological judgments and performance that indicate design efficacy of VEs. To some degree, the intuitive principle is undoubtedly true that user's perceived interactivity, immersion and presence reflect the quality of knowledge visualization a VE induce.

## 2.2 Interaction and Immersion

Interaction can be viewed as the process that takes place when a human user operates a machine [15]. 3D VEs involve user interaction with tasks and task environment directly, in a 3D spatial context. Unlike 2D synthetic environments, 3D VEs provide opportunities for new kinds of experience in ways not possible for users of 2D. Such experience is achieved through 3D graphics and interaction techniques to map the user input (such as hand gestures captured by input devices) into the corresponding control actions and commands that function as 3D interfaces to a repository of images, sounds and haptics. In addition, research has shown physical immersion simulated via 3D I/O devices such as head and or hand tracking are the primary characteristics that make VEs so compelling to its user [9]. Therefore VEs enabling users to interact with objects and navigate in 3D space with higher degree of immersion and presence [5, 10]. This strong sense of 'being' surrounded in the virtual space (i.e. immersion) and feels in a place away from actual physical setting while experiencing a computer generated simulation [16] from the user, has shown positive impact on performance [9, 17].

## 2.3 Spatial Interaction

Spatial aspects of interaction are related to the actions and activities conducted by the users. The VE, perceived as spatial surroundings can be used to greatly enhance the interaction between the user and the world. Spatial behaviour, such as moving, orientation, position in 3D space during locomotion in which human motor capabilities permit movement is important of human-VE interaction experience. Integration of VE system components such as handheld 3D mice, data glove and 3D trackers can facilitate user's spatial interaction experience. For example, handheld 3D mice can be used to control VE object's position and orientation. Data glove as a hi-tech hand-shaped pointing device can be used to initiate commands and to navigate through the VE. 3D trackers can generate the VE's images according to the movements of the user's body (through measure real time change in a 3D object position and orientation) [18]. In VE mediated learning context, effective human-VE interaction relies on dynamic and timely system feedback to the user and rapid user adoption of various system input control devices. Although the quality of feedback and adoption can be determined by the skills of the designer to produce a system that can be respond appropriately to the user's inputs [19], the quality of the training system or program is determined by users' feedback and their perceptions of the design effectiveness. Research has shown the most important potential contribution of 3D VEs to conceptual understanding is through facilitation of spatial knowledge development [20].

### 3 System Evaluation

A virtual training simulator for object assembly operations has been developed at the Centre for Intelligent Systems Research (CISR), Deakin University, as shown in Figure 1. The system was to support the learning process of general assembly operators as well as provide an intuitive training platform to enable assembly operators to perform their learning practices, repeatedly, until they are proficient with their assembly tasks and sequences. By imitating the real physical training environments within the context of visualization and physical limitations, the system capability of providing haptic, audio, visual feedbacks to the user dynamically. To access its design efficacy on knowledge visualization, seventy six volunteers of undergraduate and postgraduate students as well as academic staff ( $N=76$ ; 56 male and 20 female) with diverse background and age-level differences: 18-24 ( $N=32$ ), 25-34 ( $N=33$ ), 33-45 ( $N=8$ ) and over 46 ( $N=3$ ), recruited from School of Engineering, Deakin University performed a series of object assembly tasks in the VE training system. This study was approved by the Deakin University Ethics Committee.



**Fig. 1.** VE Training system interface

In the present study, performances were measured both objectively on participants' real time task performance, and recognition and recall on memory-test; and subjectively on user's perceptions of audio, visual and haptics feedback in convey information to facilitate the user effective learning. A self-report user perception measure of perceived VE efficacy (PVE) was utilized to measure the individual's beliefs of the effectiveness to which the VE assisted them in learning object assembly tasks. A 7-point Likert scale was used to gather participants' rating for each item, ranging from 1 (Very strongly disagree) to 7 (Very strongly agree). Sample questions include "I have a strong sense of "being there" (sufficiently immersed) in the virtual training environment", "my experience with the virtual training seemed consistent with my real world experience" and "the virtual environment accurately replicates a real world training assembly". Higher ratings are considered to indicate higher perception of VE efficacy.

## 4 Results and Discussion

This section is organized by exploring the results pertaining to the first hypothesis regarding effect of knowledge visualization of a VE on performance and spatial learning, followed by the second hypothesis regarding user perception and affect of immersion, presence and satisfaction. These are derived from multimodal information collected using various measurement methods.

### 4.1 Task Performance and Spatial Knowledge Development

As shown in Table 1, 75 participants successfully completed one or more assembly tasks (T1 to T7) at various level of difficulty (Low to High) within 15 minutes. Overall, participants showed high level of object assembly skills after training in the VE, and mean score of task performance is 77 (SD=24, N=75). As expected accuracy was much higher for the low LOD task - T1 (N=75), in contrast to moderate or high LOD tasks - T2 to T7, with accuracy range from 27 to 67. Also user spent longer time on the task with highest difficulty level (T7). Interestingly, for assembly tasks with moderate level of difficulty, users seem to achieve mixed results in terms of efficiency. In particular, the user spent more time on task with low difficulty level (T1) than tasks at moderate level of difficulty (T2-T6). This may be due to task T1 was presented at the beginning of the test and users were getting used to the test environment and task scenario.

Memory structure of a VE may include the following dimensions – types, shapes, colours, relative locations, relative sizes, and event sequences [16]. In the present study, memory test included a list of questions and images of assembly objects and tools, they used in VE training. 19 participants responded to the memory test, and majority of them were able to recognize and recall well of learnt task procedure and tools used in VE. Overall, mean score of memory test was 72 (SD=20, N=20). 4 participants achieved full score (100). Similar to previous research (Wallet et al 2009), which found that in active learning condition recall of spatial layout was better as the subjects were able to respect perception-action coupling and to correlate motor and visual inputs. Immersive and interactive feature of the current VE and active learning approach adapted might have helped users to achieve high level task performance and performance memory test.

**Table 1.** Task performance outcomes

Object assembly tasks	Level of difficulty (LOD)	Accuracy (N=75)	Time on Task (in seconds)
1Fix radio box	Low	75	36.21
2 Drill in screw A	Moderate	71	27.69
3 Drill in screw B	Moderate	67	20.29
4 Drill in screw C	Moderate	64	12.97
5 Drill in screw D	Moderate	64	25.86
6 Fix stereo	Moderate	53	31.88
7 Fix power connector	High	27	152.04

## 4.2 User Perceptions and Affect

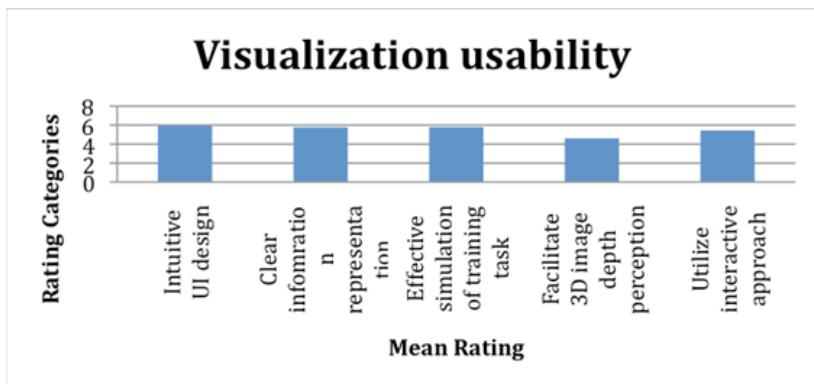
User perception refers to how the user feels regarding the VE in supporting their learning, understanding and master of trained skills. Affect is the arousal of feelings or emotions that it brings to the user. It is considered as an important aspect of interactivity [19]. Laurel [21] refers affective side of interactivity as a sense of ‘first-personeness’. Positive user affect has been associated with well designed user interfaces that engage the user’s attention and sustain the user’s sense of immersion.

*Visualization usability.* As Figure 2 illustrates, user perceived visualization usability of the VE is at high level, with mean rating close to 6 (Strongly Agree) among 5 response categories. In particular, high rating on intuitive user interface (UI) design ( $M=5.9$ ), effective simulation of training tasks ( $M=5.8$ ) and clear information representation ( $M=5.8$ ) were found that support such claim. With respect to HMD for 3D image depth perception in virtual training, the user also reported positive feedback.

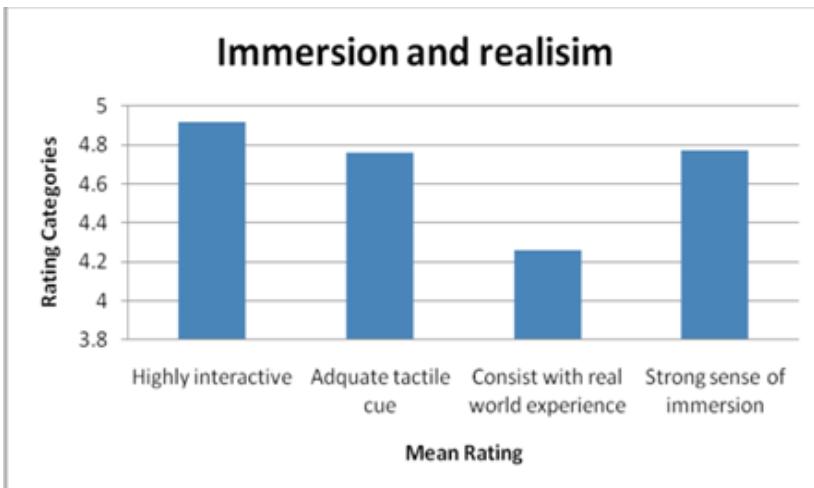
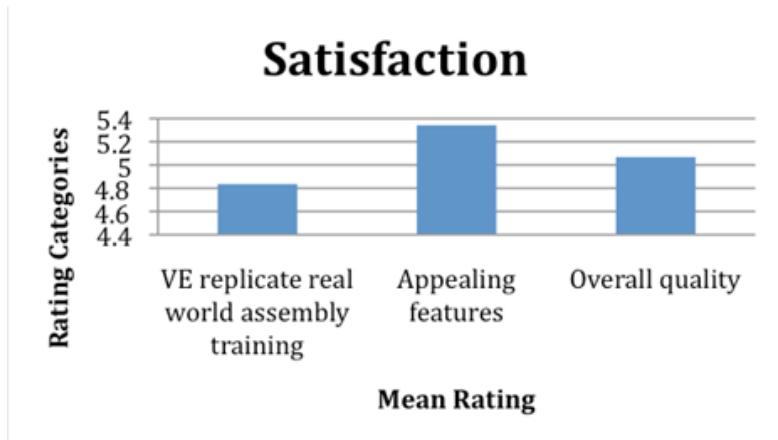
*Immersion and presence.* Figure 3 shows that users seem unsure about immersion and realism induced by the VE, with overall rating range between 4 (Nature) and 5 (Agree). Nevertheless, some indeed enjoyed different experiences that the VE brought to them and indicate this is due to unique technological characteristics that make them experience ‘reality’ feels. Interview scripts also suggest that more time on training for users to be more familiar with the task environment and multimodal information displays may enhance their feeling of immersion and presence.

*Satisfaction.* Moderately high level of user satisfaction was found in this study as Figure 4 illustrates. Users in particular seem satisfied with the overall design of the VE and the report strong satisfaction of appealing features the VE posses. Although not all users agree the VE replicates real world assembly training with mean rating slightly lower than the result of other satisfaction categories, high level of visualization usability, task performance as well as memory shown effectiveness of the overall design.

*Interactivity.* Interactivity utilises direct human feedback to stimulate the reflection by learners’ on their learning [19]. High level of interaction implies minimal latency, a sense of presence, and the ability to both access and modify content [22]. Overall, participants had positive perceptions of interactivity (see Figure 5) toward the VE.

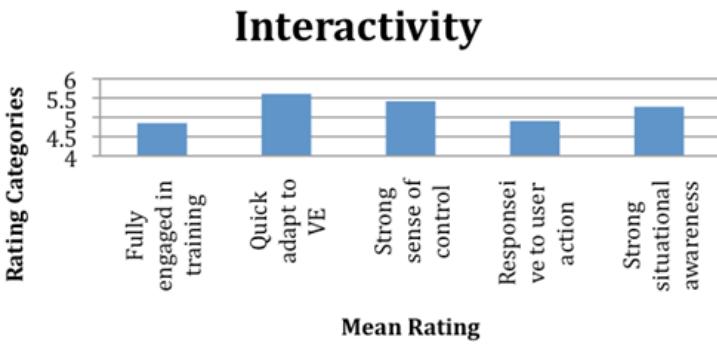


**Fig. 2.** Visualization usability

**Fig. 3.** Immersion and realism**Fig. 4.** Satisfaction

Feeling of in control is another indicator of one's confidence in performing task effectively. Research implies that higher level of feeling of in control is associated with one's performance outcome [19], which is also found in this study.

The user perceived efficacy of knowledge transfer from the VE was measured on three questions: 1) "The virtual environment helped me increase my understanding of the required assembly tasks", 2) I was able to focus my attention on learning assembly procedures rather than the input control tools (e.g. haptics device)" and 3) "It was easy to self direct my learning experience in the virtual environment". User rating shown satisfying results with mean ratings all above 5 (Agree).

**Fig. 5.** Interactivity

## 5 Conclusion

In this study, spatial knowledge and technical skill acquisition, knowledge visualization and user perceptions in a machine assembly training scenario were explored. Analysis of the data provide support for the established hypotheses, which suggest 3D VE is effective in integrating multimodal system feedback and present information appropriate to user input. In addition, support cognition and performance, as well as induce positive user perception and affect. More investigation is required to explore relationships between performance, perception and spatial knowledge learning. Future research may also need to address information due to simultaneous multimodal information feedback. This has wider implications for effective design of the VE training systems for aerospace and automotive industry and benefits research community in other fields such as medical and education.

## References

1. Clark, R.C., Chopeta, L.: *Graphics for Learning. Proven Guidelines for Planning, Designing, and Evaluating Visuals in Training Materials*. Pfeiffer, San Francisco (2004)
2. Burkhard, R.A.: Towards a Framework and a Model for Knowledge Visualization: Synergies Between Information and Knowledge Visualization. In: Tergan, S.-O., Keller, T. (eds.) *Knowledge and Information Visualization*. LNCS, vol. 3426, pp. 238–255. Springer, Heidelberg (2005)
3. Chen, C.: *Information Visualization: Beyond the Horizon*. Springer, London (2004)
4. Card, S.K., Mackinlay, J.D., Shneiderman, B.: *Readings in Information Visualization: Using Vision to think*. Morgan Kaufmann, Los Altos (1999)
5. Brough, J.E., Schwartz, M., Gupta, S.K., Anand, D.K., Kavetsky, R., Pettersen, R.: Towards the development of a virtual environment-based training system for mechanical assembly operations. *Virtual Reality* 11, 189–206 (2007)
6. Abate, A., Guida, M., Leoncini, P., Nappi, M., Ricciardi, S.: A haptic-based approach to virtual training for aerospace industry. *Journal of Visual Languages and Computing* 20, 318–325 (2009)

7. Turk, M., Robertson, G.: Perceptual user interfaces. *Communications of the ACM* 43, 33–34 (2000)
8. Datey, A.: Experiments in the Use of Immersion for Information Visualization. In: Faculty of Virginia Polytechnic Institute. Master of Science Virginia: State university (2000)
9. Raja, D., Bowman, D.: Exploroing the benefits of immersion in abstract information visualization. In: 8th International Immersive Projection Technology Workshop (2004)
10. Himberg, H., Motai, Y.: Head Orientation Prediction: Delta Quaternions Versus Quaternions. *IEEE Transactions on Systems, Man, and Cybernetics-PartB: Cybernetics* 39, 1382–1392 (2009)
11. Abate, A.F., Guida, M., Leoncini, P., Nappi, M., Ricciardi, S.: A haptic-based approach to virtual training for aerospace industry. *Journal of Visual Languages & Computing* 20, 318–325 (2009)
12. O’Malley, M.K., Gupta, A.: Haptic Interfaces. In: L HCI Beyond the GUI, pp. 25–73. Morgan Kaufmann, San Francisco (2008)
13. Muller, P., Joseph, L., Cdr, C., Schmorow, D., Stripling, R., Stanney, K., Milham, L., Whitton, M.C., Fowlkes, J.E.: The Fidelity Matrix: Mapping System Fidelity to Training Outcome. In: Interservice/Industry Training, Simulation, and Education Conference, I/ITSEC (2006)
14. Waller, D., Hunt, E., Knapp, D.: The Transfer of Spatial Knowlege in Virtual Environment Training. *J. Presence:Teleoperators and Virtual Environments* 7, 129–143 (1998)
15. Jensen, J.F.: Film Theory Meets 3D: A Film Theoretic Approach to the Design and Analysis of 3D Spaces. In: Qvortrup (ed.) *Virtual Interaction: Interaction in Virtual Inhabited 3D Worlds*, pp. 311–328. Springer, London (2001)
16. Lin, J.J.W., Dhu, H.B.L., Parker, D.E., Abi-Rached, H., Furness, T.A.: Effects of Field of View on Presence, Enjoyment, Memory, and Simulator Sickness in a Virtual Environment. In: IEEE Virtual Reality, pp. 164–171 (2001)
17. Slater, M., Linakis, V., Usoh, M., Kooper, R.: Immersion, Presence, and Performance in Virtual Environments: An Experiment with Tri-Dimensional Chess. In: ACM Symposium on Virtual Reality Software and Technology, Hong kong, pp. 163–172 (1996)
18. Burdea, G., Coiffet, P.: *Virtual Reality Technology*. John Wiley & Sons, Inc., Chichester (2003)
19. Pearce, J.M.: An investigation of interactivity and flow: student behaviour during online instruction. Department of Information Systems. Doctor of Philosophy Melbourne: The University of Melbourne, p. 342 (2004)
20. Dalgarno, B., Hedberg, J., Harper, B.: The Contribution of 3D Environments to Conceptual Understanding. In: Proceedings Of the 19th annual conference of the Australasian Society for Computers in Learning in Tertiary Education, Auckland, NZ, pp. 145–158 (2002)
21. Laurel, B.: Computer as Theatre. Addison-Wesley Perfessional, Reading (1993)
22. Brutzman, D.: Graphics Internetworking: Bottlenecks and Breakthroughs. In: Dodsworth, C. (ed.) *Digital Illusion: Entertaing the Future with High Technology*, pp. 61–95. ACM Press, New York (1998)

# Proximity-Based Federation of Smart Objects: Liberating Ubiquitous Computing from Stereotyped Application Scenarios

Yuzuru Tanaka

Meme Media Laboratory, Hokkaido University  
N13, W8, Sapporo, 060-8628 Japan  
tanaka@meme.hokudai.ac.jp

**Abstract.** This paper proposes three new formal models of autonomic proximity-based federation among smart objects with wireless network connectivity and services available on the Internet. Each smart object is modeled as a set of ports, each of which represents an I/O interface for a function of this smart object to interoperate with some function of another smart object. This paper first proposes our first-level formal modeling of smart objects and their federation, giving also the semantics of federation defined in a Prolog-like language. Then it proposes the second-level formal modeling using graph rewriting rules for developing application frameworks using a dynamically changing single federation, and finally, proposes the third-level formal modeling based on collectively autocatalytic sets for the development of complex application scenarios in which many smart objects are involved in mutually related more than one federation.

**Keywords:** Ubiquitous Computing, Pervasive Computing, Service Federation, Smart Object.

## 1 Introduction

Information system environments today are rapidly expanding their scope of subject resources, their geographical distribution, their reorganization, and their advanced utilization. Currently, this expansion is understood only through its several similar but different aspects, and referred to by several different stereotyped terms such as ubiquitous computing, pervasive computing, mobile computing, and sensor networks. No one has clearly defined this expansion as a whole. Recently it is often pointed out that the lack of a formal computation model capable of context modeling to cover this diversity as a whole is the main reason why most applications of ubiquitous computing are still within the scope of the two stereotyped scenarios [1, 2], i.e., the location-transparent service continuation, and the location- and/or situation-aware service provision. Some researchers are trying to extend the application target of formal computation models of process calculi, which were originally proposed to describe dynamically changing structures and behaviors of interoperating objects, from sets of software process objects to sets of mobile physical computing objects [1]. Such formal computation models of process calculi include Chemical Abstract Machine [3], Mobile Ambients [4], P-Systems [5], Biographical Reactive System [6], Seal Calculus

[7], Kell Calculus [8], and LMNtal [9]. These trials mainly focus on mathematical description and inference of the behavior of a set of mobile objects, but not those of the dynamically changing interconnection structures among mobile physical objects based on abstract description of their interfaces. For this reason, their formal computation models are not sufficient to develop innovative application frameworks. As to the modeling and analysis of dynamically changing topology of *ad hoc* networks, there have been lots of mathematical studies on network reconfiguration and rerouting for energy saving, for improving quality of service, and/or for maintaining connectivity against mobility [10, 11]. They focus on physical connectivity among nodes, but not on their logical or functional connectivity. These models cannot describe application frameworks. Some studies on mobile *ad hoc* networks are inspired by biological systems that share such similar features as complexity, heterogeneity, autonomy, self-organization, and context-awareness. These approaches are sometimes categorized as studies on bio-inspired networking [12]. The latter half of this paper is also bio-inspired, especially by DNA self-replication and RNA transcription mechanisms.

In expanding information environments of ubiquitous and/or pervasive computing, some resources are accessible through the Web, while others are accessible only through peer-to-peer *ad hoc* networks. Any advanced utilization of some of these resources needs a way to select them, and a way to make them interoperable with each other to perform a desired function. Here we use the term ‘federation’ to denote the definition and execution of interoperation among resources that are accessible either through the Internet or through peer-to-peer *ad hoc* communication. This term was probably first introduced to IT areas by Dennis Heimbigner in the context of a federated database architecture [13], and then secondarily in late 90s, by Bill Joy in a different context, namely, federation of services [14]. Federation is different from integration in which member resource objects involved are assumed to have previously designed standard interoperation interface. The current author has already proposed federation architectures for resources over the Web [15-18], and extended their targets to cover sensor networks using ZigBee Protocol, and mobile phone applications. These architectures are, however, still within the framework of Web-based federation.

This paper will focus on the proximity-based federation of intellectual resources on smart objects. Proximity-based federation denotes federation that is autonomously activated by the proximity among smart objects. Smart objects denote computing devices such as RFID tag chips, smart chips with sensors and/or actuators, mobile phones, mobile PDAs, intelligent electronic appliances, embedded computers, and access points with network servers.

This paper will propose three new formal models of autonomic proximity-based federation among smart objects including both physical smart objects with wireless network connectivity and software smart objects such as services on the Web. These three formal models focus on different levels, i.e., federation and interoperation mechanisms, dynamic change of federation structures, and complex application scenarios with mutually related more than one federation.

Our first-level formal modeling focuses on the federation interface of smart objects, hiding any details on how functions of each smart object are implemented. Each smart object is modeled as a set of ports, each of which represents an I/O interface of a service provided by this smart object. We consider the matching of a service-requesting query and a service-providing capability as the matching of a service-requesting port and a service-providing port. In the preceding research studies, federation mechanisms were

based on the matching of a service-requesting message with a service-providing message, and used either a centralized repository-and-lookup service as in the case of Linda [19] or multiple distributed repository-and-lookup services each of which is provided by some mobile smart object as in the case of Lime [20]. Java Space [21] and Jini [22] are Java versions of Linda and Lime middleware architectures. A recent survey on such middleware architectures can be found in [23]. In these architectures, messages to be matched are issued by program codes, and therefore the dynamic change of federation structures by message matching cannot be discussed independently from the codes defining the behavior of the smart objects.

Our first-level formal modeling allows us to discuss applications from the view point of their federation structures. This enables us to extract a common substructure from similar applications as an application framework. Our second-level formal modeling based on graph rewriting rules focuses on developing application frameworks each of which uses a dynamically changing single federation, while our third-level formal modeling focuses on developing complex application scenarios in which many smart objects are involved in mutually related more than one federation, and describes them as collectively autocatalytic sets proposed by Stuart Kauffman in the context of complex systems [24].

This paper will show how our three formal models of federation enable us to describe application frameworks not only for stereotyped applications, but also novel applications including those inspired by molecular-biological mechanisms.

## 2 Smart Object and Its Formal Modeling

Each smart object communicates with another through a peer-to-peer communication facility, which is either a direct cable connection or a wireless connection. Some smart objects may have WiFi communication and/or cellular phone communication facilities for their Internet connection. These different types of wireless connections are all proximity-based connections, i.e., each of them has a distance range of wireless communication. We model this by a function  $\text{scope}(o)$ , which denotes a set of smart objects that are currently accessible by a smart object  $o$ .

For a smart object to request a service running on another smart object, it needs to know the id and the interface of the service. We assume that each service is uniquely identified by its service type in its providing smart object. Therefore, each service can be identified by the concatenation of the object id of its providing smart object and its service type. The interface of a service can be modeled as a set of attribute-value pairs without any duplicates of the same attribute. We call each attribute and its value respectively a signal name and a signal value.

Pluggable smart objects cannot specify its access to another or its service request by explicitly specifying the object id or the service id. Instead, they need to specify the object by its name or the service by its type. The conversion from each of these three different types of reference to the object id or the service id is called ‘resolution’. These are respectively called object-name resolution and service type resolution.

When a smart object can access the Internet, it can ask a central repository-and-lookup service to perform each resolution. When a smart object can access others only through peer-to-peer network, it must be able to ask each of them to perform each resolution. Here we assume that every smart object performs required resolution for its own services.

When a service-requesting smart object  $o$  requests a service, it sends an object id  $oid'$ , an object name  $oname$ , or a service type  $stype$  to each smart object with  $oid$  as its object id in its proximity represented by  $scope(o)$ . Each recipient smart object with  $oid$ , when receiving  $oid'$ ,  $oname$  or  $stype$ , respectively performs object-id resolution, object-name resolution, or service-type resolution. Object-id resolution returns the input  $oid'$  if it is equal to  $oid$ , or ‘nil’ otherwise. Object-name resolution returns  $oid$  if the recipient has  $oname$  as its name, or ‘nil’ otherwise. Service-type resolution returns  $oid$  if the recipient provides a service of type  $stype$ , or ‘nil’ otherwise. After obtaining  $oid$ , the service-requesting smart object can directly request the object with  $oid$  for a service of type  $stype$ . The object with  $oid$  can acknowledge this request if it provides such a service. Otherwise it returns ‘nil’.

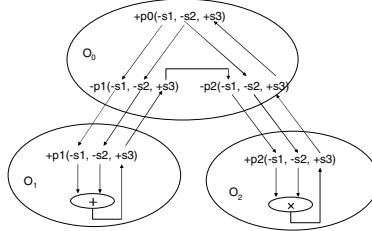
Our model represents each resolution mechanism as well as the corresponding resolution request (, namely, the corresponding access request,) as a port. Each smart object is modeled as a set of ports. Each port consists of a port type and its polarity, i.e., either a positive polarity ‘+’ or a negative polarity ‘-’. Each resolution mechanism is represented by a positive port, while each resolution request is represented by a negative port. A smart object with  $oid$  as its identifier has a port  $+oid$ . If it exposes its name  $oname$ , namely, if it allows its reference by its name, then it has a port  $+oname$ . A smart object has ports  $-oid$  and/or  $-oname$  if it requests another smart object identified by  $oid$  and/or  $oname$ . A smart object that provides a service of type  $stype$  has a port  $+stype$ . A smart object has a port  $-stype$  if it requests a service of type  $stype$ . A smart object with  $oid$  and  $oname$  may have  $+oid$  and  $+oname$  as its ports, but neither of them is mandatory. Some smart objects may hide one or both of them.

Federation of a smart object  $o$  with another smart object  $o'$  in its scope  $scope(o)$  is initiated by a program running on  $o$  or on some other activating smart object that can access both of these objects. This program detects either a specific user operation on  $o$  or the activating object, a change of  $scope(o)$ , or some other event on  $o$  or the activating object as a trigger to initiate federation. The initiation of federation with  $o'$  by a smart object  $o$  or by some other activating object performs the port matching between the ports of  $o$  and the ports of  $o'$ . As its result, every port  $-p$  (or  $+p$ ) in  $o$  is connected with a port  $+p$  (or  $-p$ ) in  $o'$  by a channel identified by their shared port type  $p$ . We assume that ports are not internally matched with each other to set any channel within a single object.

The same smart object may be involved in more than one different channel. The maximum number of channels in which the same port can be involved is called the arity of this port. Unless otherwise specified, we assume in all the examples described in this paper that the arity of each port is one.

### 3 Semantics of Federation and Software Smart Objects

Let us consider a federation among three smart objects  $O_0$ ,  $O_1$ , and  $O_2$  as shown in Figure 1, respectively having the following port sets,  $\{+p0, -p1, -p2\}$ ,  $\{+p1\}$ , and  $\{+p2\}$ . Each of these three ports has three IO signals represented by  $s1$ ,  $s2$ , and  $s3$ . The polarity  $-s$  or  $+s$  denotes that it works respectively as an input or as an output. The smart objects  $O_1$  and  $O_2$  respectively perform addition  $s3:=s1+s2$  and multiplication  $s3:=s1\times s2$ , while the service  $p0$  provided by  $O_0$  combines these two functions provided by  $O_1$  and  $O_2$  to calculate  $(s1+s2)\times s2$  as the value of  $s3$ .



**Fig. 1.** Example federation among three smart objects

Here we will describe the federation semantics using a Prolog-like description of the function of each smart object:

Example 1:

```

 $O_0: p0(x, y, z) \leftarrow \text{ext}(p1(x, y, w)), \text{ext}(p2(w, y, z))$ 
 $O_1: p1(x, y, z) \leftarrow [z:=x+y]$ 
 $O_2: p2(x, y, z) \leftarrow [z:=x\times y]$ 

```

Each literal on the left-hand side of a rule corresponds to a service-providing port, while each literal on the right-hand side corresponds to either a program code or a service-requesting port. Each literal  $\text{ext}(L)$  denotes that  $L$  is evaluated externally by some other accessible smart objects. When the service-providing port  $p0$  of  $O_0$  is accessed with two signal values ‘a’ and ‘b’, the smart object  $O_0$  begins to evaluate a goal  $\leftarrow p0(a, b, z)$ . The evaluation of this goal finally obtains that  $z=(a + b) \times b$ .

Our model treats each WiFi access point as a smart object. Once a smart object federates with some access point *apoint*, it can access whatever Web services this access point is given permission to access. Such an access point *apoint* provides the following service:

```
ResDelegation(x, y)  $\leftarrow$  isURL(x), permitted(apoint, x), [WebEval(x, y)].
```

The procedure  $\text{WebEval}(x, y)$  invokes the web service  $x$  with signals  $y$ . In the proximity of *apoint*, any smart object  $o$  with a port  $-\text{ResDelegation}$  can request this service with a Web service URL  $url$  and a list of parameters  $v$ . The access point *apoint* delegates this access request to the target Web service at  $url$  together with the parameter list  $v$ , which enables  $o$  to utilize this Web service through *apoint*.

A smart object may presume a standard API to request a service of another smart object that does not provide the compatible API but provides a downloadable driver to access this service through the presumed standard API. Such a mechanism is described as follows. Suppose that a smart object has a service to download a software smart object  $y$  satisfying the query  $x$ . This smart object has a port  $-\text{SOdownload}$  defined as follows:

```
SOdownload(x, y)  $\leftarrow$  [find(x, y)]
```

Here the procedure  $\text{find}(x, y)$  finds the software smart object  $y$  satisfying the query  $x$  specified in the list format of attribute-value pairs  $((\text{attr}1, v1), \dots, (\text{attr}k, v_k))$ . If there are more than one such smart object, it returns an arbitrary one of them. A requesting smart object with a port  $-\text{SOdownload}$  can ask this smart object to download a software smart object  $y$  that satisfies a query  $x$ , and install this software smart object  $y$  to

itself. The evaluation of the following by the requesting smart object performs both the downloading and the installation.

$$\text{SOdownloadInstall}(x) \leftarrow \text{ext}(\text{SOdownload}(x, y)), [\text{install}(y)]$$

The installation of a software smart object  $y$  by a requesting smart object  $o$  adds  $y$  in the scope of  $o$ , initiates a federation between  $o$  and  $y$ .

If a downloaded smart object is a proxy object to a Web service or another smart object, the recipient smart object can access this remote service through this proxy smart object.

## 4 Describing Some Basic Applications

### 4.1 Location-Transparent Service Continuation

Let us consider the following example (Example 2). Suppose that PDA1 has federated with Office, an access point with a server. Office provides DB service and print service that are available at the user's office. Let Home be another access point with a server providing a print service available at his or her home. The location transparent continuation of services means that he or she can continue the job that was started at the office using PDA1 even after he or she goes back home carrying PDA1. This is realized by the following mechanism. When he or she carries out PDA1 from Office environment, he or she just needs to make PDA1 download the proxy smart object of Office. This operation is called federation suspension. When arriving at home, PDA1 is WiFi connected to Home, and then federates with Home. At the same time, the proxy smart object installed in it resumes the access to Office. Therefore, they set up two channels for print services, to the one available at home and to the one at the office, and one more channel for the DB service to access the Office DB service from home. Now the PDA can access the database service of Office, and the two printing services of Office and Home. When PDA1 requests a print service, it asks its user or the application which of these services to choose. This enables the PDA user to restart his work with the same service accessibility after moving from the office to home.

This downloadable software proxy smart object can be defined as follows:

$$\begin{aligned} \text{DB}(x) &\leftarrow \text{remoteEval}(\text{DB}(x), \text{Office}) \\ \text{Print}(x) &\leftarrow \text{remoteEval}(\text{Print}(x), \text{Office}) \\ \text{suspend}(\ ) &\leftarrow [\text{disconnect}(\text{Office})] \\ \text{resume}(\ ) &\leftarrow [\text{connect}(\text{Office})] \end{aligned}$$

Here,  $\text{remoteEval}(p(x), y)$  denotes an evaluation of  $p(x)$  in a remote object  $y$  for which this proxy object works as the proxy. The last two rules denote the disconnection and connection of this proxy object from and to the smart object Office.

### 4.2 Confederation of Smart Objects

Let us consider the following example (Example 3). When a rescue center receives an emergency call, it mobilizes a rescue team. Each rescue worker of the team needs to pick up some rescue equipments necessary for the mission. In a near future, those equipments may include a wearable computer, a GPS module, a handsfree mobile

phone, a head-mount display, a small reconnaissance camera robot, and its remote controller. The rescue center has a sufficient number of equipments of each type. Depending on each mission, each worker picks up one from the stock of each different type of equipment. The set of picked-up equipments may differ for different missions. These equipments are advanced IT devices, and can interoperate with each other. It is necessary to set up instantaneously all the necessary federation channels among those equipments picked-up by each worker. These federations should be able to avoid cross-talks between equipments picked up by different workers. Suppose each equipment A of a worker P needs to interoperate with his another equipment B, A and B should not interoperate with B and A of another worker P' even if P and P' works within their proximities. We need a new mechanism for the instantaneous setting-up of such federations among a set of equipments for each of more than one worker. We call such a mechanism a ‘confederation’ mechanism.

We define confederation in general as follows. Consider an  $n$ -tuple of smart objects  $o_1, o_2, \dots, o_n$ . Initially, they are independent from each other. Let  $\text{type}(o)$  denote the type of the smart object  $o$ . The type of a tuple of smart objects  $(o_1, o_2, \dots, o_n)$  is defined as  $(\text{type}(o_1), \text{type}(o_2), \dots, \text{type}(o_n))$ . Suppose we have more than one tuple of the same type. Confederator is a smart object or a complex of smart objects that sets up the same set of federations among the smart objects of each of more than one tuple of the same type. The setting-up should be performed by proximity-based federation between the confederator and each tuple of smart objects.

The above example of a rescue team shows a potential application of a confederator mechanism.

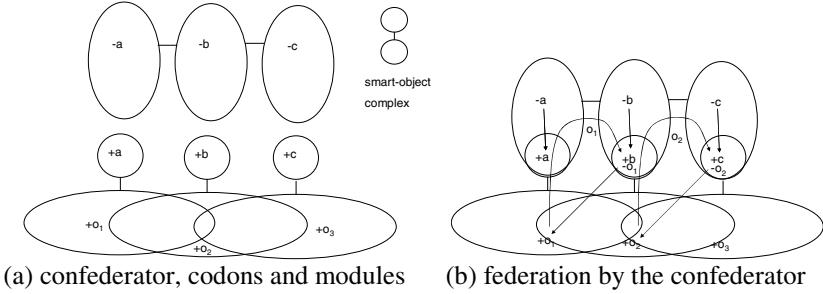
Figure 2 (a) proposes a confederator framework. Suppose we have 3 types of modules  $o_1, o_2$ , and  $o_3$  to be federated with each other. We use special smart objects called Confederator object and codon objects. Each Confederator object has a confederation requesting port, -a, -b, or -c, whereas each codon object has a confederation providing port +a, +b, or +c. Confederator objects are connected together by cables or by radio to form a smart-object complex. In order to provide each module object with an automatic confederation capability, we attach a codon object to each module object in advance. Geographically, confederation objects may be far away from each other if each module object has wide-range connectivity such as mobile phone connectivity.

When every codon object simultaneously enters the proximity of its partner confederator object, the port matching mechanism establishes three channels a, b, and c as shown in Figure 2 (b). Then the confederator complex can use the channel a to get the oid  $o_1$  of the module object, and then use the channel b to make the codon object with +b create a new port - $o_1$ . Similarly, using the channels b and c, the confederator complex can make the codon with +c create a new port - $o_2$ . These newly created two ports then establish two new channels  $o_1$  and  $o_2$ , which allows three modules interoperate with each other through their linear connection.

The above mechanism can be implemented as follows. For each confederation setting-up, the confederator complex can just evaluate the following program:

```
Conf( ) ← ext(a(x1, nil)), ext(b(x2, x1)), ext(c(-, x2))
```

Each codon is defined as follows, where  $\text{report}(oid)$  denotes a requesting port  $-oid$ , and  $\text{report}(nil)$  is nil.



**Fig. 2.** A confederation framework

```
a(x, y) ← [x←oid(child), createPortPairs((+partnerOid, reqPort(y)))]
b(x, y) ← [x←oid(child), createPortPairs((+partnerOid, reqPort(y)))]
c(x, y) ← [x←oid(child), createPortPairs((+partnerOid, reqPort(x)))]
```

The respective evaluation of `createPortPairs((+partnerOid, -oid))` and `createPortPairs((+partnerOid, nil))` adds the following two rules:

```
partnerOid(x, y) ← ext(oid(x, y)),
partnerOid(x, y) ← [y:=x].
```

The followings are example codes for the three module objects. In this example, the execution of the application program `Appl` by the third object invokes a program of the second, which in turn invokes another in the first.

```
o1: o1(x1, y1) ← [prep1(x1, x3)], partnerOid(x3, y3), [postp1(y3, y1)]
o2: o2(x2, y2) ← [prep2(x2, x1)], partnerOid(x1, y1), [postp2(y1, y2)]
o3: Appl(x, y) ← o3(x, y)
o3(x3, y3) ← [prep3(x3, x2)], partnerOid(x2, y2), [postp3(y2, y3)]
```

The interoperation structure can be arbitrarily designed by appropriately defining the application program of the confederator. The execution of the following in the Confederator establishes a ring connection among three objects.

```
Conf( ) ← ext(a (x1, nil)), ext(b(x2, x1)), ext(c(x3, x2)), ext(a(-, x3))
```

The execution of the application program `Appl` in the third object now invokes a program in the second object, which in turn invokes one in the first, which finally invokes one in the third again. This program can be terminated by one of the preprocessing and post processing procedures `prepi` and `postpi`. Here it should be noticed that programs in codons and objects need not be changed.

## 5 Modeling with Rewriting Rules

In addition to the first-level formal modeling, here we propose the second-level formal modeling of proximity-based federation. The former focuses on semantics of federation, while the latter focuses on the dynamic change of interconnection structures among smart objects in a single complex federation. This model describes a

system of smart objects as a directed graph in which each node represents either a smart object or a port, and each directed edge represents either a channel or a proximity relationship. A channel and a proximity relationship are represented respectively by a black arrow and a gray arrow. A port node with an outgoing (or incoming) channel edge  $p$  to (from) an object node  $o$  denotes that  $o$  has a service-providing (service-requesting) port  $+p$  ( $-p$ ). Each object node has its state and its type. Smart objects of the same type share the same port set and the same functions.

As mentioned in Chapter II, federation of a smart object  $o$  with another smart object  $o'$  in its scope  $\text{scope}(o)$  is initiated by a program running on  $o$  or on some other activating smart object that can access both of these objects. The formalization with graph rewriting rules aims to describe the dynamic change of the channel connections among smart objects through the activation of federation rules.

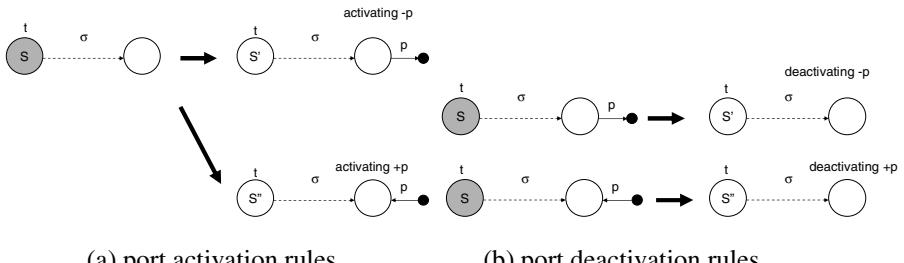
Each rewriting rule is specified as a combination of the following four types of rules, i.e., port activation/deactivation rules, state setting rules, channeling rules, and channel dependency rules. In each of the following rules, its gray node denotes that this rule is stored in this node and executed by this node. Each type of rules is designed to satisfy reasonable hardware and performance constraints of the smart objects of our concern so that it can be executed locally without any global information about the current overall federation structure.

Port activation and deactivation rules have the forms in Figure 3 (a) and (b). A dotted arrow  $\sigma$  denotes a channel path, i.e., a sequence of channels of the same direction whose length may be zero. The gray node of type  $t$  can activate a specified port of the right node through the channel  $\sigma$ , and change the state of itself. The two black smaller nodes denote port nodes.

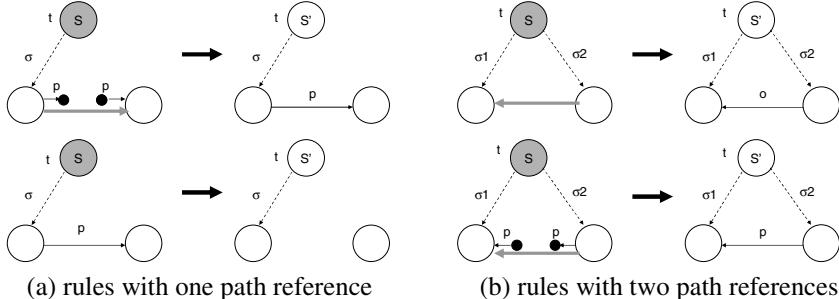
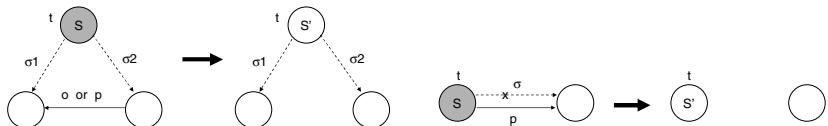
State setting rules have the form in Figure 4, where  $S' = T$  if the length of  $\sigma$  is zero.

Figure 5 shows the form of channeling rules for setting channels. In each rule in Figure 5 (a), the activation node (i.e., the gray node) can activate or deactivate a specified port of the left node through the channel  $\sigma$  to establish or to break the corresponding channel with its neighboring node. The length of  $\sigma$  may be zero. In each rule in Figure 5 (b), the activation node (i.e., the gray node) can read the oid of the left node through the channel  $\sigma_1$ , and ask the right node to create the corresponding oid-requesting port in itself. The length of either  $\sigma_1$  or  $\sigma_2$  may be zero.

Figure 6 shows the form of channeling rules for breaking channels. The activation node (i.e., the gray node) can break a specified channel between the left node accessible through  $\sigma_1$  and the right node accessible through  $\sigma_2$ . The length of either  $\sigma_1$  or  $\sigma_2$  may be zero.



**Fig. 3.** Port activation/deactivation rules

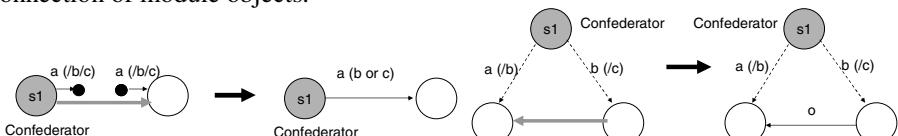
**Fig. 4.** State setting rules**Fig. 5.** Channeling rules for setting channels**Fig. 6.** Channeling rules for breaking channels      **Fig. 7.** Channel dependency rule

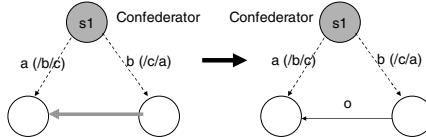
Channel dependency rules have the form in Figure 7, where the channel  $p$  is assumed to depend on the channel path  $\sigma$ . Whenever  $p$  is to be used, the existence of the channel path  $\sigma$  is checked. If  $\sigma$  is broken, the channel  $p$  is also broken.

## 6 Application Frameworks Based on Rewriting Rules

### 6.1 Confederation of Smart Objects Revisited

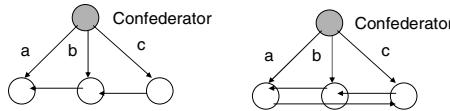
The linear connection of module objects by a confederator in Example 3 can be reformulated by using the rewriting rules as shown in Figure 8, where the notation such as a  $(/b/c)$  is used to summarize three rules into a single rule. All these rules are executed by the confederator to set up a linear connection from  $o_3$  to  $o_2$  to  $o_1$ . Here the confederator as a compound object is represented as a single node for simplicity. You may also replace the second set of rules with the set of rules in Figure 9 for the cyclic connection of module objects.

**Fig. 8.** The rewriting rules for the linear connection in Example 3



**Fig. 9.** The rewriting rules for the cyclic connection in Example 3

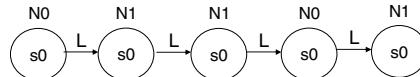
All these rules are executed by the confederator to set up either a linear connection from o<sub>3</sub> to o<sub>2</sub> to o<sub>1</sub> (in case of the first set of rules) or a cyclic connection from o<sub>3</sub> to o<sub>2</sub> to o<sub>1</sub> to o<sub>3</sub> (in case of the second set of rules) as shown in Figure 10. The three channels between the confederator and the three objects will be naturally broken when the confederator departs from them.



**Fig. 10.** Different rule sets in the confederator can set up different connections among module objects, for example, a linear connection and a cyclic connection

## 6.2 Self-replication of DNA Strand

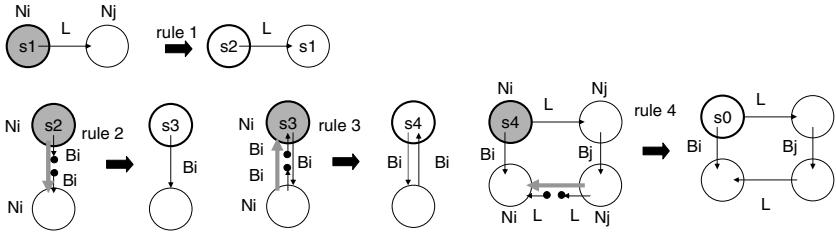
In order to show the potentiality of our formal modeling based on rewriting rules, let us simulate the self-replication of a DNA strand as shown in Figure 11. Here, nucleotides are modeled as smart objects of only two different types N0 and N1 instead of four, i.e., A, T, G, and C. Each node of type N1 has both +B0 and -B0 ports, while each node of type N1 has both +B1 and -B1 ports.



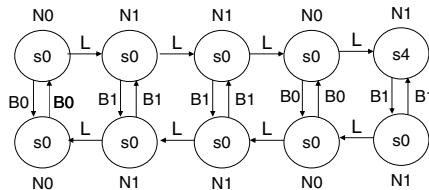
**Fig. 11.** A DNA strand using nucleotide smart objects

Each nucleotide can perform the four rules as shown in Figure 12, where the index *i* is either 0 or 1. Here we assume that there are lots of nucleotide smart objects of both types at state s<sub>0</sub> in the proximity of the original strand. If the left most node of the original DNA strand changes its state to s<sub>1</sub>, then the application of these rules will result in the double strands of DNA with opposite directions as shown in Figure 13.

This process can be also formally described at the semantic level using our Prolog-like modeling of federation. The more complex biomolecular processes can be also similarly simulated using our formal modeling of smart object federation. Such complex processes include the replication of a portion of DNA sequence to an mRNA, and the transcription from an mRNA to a sequence of amino acids with the help of tRNAs to create a protein.



**Fig. 12.** Four rules of nucleotide smart objects



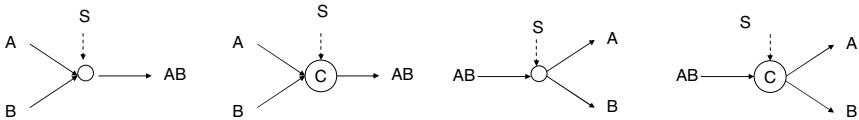
**Fig. 13.** Double strands of DNA with opposite directions obtained by a self-replication process

## 7 Modeling as a Collectively Autocatalytic Set

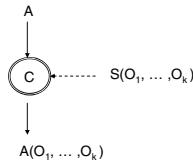
Stuart A. Kauffman believes that a collectively autocatalytic set is one of the essential mechanisms for the self-organization of life. The current author believes that complex application scenarios of autonomic federation of smart objects can be modeled as collectively autocatalytic sets. A collectively autocatalytic set is a network of catalytic reactions in which a product of each reaction may work as a source material of another reaction or as a catalyst to enhance or to repress another reaction. Each reaction is either a composition to produce a compound product from more primitive source materials, or a decomposition to decompose a source material into its component materials. These two types of reactions are shown in Figure 14, where a stimulus  $S$  and a context  $C$  such as a substrate both work as catalysts.

In smart object federation, source materials  $A$  and  $B$  are considered as smart objects, while a compound product  $AB$  corresponds to a federation of  $A$  and  $B$ . Strictly speaking, since the way they federate with each other may depend on the reaction type with  $C$  and  $S$ , the compound product  $AB$  needs to be suffixed with  $C$  and  $S$  or with the reaction type. Composition and decomposition reactions respectively correspond to federalization and defederalization actions. The second composition reaction in Figure 14 can be interpreted for example as a federation of two cars  $A$  and  $B$  in an intersection area  $C$  to avoid their collision in case of road surface freezing that is detected by a sensor  $S$ . During their federation, two cars communicate with each other to automatically maintain a safe distance from each other.

In addition to the above four primitives, we add one more primitive as shown in Figure 15 for the downloading of software smart objects, where  $O(O_1, \dots, O_k)$  denotes that a smart object  $O$  has software smart objects  $O_1, \dots, O_k$  in itself.



**Fig. 14.** Composition and decomposition catalytic reactions



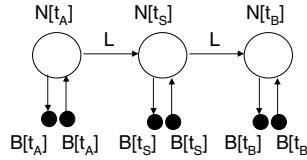
**Fig. 15.** The downloading of software smart objects

This primitive may have the following application examples. The first one is an example of the situation-aware service provision that is one of the two stereotyped application scenario types of the ubiquitous computing. The other one is the location-transparent service continuation scenario type to which Example 2 in Section 4.1 belongs.

When a mobile phone which is already set to its exhibition explanation mode comes closer to some exhibition object, its explanation program is automatically downloaded to this phone and executed there.

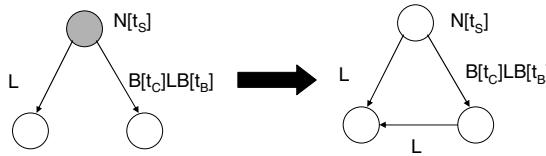
The next example downloads multiple software smart objects. A cellular phone A with its mode set to ‘house-moving’ mode can automatically download the proxy objects  $O_1, \dots, O_k$  from the server S to access the several required registration forms such as the moving-in notification form, the telephone connection form, and town gas and water supply request forms, as soon as its owner arrives at the local station of the destination town. Each  $O_i$  accesses a specific registration form and automatically fills in this form to complete a required procedure.

Now we need to show how these primitive reactions can be implemented using a generic framework based on our graph rewriting rule modeling of smart object federation. Our generic framework uses an extension of the nucleotide smart object described in Section 6. Instead of using only two types of nucleotides, we use a sufficiently large number of different nucleotide types whose indices are considered as tag codes that identify different types of primitive materials, i.e., different types of smart objects. In order to clarify this, we use the notation  $N[i]$  and  $B[i]$  instead of  $N_i$  and  $B_i$ , and furthermore, for simplicity, we identify each smart object type with its code used as an index. Therefore, we can use notations  $N[t_A]$ ,  $N[t_B]$ ,  $N[t_C]$ , and  $N[t_S]$  to denote nucleotide smart objects that are used as tagging objects to identify our smart object types  $t_A$ ,  $t_B$ ,  $t_C$ , and  $t_S$ . In our collectively autocatalytic set framework, each source material smart object or stimulus smart object of type  $t$  is generically implemented as a compound object with its corresponding tag object  $N[t]$ , while a context of type  $t_C$  accepting source material smart objects of type  $t_A$  and  $t_B$  together with a stimulus smart object of type  $t_S$  is generically implemented as shown in Figure 16.



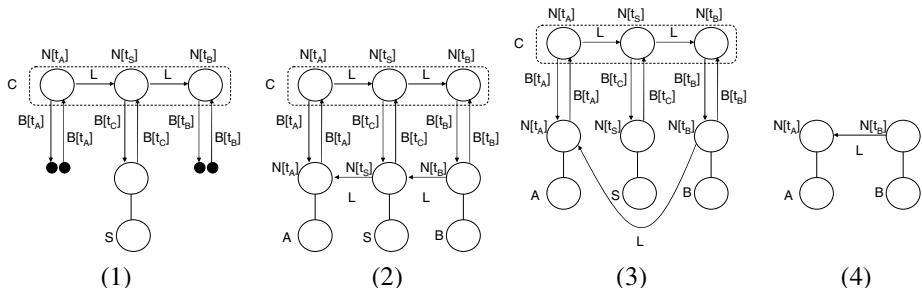
**Fig. 16.** The generic implementation of a context C accepting source material smart objects of type  $t_A$  and  $t_B$  together with a stimulus smart object of type  $t_S$

This framework uses the same rewriting rules as those shown in Section 6.2, and one additional rule in Figure 17 for a stimulus S. This rule is used to establish a channel connection from B to A in the product AB. You may change the direction of the left arrow L in this rule to establish a channel connection from A to B in the product AB. This rule is also generic because it is independent from A, B, and C.



**Fig. 17.** One additional rule for the collectively autocatalytic set framework

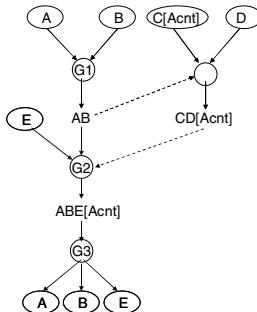
First, a stimulus smart object S with its tag object enters the proximity of the context C to form a federation with C (Figure 18 (1)). Then, the two source material smart objects A and B with their tag objects enter the proximity of C and S, and forms a federation (Figure 18 (2)). The stimulus object executes the last rewriting rule to form the federation as shown in Figure 18 (3). Finally, the two compound smart objects A and B together with their tagging objects leave the context, maintaining their federation as shown in Figure 18 (4). Using the federation (4) and a channel dependency rule in B, the smart object B can establish a channel connection of an arbitrary service type to A. Therefore, this framework gives a generic implementation of a catalytic reaction.



**Fig. 18.** A generic architecture for a catalytic composition reaction of smart objects

We can also define a generic architecture for a catalytic decomposition reaction of smart objects in a similar way.

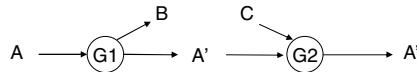
In Figure 19, we show an example scenario using more than one catalytic reaction. This scenario is not an innovative one but may be sufficient to explain how to model a smart object federation application scenario as a collectively autocatalytic set.



**Fig. 19.** Modeling a smart object federation in a 3D interactive movie theater as a collectively autocatalytic set

A user with a smart social ID card A and a smart member card B passes a check-in gate G1 of an interactive 3D movie theater. This gate sets up a federation between A and B to check if he or she is a registered member of this theater. He or she can pick up a stereoscopic pair of glasses D with an overlaid information display function, and an interactive controller C with an accounting software smart object Acnt. These two smart objects D and C[Acnt] are automatically federated to a compound smart object CD[Acnt] with the help of the federation AB as a security key so that C may output information on the display D. No user operation is necessary to set up the necessary connection between C[Acnt] and D. While viewing a movie, he or she can interactively issue purchase orders of items appearing in the movie. The software smart object Acnt records these orders. After viewing the movie, he or she passes through a gate G2 with CD[Acnt], which federates a mobile phone E with AB and downloads Acnt from CD to the compound object ABE to change it to ABE[Acnt], which enables the mobile phone to ask the user to check the purchase order records in Acnt, and then to automatically send all these records as well as the payment information. Then finally, the exit gate G3 decomposes the federation ABE[Acnt] into A, B, and E, and deletes Acnt.. The federation CD[Acnt] is decomposed just by separating C and D.

Figure 20 shows another application scenario modeled as a collectively autocatalytic set. This shows a freeway entrance G1 and a freeway exit G2. Each of the smart objects A, A' and A'' is a compound smart object that linearly federate all the cars in each freeway interval. If a car B exits at G1 then G1 defederalizes B from A to obtain A'. If a car C enters at G2 then G2 federalizes C to A' to obtain A''. In each of the federation A, A' and A'', all the cars in each federation communicate with each other to avoid their collision as well as to keep the traffic speed within a safe and efficient range by externally accelerating or deaccelerating each car when necessary.



**Fig. 20.** Modeling a smart object federation on a freeway as a collectively autocatalytic set

## 8 Concluding Remarks

This paper has proposed three new formal models of autonomic proximity-based federation among smart objects with wireless network connectivity and services available on the Internet. These three formal models focus on different levels, i.e., federation and interoperation mechanisms, dynamic change of federation structures, and complex application scenarios with mutually related more than one federation. This paper has shown how simple federation frameworks and complex federation scenarios can be described in these models. Each smart object is modeled as a set of ports, each of which represents an I/O interface for a function of this smart object to interoperate with some function of another smart object. Our models focus on the matching of service-requesting queries and service-providing capabilities that are represented as service-requesting ports and service-providing ports, instead of the matching of a service requesting message with a service-providing message. This enables us to discuss federation structures, their dynamic changes, and their mutual relationships, independently from the IO and internal processing behaviors of smart objects and compound smart objects. This enables us to extract a common substructure from similar applications or scenarios as an application or scenario framework, which further enable us to design, to discuss, and to analyze application and scenario frameworks.

In our first-level formal modeling, we have modeled smart objects and their federation as objects with ports and as a port matching mechanism. We have defined the semantics of federation in a Prolog-like language. In the second-level formal modeling, we have proposed a graph rewriting model for describing application frameworks using a dynamically changing single federation. Finally, we have proposed the formal modeling of complex application scenarios, in which many smart objects are involved in mutually related more than one federation, as collectively autocatalytic sets.

Some example applications and scenarios given in this paper are beyond the scope of stereotyped applications of ubiquitous, pervasive, and/or mobile computing. We have shown that molecular-biology may be a gold mine of ideas for developing innovative applications and scenarios of proximity-based federation of smart objects.

## References

1. Milner, R.: Theories for the Global Ubiquitous Computer. In: Walukiewicz, I. (ed.) FOS-SACS 2004. LNCS, vol. 2987, pp. 5–11. Springer, Heidelberg (2004)
2. Henricksen, K., Indulska, J., Rakotonirainy, A.: Modeling Context Information in Pervasive Computing Systems. In: Mattern, F., Naghshineh, M. (eds.) PERVASIVE 2002. LNCS, vol. 2414, pp. 167–180. Springer, Heidelberg (2002)
3. Berry, G., Boudol, G.: The Chemical Abstract Machine. In: Proc. POPL 1990, pp. 81–94. ACM, New York (1990)

4. Cardelli, L., Gordon, A.D.: Mobile Ambients. In: Nivat, M. (ed.) FOSSACS 1998. LNCS, vol. 1378, pp. 140–155. Springer, Heidelberg (1998)
5. Păun, G.: Computing with Membranes. *J. Comput. Syst. Sci.* 61(1), 108–143 (2000)
6. Milner, R.: Biographical Reactive Systems. In: Larsen, K.G., Nielsen, M. (eds.) CONCUR 2001. LNCS, vol. 2154, pp. 16–35. Springer, Heidelberg (2001)
7. Castagna, G., Vitek, J., Zappa Nardelli, F.: The Seal Calculus. *Information and Computation* 201(1), 1–54 (2005)
8. Schmitt, A., Stefani, J.-B.: The Kell Calculus: A Family of Higher-Order Distributed Process Calculi. In: Priami, C., Quaglia, P. (eds.) GC 2004. LNCS, vol. 3267, pp. 146–178. Springer, Heidelberg (2005)
9. Ueda, K., Kato, N.: LMNtal: A Language Model with Links and Membranes. In: Mauri, G., Păun, G., Jesús Pérez-Jímenez, M., Rozenberg, G., Salomaa, A. (eds.) WMC 2004. LNCS, vol. 3365, pp. 110–125. Springer, Heidelberg (2005)
10. Santi, P.: Topology Control in Wireless and Ad Hoc Sensor Networks. *ACM Computing Surveys* 37(2), 164–194 (2005)
11. Chinara, S., Rath, S.K.: A Survey on One-Hop Clustering Algorithms in Mobile Ad Hoc Networks. *Journal of Network and Systems Management* 17(1-2), 183–207 (2009)
12. Dressler, F., Akan, O.B.: A Survey on Bio-Inspired Networking. *Computing Networks* 54(6), 881–900 (2010)
13. Heimbigner, D., McLeod, D.: A federated architecture for information management. *ACM Transactions on Information Systems (TOIS)* 3(3), 253–278 (1985)
14. Edwards, W.K., Joy, B., Murphy, B.: Core JINI. Prentice Hall Professional Technical Reference (2000)
15. Tanaka, Y., Fujima, J., Ohigashi, M.: Meme media for the knowledge federation over the web and pervasive computing environments. In: Maher, M.J. (ed.) ASIAN 2004. LNCS, vol. 3321, pp. 33–47. Springer, Heidelberg (2004)
16. Tanaka, Y., Ito, K., Fujima, J.: Meme media for clipping and combining web resources. *World Wide Web* 9(2), 117–142 (2006)
17. Tanaka, Y.: Knowledge federation over the web based on meme media technologies. In: Jantke, K.P., Lunzer, A., Spyros, N., Tanaka, Y. (eds.) Federation over the Web. LNCS (LNAI), vol. 3847, pp. 159–182. Springer, Heidelberg (2006)
18. Tanaka, Y.: Meme Media and Meme Market Architectures: Knowledge Media for Editing, Distributing, and Managing Intellectual Resources. Wiley-IEEE Press (2003)
19. Gelernter, D.: Generative communication in linda. *ACM Trans. Program. Lang. Syst.* 7(1), 80–112 (1985)
20. Picco, G.P., Murphy, A.L., Roman, G.C.: Lime: Linda meets mobility. In: ICSE 1999: Proceedings of the 21st international conference on Software engineering, pp. 368–377. IEEE Computer Society Press, Los Alamitos (1999)
21. Sun Microsystems: Javaspaces service specification, version 1.2 (2001)
22. Sun Microsystems: Jini technology core platform specification, version 1.2 (2001)
23. Collins, J., Bagrodia, R.: Programming in Mobile Ad Hoc Networks. In: Proc. the 4th Conf. on Wireless Internet. WICON 2008, article no. 73 (2008)
24. Kauffman, S.: Investigations. Oxford University Press, Oxford (2000)

# A Neural Network Model to Develop Urban Acupuncture

Leandro Tortosa<sup>1</sup>, José F. Vicent<sup>1</sup>, Antonio Zamora<sup>1</sup>, and José L. Oliver<sup>2</sup>

<sup>1</sup> Departamento de Ciencia de la Computación e Inteligencia Artificial  
Universidad de Alicante

Ap. Correos 99, E-03080, Alicante, Spain

<sup>2</sup> Departamento de Expresión Gráfica y Cartografía  
Universidad de Alicante  
Ap. Correos 99, E-03080, Alicante, Spain

**Abstract.** The urban world of the 21st century is composed of numerous nodes, streams and webs, which create a new landscape of globalization and impose different logic of space and time perception. Therefore, the urban infrastructure is updated and its networks are continuously multiplied. A method known as urban acupuncture on the one hand tests the local effects of every project, and on the other hand establishes a network of points or dots to act upon. The main objective of this paper is to relate the concept of urban acupuncture with the use of a neural network algorithm to determine those points where developing actions in order to improve the quality life in cities. We apply the neural network model GNG3D to the design of a simplified network in a real city of our surrounding.

## 1 Introduction

The city is, first of all, a place for life and relationships, and public spaces are where these relationships develop under everybody's view. The quality of this space, independently from the quality of the built environment, somehow influences the quality of relationships.

The urban design concerns primarily with the design and management of public space in towns and cities, and the way public places are experienced and used. We can see [3][5][13] as general references to introduce general ideas and concepts related to urban development.

Jaime Lerner, the three-time former mayor of Curitiba, Brazil, a city best known for its innovative approaches to urban planning, is calling for what he terms *urban acupuncture* to bring revitalization and sustainability to the world's metropolitan areas. Lerner thinks that tackling urban problems at appropriate *pressure points* can cause positive ripple effects throughout entire communities (see [15]). Lerner noted that even the poorest cities can boost their standards of living by using techniques like bus rapid transit, designing multiuser buildings, and encouraging residents to live closer to their workplaces. Although many cities spend decades building underground rail systems or other costly long-term projects, every city can improve its quality of life in 3 to 4 years.

In urban centers, the only possible intervention is through pointed operations, or networks of points, by trying to create a system, through small seams and interventions of substitution (see [20]). The problem we face is where to place the *pressure points* or interventions in the urban area. Here is where we introduce neural networks algorithms based on self-organizing learning.

Self-organizing networks are able to generate interesting low-dimensional representations of high-dimensional input data. The most well known of these models is the Kohonen Feature Map [11][12]. It has been used in the last decades to study a great variety of problems such as vector quantization, biological modeling, combinatorial optimization and so on. We may also use algorithms based on self-organizing learning to simplify 2D original meshes. Then, it may be interesting for developing urban acupuncture actions to obtain simplified meshes from the original, bearing in mind that these new meshes have a small number of nodes and its shape is similar to the original.

The neural network model we are going to use is the Growing Neural Gas 3D (GNG3D) [12], an unsupervised incremental clustering algorithm which is able to produce high quality approximations of polygonal models. In [19] we have a detailed description of the efficiency of the model and some examples.

Therefore, the main objective of this work is to apply a neural network model to the field of urban acupuncture, so that we can use a self-organizing algorithm to determine a network of points on which to develop possible actions to improve or revitalize certain urban areas.

## 2 The Neural Network Model

The model GNG3D has been designed taking as a basis the Growing Neural Gas model (GNG) [9][10], with an outstanding modification consisting on the possibility to remove some nodes or neurons that do not provide us relevant information about the original model. Besides, it has been added a reconstruction phase in order to construct the faces of the optimized mesh.

The GNG3D model consists of two distinct phases: a self-organizing algorithm and a reconstruction phase. The self-organizing algorithm is developed by applying an extension algorithm of the GNG model, and the primary goal of this algorithm is to obtain a simplified set of vertices representing the best approximation of the original mesh. In the reconstruction phase we use the information provided by the algorithm to reconstruct the faces obtaining the optimized mesh as a result.

As our objective is to relate the model GNG3D with the problem of urban acupuncture, we focus on the self-organizing algorithm, since we can use it to obtain the optimum positions of the vertices in the final simplified mesh.

In the following, we consider networks consisting of

- a set  $A = \{n_1, n_2, \dots, n_N\}$  of nodes (it is equivalent to use the term vertices or neurons to refer to the nodes of the network),
- a set  $E = \{e_1, e_2, \dots, e_L\}$  of connections or edges among node pairs.

Then, the self-organizing algorithm can be summarized as:

INIT: Start with two nodes  $a$  and  $b$  at random positions  $w_a$  and  $w_b$  in  $R^n$ . Initialize the error variable to zero.

1. Generate an input signal  $\xi$  according to  $P(\xi)$ .
2. Find the nearest node  $s_1$  and the second nearest  $s_2$  to the input signal.
3. Increment the age of all edges emanating from  $s_1$ . If the age of any edge is greater than  $a_{max}$ , then mark it in order to be eliminated afterwards.
4. Increment the local activation counter variable of the winner node. Add the square distance between the input signal and the nearest node in input space to a local error variable:

$$\Delta error(s_1) = \|w_{s_1} - \xi\|^2$$

Store the nodes with the highest and lowest value of the local counter variable

5. Move  $s_1$  and its direct topological neighbors towards  $\xi$  by fractions  $\epsilon_b$  and  $\epsilon_n$ , respectively, of the total distance:

$$\Delta w_{s_1} = \epsilon_b(\xi - w_{s_1}),$$

$$\Delta w_{s_n} = \epsilon_n(\xi - w_n),$$

where  $n$  represents all direct neighbors of  $s_1$ .

6. If  $s_1$  and  $s_2$  are connected by an edge, set the age of this edge to zero. If such an edge does not exist, create it.
7. Remove edges with an age larger than  $a_{max}$ . If this results in nodes having no emanating edges, remove them as well.
8. Decrease the error variables of all the nodes by multiplying with a constant  $d$ .
9. Repeat steps 1 to 8  $\lambda$  times, with  $\lambda$  an integer.
  - If the maximum number of nodes has not been reached then insert a new node as follows:
    - Determine the node  $q$  with the maximum accumulated error.
    - Insert a new node  $r$  halfway between  $q$  and its neighbor  $f$  with the largest error variable:

$$w_r = 0.5(w_q + w_f).$$

- Insert edges connecting the new node  $r$  with nodes  $q$  and  $f$ , and remove the original edge between  $q$  and  $f$ .
- Decrease the error variables of  $q$  and  $f$  by multiplying them with a constant  $\alpha$ . Initialize the error variable and the local counter of node  $r$  with the new value of the error variable and local counter of  $q$ , respectively.
- If the maximum number of nodes has been reached then remove a node as follows:
  - Let  $k$  be the stored node with the lowest error variable.
  - Remove node  $k$  and all the edges emanating from  $k$ .

10. If  $N$  is the total number of nodes, every  $\mu \cdot N$  iterations of steps 1 to 8 remove all the nodes that have not been used (local activation counter equal to zero) and all the edges emanating from them. Reset the local counter of all the nodes to zero.

As it is happen in all the self-organizing algorithms, there is a set of parameters that control the whole process. Although we do not want to go into details, it is necessary to clarify some essential aspects of the algorithm presented above.

### 3 Some Highlights of the Model

In general, the self-organizing algorithm can be seen as a training process based on neural networks. At the end of this process a set of nodes, which represent the new vertices of the optimized mesh is computed. The edges connecting these nodes show the neighboring relations among the nodes generated by the algorithm.

The parameters involved in the algorithm described in Sect. 2 are:

- $a_{max}$ , the maximum age for the edges,
- $\epsilon_b$ , related to the displacement of the winner node in the space,
- $\epsilon_n$ , related to the displacement of the neighbors nodes in the space,
- $d$ , a constant to decrease the error variables,
- $\lambda$ , an integer to determine when to create a new node,
- $\alpha$ , a constant to decrease the error variables of the nodes after adding a new one,
- $\mu$ , a constant to know when to remove the nodes that have not been referenced in successive iterations.

The  $\mu$  parameter is introduced in the step 11 of the algorithm and is used to determine when the non-referenced nodes must be removed. Moreover, the step 11 must be repeated every  $\mu * N$  iterations, where  $N$  is the total number of nodes. The reason for this periodicity is due to the fact that when increasing the number of nodes of the generated mesh the probability that a particular node was activated during  $\mu$  iterations decreases. Therefore, it is convenient to take  $\mu \geq \lambda$ .

There is no theoretical method to obtain a set of parameters which produces the best results for all the meshes that we may need to simplify. The way to obtain them is by experimentation. So, it has been implemented the possibility to change the values of these parameters. The set of parameters that we have used to simplify some examples are shown in Table I.

**Table 1.** Typical set of parameters included in the self-organizing algorithm

$a_{max}$	$\epsilon_b$	$\epsilon_n$	$d$	$\lambda$	$\alpha$	$\mu$
30	0.6	0.01	0.2	10	0.8	30

## 4 Real Examples

### 4.1 Some Real Examples of Urban Acupuncture

Curitiba is a large provincial capital city in southeastern Brazil with a population of roughly 2.4 million inhabitants. It is not known for any exceptional landmark. Rather than becoming an urban metropolis overrun with poverty, unemployment, inequity, and pollution over the past half-century, Curitiba and its citizens have instead seen a continuous and highly significant elevation in their quality of life. Though starting with the dismal economic profile typical of its region, in nearly three decades the city has achieved measurably better levels of education, health, human welfare, public safety, democratic participation, political integrity, environmental protection, and community spirit than its neighbors. The miracle has a name, Jaime Lerner and the application of a theory: urban acupuncture (see [17]).

Jaime Lerner, therefore, sees the city as a living organism composed of a network of energetic centers, each of which serve as potential leverage points for catalyzing the revitalization of the entire system. Much of Lerner's work, therefore, revolves around the work of identifying these potential leverage points and making appropriate interventions that can catalyze or awake the entire system into working at a different and higher order level of health. It appears evident that Jamie Lerner, himself, views urban centers as interconnected networks of living energy centers.

We can see other references for urban acupuncture actions, as for example [14,18].

### 4.2 Using the Neural Network Model to Design an Urban Network

Lerner and his team developed some actions in Curitiba: created public spaces and transportation systems that enabled people to more freely move about the city and be drawn out of their private homes to engage in public interactions and events. And this is exactly what we want to show in our example. We want to create a transport network in the downtown of a city and, to perform this task, we will use the neural network algorithm described in Sect. 2. Note that there is an extensive bibliography on the topic of the design and implementation of urban transport networks, as for example [4,7,8,16].

Let us apply the theory exposed to a concrete case of a real city in our surrounding, Elche (Spain).

Functional rather than structural, the urban acupuncture approach proposes to act with specific projects in a selection of hot points in the historic center of the city of Elche (Fig. 1). This is a soft approach, which takes care of the context, whose purpose is to drive the development rather than to control it. The urban acupuncture wishes to maintain this energy of the city, and use potentials and renewals instead of moving things heavily and it is all about punching the proper project to the right place. Strategic points to act on can be connection points for different flows: water, transport, people.



**Fig. 1.** Downtown of the city (Elche, Spain)

Let us assume that we want to create a network consisting of 30 nodes or points on which to develop possible actions of urban acupuncture. These nodes may represent, following the example of Curitiba, a urban network transport.

The map of the city where we are going to determine hot points is shown in Figure 1. We identify each of the blocks of houses or buildings with a node in the mesh and perform a triangulation process with these nodes, obtaining a two-dimensional grid made up of planar triangles, as we can see in Fig. 1. This initial mesh has 367 vertices or nodes (houses) and 950 edges.

Now, the objective is, by means of the self-organizing algorithm described in Sect. 2, determine a set of positions in the urban area in which to place some basic points on which to develop actions following the model of urban acupuncture. Let us assume, therefore, that we want to obtain a simplified mesh with 30 nodes and, what is more important, we need that this reduced grid looks like as possible to the original mesh.

We run the self-organizing algorithm starting from two nodes (or neurons) and stopping when the 30 nodes are created. The resulting nodes are shown in Fig. 2 and their positions are detailed in Table 2.

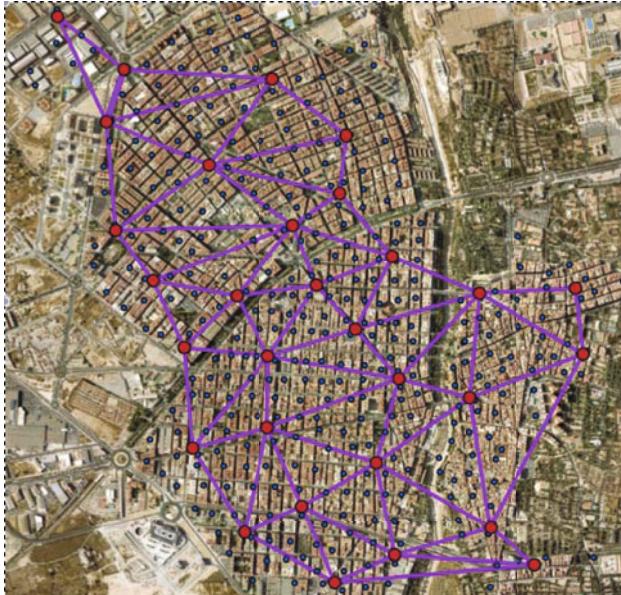
Then, in Fig. 2 we have a distribution of new nodes or *hot points* where we could implement urban acupuncture. It is important to note that each of these



**Fig. 2.** Positions of the nodes after running the self-organizing algorithm

**Table 2.** Final position of the nodes in the mesh shown in Fig. 3

Final nodes		Final nodes	
Node	Position	Node	Position
$k_1$	(9.91, 10.74)	$k_2$	(11.19, 2.06)
$k_3$	(15.57, 8.47)	$k_4$	(10.78, 4.26)
$k_5$	(7.42, 8.27)	$k_6$	(13.50, 2.71)
$k_7$	(9.35, 8.53)	$k_8$	(11.16, 9.22)
$k_9$	(8.29, 13.47)	$k_{10}$	(6.75, 11.42)
$k_{11}$	(6.16, 7.03)	$k_{12}$	(8.17, 5.11)
$k_{13}$	(4.72, 13.70)	$k_{14}$	(5.42, 8.64)
$k_{15}$	(3.13, 14.97)	$k_{16}$	(9.76, 1.41)
$k_{17}$	(13.25, 8.33)	$k_{18}$	(6.36, 4.61)
$k_{19}$	(10.28, 7.47)	$k_{20}$	(14.59, 1.83)
$k_{21}$	(9.00, 3.21)	$k_{22}$	(15.75, 6.88)
$k_{23}$	(8.78, 9.97)	$k_{24}$	(11.32, 6.28)
$k_{25}$	(13.01, 5.82)	$k_{26}$	(10.06, 12.12)
$k_{27}$	(4.30, 12.44)	$k_{28}$	(8.19, 6.82)
$k_{29}$	(4.52, 9.82)	$k_{30}$	(7.61, 2.60)



**Fig. 3.** Reconstruction of the final mesh

nodes represents a hub or a hot point in the network, that is, represents more than a simple node in a mesh.

Note that the number of vertices of the final mesh is much smaller than the original mesh; nevertheless, the algorithm provides an efficient distribution of the vertices so that covers as much area as possible on the original mesh. This aspect must be pointed out as one of the great advantages of using neural network algorithms in this problem, since they are able to learn the shape of the initial object and, therefore, creates copies that are very similar to the original.

Once we have obtained the points for the final mesh, it is possible to carry out the triangulation process from the information provided by the self-organizing algorithm. The basis of triangulation process is the comparison between the original nodes of the network and the new nodes. The reconstruction of the simplified mesh is shown in Fig. 3.

## 5 Conclusion

The urban acupuncture wishes to maintain the energy of the city, and use potentials and renewals instead of moving things heavily and it is all about punching the proper project to the right place. Strategic points to act on can be connection points for different flows: water, transport, people. In this paper we have applied a self-organizing algorithm based on GNG3D model to the problem of determining those points at which specific actions must be taken. The systems of

public transportation, systems of business/tourist flows and systems for information interchange (telecommunication networks) open the numerous possibilities for urban integration. Following the current Curitiba Integrated Transportation Network (ITN), we apply the neural network model to the design of a transport network in a concrete real city, obtaining a simplified network where each of the points of the network represents a hub or a hot point in the urban acupuncture strategy.

## References

1. Alvarez, R., Noguera, J., Tortosa, L., Zamora, A.: GNG3D - A Software Tool for Mesh Optimization Based on Neural Networks. In: Proceedings of the IJCNN 2006, pp. 4005–4012 (2006)
2. Alvarez, R., Noguera, J., Tortosa, L., Zamora, A.: A mesh optimization algorithm based on neural networks. *Information Sciences* 177, 5347–5364 (2007)
3. Barnett, J.: An introduction to urban design. Harper and Row, New York (1982)
4. Beltran, B., Carrese, S., Cipriani, E., Petrelli, M.: Transit network design with allocation of green vehicles: A genetic algorithm approach. *Transportation Research Part C: Emerging Technologies* 17(5), 475–483 (2009)
5. Busquets, J.: Barcelona- The Urban Evolution of a Compact City, Harvard, MA (2005)
6. Castello, P., Sbert, M., Chover, M., Feixas, M.: Viewpoint-based simplification using f-divergences. *Information Sciences* 178(11), 2375–2388 (2008)
7. Dunphy, R., Cervero, R., Dock, F., McAvey, M., Porter, D., Swenson, C.: Developing Around Transit: Strategies and Solutions that Work. Urban Land Institute, Washington, DC (2004)
8. Fan, W., Machemehl, R.: A Tabu search based heuristic method for the transit route network design problem. In: 9th International Conference on Computer-Aided Scheduling of Public Transport, CASPT (2004)
9. Fritzke, B.: Growing cell structures - a self-organizing network for unsupervised and supervised learning. *Neural Networks* 7(9), 1441–1460 (1994)
10. Fritzke, B.: A growing neural gas network learns topology. In: Tesauro, G., Touretzky, D.S., Leen, T.K. (eds.) *Advances in Neural Information Processing Systems*, vol. 7, pp. 625–632. MIT Press, Cambridge (1995)
11. Kohonen, T.: Self-Organizing formation of topologically correct feature maps. *Biological Cybernetics* 43, 59–69 (1982)
12. Kohonen, T.: The Self-Organizing Map. *Proceedings of the IEEE* 76(9), 1464–1480 (1990)
13. Larice, M., MacDonald, E. (eds.): *The Urban Design Reader*. Routledge, New York (2007)
14. Leonardo, S.: Urban acupuncture as a strategy for Sao Paulo. Ph. Thesis, Massachusetts Institute of Technology (2006)
15. Lerner, J.: *Acupuntura Urbana*. Record, Rio de Janeiro, Brasil (2003)
16. Lownes, N.E., Machemehl, R.: Exact and heuristic methods for public transit circulator design. *Transportation Research Part B Methodological* 44 (2), 309–318 (2009)
17. MacLeod, K.: Orienting urban planning to sustainability in Curitiba, Brazil (2005), <http://www3.iclei.org/localstrategies/pdf.curitiba.pdf>

18. Marzi, M., Ancona, N.: Urban acupuncture, a proposal for the renewal of Milan's urban ring road, Milan, Italy. In: Proceedings of 40th ISoCaRP Congress 2004, Geneva, Switzerland, September 18-22 (2004)
19. Noguera, J., Tortosa, L., Zamora, A.: Analysis and efficiency of the GNG3D algorithm for mesh simplification. *Applied Mathematics and Computation* 197(1), 29–40 (2008)
20. Stupar, A., Savcic, V.: The new urban acupuncture: intermodal nodes between theory and practice. In: Proceedings of the REAL CORP 2009, pp. 499–505 (2009)

# Discovering Process Models with Genetic Algorithms Using Sampling

Carmen Bratosin, Natalia Sidorova, and Wil van der Aalst

Department of Mathematics and Computer Science

Eindhoven University of Technology

P.O. Box 513, 5600 MB Eindhoven, The Netherlands

{c.c.bratosin,n.sidorova,w.m.p.v.d.aalst}@tue.nl

**Abstract.** Process mining, a new business intelligence area, aims at discovering process models from event logs. Complex constructs, noise and infrequent behavior are issues that make process mining a complex problem. A genetic mining algorithm, which applies genetic operators to search in the space of all possible process models, deals with the aforementioned challenges with success. Its drawback is high computation time due to the high time costs of the fitness evaluation. Fitness evaluation time linearly depends on the number of process instances in the log. By using a sampling-based approach, i.e. evaluating fitness on a sample from the log instead of the whole log, we drastically reduce the computation time. When the desired fitness is achieved on the sample, we check the fitness on the whole log; if it is not achieved yet, we increase the sample size and continue the computation iteratively. Our experiments show that sampling works well even for relatively small logs, and the total computation time is reduced by 6 up to 15 times.

**Keywords:** Genetic algorithms, business intelligence, process mining, sampling.

## 1 Introduction

In recent years, *process mining*, also known as *automatic process discovery*, has emerged as a business intelligence area focused on the analysis of systems and their behavior based on *event logs* [2]. An event log records information on the execution of instances of the same process. Unlike other data mining domains the focus of process mining is on concurrent processes. Process mining can be applied to a wide spectrum of systems ranging from information systems (e.g., Enterprise Resource Planning Systems) to systems where hardware plays a more prominent role (e.g., embedded systems, sensor networks). Real life case studies performed for Phillips Medical Systems (PMS) [7] and ASML [13] proved process mining algorithms valuable in providing insight into processes, discovering bottlenecks or errors and assist in improving processes.

*Process Mining Algorithms* (PMAs) [2][15] allow to build process models that capture the processes as they have been executed. Process models graphically depict the flow of work using languages such as Petri Nets or BPMN. Most of the PMAs [2][15] use heuristic approaches to retrieve dependencies between activities based on event patterns. Heuristic algorithms often fail to capture complex process structures, e.g. choices depending on combinations of earlier events, and they are not robust to noise, i.e., random and rare deviations of activities from the intended behavior. In [34], Alves de

Medeiros et al. proposed a *Genetic Mining Algorithm* (GMA) that uses genetic operators to overcome these shortcomings. GMAs evolve populations of graph-based process models towards a process model that fulfills the fitness criteria: the process model manages to replay all the behaviors observed in the event log and does not allow additional ones. An empirical evaluation in [4] confirms that GMA achieves the goal to discover better models than other PMAs. Commercial business intelligence tools *Futura Reflect* (<http://www.futura-tech.nl>) and *BPMone* (<http://www.pallas-athena.com>) provide process mining facilities based on this GMA.

Although GMA prevails against other algorithms in terms of model quality, heuristic-based algorithms proved to be significantly more time efficient [4]. The high time consumption of GMA is mainly due to the fitness computation time. Individuals (process models) are evaluated against all the process instances from the event log. Therefore, the *fitness computation time* is linearly dependent on the number of process instances in the event log. Note that other factors such as the average length of the traces or the quality of the process model under fitness evaluation may influence the execution time.

In this paper, we improve the time efficiency of GMA by *random sampling* of the process instances from the event log. We create the initial population using a smart and fast heuristics based on the whole log, but then we use a random sample of the log for fitness computations until the desired fitness is achieved on this sample. The use of larger sample increases the chance that this sample is representative for the whole log, i.e. we get the desired fitness on the whole log as well. However, a larger sample size reduces the time advantage of sampling. To balance between the two objectives (mining quality and time efficiency) we use an *iterative algorithm* (denoted *iGMA*) that adds a new part to the sample until the discovered process model has the required quality for the entire event log. After increasing the sample, we use the already mined population of process models as initialization.

Sampling exploits the redundancy in the event log. A process structure is in fact a composition of multiple control-flow patterns, including choice, parallel composition, iteration [1]. Different instances can contain e.g. different combinations of choices made, different numbers of iterations taken, or different interleavings of events from parallel branches. Even when all instances in the log give different execution traces, many of them represent the “same” behavior with respect to some pattern from the process structure. Since the process structure is unknown, this redundancy present in the log is not directly visible, and we use random sampling instead of some smart sampling strategies.

The degree of redundancy of an event log is closely related to the *completeness* of an event log. A log is complete if it contains enough behavior in order to discover the original process model. A high degree of redundancy in the event log increases the confidence that we observed enough process behavior. Note that the iGMA algorithm converges even if the event log contains no redundancy due to its incremental nature, i.e., at the last iteration the sample includes the entire log. In this case, we cannot be confident that the model we mined is the truthful representation of the original process.

We empirically assess the algorithm for different event logs and we show that the algorithm converges significantly faster than the original GMA. Interestingly, in many cases only a small fraction of the log is sufficient for achieving 0.8 fitness.

**Related Work.** Sampling is a common practice in business intelligence domains such as data mining and knowledge discovery. [10][11][14] show how sampling can be used in order to create time efficient and accurate algorithms in these domains. Kivinen and Manilla [10] use small amount of data to discover rules efficiently and with reasonable accuracy. Tan [14] discusses the challenges in using sampling such as choosing the sample size and he proposes to increase the sample size progressively.

Reducing the fitness computation time is one of the main topics in *genetic algorithms* literature such as: approximating the fitness function by a meta-model or a surrogate [8][9], replacing the problem with a similar one that is easier to solve [9], or, inheriting the fitness values [5]. The techniques used in [6][12] are close to the one we use, although their motivation and goals are different: they analyze the effect of sampling when the fitness function is noisy. Fitzpatrick and Grefenstette [6] show that genetic algorithms create “more efficient search results from less accurate evaluations”.

The paper is organized as follows: Section 2 presents our approach. We analyze the performance for the three event logs in Section 3. We give conclusions and describe future work in Section 4.

## 2 Genetic Process Mining Exploiting Data Structure

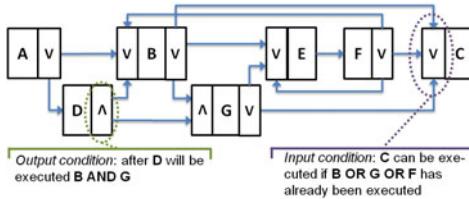
In this section we present the process mining domain and the genetic mining algorithm. We show how we integrate sampling into the genetic mining algorithm in order to create a more time efficient GMA.

### 2.1 Process Mining and Event Logs Characteristics

Process mining aims to discover process models from event logs, thus recording (parts of) the actual behavior. An *event log* describes previous executions of a process as sequences of *events* where each event refers to some activity. Table 1 presents a very much simplified event log inspired by a real-world process: a document issue process for a Dutch governmental organization. This log contains information about six process instances (individual runs of a process). One can notice that each process instance is uniquely identified and has an associated *trace*, i.e., the executed sequence of activities. The log shows that seven different activities occurred: A, B, C, D, E, F, and G. Each process instance starts with the execution of activity A, ends with the execution of activity C and contains activity B. Process instances with id 2, 4 and 6 also contain activities D and G suggesting that there is a relation between their occurrences. The activities E, F and G appear always after B and activity D occurs prior to B. Moreover, the occurrence of F is always preceded by the occurrence of E. Instances 5 and 6 show that loops are possible in the process.

Different traces of an event log might contain information about certain dependencies between activities which can be deducted from other traces. In Table 1, trace 6 does not add any information about the process structure to traces 1-5. For example, both loops BEF and EF in trace 6 can be identified from trace 5.

Note that there are multiple models that can reproduce a particular log. The quality of a model is given by its ability to balance between underfitting and overfitting.



**Fig. 1.** Process model representing the behavior in the event log from Table 1

**Table 1.** A simplified event log

Process instance id	Trace
1	A B C
2	A D B G C
3	A B E F C
4	A D B G E F C
5	A B E F E F B E F C
6	A D B G E F B E F E F C

**Table 2.** Causal Matrix

Activity	Input set	Output set
A	-	B ∨ D
B	A ∨ D ∨ F	C ∨ E ∨ G
C	B ∨ F ∨ G	-
D	A	B ∧ G
E	B ∨ G ∨ F	F
F	E	B ∨ C ∨ E
G	B ∧ D	C ∨ E

An underfitting model allows for too much behavior while an overfitting model does not generalize enough. Figure 1 shows a process model with fitness 0.98 for the example event log from Table 1. The process model is a graph model that expresses the dependencies between activities. Each node in the graph represents an activity. Each activity has an *input* and *output set*. *Causal matrices* [3] are used to represent the dependencies between activities in a compact form. The causal matrix for the process model in Figure 1 is shown in Table 2. Since A is the start activity its input condition is empty and activity A is enabled. After the execution of activity A the output condition  $\{B \vee D\}$  is activated. Further on, activity B is enabled because the input condition of B requires that at least one of the activities A, D or F is activated before B. The input condition of activity D requires only A to be activated before activity D. If we assume that activity D is executed, we observe that B is automatically disabled because the output condition of A is no longer active. If activity D is executed, activity B remains enabled because the output condition of D,  $\{B\}$ , is activated. For more details on the semantics of causal matrices we refer to [34][15].

The computational complexity of a particular PMA depends on the log characteristics. The following parameters give the basic characteristics of a log: the *size* of a log (the sum of the lengths of all traces); the *number of traces* (influencing the confidence in the obtained process models); and the *number of different activities* (defining the search space for the model). In our example, the event log size is 42, the number of traces is 6 and the number of different activities is 7.

Many heuristics algorithms, such as the  $\alpha$ -algorithm [2] are linear in the size of the log. However, such algorithms do not perform well on real-life data [7][13] and therefore more advanced process mining algorithms are needed. For example, the  $\alpha$ -algorithm does not capture the dependency between the activities D and G from the example event

---

**Algorithm 1.** iGMA algorithm

---

**Input:**  $\text{Log}$ ,  $\text{StopCondition}$ ,  $n$  /\* the sample size is  $1/n$  Log size

**Output:**  $\text{ProcessModel}$

```

PartitionTheLog( $\text{Log}, \text{Log}_1, \text{Log}_2, \dots, \text{Log}_n$ )
 $P = \text{BuildInitialPopulation}(\text{Log})$ 
 $\text{SampleLog} = \{\}$ 
 $m = 1$ 
repeat
     $\text{SampleLog} = \text{SampleLog} \cup \text{Log}_m$ 
     $\text{Fitness} = \text{ComputeFitness}(P, \text{SampleLog})$ 
    while  $\text{StopCondition}(\text{Fitness}) == \text{false}$  do
         $P = \text{ComputeNextPopulation}(P, \text{Fitness})$ 
         $\text{Fitness} = \text{ComputeFitness}(P, \text{SampleLog})$ 
    end while
     $m = m + 1$ 
     $\text{Fitness} = \text{ComputeFitness}(P, \text{Log})$ 
until  $\text{StopCondition}(\text{Fitness}) == \text{false}$ 
 $\text{ProcessModel} = \text{bestIndividual}(P)$ 

```

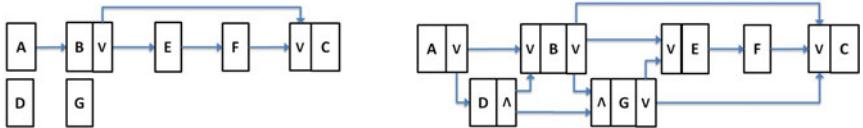
---

log which results in underfitting the event log. GMA [34] applies genetic operators on a population of process models in order to converge to models that represent the event log behavior precisely. The main advantages of GMA are the ability to discover non-trivial process structures and its robustness to noise as demonstrated by [4]. Its main drawback is the time consumption, this is due to: 1) the time required to compute the fitness for an individual and 2) the large number of fitness evaluations needed.

## 2.2 Iterative Genetic Process Mining Algorithm (iGMA)

In this subsection we present a new genetic process mining algorithm that we call *iterative Genetic Mining Algorithm* (iGMA). iGMA is based on [34]. The algorithm improves the overall execution time of existing GMA by incrementally learning from samples of traces from the event log. Algorithm 1 presents the main iGMA steps. iGMA uses the same genetic operators and fitness computation as GMA [34]. The main difference is in handling input data. The algorithm starts with  $\text{PartitionTheLog}(\text{Log}, \text{Log}_1, \text{Log}_2, \dots, \text{Log}_n)$  that divides the  $\text{Log}$  into  $n$  random samples of "equal" size. At each iteration  $m$ , genetic operators are applied using a sample from the event log corresponding to the union of the first  $m$  samples. The iteration stops when  $\text{StopCondition}$ , i.e. reaching a given quality of the population, is fulfilled. Note that the  $\text{StopCondition}$  of the internal *while* loop is evaluated on the  $\text{SampleLog}$  and not on the entire  $\text{Log}$ . The algorithm stops when the same  $\text{StopCondition}$  is satisfied for the *entire event log*.

The iGMA individuals are graph models, such as the one presented in Figure 1. The individuals are encoded as a set of activities and their corresponding input and output sets, i.e., a *causal matrix*. Note that it is trivial to construct the graph representation from Figure 1 based on the compact representation from Table 2. Each process model contains all the activities in the log; hence, the search space dimension depends on the number of activities in the log.



(a) Results after the first iteration. Fitness on the sample is 0.98; fitness on the log is 0.77      (b) Results after the second iteration. Fitness on the sample is 0.98; fitness on the log is 0.87

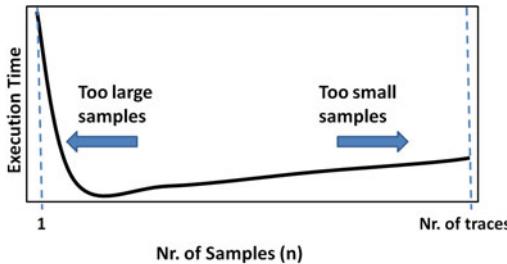
**Fig. 2.** Process models for the example event log (Table I)

*BuildInitialPopulation(Log)* generates individuals from the search space using an heuristic approach. This heuristic uses the information in the log to determine the probability that two activities have a dependency relation between them: the more often an activity  $A$  is directly followed by an activity  $B$ , the higher the probability is that the individuals are built with a dependency between  $A$  and  $B$ .

*ComputeFitness( $P$ , (Sample)Log)* assesses each individual against the (Sample)Log. *Fitness* values reflect how well each individual represents the behavior in the log with respect to *completeness*, measuring the ability of the individual to replay the traces from the log, and *preciseness*, quantifying the degree of underfitting the log [34]. The *completeness* of an individual is computed by parsing the log traces. When an activity of a trace cannot be replayed, i.e. the input condition is not satisfied, a penalty is given. In order to continue, the algorithm assumes that the activity was executed and it tries to parse the next activity from the trace [34]. Additional penalties are given when the input/output conditions enable an activity incorrectly, e.g. if in the input condition, activity  $A$  and activity  $B$  are in an AND relation and in the trace only activity  $B$  is executed. In the end, the fitness quantifies all the penalties and compares them with the number of correctly parsed activities. The fitness *preciseness* is computed by comparing the individuals from the current population against each other. The idea is to penalize the individuals that allow more behavior than their siblings. The exact formula is out of the scope of this paper but can be found in [34].

*ComputeNextPopulation( $P$ , Fitness)* applies the genetic operators (*selection*, *mutation* and *crossover*) to generate a new population. The selection operator ensures that the best individuals are carried forward from the current population to the next one. The mutation modifies the input/output conditions of a randomly selected activity by insertion, removal or exchanging the AND/OR relations of the activities. The crossover exchanges the input/output conditions of a selected activity between two individuals.

The convergence of the iGMA algorithm is ensured by the fact that at each iteration a larger sample is taken together with convergence of the traditional GMA. If new activities or dependencies appear in the newly added sample, they are discovered by exploring the space using the genetic operators. Basically, the algorithm is learning at each iteration by increasing its knowledge. Figure 2 shows one variant of intermediary process models when iGMA is applied to the example event log (Table I). We consider the log divided in three subsets: {1, 3}, {2, 4}, and {5, 6}. The first result does not contain activities  $D$  and  $G$  connected which does not harm the fitness value since the activities do not appear in the first sample. At the second iteration, when using the first and the second sample the activities  $D$  and  $G$  are integrated into the process model but the loops



**Fig. 3.** Execution Time Variation

are not present. After the last iteration we obtain the process model presented in Figure 11. Note that the same process model is obtained if the log would only contain traces 1–5. We observe that the algorithm converges by adding new dependencies to the previously obtained model.

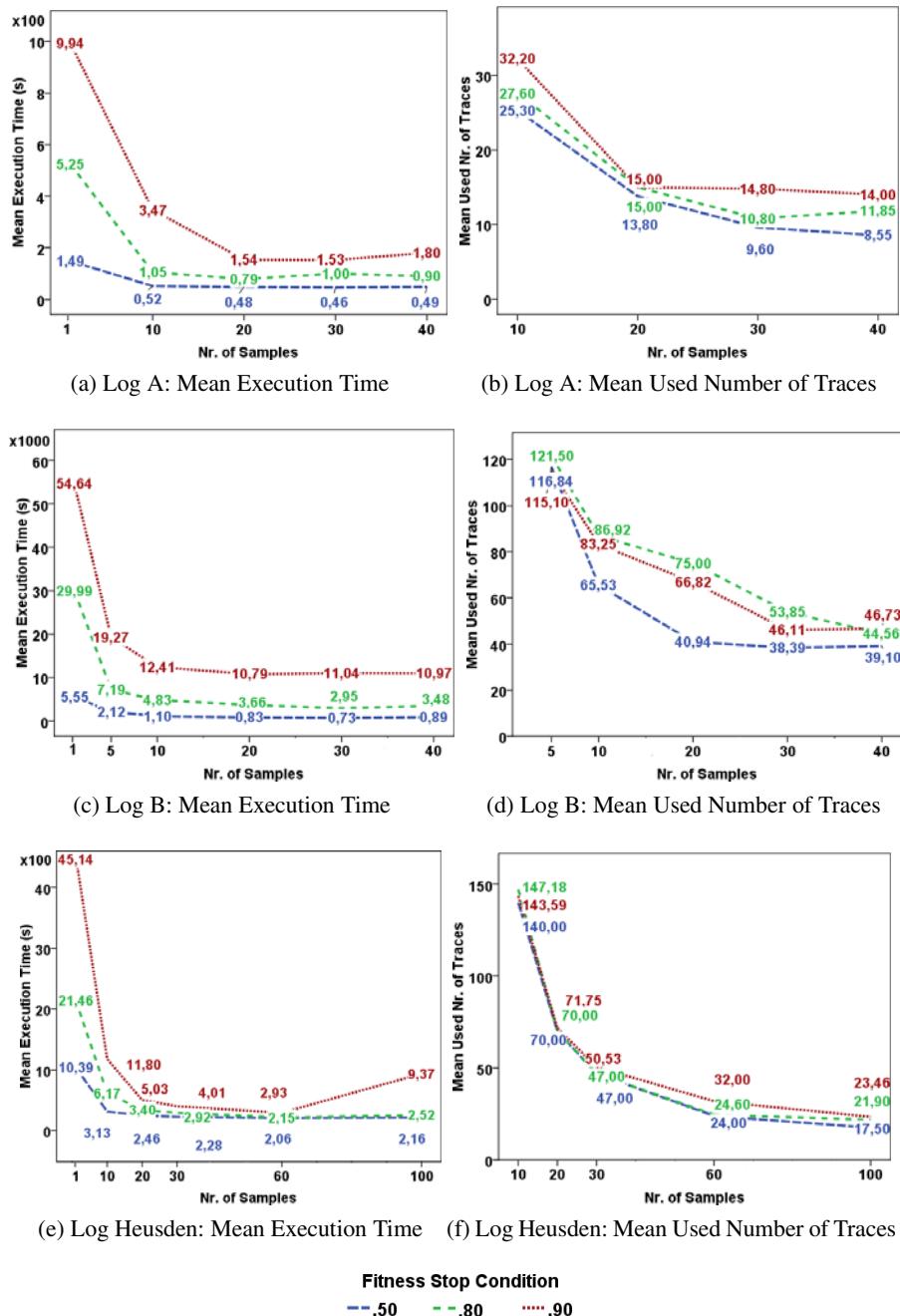
A very important parameter for the iGMA is the *sample size*. Note that for simplicity the algorithm is parameterized by the number of samples  $n$ , that is also the maximal number of iterations. Figure 3 shows the execution time as a function of  $n$ . Too a large sample results in high execution time due to high fitness computation time. On the other hand, too a small sample has very low probability to be representative for the log and therefore multiple iterations are needed. The execution time slope decreases when more samples are used because only the fitness component of the execution time is reduced. In the next section, empirical results for three different logs show that the execution time follows the curve from Figure 3.

### 3 Experiments

In this section we compare the iGMA performance against the GMA on three different logs (*A*, *B* and *Heusden*). Our experiments show that there is a connection between the log characteristics and the optimal sample size.

Table 3 presents the parameters *number of activities*, *number of traces* and *size* for the three logs. The first two logs are generated by students as part of their assignment for the process mining course. The third log is a real life log for the process of handling objections against the real-estate property valuation at the Municipality of Heusden. Underlying processes for *A* and *Heusden* turn to have quite a simple structure, which makes them "easy" to mine. The process underlying in *B* has a complex structure, which makes *B* a difficult log. The challenge of log *Heusden* is in the large number of traces. To compute the fitness, we need to assess each individual, i.e. process model, against all traces, which implies longer fitness computation time for *Heusden*.

We performed the experiments on a Unix machine with the following configuration: eight Intel(R) Xeon processors with a frequency of 2.66 GHz and 16 Gb RAM. We assess the results in terms of *Mean Execution Time* (MET) and *Mean Used Number of Traces* (MUNT), i.e. the average of number of traces necessary for iGMA to converge, for three fitness stop conditions: 0.5, 0.8 and 0.9. Note that the highest possible fitness



**Fig. 4.** Mean Execution Time and Mean Used Number of Traces for the test three logs

**Table 3.** Logs characteristics

Name	Number of activities	Number of traces	Size
Log A	24	229	3994
Log B	25	300	9054
Log Heusden	18	1394	11467

**Table 4.** MET improvement

Name	Required Fitness Value		
	0.5	0.8	0.9
Log A	3	5	6
Log B	7	10	6
Log Heusden	5	10	15

value is 1. The execution time and the used number of traces are averaged over 40 independent runs. For the visualization and analysis of results we use SPSS.

Figure 4 presents the results of our experiments. We observe that MET for all our experiments follow the curve from Figure 3. Table 4 presents the maximum improvement, i.e. the ratio between the MET for GMA and the minimum MET for iGMA, obtained for the three logs. We observe that the best times are obtained for log Heusden, that needs a sample size of only 1% of the log. Log A has a lower speed-up than the Heusden log since 6% of log A is used. Log B has a more difficult structure and more generations are needed to converge; this leads to high improvement in spite of the high used number of traces.

The number of iterations is mostly one for logs Heusden and A since the initial sample is already sufficient to find a good model in most of the experiments, see Figures 4b and 4f. This suggest that these two logs contain a high degree of redundancy and, thus, the *logs are complete*. Also for log B (Figure 4d), for all our experiments the sample size is low enough to conclude that log B is complete.

The “optimal” MUNT, that minimizes the MET, is correlated with the log characteristics, the model difficulty and the stop fitness value. The number of samples  $n$  is correlated with the number of traces: the higher is the number of traces, the more samples we can create. “Easy” event logs imply that the number of dependencies in the process is far lower than the total number of dependencies (that is  $N_A^2$ , where  $N_A$  is the number of activities) and thus the necessary number of traces to capture the whole event log behavior is small (e.g., 14 traces for logs A and Heusden). The increase in the log difficulty results in the need for a higher number of traces (e.g., more than 60 traces in the case of log B). As one can expect, MUNT decreases when lower accuracy is required. However, MUNT does not differ significantly between the results for stop fitness equal to 0.9 and the ones for a 0.8 stop fitness. For example, for log B with  $n = 20$  the t-test result between MUNT for 0.8 and MUNT for 0.9 is  $t(70) = 0.5$ ,  $p = 0.615$  that means that there is no evidence of a difference between the means.

## 4 Conclusion and Future Work

In this paper, we proposed a new approach for mining large event logs based on genetic algorithms using sampling. The method relies on the genetic operators and the fitness function proposed by [34]. We *improve the time efficiency* of GMA by more than 5 times for a stop condition of 0.8 by exploiting the sampling in an incremental manner. The presence of redundancy in the event logs allows to use a small amount of traces for guiding the search of the solution. If the initial sample is not representative for the entire

event log, the iterative increase of the sample ensures that the algorithm will converge to a process model that describes the overall event log. A side-advantage of our algorithm is in determining the *completeness* of an event log: if a small fraction of the traces is sufficient, we get confidence that the event log contains enough information to mine the process model.

We validate our approach on three different logs and we show that our method converges much faster than the original GMA. Our experiments provide clear evidence that the sample size is strongly correlated with the logs characteristics and their level of difficulty from the mining point of view. The optimal choice of the sample size is still an open research question which we plan to investigate in our future work. Moreover, currently we are working towards a parallel version of iGMA.

## References

1. van der Aalst, W.M.P., Ter Hofstede, A.H.M., Kiepuszewski, B., Barros, A.P.: Workflow patterns. *Distrib. Parallel Databases* 14(1), 5–51 (2003)
2. van der Aalst, W.M.P., Weijters, A.J.M.M., Maruster, L.: Workflow Mining: Discovering Process Models from Event Logs. *IEEE Transactions on Knowledge and Data Engineering* 16(9), 1128–1142 (2004)
3. Alves de Medeiros, A.K.: Genetic Process Mining. PhD thesis, Technische Universiteit Eindhoven, Eindhoven, The Netherlands (2006)
4. Alves de Medeiros, A.K., Weijters, A.J.M.M., van der Aalst, W.M.P.: Genetic process mining: An experimental evaluation. *Data Mining and Knowledge Discovery* 14(2), 245–304 (2007)
5. Chen, J.-H., Goldberg, D.E., Ho, S.-Y., Sastry, K.: Fitness inheritance in multi-objective optimization. In: GECCO, pp. 319–326 (2002)
6. Fitzpatrick, J.M., Grefenstette, J.J.: Genetic algorithms in noisy environments. *Machine Learning* 3, 101–120 (1988)
7. Günther, C.W., Rozinat, A., van der Aalst, W.M.P., van Uden, K.: Monitoring deployed application usage with process mining. Technical report, BPM Center Report BPM-08-11, BPMcenter.org (2008)
8. Jin, Y.: A comprehensive survey of fitness approximation in evolutionary computation. *Soft Computing* 9(1), 3–12 (2005)
9. Jin, Y., Branke, J.: Evolutionary optimization in uncertain environments-a survey. *IEEE Trans. Evolutionary Computation* 9(3), 303–317 (2005)
10. Kivinen, J., Mannila, H.: The power of sampling in knowledge discovery. In: PODS 1994: Proceedings of the thirteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, pp. 77–85. ACM, New York (1994)
11. Lee, S.D., Cheung, D.W., Kao, B.: Is sampling useful in data mining? a case in the maintenance of discovered association rules. *Data Min. Knowl. Discov.* 2(3), 233–262 (1998)
12. Miller, B.L.: Noise, Sampling and Efficient Genetic Algorithms. PhD thesis, Department of Computer Science, University of Illinois, USA (1997)
13. Rozinat, A., de Jong, I., Günther, C., van der Aalst, W.: Process Mining Applied to the Test Process of Wafer Scanners in ASML. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* 39(4), 474–479 (2009)
14. Tan, P.-N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addison-Wesley Longman Publishing Co., Inc., Boston (2005)
15. Weijters, A.J.M.M., van der Aalst, W.M.P.: Rediscovering workflow models from event-based data using little thumb. *Integr. Comput.-Aided Eng.* 10(2), 151–162 (2003)

# A Multi-Objective Evolutionary Approach for the Antenna Positioning Problem

Carlos Segura, Yanira González, Gara Miranda, and Coromoto León

Dpto. Estadística, I.O. y Computación. Universidad de La Laguna  
La Laguna, 38271, Santa Cruz de Tenerife, Spain  
[{csegura,ygonzale,gmiranda,cleon}@ull.es](mailto:{csegura,ygonzale,gmiranda,cleon}@ull.es)

**Abstract.** Antenna Positioning Problem (APP) is an NP-Complete Optimisation Problem which arises in the telecommunication field. It consists in identifying the infrastructures required to establish a wireless network. Several objectives must be considered when tackling APP: minimise the cost, and maximise the coverage, among others. Most of the proposals simplify the problem, converting it into a mono-objective problem. In this work, multi-objective evolutionary algorithms are used to solve APP. In order to validate such strategies, computational results are compared with those obtained by means of mono-objective algorithms. An extensive comparison of several evolutionary algorithms and variation operators is performed. Results show the advantages of incorporating problem-dependent information into the evolutionary strategies. Also, they show the importance of properly tuning the evolutionary approaches.

## 1 Introduction

Engineering of mobile telecommunication networks evolves two major problems [14]: the antenna positioning problem (APP), and the assignment frequency problem (AFP). APP consists in positioning base stations (BS) or antennas on potential sites, in order to fulfil some objectives and constraints. AFP sets up frequencies used by such antennas with the aim of minimising interferences, i.e. maximising the offered quality of service. Both problems play a major role in various engineering, industrial, and scientific applications because its outcome usually affects cost, profit, and other heavy-impact business performance metrics. This means that the quality of the applied approaches has a direct bearing on industry economic plans.

In this paper we address the APP problem. APP is referred in the literature using several names, mainly: Radio Network Design (RND) and Base Station Transmitters Location Problem (BST-L). APP is an NP-complete problem [12]. The number of antennas that should be used in a network, and their corresponding locations, must be determined. In some cases, several kinds of antennas are supported. In such cases, the kind of antenna that must be used in each location must also be established. Generally, a set of potential locations are given, i.e. there are restrictions about the places in which antennas can be located. APP and AFP are together

analysed in some cases [1]. In other cases, they are studied as independent problems [1][13], solving first the APP, and afterward, the AFP. When both problems are solved simultaneously, interactions between them are considered, so better solutions can potentially be reached [9]. However, the search space hugely increases. Thus, solving the problem as a whole, requires more complex algorithms, and more computational resources.

Generally, a wave propagation model is incorporated into a network simulator in order to solve the problem. Among other, the free space, Hokumara-Hata and Walfish-Ikegami models can be used [15][16]. Another alternative relies on constructing an interference matrix by using a set of Mobile Measurement Reports (MMR) [1], i.e. measure the interferences that appear when several antennas are used simultaneously. Measuring the interferences in every potential location is very expensive and time-consuming. Thus, in this project we make use of a simplified model [2][9] in order to determine the coverage of a network configuration. Afterwards, MMRs can be obtained for the selected locations in order to solve the AFP. The advantages of such an approach reside in the cost saving when compared to those alternatives based on taking measures on each potential location. Moreover, since the last step of the design can be performed with the information given by the MMRs, the process is more realistic than those based in propagation models.

Many real-world engineering problems involve simultaneous optimisation of more than one objective function. The multiple objectives are typically conflicting but must be simultaneously optimised. In this kind of Multi-Objective Optimisation Problems (MOPs), a solution optimising every objective usually does not exist. However, there are some preferable solutions that are at least as good as others in all objectives and better for at least one objective. This relation is commonly known as Pareto dominance. The set of non-dominated solutions is called the Pareto Front. Solving a MOP consists in searching a non-dominated solution set, as close as possible to the Pareto Front. MOPs can be simplified by using several methods. In order to avoid the complexity of dealing with all the objectives at the same time, some simplifications have been proposed. A reasonable approximation is to convert the original problem into a single-objective one. This can be done by evaluating the objectives with a vector of weight, or any other fitness function. Other method consists in turning some objectives into constraints. Both alternatives have some drawbacks. On one hand, it is difficult to design a suitable fitness function. On the other hand, such methods only obtain one solution instead of a set of them, so there is a lost of diversity. In order to maintain the diversity of solutions, many approaches have been designed to directly solve multi-objective problems.

Several objectives can be considered when designing a network. Most typical considered objectives are: minimise the number of antennas, maximise the amount of traffic held by the network, maximise the quality of service, and/or maximise the covered area. Therefore, APP can be tackled with multi-objective strategies. Most of the proposed models for APP in the literature are mono-objective [4][13]. In such cases, several objectives are integrated into a fitness

function. In [17] APP is also tackled as a mono-objective function by translating the other considered objectives into restrictions. In [16] several objectives are considered simultaneously and multi-objective strategies are applied. The main advantage of the multi-objective approaches is the improvement in the diversity of the obtained solutions. Moreover, the involved decision making process is performed in the last step of the optimisation, so much more information can be considered. Many strategies have been applied for both, the mono-objective and multi-objective problem. In [11,17] ad-hoc heuristics have been applied. In [18] ad-hoc heuristics have been integrated into a Tabu Search based algorithm. The aforementioned strategies use a lot of problem-dependent information, so it is difficult to adapt them to other variations of the problem. Metaheuristics can be considered as high-level strategies that guide a set of simpler heuristic techniques in the search of an optimum [3]. They are more general than ad-hoc heuristics. In [12] several metaheuristics were applied to the mono-objective APP and extensively compared. In [2] the same definition of APP is solved by means of genetic algorithms. APP has also been solved by incorporating problem-dependent mutation operators [19] inside an evolutionary approach.

In this work an evolutionary approach is applied to a multi-objective definition of APP. The main contributions of the paper are the following: an extensive comparison among several evolutionary approaches is performed, problem-dependent and problem-independent crossover operators are analysed, and advantages and drawbacks of multi-objective approaches when compared to the best-known mono-objective strategies are analysed. The remaining content is structured in the following way: the mathematical formation of the multi-objective APP is given in Section 2. Section 3 describes the applied evolutionary approaches and variation operators. The computational study is presented in section 4. Finally, the conclusions and some lines of future work are given in section 5.

## 2 APP: Mathematical Formulation

APP is defined as the problem of identifying the infrastructures required to establish a wireless network. It comprises the maximisation of the coverage of a given geographical area while minimising the *base stations* - BS - deployment. Thus, it is an intrinsically multiobjective problem. A BS is a radio signal transmitting device that irradiates any type of wave model. The region of the area covered by a BS is called a cell. In our definition of APP, BS can only be located in a set of potential locations. APP mathematical formulation derives from the mono-objective one proposed in [2,19]. In such cases, two objectives are considered: the maximisation of the coverage (*CoverRate*), and the minimisation of the BS or transmitters deployed (*Transmitters*). In [2,19] the problem is simplified by defining a fitness function which converts the problem into a mono-objective one:

$$f(solution) = \frac{CoverRate^\alpha}{Transmitters} \quad (1)$$

In the previous scheme a decision maker must select a value for  $\alpha$ . It is tuned considering the importance given to the coverage, in relation with the number of deployed BS. However, since usually there is no information about the quality of the solutions which can be achieved, such a task is very difficult. Thus, obtaining the desired solutions is hard, and probably several values of  $\alpha$  must be tested. The parameterisation can be avoided by defining APP as a MOP. With such a definition, strategies attain an approximation of the Pareto Front.

The geographical area  $G$  on which a network is deployed is discretized into a finite number of points or locations.  $Tam_x$  and  $Tam_y$  are the number of vertical and horizontal subdivisions, respectively. They are selected by communications experts, depending on several characteristics of the region and transmitters.  $U$  is the set of locations where BS can be deployed:  $U = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ . Location  $i$  is referred using the notation  $U[i]$ . The  $x$  and  $y$  coordinates of location  $i$  are named  $U[i]_x$  and  $U[i]_y$ , respectively. When a BS is located in position  $i$  its corresponding cell is covered. The cell is named  $C[i]$ . In our definition we use the canonical APP problem formulation, i.e., an isotropic radiating model is considered for the cell. The set  $P$  determines the locations covered by a BS:  $P = \{(\Delta x_1, \Delta y_1), (\Delta x_2, \Delta y_2), \dots, (\Delta x_m, \Delta y_m)\}$ . Thus, if BS  $i$  is deployed, the covered locations are given by the next set:  $C[i] = \{(U[i]_x + \Delta x_1, U[i]_y + \Delta y_1), (U[i]_x + \Delta x_2, U[i]_y + \Delta y_2), \dots, (U[i]_x + \Delta x_m, U[i]_y + \Delta y_m)\}$ .

Being  $B = [b_0, b_1, \dots, b_n]$  the binary vector which determines the deployed BS, APP is defined as the MOP given by the next two objectives:

$$\begin{aligned} f_1 &= \sum_{i=0}^n b_i \\ f_2 &= \sum_{i=0}^{tam_x} \sum_{j=0}^{tam_y} \text{covered}(i, j) \end{aligned}$$

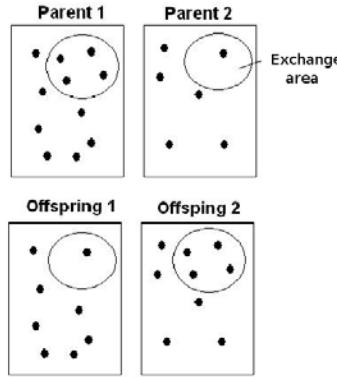
where:

$$\text{covered}(x, y) = \begin{cases} 1 & \text{If } \exists i / \{(b_i = 1) \wedge ((x, y) \in C[i])\} \\ 0 & \text{Otherwise} \end{cases} \quad (2)$$

Objective  $f_1$  is the number of deployed BS, so it must be minimised. Objective  $f_2$  is a measure of the covered area, so it must be maximised.

### 3 Multi-Objective Strategies

Several multi-objective approaches have been designed with the aim of obtaining an approximation of the MOP Pareto front. *Multi-objective evolutionary algorithms* (MOEAs) are one of the most widely used metaheuristic to deal with MOPs. Evolutionary algorithms have shown great promise for calculating solutions to large and difficult optimisation problems and have been successfully used across a wide variety of real-world applications [6]. They are population-based algorithms inspired on the biological evolution. Several MOEAs have been proposed in the literature. Although several comparisons have been performed among them, their behaviour depends on many problem-dependent characteristics. Thus, usually several approaches are applied to a problem, in order to determine the most suitable one. The following MOEAs have been used in this work: Non-Dominated



**Fig. 1.** Creation of offspring with GC

Sorting Genetic Algorithm II (NSGA-II) [5], Strength Pareto Evolutionary Algorithm 2 (SPEA2) [21] and the adaptive and non-adaptive Indicator-Based Evolutionary Algorithm (IBEA) [20].

In order to apply the previous MOEAs, an encoding for the individuals must be defined. Tentative solutions are represented as binary strings with  $n$  elements. Each gene determines whether the corresponding BS is deployed. Also, a set of variation operators must be defined in order to employ MOEAs. In [19] a comparison between several mutation operators is performed. Random mutation operators, as well as, directed mutation operators are analysed. Directed mutation operators include problem-dependent information. Advantages of directed mutation operators are not very clear, but a combination of random and directed operators seems suitable. In this paper we propose a similar comparison, but using crossover operators instead of mutation operators. The mutation operator applied was a random operator: Bit-Inversion Mutation(BIM). Each gene is inverted with a probability  $p_m$ . Two different crossover operators - one random, and one directed - were applied. The crossover operators are the following:

- One-Point Crossover (OPC) [9]: it chooses a random gene, and then splits both parents at this point and creates the two offspring by exchanging the tails.
- Geographic Crossover(GC) [16] (Figure 1): it exchanges the BS that are located within a given radius ( $r$ ) around a randomly chosen BS. The main purpose is to develop a non-destructive operator.

The aforementioned techniques were incorporated into the tool METCO [10] (*Metaheuristic-based Extensible Tool for Cooperative Optimisation*). METCO is a plugin-based tool which incorporates a set of multi-objective schemes to tackle MOPS. It provides both, sequential and parallel execution schemes. In order to incorporate the APP and the variation operators, a new C++ plugin was developed.

**Table 1.** Statistical comparison of MOEAs

CONFIGURATION	SPEA2	ADAPT. IBEA	IBEA	NSGA-II
SPEA2	0	15	15	15
Adapt. IBEA	0	0	14	11
IBEA	0	0	0	10
NSGA-II	0	3	4	0

## 4 Experimental Evaluation

In this section the experiments performed with the different optimisation schemes depicted in Section 3 are described. Tests have been run on a Debian GNU/Linux cluster of 8 HP nodes, each one consisting of two Intel(R) Xeon(TM) at 3.20GHz and 1Gb RAM. The compiler and MPI implementation used were *gcc 3.3* and *MPICH 1.2.7*. A real world-world-sized problem instance [8] was used. It is defined by the geographical layout of the city of Malaga (Spain). This instance represents an urban area of 27.2 Km<sup>2</sup>. The terrain has been modelled using a 450 x 300 grid, where each point represents a surface of approximately 15 x 15 m. This fine-grained discretization enables us to achieve highly accurate results. BS are modelled as omnidirectional isotropic antennas, with a radius of approximately one half kilometre. The dataset contains 1000 candidate sites for the BS.

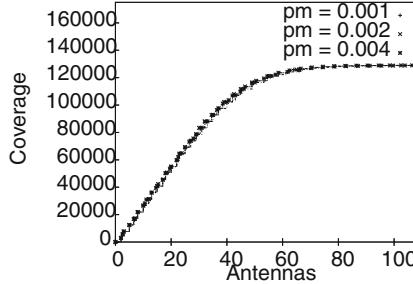
Since we are dealing with stochastic algorithms, each execution was repeated 30 times. In order to provide the results with confidence, comparisons have been performed following the next statistical analysis. First, a Kolmogorov-Smirnov test is performed in order to check whether the values of the results follow a normal (gaussian) distribution or not. If so, the Levene test checks for the homogeneity of the variances. If samples have equal variance, an ANOVA test is done; otherwise a Welch test is performed. For non-gaussian distributions, the non-parametric Kruskal-Wallis test is used to compare the medians of the algorithms. A confidence level of 95% is considered, which means that the differences are unlikely to have occurred by chance with a probability of 95%. The analysis is performed using the hypervolume [22] metric, and attainment surfaces [7]. In every case the attainment surface 15 is used.

In our first experiment a comparison among the tested MOEAs is carried out. Each MOEA is executed with a stopping criterion of 2 hours. The analysis is performed in terms of the obtained hypervolume. For each MOEA, 15 parameterisations were analysed. They were constituted by combining the BIM operator with  $p_m = \{0.001, 0.002, 0.004\}$  and the OPC operator with  $p_c = \{0, 0.25, 0.5, 0.75, 1\}$ . Table II shows the number of column configurations which are worse than the corresponding row configuration. For the analysed instance, SPEA2 is clearly the best strategy in terms of the achieved hypervolume. Results show that SPEA2 is better than any other algorithm, with any of the tested parameterisations.

The variation process in MOEAs is performed by combining mutation and crossover operators. The probability of mutation is usually fixed as  $1/n$ , being  $n$  the number of genes. The next experiment tests the robustness of the mutation

**Table 2.** Advantages of incorporating problem-dependent information

CONFIGURATION	OPC, $p_c = 1$	GC, $p_c = 0.25$	GC, $p_c = 0.5$	GC, $p_c = 0.75$	GC, $p_c = 1$
OPC, $p_c = 1$	$\leftrightarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
GC, $p_c = 0.25$	$\uparrow$	$\leftrightarrow$	$\leftrightarrow$	$\downarrow$	$\downarrow$
GC, $p_c = 0.5$	$\uparrow$	$\leftrightarrow$	$\leftrightarrow$	$\leftrightarrow$	$\downarrow$
GC, $p_c = 0.75$	$\uparrow$	$\uparrow$	$\leftrightarrow$	$\leftrightarrow$	$\leftrightarrow$
GC, $p_c = 1$	$\uparrow$	$\uparrow$	$\uparrow$	$\leftrightarrow$	$\leftrightarrow$

**Fig. 2.** Attainment Surfaces with several  $p_m$  values

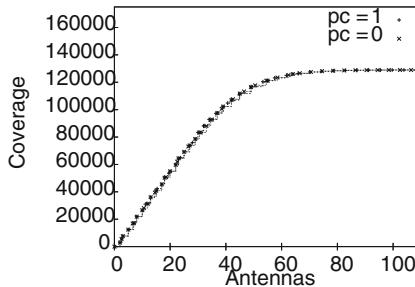
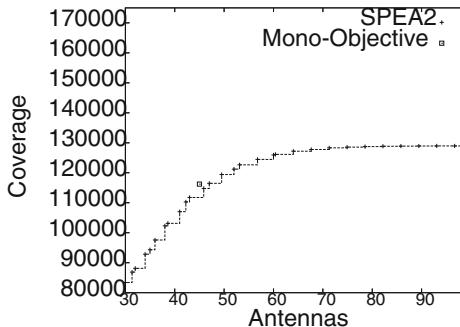
operator. Figure 2 shows the attainment surfaces for SPEA2 with BIM using  $p_m = \{0.001, 0.002, 0.004\}$ , and  $p_c = 0$ . We can note that changing  $p_m$  in the selected range do not produce a noticeable worsening in the achieved results. On the other hand, according to [16], it is necessary to include problem-dependent information in the crossover, in order to develop a non-destructive operator. However, they not provide any results with random operators. Figure 3 shows the attainment surfaces for SPEA2, with  $p_c = \{0, 0.5, 1\}$ , and  $p_m = 0.001$ . It confirms that OPC is not providing any improvement in the achieved results.

Next experiment tests the advantages of including problem-dependent information inside the crossover operator. Operator GC is tested with different crossover probabilities, and compared with OPC ( $p_c = 1$ ). Table 2 shows whether the row configuration is statistically better ( $\uparrow$ ), not different ( $\leftrightarrow$ ), or worse ( $\downarrow$ ), than the corresponding column configuration. It shows the benefits of incorporating problem information into de operator. Moreover, the best results are achieved by fixing high probabilities to the crossover operator.

GC uses a parameter which represents the radius ( $r$ ) of the exchange area. It is interesting to test the behaviour of such an operator with different values of  $r$ . SPEA2 was executed with GC fixing the next values for the radius:  $r = \{15, 30, 45, 60\}$ . Table 3 shows the statistical comparison with the same meaning as table 2. The best behaviour is obtained by using the parameterisation  $r = 30$ . This make sense because, since the antenna radius is also 30, the exchange area and the area influenced by the randomly selected antenna coincide.

**Table 3.** Statistical comparison of GC with different radius

CONFIGURATION	GC, $r = 15$	GC, $r = 30$	GC, $r = 45$	GC, $r = 60$
GC, $r = 15$	$\leftrightarrow$	$\downarrow$	$\downarrow$	$\downarrow$
GC, $r = 30$	$\uparrow$	$\leftrightarrow$	$\uparrow$	$\uparrow$
GC, $r = 45$	$\uparrow$	$\downarrow$	$\leftrightarrow$	$\uparrow$
GC, $r = 60$	$\uparrow$	$\downarrow$	$\downarrow$	$\leftrightarrow$

**Fig. 3.** Attainment Surfaces with several  $p_c$  values**Fig. 4.** Comparison of multi-objective and mono-objective approaches

Finally, Figure 4 compares the results obtained by multi-objective and mono-objective approaches. It shows the attainment surface achieved with the best tested configuration of SPEA2, as well as the best solution obtained by the best known mono-objective strategy [12]. On one hand, the multi-objective schemes are obtaining many solutions which are non-dominated by the best mono-objective solutions (in average the 97% of the solutions are non-dominated). Thus, the schemes are providing a large number of high-quality solutions. On the other hand, since the mono-objective solution dominates about 3% of the achieved solutions, there is still some room for improvement for the multi-objective schemes.

## 5 Conclusions and Future Work

In this paper we have designed and tested the ability of some multi-objective approaches to solve the APP. Several evolutionary algorithms have been analysed. In terms of hypervolume, SPEA2 has provided the highest quality solutions. Problem-dependent and problem-independent variation operators have been tested. Computational results show the advantages of incorporating problem-dependent information into the variation operators. Moreover, it shows the great importance of tuning such operators. A comparison between multi-objective and mono-objective approaches have been performed. Most of the multi-objective solutions are not dominated by the best mono-objective solution. However, there is still some room for improvement.

In order to improve the quality of the results, several modifications can be performed. On one hand, more problem-dependent information can be included into the strategy. For instance, by incorporating a local search scheme, or by applying some directed mutation operators. On the other hand, MOEAs can be parallelised easily. MOEAs parallelisation aims not only to achieve time saving by distributing the computational effort but also to get benefit from the algorithmic aspect by the cooperation between different populations. Although high quality solutions were achieved, a tuning step of the algorithms was required. In order to avoid such a step, parallel hyperheuristic-based island schemes [10] are an alternative.

## Acknowledgements

This work was supported by the EC (FEDER) and the Spanish Ministry of Science and Innovation as part of the 'Plan Nacional de I+D+i', with contract number TIN2008-06491-c04-02 and by Canary Government project number PI2007/015. The work of Carlos Segura was funded by grant FPU-AP2008-03213.

## References

1. Akella, M.R., Batta, R., Delmelle, E.M., Rogerson, P.A., Blatt, A., Wilson, G.: Base station location and channel allocation in a cellular network with emergency coverage requirements. European Journal of Operational Research 164(2), 301–323 (2005)
2. Alba, E.: Evolutionary algorithms for optimal placement of antennae in radio network design. In: International Parallel and Distributed Processing Symposium, vol. 7, p. 168 (2004)
3. Blum, C., Roli, A.: Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison. ACM Computing Surveys 35(3), 268–308 (2003)
4. Calégari, P., Guidec, F., Kuonen, P., Kobler, D.: Parallel island-based genetic algorithm for radio network design. J. Parallel Distrib. Comput. 47(1), 86–90 (1997)
5. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation 6, 182–197 (2002)
6. Eiben, A.E.: Handbook of Evolutionary Computation. IOP Publishing Ltd/Oxford University Press (1998)

7. Fonseca, C., Fleming, P.J.: On the performance assessment and comparison of stochastic multiobjective optimizers. In: Ebeling, W., Rechenberg, I., Voigt, H.-M., Schwefel, H.-P. (eds.) PPSN 1996. LNCS, vol. 1141, pp. 584–593. Springer, Heidelberg (1996)
8. Gmez-Pulido, J.: Web site of net-centric optimization,  
<http://oplink.unex.es/rnd>
9. Holland, J.H.: Adaptation in natural and artificial systems. MIT Press, Cambridge (1992)
10. León, C., Miranda, G., Segura, C.: METCO: A Parallel Plugin-Based Framework for Multi-Objective Optimization. International Journal on Artificial Intelligence Tools 18(4), 569–588 (2009)
11. Luna, F., Estébanez, C., León, C., Chaves-González, J.M., Alba, E., Aler, R., Segura, C., Vega-Rodríguez, M.A., Nebro, A.J., Valls, J.M., Miranda, G., Gómez-Pulido, J.A.: Metaheuristics for solving a real-world frequency assignment problem in GSM networks. In: GECCO 2008: Proceedings of the 10th annual Conference on Genetic and Evolutionary Computation, Atlanta, GA, USA, pp. 1579–1586. ACM, New York (2008)
12. Mendes, S.P., Molina, G., Vega-Rodrguez, M.A., Gomez-Pulido, J.A., Sez, Y., Miranda, G., Segura, C., Alba, E., Isasi, P., Len, C., Snchez-Prez, J.M.: Benchmarking a Wide Spectrum of Meta-Heuristic Techniques for the Radio Network Design Problem. IEEE Transactions on Evolutionary Computation, 1133–1150 (2009)
13. Mendes, S.P., Pulido, J.A.G., Rodriguez, M.A.V., Simon, M.D.J., Perez, J.M.S.: A differential evolution based algorithm to optimize the radio network design problem. In: E-SCIENCE 2006: Proceedings of the Second IEEE International Conference on e-Science and Grid Computing, p. 119. IEEE Computer Society, Washington (2006)
14. Meunier, H., Talbi, E.G., Reininger, P.: A multiobjective genetic algorithm for radio network optimization. In: Proceedings of the 2000 Congress on Evolutionary Computation, pp. 317–324. IEEE Press, Los Alamitos (2000)
15. Seybold, J.: Introduction to RF Propagation. Wiley Interscience, Hoboken (2005)
16. Talbi, E.G., Meunier, H.: Hierarchical parallel approach for gsm mobile network design. J. Parallel Distrib. Comput. 66(2), 274–290 (2006)
17. wan Tcha, D., Myung, Y.S., hyuk Kwon, J.: Base station location in a cellular cdma system. Telecommunication Systems 14(1-4), 163–173 (2000)
18. Vasquez, M., Hao, J.K.: A heuristic approach for antenna positioning in cellular networks. Journal of Heuristics 7(5), 443–472 (2001)
19. Weicker, N., Szabo, G., Weicker, K., Widmayer, P.: Evolutionary multiobjective optimization for base station transmitter placement with frequency assignment. IEEE Transactions on Evolutionary Computation 7(2), 189–203 (2003)
20. Zitzler, E., Künzli, S.: Indicator-Based Selection in Multiobjective Search. In: Yao, X., Burke, E.K., Lozano, J.A., Smith, J., Merelo-Guervós, J.J., Bullinaria, J.A., Rowe, J.E., Tiño, P., Kabán, A., Schwefel, H.-P. (eds.) PPSN 2004. LNCS, vol. 3242, pp. 832–842. Springer, Heidelberg (2004)
21. Zitzler, E., Laumanns, M., Thiele, L.: SPEA2: Improving the Strength Pareto Evolutionary Algorithm for Multiobjective Optimization. In: Evolutionary Methods for Design, Optimization and Control, pp. 19–26 (2002)
22. Zitzler, E., Thiele, L.: Multiobjective Optimization Using Evolutionary Algorithms - A Comparative Case Study. In: Eiben, A.E., Bäck, T., Schoenauer, M., Schwefel, H.-P. (eds.) PPSN 1998. LNCS, vol. 1498, pp. 292–301. Springer, Heidelberg (1998)

# CLONAL-GP Framework for Artificial Immune System Inspired Genetic Programming for Classification

Hajira Jabeen and Abdul Rauf Baig

National University of Computer and Emerging Sciences, Islamabad, Pakistan  
`{hajira.jabeen, rauf.baig}@nu.edu.pk`

**Abstract.** This paper presents a novel framework for artificial immune system (AIS) inspired evolution in Genetic Programming (GP). A typical GP system uses the reproduction operators mimicking the phenomena of natural evolution to search for efficient classifiers. The proposed framework uses AIS inspired clonal selection algorithm to evolve classifiers using GP. The clonal selection principle states that, in human immune system, high affinity cells that recognize the invading antigens are selected to proliferate. Furthermore, these cells undergo hyper mutation and receptor editing for maturation. In this paper, we propose a computational implementation of the clonal selection principle. The motivation for using non-Darwinian evolution includes avoidance of bloat, training time reduction and simpler classifiers. We have performed empirical analysis of proposed framework over a benchmark dataset from UCI repository. The CLONAL-GP is contrasted with two variants of GP based classification mechanisms and results are found encouraging.

**Keywords:** Artificial Immune Systems, Genetic Programming, Classification.

## 1 Introduction

Genetic Programming was originally introduced as an extension of Genetic Algorithm to automatically evolve computer programs. It posses several outstanding features when compared to other traditional evolutionary algorithms. These include: variable sized solution representation, ability to work with little or no knowledge about the solution structure, transparency, data independence and efficient data modeling ability. These features make GP readily applicable to evolve classifiers. This property of GP has been recognized since its inception [1]. Numerous researchers have developed different techniques to solve classification problems using GP. One of the profound techniques [2] [3] is evolution of arithmetic expressions as discriminating function between different classes. An arithmetic classifier expression (ACE) is created using numerical attributes of the data and some random constants. The value of expression is evaluated for every instance of the data where the output is a real value. This real output is mapped onto different classes of the data.

We have investigated the proposition to evolve the ACE using AIS principles. AIS [4]mimic the principles of human immune system, and are capable of performing many tasks in various areas. In this work, we will review the clonal selection concept, together with the affinity maturation process, and demonstrate that these biological

principles can lead to the development of powerful computational tools. AIS operates on the principle of human immune system [5], and, is capable of performing many tasks in various areas [6]. In real life when a human is exposed to an antigen (Ag), some subpopulation of its bone marrow derived cells (B cells) respond by producing antibodies (Ab). Each cell secretes a single type of antibody, which is relatively specific for the antigen. The antigen simulates the B-cell by binding to antibodies and to divide and mature into final (nondividing) secreting cells known as Plasma cells. These cells are most active antibody secretors. On the other hand B lymphocytes, which divide rapidly, also secrete antibodies, at a lower rate. T cells play a central role in regulating B cell response, but will not be explicitly accounted for the development of our model. Lymphocytes, in addition to proliferating and/or differentiating into plasma cells, can differentiate into long-lived B memory cells. Memory cells circulate through the blood, lymph and tissues, and when exposed to a second antigenic stimulus commence to differentiate into large lymphocytes capable of producing high affinity antibodies, pre-selected for the specific antigen that had stimulated the primary response.

The main features of the clonal selection explored in this paper are:

- Proliferation and differentiation on stimulation of cells with antigens
- Generation of new random genetic changes, subsequently expressed as diverse antibody patterns, by a form of depth Limited mutation (a process called affinity maturation)
- Elimination of newly differentiated lymphocytes carrying low affinity antigenic receptors.

## 2 GP for Data Classification

GP has been applied for classification in various ways. One of these is evolution of classification algorithms e.g. ‘decision trees’ [7] or evolution of rules for classifier evolution [8]. In [9] the grammar to GP evolution is evolved in such a way that ultimate result is a feasible classification algorithm. For algorithm evolution GP use some type of constrained syntax/grammar so that the trees transform into an algorithm and remain valid after application of evolutionary operators. Another method is evolution of classification rules [10] [11] [12] [13]. In this type of methods GP trees are evolved as logical rules, these methods are applied on numerical and nominal data. A newer and somewhat GP specific method evolution of classifiers is in the form of mathematical expressions [2] [3] [14] [15] [16]. In this type of classification the classifiers are evolved in the form mathematical discriminating expressions. Expressions having attribute values as terminals and arithmetic operators as functions are evolved such that the real valued output is used to extract the classification result.

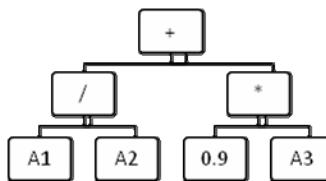
All above mentioned algorithms have used the principle of biological evolution to search for fitter solutions. Next section presents the proposed AIS inspired evolutionary framework. We have tested the performance of proposed framework on data classification problem.

### 3 Proposed Hybrid Framework

As mentioned in the previous section, the outstanding feature of GP is its ability to represent and evolve variable length solutions. This ability also introduces a drawback of increase in average tree size during evolution. We present a novel artificial immune system inspired evolution that controls bloat through depth limited receptor editing operator.

#### 3.1 Population Representation

An ACE is represented as expression trees where the operands are (+,-,\*,/.) and operators are attributes of data and a few ephemeral constants. Where ephemeral constants are randomly generated constants that remain unchanged once inserted into the tree node. For example consider a four attribute data having [A1, A2, A3, A4] as attributes for a binary classification problem. Figure 1 shows an example ACE .



**Fig. 1.** Arithmetic Classifier Expression

#### 3.2 Population Initialization

There are three well known initialization methods in GP literature. The *full* scheme creates full trees till the maximum depth allowed. It populates the tree with function nodes until maximum depth, beyond that, only terminal nodes are selected. On the other hand *grow* method randomly selects nodes from function or terminal set until maximum depth is reached and allows more diverse size and shape. The initialization method we used for ACE evolution is the well known *Ramped half and half* method [1]. The ramped half and half method utilizes advantages of both full and grow initialization schemes with equal probability. It makes use of different depths for full and grow method ranging from some minimum depth to the maximum allowed depth. This method has been widely used for initialization in many classification problems [2] [3] .

#### 3.3 Affinity Measure

For the classification purpose the antigen population **Ag** to recognize is the set of training samples. Two possible instances are given below for the tree mentioned in Figure 1.

$$I1 = [1 \ 2 \ 3 \ 4] \in C1$$

$$I2 = [1 \ 2 \ -3 \ 4] \in C2$$

For a classifier to recognize instances of particular class, we must train it to output different responses for different classes of data. We can train the expression to output a positive signal for class C1 and negative signal for C2 for a binary classification problem. In case of example expression from Figure 1, we will get the response 3.2 for I1 and -2.2 for I2. This will increase the affinity of **Ab** by 2. In this way we must calculate response for all the instances of training data to estimate the affinity measure of a particular **Ab** against antigens **Ag**. Similarly affinity measures for all the **Ab** population must be calculated. The above mentioned measure will only return the count of instances for which a classifier has output correct response. For classification we need more than correct output count.

To resolve this problem, we can see that our **Ag** population can be divided into two types **Ag+** for which **Ab** should output a positive response and **Ag-**, for which the **Ab** should output a negative response. Given these definitions we can define the new affinity measure as

$$\text{Affinity of } \mathbf{Ab} = \frac{\text{number of correct positive responses}}{\text{Number of } \mathbf{Ag} +} + \frac{\text{number of correct negative responses}}{\text{Number of } \mathbf{Ag} -}$$

This affinity measure is can also be seen as area under the convex hull(AUCH) for one threshold(0). This would ensure that the **Ab** with better discriminative power for two different **Ag** is assigned better affinity.

### 3.4 Proliferation

The proliferation of affinity or increase in average fitness during the evolution process is essential for efficient search of desired solutions. In case of AIS some highest affinity individuals are selected from population and cloned in proportion to their affinity. Fitter individuals will tend to create more clones as compared to less fit members of population.

### 3.5 Clone Maturation (Depth Limited Mutation)

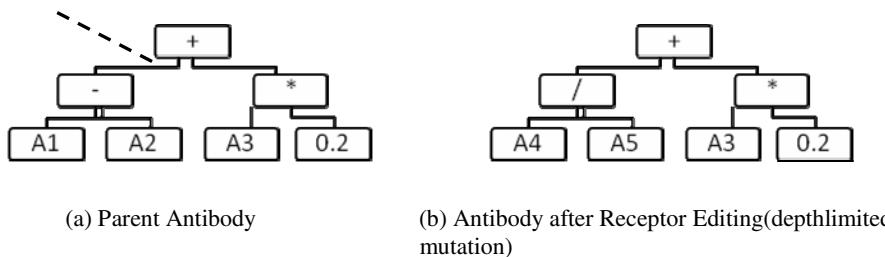
In biological immune system, the antigen activated B cell population is diversified by two mechanisms. These mechanisms are hyper mutation and receptor editing [17] [18]. Random changes are introduced via hyper mutation into the genes responsible for the **Ag-Ab** communication and the anticipation is that one such change can result in increase in affinity of an individual. The higher affinity variants are then collected in the memory cell repository. This hyper mutation makes sure that population is diversified and some different B cells with higher affinity may be generated. On the other hand the cells with lower affinity need elimination. [17] [19]

Some studies [18] also suggest that immune system practices molecular selection of receptors in addition to clonal selection of B cells. In such case B cell delete their low affinity receptors and developed entirely new ones. This editing operation offers the ability to escape from the local minima. In addition to hyper mutation and receptor editing, a fraction of newcomer cells from the bone marrow are added into the lymphocyte pool to ensure and maintain diversity of population. Sometimes 5% to 8% of population is replaced by newly created B Lymphocytes.

In the context of the classification problem, following terminology is adopted for abovementioned biological concepts.

Hyper mutation can be seen as point mutation where one random node from the expression is selected and replaced by its counterpart selected randomly from the primitive set. For example a function node is replaced by a new random function node and a terminal node is randomly replaced by another terminal from the terminal set.

Receptor editing means some large random change in a given B lymphocyte as opposed to small change in case of hyper mutation. In this editing operation we have proposed a new '*depthlimited*' mutation. A random node is selected from the expression and the subtree rooted at this node is replaced by a new generated tree. To ensure '*depthlimited*' mutation we have applied a restriction that the "depth" of newly created tree must be less than or equal to the replaced subtree. This phenomenon will ensure that the population of expressions does not suffer from complexity increase. This means the well known *bloat* problem is eliminated.



**Fig. 2.** Receptor editing

Newcomer cells are introduced by initializing some new expression trees and making them part of new population by discarding some unfit one from the population pool.

### 3.6 CLONAL-GP Algorithm

The overall CLONAL-GP algorithm can be described as follows

- Step 1. Begin
- Step 2. Randomly initialize population of B Lymphocytes  
*Ab*(B-cells/Antibodies)
- Step 3. While(termination condition)
  - Present Antigens *Ag* to the antibody population
  - Calculate affinity of all *Ab*
  - Select n highest affinity antibodies *Ab'*
  - Create clone population *C'* by Cloning *Ab'* proportional to their affinities (higher affinity *Ab* will generate more clones )
  - The *C'* is applied population maturation process resulting in *C''*(high affinity clones will be mutated less).
  - Determine affinity of matured clones against antigens.
  - Select high affinity individuals and compare them with memory cells.

- If the affinity of new B Lymphocytes is greater, then replaced memory cells.
  - Replace  $k$  lowest affinity antibodies by  $k$  newly generated individuals.
- Step 4 . End while  
 Step 5 . Output best memory cell  
 Step 6 . End

## 4 Results

We have experimented the proposed framework over the well know Wisconsin breast cancer dataset from the UCI repository. The experiment has been performed by applying 10 fold cross validation twice on one random sampling of data. This process is repeated five times. Therefore tenfold cross validation is repeated ten times on five different partitions of data.

**Table 1.** Traditional GP Parameters

S. No	Population size	600
1	Crossover Rate	0.50
2	Mutation Rate	0.25
3	Reproduction Rate	0.25
4	Selection for cross over	Tournament selection with size 7
5	Selection for mutation	Random
6	Selection for reproduction	Fitness Proportionate selection
7	Mutation type	Point Mutation
8	Initialization method	Ramped half and half method with initial depth 6
9	Initial Depth	4 Standard GP, Variable in DepthLimited GP

The parameters used in our empirical analysis for Traditional and DepthLimited GP are mentioned in Table 1. These parameters have been empirically selected in our previous work [20].

The parameters for CLONAL-GP framework proposed in this paper are mentioned in Table2. We have tried to keep the parameters consistent with our previous work.

Table 3 provides a comparison between classification accuracy achieved by three different variants of GP. The standard GP uses a typical classification method used various propositions [2] [16]. DepthLimitedGP [20] has been proposed to avoid bloat for classifier evolution. It is almost similar to standard GP with an exception of crossover operator that restricts crossover between subtrees of same depth. We can see that the proposed CLONAL-GP has performed better than both other GP variants for WBC dataset. The reason for this better performance may be the mutation operators that ensure and constantly enforce diversity in each evolutionary cycle.

**Table 2.** CLONAL-GP parameters for classification

S. No	Parameter	Value/Description
1	Population	600
2	Memory	10
3	Number of members selected for cloning	100
4	Number of members replaced in each population	5% (30)
5	Hypermutation	Point mutation
6	Receptor editing	Subtree mutation (depthLimited)
7	Termination Condition	100 generations or affinity =1

**Table 3.** Comparison of different GP variants

Dataset		Standard GP	DepthLimited GP	CLONAL-GP
WBC	Accuracy	94.5%	95.8 %	97.2%
	Tree Size	958.2	31.8	30.6
	Time	18629 sec	11762 sec	10324 sec

The point and depthlimited mutation ensure that the trees will not increase in their complexities during evolution. This automatically eliminates the bloat prevalence in population.

**Table 4.** Comparison with other classification methods

Classifiers	CLONAL-GP	SVM	ANN	C4.5
Accuracy	97.2%	96.7%	95.6%	96%

Table 4 compares the performance of proposed algorithm with other classification algorithms and CLONAL-GP has achieved better performance over WBC data.

## 5 Conclusion

This paper presents a novel framework for using clonal selection in GP for classification purpose. The framework offers several advantages including, elimination of bloat, simpler classifiers. All these issues tend to overburden traditional GP which is a powerful tool for classifier evolution. The proposed framework is a part of ongoing research and various problems must be addressed in the future. This includes determination of optimal parameters. Moreover, we have investigated our proposition over binary classification problem. The work can be extended to incorporate multiclass classification problems.

## References

- [1] Koza, J.R.: Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge (1992)
- [2] Kishore, J.K., et al.: Application of Genetic Programming for Multicategory Pattern Classification. *IEEE Transactions on Evolutionary Computation* (2000)
- [3] Muni, D.P., Pal, N.R., Das, J.: A Novel Approach To Design Classifiers Using GP. *IEEE Transactions of Evolutionary Computation* (2004)
- [4] Ishida, Y.: The Immune System as a Self Identification Process: A Survey and a Proposal. In: *Proceedings of the IMBS 1996* (1996)
- [5] Burnet, F.M.: Clonal Selection and After. *Theoretical Immunology*, 63–85 (1978)
- [6] Castro, L.N., Zuben, F.J.V.: Learning and Optimization Using the Clonal Selection Principle. *IEEE Transaction on Evolutionary Computation* (2001)
- [7] Koza, J.R.: Concept formation and decision tree induction using the genetic programming paradigm. In: Schwefel, H.-P., Männer, R. (eds.) *PPSN 1990. LNCS*, vol. 496. Springer, Heidelberg (1991)
- [8] South, M.C.: The Application of Genetic Algorithms to Rule Finding in Data Analysis (1994)
- [9] Pappa, G.A., Freitas, A.A.: Evolving Rule Induction Algorithms with Multiobjective Grammer based Genetic Programming. *Knowledge and Information Systems* (2008)
- [10] Engelbrecht, A.P., Schoeman, L., Rouwhorst, S.: A Building Block Approach to Genetic Programming for Rule Discovery. In: Abbass, H.A., Sarkar, R., Newton, C. (eds.) *Data Mining: A Heuristic Approach*, pp. 175–189. Idea Group Publishing, USA
- [11] Mendes, R.R.F., et al.: Discovering Fuzzy Classification Rules with Genetic Programming and Co-Evolution. In: *Genetic and Evolutionary Computation Conference* (2001) (Late Breaking Papers)
- [12] Bojarczuk, C.C., Lopes, H.S., Freitas, A.A.: Discovering Comprehensible Classification Rules using Genetic Programming: A Case Study in a Medical Domain. In: *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 953–958. Morgan Kaufmann, San Francisco (1999)
- [13] Falco, I.D., Cioppa, A.D., Tarantino, E.: Discovering Interesting Classification Rules With GP. In: *Applied Soft Computing*, pp. 257–269 (2002)
- [14] Zhang, M., Wong, P.: Genetic Programming for Medical Classification: A Program Simplification Approach. In: *Genetic Programming and Evolvable Machines*, pp. 229–255 (2008)
- [15] Zhang, M., Ciesielski, V.: Genetic Programming For Multiple Class object Detection. In: *Proceedings of the 12th Australian Joint Conference on Artificial Intelligence*, Australia, pp. 180–192 (1999)
- [16] Bojarczuk, C.C., Lopes, H.S., Freitas, A.A.: Genetic programming for knowledge discovery in chest-pain diagnosis. *IEEE Engineering in Medicine and Biology Magazine*, 38–44 (2000)
- [17] Berek, C., Ziegner, M.: The Maturation of Immune Response. *Imm Today*, 400–402 (1993)
- [18] Tonegawa, S.: Somatic Generation of Antibody Diversity. *Nature*, 575–581 (1983)
- [19] Nussenzweig, M.C.: Immune Recepto Editing; Revise and Select. *Cell*, 875–878 (1998)
- [20] Jabeen, H., Baig, A.R.: DepthLimited Crossover in Genetic Programming for Classifier Evolution. In: *Computers in Human Behaviour*. Elsevier, Ulsan (2009) (accepted)
- [21] Loveard, T., Ciesielski, V.: Representing Classification Problems in Genetic Programming. In: *IEEE Congress on Evolutionary Computation*, pp. 1070–1077 (2001)
- [22] Smart, W., Zhang, M.: Using Genetic Programming For Multiclass Classification By Simultaneously Solving Component Binary Classification Problems. *LNCS*. Springer, Heidelberg (2005)

# Solving Industrial Based Job-Shop Scheduling Problem by Distributed Micro-Genetic Algorithm with Local Search

Rubiyah Yusof, Marzuki Khalid, and Tay Cheng San

Universiti Teknologi Malaysia  
Center for Artificial Intelligence and Robotics (CAIRO),  
Jalan Semarak,  
54100 Kuala Lumpur  
rubiyah@ic.utm.my, marzuki@utm.my

**Abstract.** Genetic algorithms (GAs) have been found to be suitable for solving Job-Shop Scheduling Problem (JSSP). However, convergence in GAs is rather slow and thus new GA structures and techniques are currently widely investigated. In this paper, we propose to solve JSSP using distributed micro-genetic algorithm (micro-GA) with local search based on the Asynchronous Colony Genetic Algorithms (ACGA). We also developed a representation for the problem in order to refine the schedules using schedule builder which can change a semi-active schedule to active schedule. The proposed technique is applied to Muth and Thompson's 10x10 and 20x 5 problems as well as a real world JSSP. The results show that the distributed micro GA is able to give a good optimal makespan in a short time as compared to the manual schedule built for the real world JSSP.

**Keywords:** Genetic Algorithm, Asynchronous Colony GA (ACGA), Job-Shop Scheduling Problem (JSSP), Distributed Micro-GA.

## 1 Introduction

The JSSP is not only an NP-hard problem, but also very well known as one of the worst NP-hard problem. An indication to this is the Muth and Thompson 10 x 10 problem that was formulated in 1963 remain unsolved for more than 20 years [1]. Over the years, many researchers tried to solve JSSP using many techniques. First attempt to solve JSSP was done by Giffler and Thompson [2] in which simple JSSP was solved using mixed integer linear programming. However, the first heuristic that manage to solve complex JSSP is branch and bound algorithms [3]. Work on solving JSSP using GA can be found in [4],[5], [6],[7].

Generally, GA is employed to minimize the makespan of the scheduling jobs with quick convergence to the optimal solution, hence reducing the computational cost. However, as the problems become larger and complex, the computation time to find the optimal solution using GA will increase since it needs a bigger population. To solve this problem, Parallel GA (PGA) is explored by researchers to make the technique faster by

executing GAs on parallel computer. PGA Many researchers employed PGA in solving JSSP with varying degree of success such as [8], [9] and [10]. Another paper on JSSP that employed PGA is by Park et al. [11], which proposed an island-model PGA to prevent premature convergence, minimize the makespan and generate better solution than serial GA.

The main requirement for most real industrial based JSSP is the time needed to build the schedule. More often than not, ad hoc schedule is needed in order to cater for ad hoc changes such as machine breakdown, unavailability of maintenance personnel, changes in production requirement, etc. In order to cater for this need, we propose the use of a distributed GA with local search in order to solve a real industrial JSSP which is part of one of the local industry problem. The proposed distributed GA was based on Inoue et al. [12] which used Asynchronous Colony GA (ACGA). In ACGA, as proposed by Inoue [12], the supervisor task (parent GA task) will start a predetermined number of sub-GA tasks, which run on a single computer. The method was used on symmetrical multiprocessor (SMP) machine (which have up to sixteen CPUs) to execute the sub-GAs. Each sub-GA, consists of small populations (not more than 100 individuals per population) will communicate among themselves, sharing the information through asynchronous message passing through a host. The number of sub-GA tasks “spawned” is heavily dependent on the system speed and amount of RAM (random access memory) available. The method proved to be very fast and able to give optimal results within a short time.

However, SMP is very expensive and not applicable to be used for the job shop scheduling discussed in this paper. Therefore, we proposed to use a distributed micro GA with local search configuration on a normal personal computer (PC). As micro GA [13] evolves a very small population, the computation time will be very much reduced. One the main problems of micro GA though, is the loss of genetic diversity because of the small population. However, the proposed distributed micro GA model, ensures the retention of the genetic diversity due to the passing of individuals from one sub GAs to another. Although there are a number of researches done on JSSP, not many of them can be implemented to solve our industrial based JSSP. This is because of: most of the methods use representation to suit the reference JSSP, whereby an operation is processed by one machine. However, in our real world JSSP, an operation can be processed by more than one machine, making it difficult to be implemented. Also, in solving real world JSSP, we must also consider other constraints such as machine setting time etc. To address these issues, we develop a more flexible representation of the problem.

The organization of the paper is as follows: section 2 will be a description of the JSSP and also the industrial based JSSP which is of concern in this paper, section 3 will discuss the distributed micro GA with local search, Section 4 will describe the experiments done and some discussion on the results. The conclusion will be in Section 5.

## **2 Job Shop Scheduling**

Job-shop scheduling is an activity to allocate shared resources (normally machines) over time to competing activities (jobs which consist of operations using different machines). All the machines can only process a single operation at a time, and all the operations in a job must be processed in a technological sequence.

For a JSSP which consists of a set of  $n$  jobs, job  $J$  can be written as  $\{J_i\}_{1 \leq i \leq n}$ . These jobs will be processed on  $m$  machines, which can be written as  $\{M_r\}_{1 \leq r \leq m}$ . Each job has a set of operations or tasks that must be process in technological sequence. An operation of job  $J_i$  that processed on a machine  $M_r$  can be written as  $O_{ir}$ . Therefore, a schedule is an order of operations of all the jobs given that:

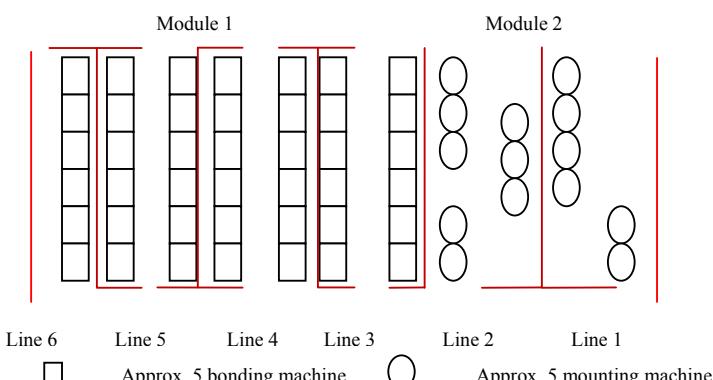
- i. All operation must be in technological sequence order, e.g. second operation of job one must be processed after first operation of job one, and so on.
- ii. All operation can only be processed on one machine at the same time.
- iii. The time needed to complete a schedule is called makespan  $L$ . For a general JSSP, the main objective of optimization is to minimize the makespan. For example in Fig 1, makespan for the schedule is 20 units of time.

## 2.1 Real-World JSSP

A production data from an integrated circuit (IC) assembly factory in Malaysia has been gathered for the purpose of this work. The data are collected from two assembly modules where each of the modules performs different operations, namely mounting and bonding. There are 2 lines in mounting module which consists of 68 mounting machines; while bonding module have 4 lines which consists of 212 bonding machines. Figure 1 shows the layout of the plant.

The raw materials are stored in a box and each box contains 4000 units of raw materials, called a lot. These boxes must be processed in module 1 before they can be processed in module 2. The production data collected contains 1093 lots of raw materials. Each raw material consists of an IC with pins on it and also a lead frame. The processing time needed for each raw material depends on the number of pins and the type of lead frame. Each set of number of pins is allocated with a type of lead frame and the number of lots. In total there are 23 combinations of the number of pins, lead frames and number of lots. And there are a total 1093 lots.

All of the machines will need different time to process each of the combination of number of pins, types of lead frames and the number of lots allocated. Therefore, all together, the mounting machines have 1564 possible processing time while bonding



**Fig. 1.** Layout of IC Assembly Plant

machines have 4876 possible processing time for a type of raw material. Other than processing time, setup time is also considered if the previous raw material type processed was different from the current raw material type. The differences between classical JSSP and real-world JSSP in our case is that the machine needed some extra time for setup before start processing a raw material if the raw material processed before has different combination and an operation can be processed in any machine.

### 3 Micro-Genetic Algorithm

Micro-Genetic Algorithm (Micro-GA) refer to GAs that works on a small set of population and with reinitialization. The idea of micro-GA is first suggested by Goldberg[14] who found that a population with size of three individuals is enough to converge without considering the length of chromosome. The basic idea of micro-GA is to work on a small population until nominal convergence, which mean until all the individuals in the population have very similar chromosome. From there, the best individual will be kept and transfer to new population which will generated randomly. Micro-GA is first implemented by K. Krishnakumar with population size of five [13]. Basically, micro-GA differs from the conventional genetic algorithms by an outer loop on top of conventional GA and it works on small population size.

In this paper, we use a random number generator to generate a set of initial population. An active schedule builder is used to transform those individuals in the population into active schedule. Roulette selection method which put more weight onto fitter individuals is used to choose individuals for crossover. The fitness function allows for fitter individuals to have a better chance to participate in the crossover and mutation process. In JSSP, makespan is used as fitness by most of the researchers. In our case, we are also using makespan as the fitness for each individual. In this problem, the objective is to maximize the fitness value which is positive. The fitness of individuals is evaluated using the following formulas:

$$\text{fitness} = f(x_i) = w - L \quad (1)$$

Where

$w$  is maximum makespan,

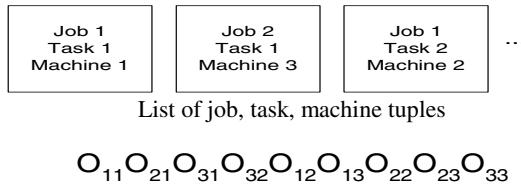
$L$  is makespan,

$x_i$  is the  $i$ th individual,

#### 3.1 Genetic Representation

In GA, choosing representation of the problem is very important to match the techniques used. Early genetic algorithms are mostly focused on bit representation, but GA can operate on any other data types, such as integer or list representation. For any representation, only minimum information should be included. This is because if the representation includes more information than needed, the operation on genetic algorithm will be slower. In this paper, because of the design of this genetic algorithm engine is not only meant to solve the JSSP problem, but more complex problem that will be applied to real production software, a more flexible or extensible representation is chosen. The list of job, task, machine tuples are used as representation of the problem. This representation and its associated 3x3 JSSP schedule is shown in Fig 2.

The advantage of using this representation is when the GA is applied to real world industrial problem, it is easier to modify. The disadvantage is it will consume more resources than the other JSSP specific representation. It is impossible to find a representation that will not yield invalid schedule in JSSP. Therefore, every time after a schedule is generated by initialization or crossover process, we use a repairing function to repair the schedule. The repair function act as a checker for constraints of the problem.



**Fig. 2.** A list of tuples representing a  $3 \times 3$  schedule

### 3.2 Schedule Builder

A completely random schedule called semi-active schedule is generated initially. Semi-active schedule is a schedule that can still be improved by applying permissible left-shift to the operations in schedule. An active scheduler performs a kind of local search, which can be used to introduce heuristic improvement into genetic search. A very good example is the Giffler-Thompson algorithm [2] used by [7] and [6].

### 3.3 Crossover

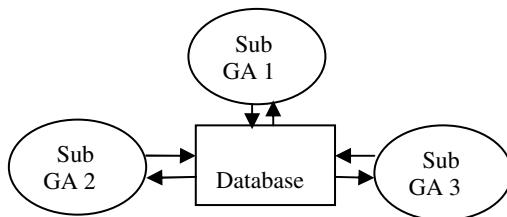
Crossover is an important operator in GA. A good crossover operator can inherit good solutions from the parents to one or more offsprings. We have chosen Generalized Order-Crossover (GOX) operator, which was introduced by Bierwirth [15], as the crossover operator. GOX uses two parents to produce an offspring. The first parent, termed as receiver will receive crossover-string from the second parent, termed as donor. Offset position of a crossover string is selected randomly using random number generator within the range of donor chromosome. The length of crossover-string is randomly chosen in between one half and one third of the total length on a chromosome. This is to ensure that both parents have almost the same amount of information in their offspring.

## 4 Distributed Micro GA

The distributed micro GA used in this paper shown in Fig 3 is similar to ACGA. In ACGA, there is a supervisor task (parent GA task) which will start a predetermined number of sub-GA tasks on a single PC. Instead of using Symmetrical Multiprocessor (SMP) machine as implemented by Inoue et al. [12], this research used a normal PC which consists of only a single CPU to execute the sub-GA. Each sub-GA consists of small populations, where they will communicate among themselves and sharing the information through asynchronous message passing.

In the normal ACGA, the supervisor GA has two responsibilities: firstly to handle its own version of GA, and secondly to store a duplicate of the best individual from its own “colonies” to a “global mailbox”. In this paper, we propose the use of a database as the exchange medium or “global mailbox” for all the sub GA processes. And all sub GA have the same responsibilities and none are termed as supervisor. One of the reasons for doing this is to avoid any of the sub GA with extra burden of comparing the solutions from other sub GAs, thus balancing the operation of the sub GAs. Also, the database will store all the good solutions from the sub GAs, thereby increasing the exploration capabilities of the GA. The number of processes that can participate in optimization is not limited. In this way, each of the sub GA process will improve the solutions on their own and at the same time exchange the solutions among each other through the database. Each of the processes will perform the steps below:

1. Run micro GA for optimization
2. After each complete micro GA loop, check the current best solutions in the database
3. If the solution in database is better than the best individual in the population, replace the worst individual in population with the solution in database; if the solution in database is worse than the best individual in population, replace the solution in database with the best individual in population; else take no action
4. Back to step 1



**Fig. 3.** Proposed Distributed Micro GA

GA is well known for being a great tool for global search but it is rather slow to converge. On the other hand, local search is a good and fast method for fine tuning a solution. Therefore, many researchers had proposed numerous ways to speed up the GAs, including combining the GAs with local search [16]. Using these methods, GA becomes the main operator that will perform the global search while local search is used to fine tune the individuals produced by crossover operator. To speed up the convergence of the micro GAs, we use a local search function after each generation. Our local search function can be described in the following steps:

1. Select a chromosome
2. Randomly find a the position of the gene or locus within the selected chromosome
3. Search through the other possible value for the selected gene
4. Return the best chromosome after search

## 5 Experimental Results and Discussions

The experiments are done in two sets. The first set of the experiment is for comparing the effectiveness of using micro GA, distributed Micro GA with local search as compared to the conventional GA with local search. The Muth and Thompson 10 x 10 and 20 x 5 datasets that was downloaded from OR library from (<http://mscmga.ms.ic.ac.uk/info.html>) are used as the basis of comparison.

Secondly, the distributed micro GA with local search is tested using the real world JSSP and the results of the schedule are compared with the one obtained by manual schedule. In implementing the JSSP using the algorithms, several GA operators have to be determined. These settings of the GA and Micro GA operators are given in Table 1.

**Table 1.** Settings for Conventional GA & Micro GA with Local Search (Datasets)

Setting	Value (GA)	Value (Micro GA)
Crossover Rate	0.8 (80%)	0.8 (80%)
Mutation Rate	0.001 (0.1%)	0
Generation	5000	50
Population Size	100	5
Maximum Cost	3000	100000
Micro GA loop		5000

All three types of GA are executed several times in order to check the consistency of the results. In each execution, the GA runs the complete 5000 generations if the minimum optimal makespan is not obtained, otherwise the GA will stop once it obtained the minimum optimal makespan for the dataset. The minimum optimal makespan is set to be 930 for the 10x10 dataset and 1170 for the 20x5 dataset. Figures 4 and 5 show the comparative results of the three configurations.

From the result of 10x10 and 20x5 datasets, it can be seen that conventional GA gives a more consistent result in terms of the time taken to obtain minimum optimal makespan as compared to the micro GA when run for 10 executions. The conventional

---

**Fig. 4.** Time taken for optimal makespan for 10x10 Muth and Thompson Dataset

GA takes almost the same amount of time to reach a minimum optimal makespan with an average time of 24.2 minutes over 10 different execution, and the variance of time taken over 10 execution runs is small at 0.55 for the 10x10 dataset. Whereas, the micro GA gives a lower average of time taken to obtain the optimal minimum makespan at 10.97 minutes, but with a larger variance of 3.62. Using the distributed micro GA helps to reduce both the variance as well as the average time taken for obtaining the minimum optimal makespan of the micro GA, with the average at 6.57 minutes and the variance reduced to 0.65. The result of the 20x5 dataset bears the same pattern as the 10x10 dataset. Table 2 summarizes the comparison of performance of the algorithms in terms of variance and average time taken for obtaining the optimal minimum makespan.

---

**Fig. 5.** Time taken for optimal makespan for 20x5 Muth and Thompson Dataset

**Table 2.** Summary of the consistency results for different execution run

<b>Time taken to Obtain Optimal Minimum Makespan(mins)</b>								
	10x10 Muth & Thompson				20x5 Muth & Thompson			
	Ave	Best	Worst	Variance	Ave	Best	Worst	Variance
Conv GA	24.24	23.12	25.3	0.55	26.1292	24.4	27.56	0.891
Micro GA	16.82	12.7	19.1	3.62	20.437	18.023	23.99	3.812
Distributed GA	10.16	9.19	11.4	0.65	15.8066	14.18	16.95	1.228

One of the reasons for the inconsistency of the micro GA is because it suffers from the loss of genetic diversity because of the small population. However, using the distributed micro GA, ensures the retention of some of the genetic diversity due to the exchange of and the passing of individuals from one sub-GA to another. In this way the exploitation characteristics of the GA is much improved. In terms of makespan, all the 3 GAs are executed over the maximum generation in order to get the best makespan possible. The distributed micro GA gives the best makespan for the 10 execution run, with an average of 928 which is 15.77% lower than using the conventional GA, and the best makespan is 914. Figure 6 shows the performance of the 3 GAs based on makespan. Apart from giving an optimal solution at a faster rate as compared to the conventional GA and the micro GA, the distributed micro GA also gave a good minimum optimal makespan based on the 10 execution run.

**Fig. 6.** Optimal makespan for 10x10 Muth and thompson Dataset

The distributed GA is applied to the JSSP using a software developed. The information of the JSSP is given in Table 3. The software consists of an interface for the distributed micro GA execution as well as a database for storing data on raw materials, raw material combinations, machines, etc. The results of the performance of the distributed micro GA is compared with a manual schedule which is developed in the industry. The makespan for the manual schedule is 5 days 18 hours 57 minutes or 8337 minutes. Table 4 gives a summary of the makespan obtained by the distributed micro GA. From Table 3, it can be seen that the distributed micro GA can reduce the makespan for the JSSP by up to 28.9% if run for more than 3 hours. The reduction in makespan for this kind of problem is very useful as it will reduce the cost of operations. Moreover, the time taken to obtain the schedule with minimal makespan or reasonable makespan is very much less as compared to doing a manual schedule. For this kind of real-world JSSP , it is very useful as generation of master schedule for daily basis are required. Also, the powerful feature of micro GA, which is fast convergence, makes it suitable to be used for any ad-hoc changes on master schedule because it can generate a reasonable schedule which is better than the manual scheduler in less than 3 minutes.

**Table 3.** Summary of Real-World JSSP Data

Parameter	Value
Lot	1093
Mount Machine	68
Bond Machine	212
Pin Type	9
Lead Frame Type	25
Pin-Lead Frame Combination	23
Mount Machine Setup Time	116.7 minutes
Bond Machine Setup Time	348.3 minutes

**Table 4.** Summary of the result of distributed Micro GA applied to real world JSSP

Execution time	Makespan	Makespan (minutes)	% improvement
<b>150 Seconds</b>	5 days 13 hrs 12 mins	7992	4%
<b>25 minutes</b>	4 days 22 hrs 29 mins	7109	17.2%
<b>3 hours</b>	4 days 11 hrs 46 mins	6466	28.9%

## 6 Conclusions and Future Work

We have shown that the use of distributed micro GA is effective in solving real world JSSP problem which are complicated and requires a lot of computation. One of the main advantages of the micro GA in distributed form is the ability to give a reasonable solution within a short time. We have also demonstrated that the distributed micro GA is able to overcome the problem of the inconsistency of the micro GA due to the loss of diversity of the solutions. The use of the micro GA in distributed form improves not only the exploitation capabilities of the GA, but also strengthen the exploration capabilities of the GA by the passing of individuals from one sub GA to the other. In future, the concept of parallel GA using multiple computers can be used to solve the JSSP more optimally.

## References

1. Muth, J., Thompson, G.: Industrial Scheduling. Prentice Hall, Englewood Cliffs (1963)
2. Giffler, B., Thompson, J.L.: Algorithms for Solving Production-Scheduling Problems. Operations Research 8, 487–503 (1960)
3. Carlier, J., Pinson, E.: An Algorithm for Solving Job-Shop Problem. Management Science 35, 164–176 (1989)
4. Kubota, A.: Study On Optimal Scheduling for Manufacturing System by Genetic Algorithms. Ashikaga Institute of Technology: Master Thesis (1995)
5. Holsapple, C., Jacob, V., Pakath, R., Zaveri, J.: A Genetics-Based Hybrid Scheduler for Generating Static Schedules in Flexible Manufacturing Contexts. IEEE Transactions on System, Man, and Cybernetics 23, 953–971 (1993)
6. Yamada, T., Nakano, R.: Genetic Algorithms for Job-Shop-Scheduling Problems. In: Proceedings of Modern Heuristic for Decision Support, UNICOM seminar, London, pp. 67–81 (1997)
7. Dorndorf, U., Pesch, E.: Evolution Based Learning in A Job Shop Scheduling Environment. Computers Ops. Res. 22, 25–40 (1995)
8. Zhang, H., Chen, R.: Research on Coarse-grained Parallel Genetic Algorithm Based Grid Job Scheduling. In: Proceedings of the Fourth International Conference on Semantics, Knowledge and Grid, pp. 505–506 (2008)
9. Kirley, M.: A Coevolutionary Genetic Algorithm for Job Scheduling Problems. In: Proceedings of the 1999 Third International Conference on Knowledge-Based Intelligent Information Engineering Systems, pp. 84–87 (1999)
10. Defersha, F.M., Chen, M.: A Coarse-Grain Parallel Genetic Algorithm for Flexible Job-Shop Scheduling with Lot Streaming. In: Proceedings of International Conference on Computational Science and Engineering, 2009, pp. 201–208 (2009)

11. Park, B.J., Choi, H.R., Kim, H.S.: A hybrid genetic algorithm for the job shop scheduling problems. *Computers & Industrial Engineering* 45, 597–613 (2003)
12. Inoue, H., Funyu, Y., Kishino, K., Jinguiji, T., Shiozawa, M., Yoshikawa, S., Nakao, T.: Development of Artificial Life Based Optimization System. In: Proceedings of the Eighth International Conference on Parallel and Distributed Systems, pp. 429–436 (2001)
13. Krishnakumar, K.: Micro-genetic algorithms for stationary and non-stationary function optimization. In: SPIE Proceedings Intelligent Control and Adaptive Systems, pp. 289–296 (1989)
14. Goldberg, D.: *Genetic Algorithms in Search Optimization and Machine Learning*. Addison-Wesley, Menlo Park (1988)
15. Bierwirth, C.: A Generalized Permutation Approach to Job Shop Scheduling with Genetic Algorithms. *OR Spektrum* 17, 87–92 (1995)
16. Merz, P., Freisleben, B.: A Genetic Local Search Approach to the Quadratic Assignment Problem. In: Proceedings of the Seventh International Conference on Genetic Algorithms, ICGA 1997 (1997)

# Data Mining via Rules Extracted from GMDH: An Application to Predict Churn in Bank Credit Cards

Nekuri Naveen<sup>1,2</sup>, V. Ravi<sup>1,\*</sup>, and C. Raghavendra Rao<sup>2</sup>

<sup>1</sup> Institute for Development and Research in banking Technology, Castle Hills Road #1,  
Masab Tank, Hyderabad – 500 057 (A P) India

<sup>2</sup> Department of Computer & Information Sciences, University of Hyderabad,  
Hyderabad – 500 046 (A P) India  
naveen.nekuri@gmail.com, rav\_padma@yahoo.com,  
crrcs@uohyd.ernet.in

**Abstract.** This paper proposes a hybrid method to extract rules from the trained Group Method of Data Handling (GMDH) neural network using Decision Tree (DT). The outputs predicted by the GMDH for the training set along with the input variables are fed to the DT for extracting the rules. The effectiveness of the proposed hybrid is evaluated on four benchmark datasets namely Iris, Wine, US Congressional, New Thyroid and one small scale data mining dataset churn prediction using 10-fold cross-validation. One important conclusion from the study is that we obtained statistically significant accuracies at 1% level in the case of churn prediction and IRIS datasets. Further, in the present study, we noticed that the rule base size of proposed hybrid is less in churn prediction and IRIS datasets when compared to that of the DT and equal in the case of remaining datasets.

**Keywords:** Group Method of Data Handling (GMDH), Decision Tree (DT), Rule Extraction, Data Mining, Classification, Churn prediction.

## 1 Introduction

Data mining has become ubiquitous in every domain like financial, bio-medical, manufacturing [1] etc. By learning from the past experience (data) and generalizing on the new data, Artificial Neural Networks (ANN) yield outstanding performance in prediction. In general, ANN achieves very high accuracies. An important drawback of the neural networks, however, is the black box stigma. It means that the knowledge learned by the neural network during training cannot be comprehended by human beings because the neural network (NN) simply does not output it. To overcome this disadvantage many researchers proposed various methods to extract knowledge from trained NN, in the form of IF-THEN rules. In applications like Churn Prediction [2], Medical Diagnosis [3], Bankruptcy Prediction, Credit approval in banks, Fraud detection [4], it is very much important to understand the knowledge gained by the NN.

---

\* Corresponding author. Ph: +91-40-2353 4981.

Since we solved the churn prediction problem in this paper, we briefly describe it as follows. Churn is an expensive phenomenon in the entire service industry which includes banking, finance, insurance, retail, health care and manufacturing. As regards churn prediction problem in banks decision makers are more conscious about the quality of service they provide to the customer because of increase in the attrition or churn of customers by the day. Loyal customers defecting from one bank to another has become common. Churn occurs due to lack of latest technology, unease of utility of the technology, customer friendly staff, etc. Thus, there is a vital need in developing a data mining model which can predict the probability that an existing loyal customer is going to churn out in the near future [2]. Churn prediction is one of the most vital activities in customer relationship management (CRM). Once potential churners are identified, management employs anti-churn strategies that are less expensive than acquiring new customers [5].

As regards making the neural networks transparent, Gallant [6] was the first to extract rules from the trained NN. Then several works appeared in literature [3, 7-17] to extract rules from NN in different ways. Most recently, Naveen et al., [18] extracted rules from differential evolution trained radial basis function network (DERBF) using GATree. Then, Farquad et al., [19] extracted rules from SVM using NBTree and applied to predict churn in bank credit cards.

Of all neural network architectures, GMDH is a relatively under explored. Srinivasan [20] used GMDH for electric energy forecasting. Then, recently, Mohanty et al. [21] used GMDH for software reliability prediction. It should be noted that Srinivasan [20] and Mohanty et al. [21] used GMDH and solved small size datasets. Like all neural network architectures, GMDH also suffers from the black box stigma. In other words, it does not let the users know the knowledge learnt by it during training. Thus, there is a pressing need to develop hybrid methods which would bring in transparency to the GMDH. Motivated by this idea, in this study, we propose a hybrid methodology comprising GMDH and C4.5 in tandem. To test the effectiveness of the hybrid methodology and the rules extracted, we analyzed both small scale benchmark datasets and a medium scale dataset taken from the area of CRM pertaining to churn prediction in bank credit cards.

The remainder of the paper is organized as follows: In Section 2, an overview of the GMDH and DT is presented. Section 3 describes the proposed hybrid. Results and discussion are presented in Section 4. Finally, conclusions are presented in Section 5.

## 2 Overview of GMDH and DT

### 2.1 Group Method of Data Handling

GMDH, a self-organizing network, was introduced by Ivakhnenko in 1966 [22] as an inductive learning algorithm for complex systems. Its main advantage is that it accurately estimates the parameters from its original structure. GMDH builds hierarchical solutions by trying for many easy models and by retaining the best and constructing on them to get the composition function.

The polynomial nodes are generally in quadratic form  $z = w_0 + w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_2^2 + w_5x_1x_2$  for the inputs  $x_1$  and  $x_2$ , weight vector  $w$  and output node  $z$ . The weights are found by solving the linear regressing equation  $z = y$ , the response vector.

The learning process of GMDH is as follows:

The GMDH develops on dataset with independent variables  $X_1, X_2, \dots, X_n$  and dependent variable  $y$  before learning processing is initiated. First the dataset is split into training and test set. During the learning process GMDH is developed as follows:

1. Input layer, as usual, consists of independent variables.
2. When constructing the hidden layer, initial population of units is created. Each unit is in the form of Ivakhnenko polynomial form:  

$$y = a + bx_1 + cx_2 + dx_1^2 + ex_1x_2 + fx_2^2 \text{ or } y = a + bx_1 + cx_2 + dx_1x_2,$$
where  $y$  is the dependent variable,  $x_1, x_2$  are independent variables and  $a, b, c, d, e, f$  are parameters. Parameters are estimated using training set.
3. The mean squared error (MSE) is calculated for the test set.
4. Units are sorted using the MSE values. The units with good approximation are taken to the next construction layers and the remaining are deleted.
5. Next layers are constructed while the MSE value decreases for the best units.
6. The response of the unit having minimum MSE value is considered as the output of GMDH.

In this network, the important input variables, number of hidden layers, neurons in each hidden layer are determined automatically. Majority of GMDH network implementations use regression analysis for solving the problem. The first step is to decide the type of polynomial that regression should find. General connection between input and output variables can be expressed by Volterra functional series, discrete analog of which is Kolmogorov–Gabor polynomial. The next step is to construct a linear combination of all of the polynomial terms with variable coefficients. The algorithm determines values of these coefficients by minimizing the squared sum (over all samples) of differences between sample outputs and model predictions.

## 2.2 Decision Tree

Decision Tree introduced by the Quinlan [23] in 1992. DT is too well known to be described in detail here. One can refer to [23] for further details.

## 3 Churn Dataset and Its Preprocessing

The dataset is taken from a Latin American bank that suffered from an increasing number of chucks with respect to their credit card customers and decided to improve its retention system. Table 1 presents the description of variables in the dataset.

The dataset consists of 21 independent variables and 1 dependent variable. It consists of 14814 records, of which 13,812 are loyal customers and 1,002 are churners, which means there are 93.24% loyal customers and mere 6.76% churners. It clearly shows that the dataset is highly unbalanced [24]. Consequently, we adopted the following methodology. First, we divided it into 80% and 20% using stratified random sampling to be used as training set and validation set respectively. During training, we performed the 10-fold cross-validation (10-FCV) on the training set. Further, we used the validation set to evaluate the effectiveness of the GMDH and the hybrid. This kind of splitting the data ensures that data leakage does not happen during training and validation [25]. Before analyzing the dataset, the training set is artificially balanced so that the classifier (GMDH or DT or hybrid) will not be adversely influenced by the majority class i.e., loyal customers. The balancing is done using synthetic minority over-sampling techniques (SMOTE) [26]. SMOTE is a technique where the minority class samples are over-sampled by taking each sample and introducing the synthetic samples along the line segment of the minority class nearest neighbors. After SMOTE is done, the training set is used for 10-fold cross-validation. One important thing to be noted here is that the SMOTE is applied on the training data only and not validation data. It is because a classifier gets trained on the conditioned (after SMOTE) dataset, and gets ready to be used in real life situation, which is presented in the validation set in the form of imbalance.

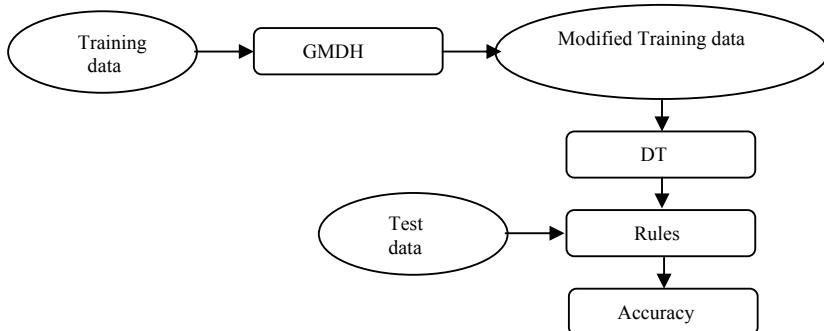
**Table 1.** Feature description of churn prediction dataset

Feature	Description	Value
<i>Target</i>	Target Variable	0-NonChurner 1-Churner
CRED_T	Credit in month T	Positive real number
CRED_T-1	Credit in month T-1	Positive real number
CRED_T-2	Credit in month T-2	Positive real number
NCC_T	Number of credit cards in months T	Positive integer value
NCC_T-1	Number of credit cards in months T-1	Positive integer value
NCC_T-2	Number of credit cards in months T-2	Positive integer value
INCOME	Customer's Income	Positive real number
N_EDUC	Customer's educational level	1 - University student 2 - Medium degree 3 - Technical degree 4 - University degree
AGE	Customer's age	Positive integer
SX	Customers sex	1 - male 0 - Female
E_CIV	Civilian status	1-Single 2-Married 3-Widow 4-Divorced
T_WEB_T	Number of web transaction in months T	Positive integer
T_WEB_T-1	Number of web transaction in months T-1	Positive integer
T_WEB_T-2	Number of web transaction in months T-2	Positive integer
MAR_T	Customer's margin for the company in months T	Real Number
MAR_T-1	Customer's margin for the company in months T-1	Real Number
MAR_T-2	Customer's margin for the company in months T-2	Real Number
MAR_T-3	Customer's margin for the company in months T-3	Real Number
MAR_T-4	Customer's margin for the company in months T-4	Real Number
MAR_T-5	Customer's margin for the company in months T-5	Real Number
MAR_T-6	Customer's margin for the company in months T-6	Real Number

## 4 Proposed Approach

The proposed hybrid consists of two phases. The block diagram of the hybrid is depicted in Fig 1. In phase 1, GMDH was trained using the training set. In phase 2, the predicted label of the output variable obtained from GMDH along with the corresponding independent variables are fed as new training set to the DT to generate rules. Since DT uses the predictions of the class label yielded by GMDH, the hybrid ensures that the rules generated eventually by DT represent the knowledge learnt by the GMDH network during training. The effectiveness of the rules thus generated were tested on the test data and also on validation set of the given classification problem.

In the proposed hybrid, we trained the GMDH using SMOTE data. Once the network is trained well, it is tested on the test data and also on the validation set. After getting high accuracies in the phase 1, we invoke phase 2. The accuracy of GMDH and the hybrid is tested against the validation set. The difference between the DT and the proposed hybrid is that with DT we can extract rules directly from dataset. But with proposed approach, we extract rules from GMDH using DT. Rules extracted using the hybrid can be used as an early warning system. This feature successfully removes the black box stigma of the GMDH.



**Fig. 1.** Flow diagram of the proposed hybrid approach

The quantities employed to measure the quality of the GMDH, DT and the proposed hybrid are sensitivity, specificity and accuracy, which are defined as follows [27]:

Sensitivity is the measure of proportion of the true positives, which are correctly identified.

$$\text{Sensitivity} = \frac{\text{true positive}}{(\text{true positive} + \text{false negative})}$$

Specificity is the measure of proportion of the true negatives, which are correctly identified.

$$\text{Specificity} = \frac{\text{true negative}}{(\text{true negative} + \text{false positive})}$$

Accuracy is the measure of proportion of true positives and true negatives, which are correctly identified.

$$\text{Accuracy} = \frac{\text{true positive} + \text{true negative}}{(\text{true positive} + \text{true negative} + \text{false positive} + \text{false negative})}$$

Fidelity of rules signifies how well they mimic the behavior of the neural network from which they are extracted.

## 5 Results and Discussion

To illustrate the effectiveness of the proposed hybrid we worked with four benchmark datasets viz. Iris, Wine, US Congressional, New Thyroid datasets, taken from UCI repository (<http://archive.ics.uci.edu/ml/datasets.html>) and a medium scale data mining problem namely churn prediction. The churn prediction dataset used for the experiments is the same as the one used in [2, 19, 28]. We employed NeuroShell 2.0 [29] and KNIME 2.0 [30] for conducting the experiments in the study.

The average results of the 10-FCV on the benchmark datasets and t-test values at 1% level of significance are presented in Table 2. In the case of Iris dataset, the proposed hybrid yielded accuracy of 93.33% on the validation set which is higher compared to that of the GMDH and DT alone. For Wine dataset, the hybrid yields accuracy of 98.00% which is very much close to that of the GMDH. In case of New Thyroid dataset hybrid yielded accuracy of 88.80%, which is less compared to that of the GMDH and DT standalone techniques on validation set. In the case of US Congressional dataset the GMDH and the hybrid yielded accuracy of 73.33% on the validation set. t-Test performed indicates that on Wine and US Congressional datasets, the difference between the hybrid and GMDH is statistically insignificant. However, since our main objective is to represent the knowledge learnt by the GMDH in the form *if-then* rules, we suggest the hybrid be used. But, in the case of Iris and New Thyroid, t-test indicates that hybrid is better than GMDH and in New Thyroid t-test indicates that GMDH is better than hybrid. Hence, our proposed approach is advantageous because we achieved transparency without sacrificing too much of accuracy. On the other hand, Fujimoto and Nakabayashi [15] also used the US congressional dataset and extracted rules from GMDH using a different method whose results cannot be compared with ours because their experimental setup in the form of dataset splitting was different and also they extracted rules directly from the GMDH.

Fidelity is also computed for all the benchmark datasets. Fidelity of the rules on validation sets of all the dataset is 91%, 96.85%, 100% and 59.53% for Iris, Wine, Us Congressional and New Thyroid datasets respectively.

As regards churn prediction dataset, we note that many business decision makers place more emphasis on the sensitivity rather than the specificity because higher sensitivity leads to more accurately identifying the churners, thereby achieving the chief objective of CRM viz., retaining the existing loyal customers. Similar kind of arguments apply to some real-world problems like fraud detection in bank credit cards and telecom services, bankruptcy prediction, cancer detection in humans based on their genetic profiles etc. Table 3 presents the average results of the 10-FCV performed on

GMDH, the proposed hybrid and the DT and t-test values at 1% level of significance. The GMDH yielded sensitivity 80.89%. The hybrid generated rules with accuracy of 81.55%. Even though the accuracy of the hybrid is reduced by 0.72% compared to the GMDH t-Test performed on sensitivity indicates the difference between GMDH and the hybrid is statistically insignificant. Further, the hybrid yielded rules which can be easily interpreted and can be used as an early warning system. DT yielded sensitivity 81.65% and accuracy 92.70% on the validation set. Even though the DT yielded higher accuracy, our main objective is to extract rules from the GMDH neural network. The rules obtained in the best fold are presented in Table 4. Strictly speaking, these results cannot be compared with that of Farquad et al., [19] because of different dataset experimental settings in the form of dataset splitting and also the fact that they dealt with unbalanced dataset. But, the fidelity on the validation set by our proposed approach got 96.86%, where as they achieved 87.58%. From the results it is observed that, our proposed hybrid performed very well.

Table 5 presents the rule base size of the DT and the proposed hybrid for the experiments conducted corresponding to the rules obtained in the best fold on the validation set. In the case of Churn Prediction dataset, the hybrid yielded 11 rules whereas the DT yielded 18 rules. From this it is clearly shown that the hybrid

**Table 2.** Average accuracies of benchmark datasets

Dataset	Techniques	Test under 10-FCV Accuracies	Validation set Accuracies	t-test value (GMDH vs. Hybrid)
IRIS	GMDH	96.66	91	3.456551
	Hybrid	95.00	<b>93.33</b>	
	DT	95.00	92.66	
WINE	GMDH	93.57	<b>98.85</b>	0.848483
	Hybrid	88.36	98.00	
	DT	89.53	97.42	
US CONGRESSIONAL	GMDH	73.27	73.33	0
	Hybrid	73.27	73.33	
	DT	97.28	<b>95.55</b>	
NEW THYROID	GMDH	100	<b>95.81</b>	4.871118
	Hybrid	91.32	88.80	
	DT	93.09	91.39	

**Table 3.** Average accuracies of churn prediction dataset

Dataset	Technique	Test under 10-FCV			Validation set			t-test value (GMDH vs. Hybrid)
		Sens*	Spec*	Acc*	Sens*	Spec*	Acc*	
CHURN PRED-ICTON	GMDH	86.06	88.12	87.09	80.89	88.74	88.21	1.709102
	Hybrid	84.96	87.78	86.37	81.55	87.92	87.49	
	DT	89.21	93.42	91.31	81.65	93.50	<b>92.70</b>	

Note: Sens\*=Sensitivity; Spec\*=Specificity; Acc\*=Accuracy;

approach reduced the number of rule base size because GMDH is learning the knowledge directly from the dataset and then the modified dataset that indicates the actual independent variables and the predicted outputs are presented to the DT to generate rules. Hence, with the modified dataset the rules generated by the DT are minimum or equal when compared to the direct applications of DT on the original dataset. In the case of Iris dataset also the number of rules is less for the hybrid compared to that of DT. In other datasets the number of rules for both the hybrid and DT are equal.

**Table 4.** Best fold rules of churn prediction dataset

- |   |
|---|
| <ol style="list-style-type: none"> <li>1. If <math>NCC\_T \leq 0.932959 \ \&amp; \ T\_WEB\_T \leq 5.977072 \ \&amp; \ CRED\_T \leq 630.21 \ \&amp; \ NCC\_T \leq 0.412727</math> then class <i>Churner</i>.</li> <li>2. If <math>NCC\_T \leq 0.932959 \ \&amp; \ T\_WEB\_T \leq 5.977072 \ \&amp; \ CRED\_T \leq 630.21 \ \&amp; \ NCC\_T &gt; 0.412727 \ \&amp; \ SX \leq 0.966218 \ \&amp; \ SX \leq 0.099431</math> then <i>Loyal customer</i>.</li> <li>3. If <math>NCC\_T \leq 0.932959 \ \&amp; \ T\_WEB\_T \leq 5.977072 \ \&amp; \ CRED\_T \leq 630.21 \ \&amp; \ NCC\_T &gt; 0.412727 \ \&amp; \ SX \leq 0.966218 \ \&amp; \ SX &gt; 0.099431</math> then <i>Churner</i>.</li> <li>4. If <math>NCC\_T \leq 0.932959 \ \&amp; \ T\_WEB\_T \leq 5.977072 \ \&amp; \ CRED\_T \leq 630.21 * NCC\_T &gt; 0.412727 \ \&amp; \ SX \leq 0.966218 \ \&amp; \ SX &gt; 0.966218</math> then <i>Loyal customer</i>.</li> <li>5. If <math>NCC\_T \leq 0.932959 \ \&amp; \ T\_WEB\_T \leq 5.977072 \ \&amp; \ CRED\_T &gt; 630.21</math> then <i>Loyal Customer</i>.</li> <li>6. If <math>NCC\_T \leq 0.932959 \ \&amp; \ T\_WEB\_T &gt; 5.977072 \ \&amp; \ CRED\_T \leq 592.937096 \ \&amp; \ NCC\_T - 2 \leq 0.499881</math> then <i>Loyal Customer</i>.</li> <li>7. If <math>NCC\_T \leq 0.932959 \ \&amp; \ T\_WEB\_T &gt; 5.977072 \ \&amp; \ CRED\_T \leq 592.937096 \ \&amp; \ NCC\_T - 2 &gt; 0.499881</math> then <i>Churner</i>.</li> <li>8. If <math>NCC\_T \leq 0.932959 \ \&amp; \ T\_WEB\_T &gt; 5.977072 \ \&amp; \ CRED\_T &gt; 592.937096</math> then <i>Loyal Customer</i>.</li> <li>9. If <math>NCC\_T \leq 0.932959 \ \&amp; \ NCC\_T &gt; 0.932959 \ \&amp; \ SX \leq 0.899029 \ \&amp; \ SX \leq 0.093913</math> then <i>Loyal Customer</i>.</li> <li>10. If <math>NCC\_T \leq 0.932959 \ \&amp; \ NCC\_T &gt; 0.932959 \ \&amp; \ SX \leq 0.899029 \ \&amp; \ SX &gt; 0.093913</math> then <i>Churner</i></li> <li>11. If <math>NCC\_T \leq 0.932959 \ \&amp; \ NCC\_T &gt; 0.932959 \ \&amp; \ SX &gt; 0.899029</math> then <i>Loyal Customer</i>.</li> </ol> |
|---|

**Table 5.** Rule Base Size of the DT and Proposed Hybrid for the best fold

Dataset	Rule Base Size	
	<i>DT</i>	<i>Proposed Hybrid</i>
<i>Churn Prediction</i>	18	11
<i>IRIS</i>	4	3
<i>WINE</i>	5	5
<i>US Congressional</i>	2	2
<i>New Thyroid</i>	6	6

## 6 Conclusions

In this paper, we presented a hybrid method to extract rules from GMDH neural network using DT to solve data mining problem like churn prediction in bank credit cards and also the benchmark datasets viz. Iris, Wine, US Congressional and New Thyroid datasets. As regards churn prediction in credit cards dataset, since identifying churner is most important from business perspective, considering sensitivity alone the proposed hybrid performed well. Also, the proposed hybrid performed well in analyzing the four benchmark datasets. In the case of churn prediction, Wine and US Congressional datasets the hybrid and GMDH yielded statistically insignificant results. On IRIS and New Thyroid datasets they outscored each other with statistically significant accuracies. The rule base size of the proposed approach is less compared to that of the stand alone DT. Future research includes the feature selection, computation of confidence and support of individual rules and application of the hybrid method to solve regression problems.

## References

1. Ravi, V., Arul Shalom, S.A., Manickavel, A.: Sputter Process Variables Prediction via Data Mining. In: proceeding of the 2004 IEEE conference on cybernetics and intelligent systems, Singapore (2004)
2. Anilkumar, D., Ravi, V.: Predicting credit card customer churn in banks using data mining. International Journal for Data Analysis, Techniques and Strategies 1(1), 4–28 (2008)
3. Andrews, R., Diederich, J., Tickle, A.B.: A survey and critique of techniques for extracting rules from trained artificial neural networks. Knowledge Based Systems 8(6), 373–389 (1996)
4. Senator, T., Goldberg, H.G., Wooton, J., Cottini, M.A., Umarkhan, A.F., Klinger, C.D., Llamas, W.M., Marrone, M.P., Wong, R.W.H.: The financial crimes enforcement network AI Systems (FAIS): Identifying potential Money Laundering from reports of large cash transaction. AI Magazine 16(4), 21–39 (1995)
5. Chu, B.H., Tsai, M.-S., Ho, C.-S.: Toward a hybrid data mining model for customer retention. Knowledge-Based Systems 20(8), 703–718 (2007)
6. Gallant, S.: Connectionist expert systems. Communications of the ACM 31(2), 152–169 (1998)
7. Kahramanli, H., Allahverdi, N.: Extracting rules for classification problems: AIS based approach. Expert Systems with Application 36, 10494–10602 (2009)
8. Fu, L.M.: Rule generation from neural networks. IEEE Transaction on Systems, Man and Cybernetics 24(8), 1114–1124 (1994)
9. Towlell, G.G., Shavlik, J.W.: The extraction of refined rules from knowledge based neural networks. Machine Learning 13(1), 71–101 (1993)
10. Arbatli, A.D., Akin, H.L.: Rule extraction from trained neural networks using genetic algorithms. Nonlinear Analysis, Theory, Methods and Applications 30(3), 1639–1648 (1997)
11. Fan, Y., James Li, C.: Diagnostic rule extraction from trained feed forward neural network. Mechanical Systems and Signal Processing 16(6), 107–1081 (2002)
12. Krishnan, R., Sivakumar, G., Bhattacharya, P.: A Search Technique for rule extraction from trained neural networks. Pattern Recognition Letters 20, 273–280 (1999)

13. McGarry, K.H., Tait, J., Wermter, S., MacIntyre, J.: Rule extraction from radial basis function networks. In: International conference on Artificial Neural Networks, Edinburgh (1999)
14. Sato, M., Tsukimoto, H.: Rule extraction from neural networks via Decision Tree Induction, pp. 1870–1875. IEEE, Los Alamitos (2001)
15. Fujimoto, K., Nakabayashi, S.: Applying GMDH algorithm to extract rules from examples. Systems Analysis Modelling Simulation 43(10), 1311–1319 (2003)
16. Campos, P.G., Ludermir, T.B.: Literal and ProRulext: Algorithms for rule extraction of ANNs. In: Proceedings of the fifth international conference on Hybrid Intelligent Systems, HIS 2005 (2005)
17. Aliev, R.A., Aliev, R.R., Guirimov, B., Uyar, K.: Dynamic data mining technique for rules extraction in a process of battery charging. Applied Soft Computing 8, 125–1258 (2008)
18. Naveen, N., Ravi., V., Raghavendra Rao, C.: Rule extraction from differential evolution trained radial basis function network using genetic algorithms. In: Fifth Annual IEEE conference on Automation Science and Engineering, Bangalore, India, pp. 152–157 (2009)
19. Farquad, M.A.H., Ravi, V., Bapi, R.S.: Data mining using rules extracted from SVM: an application to churn prediction in banks credit cards. In: Sakai, H., Chakraborty, M.K., Hassanien, A.E., Ślęzak, D., Zhu, W. (eds.) RSFDGrC 2009. LNCS, vol. 5908. Springer, Heidelberg (2009)
20. Srinivasan, D.: Energy demand prediction using GMDH networks. Neurocomputing 72(1-3), 625–629 (2008)
21. Mohanty, R., Ravi, V., Patra, M.R.: Software Reliability Prediction Using Group Method of Data Handling. In: Sakai, H., Chakraborty, M.K., Hassanien, A.E., Ślęzak, D., Zhu, W. (eds.) RSFDGrC 2009. LNCS, vol. 5908, pp. 344–351. Springer, Heidelberg (2009)
22. Ivakhnenko, A.G.: The group method of data handling- a rival of the method of stochastic approximation. Soviet Automatic Control 13(3), 43–55 (1966)
23. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo (1992)
24. Business Intelligence Cup-2004: Organized by the University of Chile (2004), [http://www.tis.cl/bicup\\_04/text-bicup/BICUP/202004/20public/20data.zip](http://www.tis.cl/bicup_04/text-bicup/BICUP/202004/20public/20data.zip)
25. Ravi, V., Kurniawan, H., Thai, P.N.K., Ravikumar, P.: Soft computing system for bank performance prediction. Applied soft computing 8, 305–315 (2008)
26. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)
27. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874 (2006)
28. Naveen, N., Ravi, V., Anilkumar, D.: Application of fuzzyARTMAP for churn prediction in bank credit cards. International Journal of Information and Decision Sciences 1(4), 428–444 (2009)
29. NeuroShell 2.0, <http://www.wardsystems.com>
30. Knime 2.0, <http://www.knime.org/>

# Sensitivity Analysis and Automatic Calibration of a Rainfall-Runoff Model Using Multi-objectives

Fan Sun and Yang Liu

School of Management and Language &  
School of the Built Environment,  
Heriot-Watt University, Edinburgh, EH14 4AS, UK  
[{fs82,y.liu}@hw.ac.uk](mailto:{fs82,y.liu}@hw.ac.uk)

**Abstract.** The practical experience with sensitivity analysis suggests that no single-objective function is adequate to measure the ways in which the model fails to match the important characteristics of the observed data. In order to successfully measure parameter sensitivity of a numerical model, multiple criteria should be considered. Sensitivity analysis of a rainfall-runoff model is performed using the local sensitivity method (Morris method) and multiple objective analysis. Formulation of SA strategy for the MIKE/NAM rainfall-runoff model is outline. The SA is given as a set of Pareto ranks from a multi-objective viewpoint. The Nondominated Sorting Differential Evolution (NSDE) was used to calibrate the rainfall-runoff model. The method has been applied for calibration of a test catchment and compared on validation data. The simulations show that the NSDE method possesses the ability to finding the optimal Pareto front.

**Keywords:** Sensitivity Analysis, Optimisation, Multi-objective Optimisation, Calibration and Validation.

## 1 Introduction

Sensitivity analysis (SA) is the study of how the variation in the output of a numerical model can be apportioned, qualitatively or quantitatively, to different sources of variation in the input of a model [1]. It is normally used to analyze how sensitive a system is with respect to the change of parameters. It is particularly useful for complex hydrological model that involves a large number of model parameters. Sensitivity analysis methods can be classified into local SA and global SA. Local SA provides information on the effect of a small change in each input parameter individually and global SA describes the effect of simultaneous arbitrary variations of multiple parameters [2]. Sensitivity analysis of models is important in many real-world problems. In SA, there are generally a number of different objectives, which may be in conflict. The general number of different objectives, arises in many fields, and is often studied through multi-objective analysis. By definition, the multi-objective problem has a very different nature from that of single-objective problem. Unlike single-objective problem where only one solution is ranked as the most sensitive parameter, a typical multi-objective problem produces a set of solutions which are superior to the rest of the solutions with respect to all objective criteria but are inferior to other solutions in one or more objectives. These solutions are known as non-dominated solutions. In absence of

additional information, it is not possible to distinguish any one of the non-dominated solutions as being objectively better than any others with respect to all the objectives concerned (i.e. there is no uniquely “best” solution); therefore, any one of them is an acceptable solution [3, 4, 5, 6, 7, 8, 9]. Most multi-objective techniques attempt to identify a set of optimal solutions which represent the trade-off surface between conflicting criteria. Practical experience with model parameter sensitivity analysis suggests that no single objective function is adequate to measure properly the parameter of all the important characteristics of the system. Therefore, there is a need to consider multiple objectives for model parameter SA to be effective. This work is dedicated to an application of the Pareto ranking method and local SA (Morris method) to a real test case, a rainfall-runoff model here [10].

Calibration is the process of modifying the input parameters to a numerical model until the output from the model matches an observed set of data. In automatic calibration, parameters are adjusted automatically according to a specified search scheme and numerical measures of the goodness-of-fit [4]. Compared to manual calibration, automatic calibration is faster while being objective and relatively easy to implement. It was found that performance of model calibration depends on choice of the objective function considered. For a more detailed description about performance measures for comparing model predictions and observations using different objective functions the reader is referred to [7, 8]. Many real-world optimisation problems, especially numerical calibration situations, require the process of simultaneous optimisation of possibly conflicting multiple objectives, and this is termed multi-objective optimisation. Differential Evolution (DE) is a new population based algorithm proposed by Storn and Price [11] for optimisation problems over a continuous domain and it has been extended to solve discrete problems. The main operators that control the evolutionary process are the mutation and selection operators. The NSDE combines the advanced operations (fast ranking of non-dominated solutions, crowding distance ranking, elitist strategy of combining parent population and offspring population together, selection and mutation operations) with a DE. Simulations for the numerical model show that the NSDE method possesses the ability to finding the optimal Pareto front. The parameters of the conceptual rainfall-runoff models cannot, in general, be determined directly from physical catchment characteristics, and hence the parameter values must be estimated by calibration against observed data. In this study, a new multi-objective sensitivity analysis of the rainfall-runoff model is performed using the multiple objective Morris method.

## 2 Sensitivity Analysis Using Multi-objective Analysis

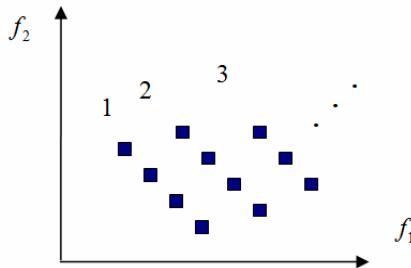
Sensitivity coefficients, which are the partial derivatives of the model states with respect to the model parameters, play an important role in parameter estimation, uncertainty analysis, and model reduction, etc, hydrological model. The sensitivity coefficient  $S_i$  is calculated from the difference of the nominal and perturbed solutions. The method is straightway and repeated the model is required  $M$  times. In Morris [10], two sensitivity measures were proposed for each factor:  $m$ , an estimate of the mean of the distribution  $S_i$ , and  $d$ , an estimate of the standard deviation of  $S_i$ . A high value of  $m$  indicates an input factor with an important overall influence on the output.

A high value of  $d$  indicates a factor involved in interaction with other factors or whose effect is nonlinear.

$$m = S_i = \frac{\partial F(\theta_i)}{\partial \theta_i} = \frac{1}{M} \sum_{j=1}^M S_i(j) = \frac{1}{M} \sum_{j=1}^M \left| \frac{F(\theta_i + \Delta\theta_i) - F(\theta_i)}{\Delta\theta_i} \right| \quad (1)$$

$$d_i = STD(S_i) \quad (2)$$

where  $i$  is the numerical model parameter index. There is a trade-off between the two objectives ( $m$  and  $d$ ) implying that traditional SA is inappropriate. The Pareto ranking method can help the designer visualize the trade-offs between different objectives and select an appropriate compromise design. The domination between two solutions can be defined as follows: A solution  $x_1$  is said to dominate the other solution  $x_2$ , if both the following conditions are true: (1) The solution  $x_1$  is no worse than  $x_2$  in all objectives, and (2) The solution  $x_1$  is strictly better than  $x_2$  in at least one objective. The special sorting is called “Pareto ranking” [12]. The procedure begins by identifying all the non-dominated individuals in the population. They are given the rank 1 and temporarily removed from the population. Then the non-dominated individuals identified in the remaining population are given the rank 2, and then they are also removed from the population. This procedure of identifying non-dominated sets of individuals is repeated until every individual has been assigned a rank, as depicted in Fig.1. This sorting procedure essentially assigns equivalent rankings to all points that lie on the same Pareto frontier.



**Fig. 1.** The relative value of the  $s$  metric upon an arbitrary choice of reference point

We call this new SA algorithm PR-SA. The PR-SA method is a basic SA strategy designed to be effective and efficient for a broad class of problem. The method is based on notion of Pareto ranking method and local SA (Morris Method).

### 3 Non-dominated Differential Evolution Algorithm

Differential Evolution (DE) is a population-based direct-search algorithm for global optimisation [11]. The standard DE works as follows: for each vector  $x_{i,G}$ ,  $i = 0, 1, 2, \dots, NP - 1$ , a trial vector  $v$  is generated according to

$$v = x_{r_1,G} + F \bullet (x_{r_2,G} - x_{r_3,G}) \quad (3)$$

With  $r_1, r_2, r_3 \in [0, NP-1]$ , integer and mutually different,  $F > 0$ , and  $r_1 \neq r_2 \neq r_3 \neq i$ .  $F$  is a real and constant factor which controls the amplification of the differential variation  $(x_{r_2,G} - x_{r_3,G})$ . In order to increase the diversity of the parameter vectors, the following vector is adopted:

$$u = (u_1, u_2, \dots, u_{NP-1})^T \quad (4)$$

With:

$$u_j = \begin{cases} v_j, & \text{if } x \in [0,1] \leq CR \text{ or } j = i; \\ x_{i,j,G}, & \text{otherwise.} \end{cases} \quad (5)$$

Where  $i, j = 0, 1, 2, \dots, NP-1$ ; and CR is a user-defined crossover rate.  $F$  and CR are both generally in the range [0.5, 1.0]. In order to decide whether the new vector  $u$  shall become a population member at generation  $G+1$ , it is compared to  $x_{i,G}$ . If vector  $u$  yields a smaller objective function value than  $x_{i,G}$ ,  $x_{i,G+1}$  is set to  $u$ , otherwise the old value  $x_{i,G}$  is retained. The multi-objective hybrid algorithm was to combine single-objective DE with NSGA-II operations without loosing performance on establishing the Pareto-front [13]. The hybrid algorithm is presented below:

**Step 1:** Generate an initial population  $P$  in a feasible space.

**Step 2:** Sort the population based on the non-domination and crowding distance ranking.

**Step 3:** Assign each individual a fitness (or rank) equal to its non-domination level (minimisation of fitness is assumed).

**Step 4:** Generate a trail vector for each individual and do crossover operation to generate the offspring.

**Step 5:** Combine the offspring and parent population to form extended population of size  $2N$ .

**Step 6:** Sort the extended population based on non-domination and fill the new population of size  $N$  with individuals from the sorting fronts starting to the best.

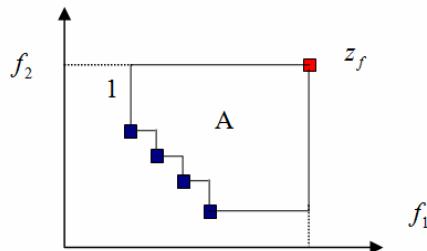
**Step 7:** Perform step (2) to (6) until the stopping criterion is met.

For a more detailed description about the hybrid optimisation the reader is referred to [12].

## 4 Performance Metrics

A number of performance metrics have been suggested in the past [14, 15]. Two goals are usually considered for a multi-objective optimisation. First, the population should converge towards the optimal Pareto front. Second, the optimal Pareto front should

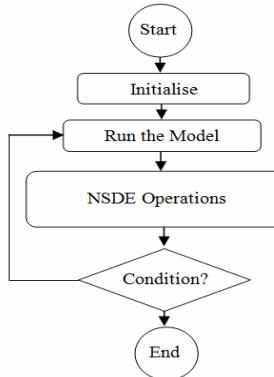
have maximum spread of solutions. Based on this notion, we adopted S metric to evaluate each of two aspects. A definition of the S metric is given in Fig. 2. The S metric calculates the hypervolume of the multi-dimensional region enclosed by A and a ‘reference point’ (see Fig. 2), hence computing the size of region A dominates. It is independent (although needs a reference point to be chosen), so it induces a complete ordering, and it is non-cardinal. Fig. 2 illustrates the Procedure of S metric calculation for an optimal Pareto front where two objective ( $f_1, f_2$ ) are to be minimized.



**Fig. 2.** The relative value of the s metric upon an arbitrary choice of reference point

## 5 Automatic Calibration Process

The general flow chart for the calibration process using NSDE for the numerical model is presented in Fig. 3. The calibration process can be performed via an automatic process. In order to do so, the user may need to write two small programs for the process. The fist program is used to change the parameters of input files of the numerical model and the second program is used to extract data and to calculate error between observed and simulated data. As the standard search progresses, the entire population tends to converge to the global Pareto front. This process is continued until a satisfied condition is met. The termination criterion for the iterations is determined according to whether the max iteration or a designed value of the fitness is reached.



**Fig. 3.** Outline of NSDE for the NAM model calibration

## 6 Application to a Rainfall-Runoff Model

The model used in this study is the NAM rainfall-runoff model that forms the rainfall-runoff module of the MIKE11 river modelling system [4]. The structure of the model is shown in Fig. 4. A brief description of calibration parameters used is given in Table 1. In order to test the validity of the proposed methodologies, the MIKE11/NAM model was applied to the Danish Tryggevælde catchment. This catchment has an area of 130km<sup>2</sup>, an average rainfall of 710 mm/year and an average discharge of 240 mm/year. The catchment is dominated by clayey soils, implying a relatively flashy flow regime. For the calibration, a 5-year period (1 Jan. 1984–31 Dec. 1988) was used where daily data of precipitation, potential evapotranspiration, mean temperature, and catchment runoff are available. The first 3 months of the calibration period were discarded in the calculation of the objective functions in order to minimise the influence from the initial conditions. In order to obtain a successful calibration by using automatic optimisation routines, it is necessary to formulate the calibration objectives. Two objective functions (each corresponding to one of the goodness-of-fit criteria) are formulated as follows [1]:

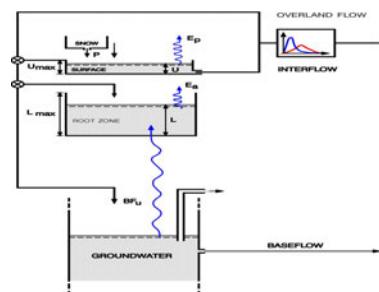
1. Average Root Mean Squared-Error (RMSE) of peak flow Events:

$$F_1(\theta) = \frac{1}{M_p} \sum_{j=1}^{M_p} \left[ \frac{1}{n_j} \sum_{i=1}^{n_j} [Q_{obs,i} - Q_{sim,i}(\theta)]^2 \right]^{1/2} \quad (6)$$

2. Average Root Mean Squared-Error (RMSE) of low flow Events:

$$F_2(\theta) = \frac{1}{M_l} \sum_{j=1}^{M_l} \left[ \frac{1}{n_j} \sum_{i=1}^{n_j} [Q_{obs,i} - Q_{sim,i}(\theta)]^2 \right]^{1/2} \quad (7)$$

In Eqs. (6)–(7),  $Q_{obs,i}$  is the observed discharge at time  $i$ ,  $Q_{sim,i}$  is the simulated discharge,  $M_p$  is the number of peak flow events,  $M_l$  is the number of low flow events,  $n_j$  is the number of time steps in peak/low event no.  $j$ , and  $\theta$  is the set of model parameters to be calibrated. Peak flow events were defined as periods with flow above a threshold value of 4.0 m<sup>3</sup>/s, and low flow events were defined as periods with flow below 0.5 m<sup>3</sup>/s.



**Fig. 4.** NAM model structure

**Table 1.** NAM model parameters

Parameter	Description	Lower Limit	Upper Limit
Umax(mm)	Maximum water content in the surface storage	5	35
Lmax(mm)	Maximum water content in the lower zone storage	50	350
CQOF	Overland flow runoff coefficient	0	1
CKIF(hour)	Time constant for interflow from the surface storage	500	1000
TIF	Threshold value for interflow	0	0.9
TOF	Threshold value for overland flow	0	0.9
TG	Threshold value for recharge	0	0.9
CK12(hour)	Time constant for overland flow and interflow routing	3	72
CKBF(hour)	Baseflow time constant	500	5000

## 7 Experimental Setup and Results

The relevant experiment parameters using the NSDE for the NAM model are listed in Table 2. For each test, a number of iterations equal to 60 were employed as a stopping criterion for NSDE when a population of  $P=100$  was used. Fig. 5 (a) shows the graphical result produced by the NSDE with respect to the S metric in Table 3. Fig. 5 (b), (c) and (d) show the simulated discharges versus observed data using the middle optimal parameter set from the optimal Pareto front obtained by NSDE for the calibration period and validation periods 1 and 2 respectively (see Table 4). From these figures, there is a good match between simulated and observed values. A value of 1% of the parameter values was used for the increment to calculate the sensitivity coefficient. Table 5 shows the rank of importance for the 9 parameters according to the Morris measures using the two objective functions ( $m$  and  $d$ ), and the Pareto rank of importance for the 9 parameters based on non-domination ranking. The number of objective functions is equal to 4 using the multi-objective SA for the study case.  $S_i^*$  and  $S_i$  are the first order derivative values of the two evaluation functions (average low flow RMSE and average peak flow RMSE) with the  $i$ th input parameter using finite difference method. The small rank indicates the parameter is less sensitive. The parameters CQOF, TIF, TOF and TG are the most sensitive parameters based on Pareto ranking method in Table 5.

**Table 2.** Experimental parameters using NSDE

Parameter	Description	Range
$F$	Control Parameter	0.5
$CR$	Crossover Rate	1
$G$	The total iterations	60
$P$	The number of particles	100

**Table 3.** Results of the s metric for NAM model using the two optimisation algorithms

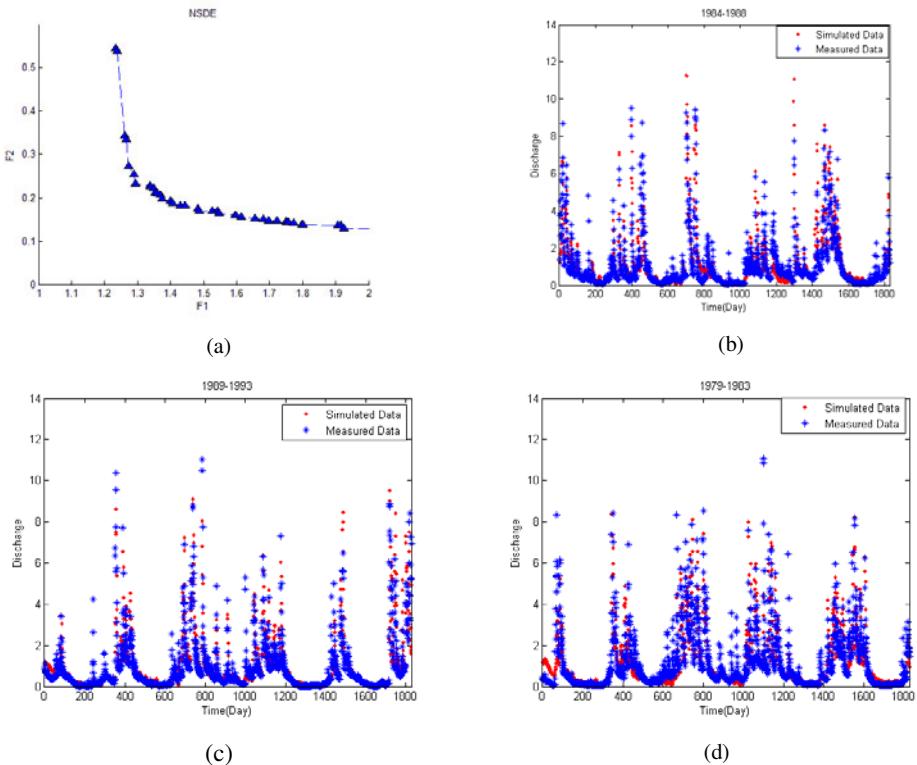
Trail ( S metric) Reference point (2, 0.6)	NSDE
1	0.3394
2	0.3263
3	0.3282
4	0.3378
5	0.3259
6	0.3203
7	0.3316
8	0.3394
9	0.3394
10	0.3263
Mean	0.3315
STD	0.0071

**Table 4.** The optimal Pareto front and function values

Umax	14.37
Lmax	242
CQOF	0.607
CKIF	892
TIF	0.311
TOF	0.9
TG	0.891
CK12	32.6
CKBF	1724
Peak flow RMSE	1.5263
Low flow RMSE	0.1693

**Table 5.** Sensitivity analysis using multi-objective Morris method

Model parameter (M=30)	$m$ $of S_i(j)$	$d of$ $S_i(j)$	$m$ $of S_i^*(j)$	$d of$ $S_i^*(j)$	Pareto rank
Umax	0.0018	0.0017	0.0232	0.0133	3
Lmax	0.0025	0.0031	0.0028	0.0018	3
CQOF	1.1708	0.7502	0.7654	0.4926	5
CKIF	0.00012	0.00017	0.00014	0.00011	1
TIF	0.6914	0.2620	0.7695	0.6028	5
TOF	0.8529	1.8397	0.7650	1.8010	5
TG	1.1746	0.6847	0.9672	0.3837	5
CK12	0.0265	0.0214	0.0159	0.0192	4
CKBF	0.0004	0.00054	0.00019	0.0002	2



**Fig. 5 (a).** Pareto front produced by the NSDE; **Fig.5 (b), (c) and (d).** Range of simulated hydrographs corresponding to the optimal parameter set shown in table 5, compared with observations.

## 8 Conclusions

The experiments showed that by using NSDE, it achieved a sufficiently accurate Pareto set and a good diversity in the obtained front. A new sensitivity analysis scheme for the NAM/MIKE 11 rainfall-runoff model has been proposed that considers the SA problem in a general multi-objective framework. The SA scheme considers numerical performance measures of four different objectives: (1) peak flows, (2) low flows, (3)  $m$ , an estimate of the mean of the distribution  $S_i$ , and (4)  $d$ , an estimate of the standard deviation of  $S_i$ . The application example demonstrated that significant trade-offs between different objectives exists, implying that single objective function is not able to evaluate all objectives simultaneously. Instead, the SA is given a set of Pareto ranks for the model parameters. The results clearly showed that the traditional concept of the SA is inappropriate. The application example revealed that the model may produce virtually equally sensitivity parameters based on non-domination ranking.

## Acknowledgement

The authors would like to thank ABP Marine Research Ltd., UK for funding the work.

## References

- [1] Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S.: Global Sensitivity Analysis. The Primer. John Wiley & Sons, Chichester (2008)
- [2] Yue, H., Brown, M., Knowles, J., Wang, H., Broomhead, D.S., Kell, D.B.: Insights into the behaviour of systems biology models from dynamic sensitivity and identifiability analysis: a case study of an NF- $\kappa$ B signaling pathway. *Molecular Biosystems* 2(12), 640–649 (2006)
- [3] Yapo, P.O., Gupta, H.V., Sorooshian, S.: Multi-objective global optimisation for hydrologic models. *Journal of Hydrology* 204, 83–97 (1998)
- [4] Madsen, H.: Automatic calibration of a conceptual rainfall-runoff model using multiple objectives. *Journal of Hydrology* 235, 276–288 (2000)
- [5] Cooper, V.A., Nguyen, V.T.V., Nicell, J.A.: Evaluation of global optimisation methods for conceptual rainfall-runoff model calibration. *Water Science and Technology* 36(5), 53–60 (1997)
- [6] Liu, Y., Khu, S.T., Savic, D.A.: A fast hybrid optimisation method of multi-objective genetic algorithm and k-nearest neighbour classifier for hydrological model calibration. In: Yang, Z.R., Yin, H., Everson, R.M. (eds.) IDEAL 2004. LNCS, vol. 3177, pp. 546–551. Springer, Heidelberg (2004)
- [7] Sorooshian, S., Gupta, H.V.: Calibration of hydrological models using multi-objectives and visualization techniques. Final Report (EAR-9418147), Department of hydrology and water resources, the University of Arizona, Tucson, AZ (1998)
- [8] Janssen, P.H.M., Heuberger, P.S.C.: Calibration of process-oriented models. *Ecological Modelling* 83, 55–66 (1995)
- [9] Liu, Y.: Automatic Calibration of a Rainfall-Runoff Model Using a Fast and Elitist Multi-objective Particle Swarm Algorithm. *Expert Systems with Applications* 36(5), 9533–9538 (2009)
- [10] Morris, M.D.: Factorial Sampling Plans for Preliminary Computational Experiments. *Technometrics* 33, 161–174 (1991)
- [11] Storn, R., Price, K.: Differential evolution: a simple and efficient adaptive scheme for global optimisation over continuous spaces. Technical report tr-95-012, international computer science institute, Berkley (1995)
- [12] Goldberg, D.E.: Genetic Algorithms in Search, Optimisation, and Machine Learning. Addison-Wesley Publishing Co., Reading (1989)
- [13] Iorio, A., Li, X.: Solving Rotated Multi-objective Optimization Problems Using Differential Evolution. In: Webb, G.I., Yu, X. (eds.) AI 2004. LNCS (LNAI), vol. 3339, pp. 861–872. Springer, Heidelberg (2004)
- [14] Deb, K., Agrawal, S., Pratap, A., Meyarivan, T.: A fast elitist non-dominated sorting genetic algorithm for multi-objective optimisation: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6(2), 182–197 (2002)
- [15] Knowles, J., Corne, D.: On metrics for comparing non-dominated Sets. In: Congress on Evolutionary Computation, pp. 711–716 (2002)

# University Course Timetabling Using ACO: A Case Study on Laboratory Exercises

Vatroslav Dino Matijaš, Goran Molnar, Marko Čupić,  
Domagoj Jakobović, and Bojana Dalbelo Bašić

Faculty of Electrical Engineering and Computing,  
Unska 3, 10000 Zagreb, Croatia  
University of Zagreb

[dinomatijas@gmail.com](mailto:dinomatijas@gmail.com), {[goran.molnar2](mailto:goran.molnar2@fer.hr),[marko.cupic](mailto:marko.cupic@fer.hr),  
[domagoj.jakobovic](mailto:domagoj.jakobovic@fer.hr),[bojana.dalbelo-basic](mailto:bojana.dalbelo-basic@fer.hr)}@fer.hr

**Abstract.** The ant colony optimisation metaheuristic has shown promise on simplified artificial instances of university course timetabling problems. However, limited work has been done applying it to practical timetabling problems. In this paper, we describe the application of the ant colony optimisation to a highly constrained real-world instance of the university course timetabling problem. We present the design of the memory-efficient construction graph and a sophisticated solution construction procedure. The system devised here has been successfully used for timetabling at the authors' institution.

## 1 Introduction

Almost every type of human organisations is occasionally faced with some form of timetabling tasks. The University Course Timetabling Problem (UCTP) and its variations are parts of the larger class of timetabling and scheduling problems. A large number of university timetabling problems have been described in the literature, and they differ from each other based on the type of institution involved, the entities being scheduled and the constraints in the definition of the problem.

Due to inherent complexity and variability of the problem, most real-world timetabling problems are NP-complete [1]. This calls for use of heuristic algorithms which do not guarantee an optimal solution but can usually generate solutions that are good enough for practical use. Due to their manageability and good performance (if properly implemented), metaheuristic techniques have shown to be particularly suitable for solving these kinds of problems.

In this work [2], we focus on the *Laboratory Exercises Timetabling Problem* (LETP), framed as an example of the university course timetabling problem. The work described is a part of the research on two different metaheuristics for

<sup>1</sup> This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia and under the grants No 036-1300646-1986 and 036-0362980-1921.

timetable construction: *genetic algorithm* [2] and *ant colony optimisation metaheuristic*. Constructing a system for solving instances of LETP is challenging both technically and administratively. Along with the efficient timetable production system, a feature-rich model of the timetable constraints is devised. It is necessary for the system to be able to account for a complex set of constraints because various departments at our institution have very different ideas about what a good timetable should look like. The definition of LETP is devised in close coordination with the departments and is constantly evolving. Therefore, a good modular design that supports changes in the problem definition is necessary.

Our approach is based on the Ant Colony Optimisation metaheuristic (ACO), proposed by Dorigo et al. [3]. It is a distributed stochastic probabilistic constructive search procedure, inspired by the foraging behaviour of ants. It utilises a problem representation in the form of a *construction graph*, on which suggested solutions are constructed by artificial ants. ACO has already shown promise for various timetabling applications. It has been successfully applied to various artificially generated UCTP datasets in [45] and in [6] its performance is compared with that of several other metaheuristics. ACO has also been applied to artificially generated instances of the examination timetabling problem, for example in [7]. However, as noted in [8],

*“A major identified weakness in the current approach to Operational Research is described as follows: a gap still remains between the output of a successful research project and what is needed for direct use by industry. In general, the area of educational timetabling is one such area.”*

Solving real such problems is a difficult task that involves modelling and handling various kinds of constraints which are usually simplified in academic problems. Moreover, only a few, relatively basic UCTP representations have been proposed in literature for use with ACO. It has been shown that these representations do not satisfy the requirements imposed by complex, large LETP instances.

This work is an effort to bridge this gap between the theory and practice of automated timetabling and to build an ACO-based system that satisfies the timetabling needs of our institution. The result of our work is an ACO-based timetabling system suitable for practical use.

The remainder of this paper is organised as follows: Section 2 introduces the actual timetabling problem, and in Section 3 we elaborate our approach. Section 4 presents the results, and Section 5 concludes the paper.

## 2 University Course Timetabling Problem

### 2.1 Problem Statement

The timetable construction problem is a combinatorial optimisation problem that consists of four finite sets: (i) a set of meetings, (ii) a set of available resources (e.g. rooms, staff, students), (iii) a set of available time-slots, and (iv) a set of constraints. The problem is to assign resources and time slots to each

given meeting, while maintaining constraints satisfied to the highest possible extent. The University Course Timetabling Problem (UCTP) is a timetabling problem where the given data consists of a set of students and sets of courses that each of the students needs to attend. A *course* is a set of events that need to take place in the timetable. The main characteristic that distinguishes the university course timetabling problem from other types of timetabling problems is the fact that students are generally allowed to choose the courses in which they wish to enrol [9]. A set of constraints is usually divided into *hard constraints* whose violation makes the timetable suggestion infeasible, and *soft constraints*, rules that improve the quality of timetables, but are allowed to be violated.

The above description of the UCTP defines a broad range of problems, whose complexity significantly depends on the specific constraints defined. Particular timetabling applications are usually focused on a more strictly defined subset of the problems, as the constraints and dimensions of the problem vary among institutions. We use the same approach, giving a detailed formal description of the problems for which our application is designed.

## 2.2 Definition of Laboratory Exercise Timetabling Problem

The *laboratory exercise timetabling problem* is defined as a six-tuple:

$$LETP = (T, L, R, E, S, C) ,$$

where  $T$  is a set of *time-quanta* in which the scheduling is possible,  $L$  is a set of *limited assets* present at the university,  $R$  is a set of *rooms*,  $E$  is a set of *events* that need to be scheduled,  $S$  is a set of attending *students*, and  $C$  is a set of *constraints*. We assume that the durations of all the events can be quantified as multiples of a fixed value of time that we call a *time-quantum*. A *time-slot* is defined as one or more consecutive time-quanta in the timetable. The duration of the quantum reflects a trade-off between the precision of scheduling and the size of the search space.

The set of limited assets (resources) shared among the different exercises is denoted  $L$ . For each resource  $l \in L$ , a fixed number of workplaces can use the resource concurrently.

Each room is defined as a set of *workplaces*, atomic room resources varying from room to room, such as seats in ordinary classrooms, computers in computer classrooms, etc. For each room  $r \in R$ , the number of workplaces, denoted  $\text{size}_r \in \mathbb{N}$  is defined. For each of the events, the desired number of students per workplace is defined. Since some rooms may not be available all the time, a set of time quanta  $T_r \subseteq T$  in which the room is available is defined for each room.

Events have the following set of properties:

- Each event  $e$  has a duration, denoted  $\text{dur}_e \in \mathbb{N}$ , a multiple of a time quantum.
- Each event  $e$  has an acceptable room set  $R_e \subseteq R$ .
- Each event  $e$  has a suitable time quanta set, denoted  $T_e \subseteq T$ .
- The set of limited assets used by the event is denoted  $L_e \subseteq L$ .

- The number of staff available for each event and the number of staff needed for event  $e$  when held in the room  $r$  are given. Because each of the departments at our institution produces staff timetables independently of our system, staff is not defined as a separate entity of the LETP.
- An ordering relation, denoted  $\succ_d$  can be defined for a pair of events. The relation  $e_2 \succ_d e_1$  is true if and only if  $e_2$  needs to be scheduled at least  $d$  days after  $e_1$ .
- The maximum number of rooms to be used concurrently for the event  $e$  can be defined.
- An event timespan can be defined to ensure that all of the time–quanta in which the event is scheduled are within a specified time interval.
- For each event  $e$ , the number of students per workplace is denoted  $spw_e \in \mathbb{N}$ .

The set  $S$  is the set of students that are to be scheduled. Each student  $s \in S$  has the following set of properties: (i) a set of time quanta  $T_s \subseteq T$  when each student  $s$  is free, and (ii) a non–empty set of events he or she needs to attend, denoted  $E_s \subseteq E$ .

The requirements of the courses are represented by a set of constraints  $C$ . The constraints are divided into hard constraints  $C_h$ , which are essential for the courses, and soft constraints  $C_s$ , which may require some manual intervention if they are not met. Hard constraints  $C_h$  are defined as follows:

- All of the properties of limited assets, rooms, events and students need to be satisfied in the timetable.
- Each room can only be occupied by one event at a time.
- Students can attend only one event at a time.
- Each event  $e$  occupies  $dur_e$  consecutive quanta of the room.
- At most  $size_r \cdot spw_e$  students can be placed in room  $r$  used for the event  $e$ .
- Enough teaching staff must be available to attend each event.

The set of soft constraints  $C_s$  contains one element: the students must attend all the events they are enrolled in. Defining this constraint as soft may seem irrational, but the reasoning behind this is as follows: 'hard' constraints are simply those that are satisfied at all times in any solution suggestion in our implementation, whereas for the 'soft' constraints this may not be the case. 'Soft' constraints are defined as such because it was not known in advance whether there even exists a solution that satisfies all the constraints (given the complex requirements). In other words, our approach tries to find the best solution within the imposed constraints and possibly to give a feedback to the course organisers if some are still severely violated. In the remainder of the text, the term 'feasible solution' denotes a solution that satisfies at least the hard constraints as defined above.

### 3 Solving LETP Using Ant Colony Optimisation

#### 3.1 Construction Graph

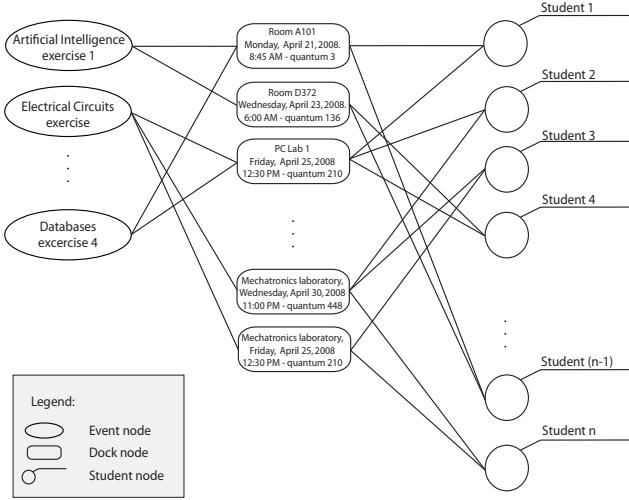
The main issue in applying ACO to a problem is to find an appropriate representation that can be used by the artificial ants to build solutions [3]. This representation

is called the *construction graph*. To ensure that the problem representation is suitable for large instances of LETP, memory-efficiency is the main design goal. The construction graph we devised can be seen in figure 11. Semantically, each of the nodes represents one of the following: a student, an event or a *dock node*. A dock is an ordered pair of the room and beginning time in which an event can be scheduled, e.g., (ComputerLab-2, (2010-09-08, 10:00)).

An edge connecting a dock node and an event node means that the event can be scheduled in that time and place. This means that the dock represents the room and time that are suitable for that event. Dock nodes are connected to student nodes as well. Student nodes are only connected to docks under the following conditions: (i) the dock is connected to at least one of the events the student needs to attend and (ii) the student is free from pre-assignments in time-quanta represented by the dock and the  $dur_{e_s}$  consecutive time quanta. The event  $e_s$  is the shortest event enrolled by the student that can be held in the aforementioned dock, and its duration is denoted  $dur_{e_s}$ . To each of the graph's edges, a pheromone concentration value  $\tau_{ij}$  and a heuristic information value  $\eta_{ij}$  are assigned. The LETP solution is a timetable with all of the events and students scheduled into the appropriate docks. A candidate solution (timetable) is represented as a subset of edges of the construction graph, connecting the timetable building blocks into a specific timetable.

For larger problem instances, the size of the construction graph can be considerable, and the efficiency of the search procedure strongly depends on the size of the graph. To reduce the size of the search space, an additional preprocessing step is performed. During that step, edges representing solution components of poor quality are removed. More precisely, an edge is removed if the number of students who can attend the corresponding event is less than 80% of the room capacity. Note that this value may vary in different problem instances. In the final step, all isolated docks are removed.

Hard constraints are included in the construction procedure through the *constraint fence* layer. This layer dynamically masks the edges of the construction graph that lead to infeasible solutions, based on each incomplete solution (partial tour) of the ants. Thus, ants moving through the graph are producing only feasible timetables. The construction graph is a very large structure, while the constraint fence is a small and expandable one. Ants access the graph exclusively through the constraint fence. They may move only on paths that are allowed by the constraint fence, and they behave as if the edges that are not allowed by the constraint fence did not exist at all. While the construction graph remains constant throughout the execution time, each ant that is moving through the graph is given its own instance of the constraint fence. The constraint fence evaluates each edge based on the desired set of timetable constraints. Each of the constraints from our library is implemented as an independent module. This makes it easy to update the constraints and to add new ones if needed. The computational cost of the constraint fence depends on the set of constraints used. If used with our current library, its computational complexity is  $O(|R|)$ , where  $R$  is the set of rooms in the LETP instance.



**Fig. 1.** LETP construction graph. Edges represent partial assignments of students and events to times and rooms. Note that the construction graph is not fully connected as some of the edges are implicit and implemented in the tour construction procedure.

### 3.2 Algorithm Description

Our approach uses a  $\mathcal{MAX} - \mathcal{MIN}$  Ant System [3], since such systems have shown great promise on various different problems, including artificially generated timetabling problems [5]. A colony of  $m$  ants is used. At each algorithm iteration, each ant constructs a complete timetable (a candidate solution). In each of the generated solutions, all of the hard constraints are satisfied.

In choosing solution components, the probability  $p_{ij}^k$  that the ant  $k$ , currently at node  $i$  will choose the node  $j$  is calculated using the *random proportional rule* given by

$$p_{ij}^k = \frac{[\tau_{ij}]^\alpha [\eta_{ij}]^\beta}{\sum_{l \in \mathcal{N}_i^k} [\tau_{il}]^\alpha [\eta_{il}]^\beta}, \quad j \in \mathcal{N}_i^k, \quad (1)$$

where  $\tau_{ij}$  is the pheromone trail value on the edge connecting node  $i$  to node  $j$ ,  $\eta_{ij}$  is the heuristic value of that edge,  $\alpha$  and  $\beta$  are the parameters which determine the relative influence of pheromone trail and the heuristic information, and  $\mathcal{N}_i^k$  is the feasible neighbourhood of ant  $k$  when it is at node  $i$ .

The pheromone trail between dock node  $i$  and event/student node  $j$  marks the desirability of scheduling that event/student in that dock. In each algorithm iteration only one of the ants updates the pheromone trails based on the quality of the constructed solution candidate. The heuristic value  $\eta_{ij}$  is controlled by the constraint fence for each ant. It is important to note that the heuristic value is dynamically modified throughout the algorithm execution. It is set to zero

when it is determined that the constraints of the problem would be violated if the edge  $(i, j)$  were included in the tour, and to one otherwise.

Solutions are constructed one event at a time and the scheduling of each event is performed in two phases. In the first phase, the event is placed into suitable rooms and time-slots. This is done by assigning a sufficient number of docks to the event, one dock at a time, until a sufficient amount of room is reserved for the event. The decision of which dock to assign to the event is a biased random choice that is influenced by the pheromone trail in the feasible neighbourhood of the event node. To determine the probability that an edge will be selected, the random proportional rule is used. In the second phase of the scheduling of the event, students are assigned to the docks chosen in the first phase. The decision of which student is to be placed in which dock is also biased by the pheromone trails deposited by previous ant generations, using the random proportional rule. After a partial timetable for a single event is constructed, the ant constructs partial timetables for all of the events that have not yet been scheduled. The order in which the ants schedule the events is heuristically determined at the start of each iteration of the algorithm, so that the events with more unscheduled students in the best tour of the previous iteration have greater priority. Each of the solutions proposed by the ants is improved using the local search routine. The local search rearranges the students among the docks assigned to the same event. It tries to find suitable docks for students that are not scheduled to the event they need to attend. The search is performed by checking whether room can be made by for an unscheduled student by switching one of the scheduled students into another dock.

After the solution construction and local search is finished for all ants, each of the solutions is evaluated. We define the *penalty function* as the total number of unscheduled obligations in a candidate solution (the exact definition of an obligation is given in Section 3.3). The number of unscheduled obligations is also used as the optimised variable. After the evaluation, the pheromone updating step is performed as described in the following subsection.

### 3.3 Pheromone Update

The pheromone trail value is updated at the end of each algorithm iteration. The pheromone is updated by the best-so-far ant or the iteration-best ant. The probability that the best-so-far ant is allowed to update the pheromone trail is 5%. Unlike usual ACO approaches, the pheromone gain value in our approach is not the same for all of the edges of a single tour. Instead, the pheromone gain is defined for each event individually. The quality value is assigned to each event of the timetable by counting the number of obligations left unscheduled for that event. The intention is to improve the search process by using the quality of the solution components to determine the pheromone gain value for individual edges in the solution suggestion.

Usually, more than one event can be scheduled in a single dock. Nevertheless, after event  $e$  is scheduled in dock  $d$ , no other event can be scheduled into  $d$ . Therefore, the quality of a partial schedule of a single event cannot be measured without considering its influence on other events. For example, suppose that event  $e_1$  can

be scheduled in the set of docks  $\{d_0, d_1, d_2\}$  and event  $e_2$  can be scheduled only in dock  $d_1$ . Dock  $d_1$  may seem to be appropriate for event  $e_1$  since it would leave zero students that need to attend  $e_1$  unscheduled. However, this is a very poor choice considering that  $d_1$  is the only suitable dock for  $e_2$ , so assigning it to  $e_1$  would leave all of the students who need to attend  $e_2$  unscheduled. The level of influence between two events is modelled by the influence function  $f : E \times E \rightarrow [0, 1]$ . When  $f_{a,b} = 0$ , event  $a$  has no influence on event  $b$ , while  $f_{a,b} = 1$  means that event  $a$  is scheduled in the entire set of docks suitable for the event  $b$ . More formally, the influence of event  $a$  on  $b$ , denoted  $f_{a,b}$ , is defined as:

$$f_{a,b} = \frac{|chosenD_a \cap D_b|}{|D_b|} ,$$

where  $chosenD_a \subseteq D_a$  is a subset of docks in which event  $a$  is scheduled in a given suggested solution, and  $D_b \subseteq D$  is the set of docks suitable for the event  $b$ .

We use the number of unscheduled *obligations* as the optimised variable. An obligation is defined as an assignment of a given student to one of the laboratory exercises he or she needs to attend. In our problem representation, this is done by assigning students to the dock nodes during the second phase of the construction procedure for a single event, as described in Section 3.2. The solution quality function  $Q_e$  for the event  $e$  and the solution suggestion  $Sug$  is defined as:

$$Q_e = \frac{assignedObligations(e, Sug)}{numberOfObligations(e)} \cdot spaceEfficiency(e, Sug)^3 ,$$

$$spaceEfficiency(e, Sug) = \left[ \frac{assignedObligations(e, Sug)}{reservedSeats(e, Sug)} \right] .$$

The space efficiency factor ensures that better solutions use the dock space more efficiently. The pheromone gain for each event  $\Delta\tau_e$  is given by

$$\Delta\tau_e = \left( \frac{\sum_{e_i \in E} (f_{e,e_i} \cdot Q_{e_i})}{|E|} \right)^4 , \quad e \in E .$$

The pheromone gain  $\Delta\tau_e$  is deposited on the edges of the tour connecting the event  $e$  to the dock  $d$  and on the edges connecting a student  $s$  to any of the docks in which the event is scheduled.

### 3.4 $\mathcal{MAX}-\mathcal{MIN}$ Ant System Parameters

Several configurations of the metaheuristic were evaluated, and the best results were achieved using the settings in Table 1. The pheromone values are initially set to  $\tau_{max}$ . The pheromone trails are updated after each algorithm iteration, as described in Section 3.3, and evaporation is used for each edge, according to the rule  $\tau_{ij} \leftarrow (1 - \rho) \cdot \tau_{ij}$ .

**Table 1.** Ant colony optimisation algorithm parameters

Parameter	Value	Parameter	Value
number of ants ( $m$ )	5	$\rho$	0.02
$\alpha$	1.0	$\tau_{max} = 1/\rho$	50
$\beta$	1.0	$\tau_{min}$	0.5

Our system uses either 10000 iterations or  $penalty = 0$  as the stop criteria. To prevent stagnation, if the best solution found is not improved after 125 consecutive iterations, the pheromone trails are reset by setting pheromone value on each edge back to  $\tau_{max}$ .

## 4 Results

The system described here was successfully applied to the laboratory exercises timetabling problem at the authors' institution. The performance of the algorithm on several datasets is presented in Table 2. These problem instances have different durations and widely varying numbers of events and attending students. Two values,  $S_{e,s}$  and  $T_s$  are given as measures of the problem instance complexity. The *student event sum*, denoted  $S_{e,s}$ , is defined as the aggregate number of events that each of the students needs to attend. More formally,  $S_{s,e} = \sum_{s \in S} |E_s|$ , where  $E_s$  is the set of events that student  $s$  needs to attend. The *timespan* of the problem instance, denoted  $T_s$ , is defined as the number of days on which the events need to be scheduled. For each dataset, 30 independent runs were performed, and each run was limited to one hour. For the purpose of comparison, the maximum run-time limit is added to the usual stop criteria described in Section 3.4. Note that these datasets are instances of real-world timetabling problems that had appeared at our faculty. Anonymised version of these datasets is publicly available for download at <http://morgoth.zemris.fer.hr/jagenda>. To illustrate the effectiveness of our approach, the results are compared with a GRASP technique (Table 2). The tested GRASP technique uses a construction search (different from the search procedure defined for our ants) to build solutions satisfying hard constraints, after which a local search that optimises the schedule of students is performed. We used the Mann–Whitney test to check the H0 hypothesis that the distribution functions of the algorithm performances were the same for the results of both the ACO and GRASP techniques. For each of the datasets, the  $p$  values were well below 0.05. On each problem instance, with very high statistical significance, we conclude that the ACO technique performs better than the GRASP technique. Note that although the Genetic algorithm has also been applied to the LETP problem as a part of a sister project at our institution [2], the performance of these approaches cannot be directly compared since different quality measures and optimised variables are used in these approaches.

In some problem instances, a solution where all students are scheduled could not be found. This was usually caused by a constellation of conflicting events,

**Table 2.** Algorithm performance on different *laboratory exercise timetabling problem* instances

instance	$S_{s,e}$	$T_s$	ACO penalty			median comparison	
			min	max	st.dev. ( $\sigma$ )	ACO	GRASP
C1	2104	5	66	150	23.31	102	330
C2	7081	9	1	14	3.46	8	71
C4	4868	5	5	73	12.42	13	125
C8	5430	4	34	146	31.00	58	190
C12	5934	5	32	78	9.31	41	333

or events with infeasible requirements posed by the course organisers. In such instances, the system is used as a tool for identifying the problematic events. In practice, the process of scheduling is usually an iterative process of querying the system for the best results, interpreting those results, and allowing the staff to make informed decisions.

## 5 Conclusion

This paper presents a case study of applying ACO metaheuristic for solving a complex large-scale timetabling problem at our institution. Our solution uses a relatively general problem representation, suitable for different types of institutions. We present an innovative, memory-efficient problem representation that is appropriate for large problem instances. Moreover, the modular design of the constraint library facilitates the addition of new constraints. It makes the system manageable, which is extremely important for practical timetabling applications. The exact problem we are solving is formulated as the laboratory exercise timetabling problem, a subset of the university course timetabling problem.

This work arose out of the specific timetabling needs of one institution. However, the approach described here is not limited to LETP, since it shares many commonalities with other UCTP instances. Thus, it is likely that the challenges we faced, such as reducing the memory footprint of the construction graph or ensuring the ease of adaptation to problem modifications, will also be faced by other researchers. Other authors may use our approach without modifying the problem representation. The only necessary adaptation may be the implementation of additional constraints that are not supported by our current library.

Furthermore, since prior work on ant colony optimisation mainly considered artificially generated UCTP instances, our work proves that ACO can be highly successful in solving real-world timetabling problems. It is an effort to help bridge the gap between theoretical research and practical adaptation of metaheuristic techniques that is currently so prevalent in the area of automated timetabling. Our work can also be viewed as an additional confirmation that ACO is not only an interesting academic research topic, but also a manageable and efficient approach able to solve highly complex real-world problems.

## References

1. Cooper, T.B., Kingston, J.H.: The complexity of timetable construction problems. In: Proceedings of the First International Conference on the Practice and Theory of Automated Timetabling (ICPTAT 1995), pp. 511–522 (1995)
2. Bratković, Z., Herman, T., Omrčen, V., Čupić, M., Jakobović, D.: University course timetabling with genetic algorithm: A laboratory exercises case study. In: Cotta, C., Cowling, P. (eds.) EvoCOP 2009. LNCS, vol. 5482, pp. 240–251. Springer, Heidelberg (2009)
3. Dorigo, M., Stützle, T.: Ant Colony Optimization. In: Bradford Books. The MIT Press, Cambridge (2004)
4. Socha, K., Sampels, M., Manfrin, M.: Ant Algorithms for the University Course Timetabling Problem with Regard to the State-of-the-Art. In: Raidl, G.R., Cagnoni, S., Cardalda, J.J.R., Corne, D.W., Gottlieb, J., Guillot, A., Hart, E., Johnson, C.G., Marchiori, E., Meyer, J.-A., Middendorf, M. (eds.) EvoIASP 2003, EvoWorkshops 2003, EvoSTIM 2003, EvoROB/EvoRobot 2003, EvoCOP 2003, EvoBIO 2003, and EvoMUSART 2003. LNCS, vol. 2611, pp. 334–345. Springer, Heidelberg (2003)
5. Socha, K., Knowles, J., Sampels, M.: A  $\mathcal{MAX}\text{-}\mathcal{MIN}$  Ant System for the University Timetabling Problem. In: Dorigo, M., Di Caro, G.A., Sampels, M. (eds.) Ant Algorithms 2002. LNCS, vol. 2463, pp. 1–13. Springer, Heidelberg (2002)
6. Rossi-Doria, O., Sample, M., Birattari, M., Chiarandini, M., Dorigo, M., Gambardella, L., Knowles, J., Manfrin, M., Mastrolilli, M., Paechter, B., Paquete, L., Stützle, T.: A Comparison of the Performance of Different Metaheuristics on the Timetabling Problem. In: Burke, E.K., De Causmaecker, P. (eds.) PATAT 2002. LNCS, vol. 2740, pp. 329–351. Springer, Heidelberg (2003)
7. Azimi, Z.: Comparison of Methheuristic Algorithms for Examination Timetabling Problem. Applied Mathematics and Computation 16, 337–354 (2004)
8. McCollum, B.: University timetabling: Bridging the gap between research and practice. In: Rudová, H., Burke, E. (eds.) PATAT 2006—Proceedings of the 6th international conference on the Practice And Theory of Automated Timetabling, Masaryk University, pp. 15–35 (2006)
9. Gross, J.L., Yellen, J.: A Handbook of Graph Theory. In: Discrete Mathematics and Its Applications. CRC Press, Boca Raton (2003)

# Nonparametric Statistical Analysis of Machine Learning Algorithms for Regression Problems

Magdalena Graczyk<sup>1</sup>, Tadeusz Lasota<sup>2</sup>, Zbigniew Telec<sup>1</sup>, and Bogdan Trawiński<sup>1</sup>

<sup>1</sup> Wrocław University of Technology, Institute of Informatics,

Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland

<sup>2</sup> Wrocław University of Environmental and Life Sciences, Dept. of Spatial Management

Ul. Norwida 25/27, 50-375 Wrocław, Poland

mag.graczyk@gmail.com, tadeusz.lasota@wp.pl,  
zbigniew.telec@pwr.wroc.pl, bogdan.trawinski@pwr.wroc.pl

**Abstract.** Several experiments aimed to apply recently proposed statistical procedures which are recommended for analysing multiple  $1 \times n$  and  $n \times n$  comparisons of machine learning algorithms were conducted. 11 regression algorithms comprising 5 deterministic and 6 neural network ones implemented in the data mining system KEEL were employed. All experiments were performed using 29 benchmark datasets for regression. The investigation proved the usefulness and strength of multiple comparison statistical procedures to analyse and select machine learning algorithms.

**Keywords:** machine learning, statistical tests, statistical regression, decision trees, neural networks, KEEL.

## 1 Introduction

Several articles on the use of statistical tests in machine learning for comparisons of many algorithms over multiple datasets have been published recently [4],[7],[8],[20]. Their authors argue that the commonly used paired tests, i.e. parametric t-test and its nonparametric alternative Wilcoxon signed rank tests, are not adequate when conducting multiple comparisons due to the so called multiplicity effect [22]. They recommend to employ rank-based nonparametric Friedman or Iman and Davenport tests followed by proper post-hoc procedures for identifying pairs of algorithms which differ significantly.

The most frequently used statistical tests to determine significant differences between two machine learning algorithms are the t-test and Wilcoxon signed-ranks test [27]. However, the former is a parametric one and requires that the necessary conditions for a safe usage of parametric tests should be fulfilled, i.e. independence, normality, heteroscedasticity [25],[30]. It is not the case in majority of experiments in machine learning [8],[20]. Thus, the nonparametric Wilcoxon matched pairs test, which is less powerful than t-test, should be employed. But, when the researcher wants to confront a newly developed technique with a number of known algorithms or choose the best one out of a set of several algorithms, the pairwise comparisons are not proper. In such situations he loses the control over the so called familywise error

rate due to an accumulated error coming from the combination of pairwise comparisons. Therefore, he should perform tests adequate to multiple comparisons together with a set of post-hoc procedures to compare a control algorithm with other algorithms ( $1 \times n$  comparisons) or to perform all possible pairwise comparisons ( $n \times n$  comparisons).

First of all, the Friedman test [6] or its more powerful derivative the Iman and Davenport test [14] should be performed. Both tests can only inform the researcher about the presence of differences among all samples of results compared. After the null-hypotheses have been rejected, he can proceed with the post-hoc procedures in order to find the particular pairs of algorithms which produce differences. The latter comprise Bonferroni-Dunn's, Holm's, and Hochberg's procedures in the case of  $1 \times n$  comparisons, and Nemenyi's, Shaffer's, and Bergmann-Hommel's procedures in the case of  $n \times n$  comparisons.

The Bonferroni-Dunn's scheme [5] leads to the statement that the performance of two algorithms is significantly different if the corresponding average of rankings is at least as great as its critical difference. More powerful is Holm's routine [12] which checks sequentially hypotheses ordered according to their p-values from the lowest to the highest. All hypotheses for which p-value is less than the significance level  $\alpha$  divided by the number of algorithms minus the number of a successive step are rejected. All hypotheses with greater p-values are supported. The Hochberg's procedure [11] is still more powerful and it operates in the opposite direction to the former, comparing the largest p-value with  $\alpha$ , the next largest with  $\alpha/2$ , and so forth until it encounters a hypothesis it can reject.

When all possible pairwise comparisons need to be performed, the easiest is the Nemenyi's procedure [21]. It assumes that the value of the significance level  $\alpha$  is adjusted in a single step by dividing it merely by the number of comparisons performed. It is a very simple way but has little power. The Shaffer's static routine [23], in turn, follows the Holm's step down method. At a given stage, it rejects a hypothesis if the p-value is less than  $\alpha$  divided by the maximum number of hypotheses which can be true given that all previous hypotheses are false. The Bergmann-Hommel's scheme is characterized by the best performance, but it is also the most sophisticated and therefore difficult to understand and computationally expensive. It consists in finding all the possible exhaustive sets of hypotheses for a certain comparison and all elementary hypotheses which cannot be rejected. The details of the procedure are described in [3], [8] and the rapid algorithm for conducting this test is presented in [13].

All the above mentioned procedures were described in detail in [4],[7],[8],[20] and used to a series of experiments on neural network algorithms [20], genetics-based machine learning algorithms [7], decision trees and other classification algorithms [8]. These experiments were conducted using machine learning algorithms and benchmark datasets devoted to classification problems. In this paper we focus on regression algorithms and use in tests 29 benchmark datasets for regression problems.

So far, the authors of the present paper have investigated several methods to construct regression models to assist with real estate appraisal: evolutionary fuzzy systems, neural networks, decision trees, and statistical algorithms using Statistica Data Miner, SAS Enterprise Miner, KEEL, RapidMiner, and WEKA data mining systems [9],[16],[17],[18],[19]. In this paper we present the results of applying 11 regression, deterministic, and neural algorithms to 29 benchmark datasets for regression and

perform a statistical analysis of the results obtained using nonparametric test and post-hoc procedures designed especially to multiple comparisons. The main goal of the study was to investigate the potential of multiple comparison statistical procedures to analyse and select machine learning algorithms for real world application such as a system to assist with real estate appraisal.

## 2 Techniques and Algorithms Used in Experiments

All experiments were conducted using *KEEL (Knowledge Extraction based on Evolutionary Learning)*, a tool for creating, learning, optimizing and evaluating various models ranging from soft computing ones to support vector machines, decision trees for regression, and linear regression. KEEL contains several dozen of algorithms for data pre-processing, designing and conducting the experiments, data post-processing and evaluating and visualizing the results obtained, which have been bound into one flexible and user friendly system. KEEL has been developed in Java environment by a group of Spanish research centres and is available for free for non-commercial purposes [1]. KEEL is designed for different users with different expectations and provides three main functionalities: *Data Management*, which is used to set up new data, data import and export, data edition and visualization, apply data transformations and partitioning, etc.; *Experiments*, which is used to design and evaluate experiments with use of selected data and provided parameters; *Education*, which is used to run experiments step-by-step in order to display learning process.

KEEL algorithms employed to carry out the experiments are listed in Table 1, they were divided into two groups comprising deterministic and neural network techniques, respectively. The references to source articles and details of the algorithms used can be found on KEEL web site: [www.keel.es](http://www.keel.es).

**Table 1.** Deterministic and neural machine learning algorithms used in study

Group	Code	KEEL name	Description
DET	LRM	Regr-LinearLMS	Statistical linear regression
	QRM	Regr-PolQuadraticLMS	Statistical quadratic regression
	SVM	Regr-NU_SVR	Support vector machines for regression
	M5T	Regr-M5	Model tree which combines a decision tree model with statistical linear regression
	M5R	Regr-M5Rules	Based on M5 algorithm. Determines rules and functions that predict correctly the output value
ANN	MLP	Regr-MLPerceptron Conj-Grad	Multilayer perceptron for modeling
	RBF	Regr-RBFN	Radial basis function neural network for regression problems
	RBI	Regr-Incremental-RBFN	Incremental radial basis function neural network for regression problems
	RBD	Regr-Decremental-RBFN	Decremental radial basis function neural network for regression problems
	IRP	Regr-iRProp+	Multilayer perceptrons trained with the iR-Prop+ algorithm - resilient backpropagation algorithm
	SON	Regr-SONN	Self organizing modular neural networks

### 3 Data Sets Used in Experiments

Twenty nine benchmark datasets for regression were used in experiments. They were downloaded from four web sites:

1. <http://archive.ics.uci.edu/ml/datasets> [2]
2. <http://www.liaad.up.pt/~ltorgo/regression/datasets.html> [26]
3. <http://sci2s.ugr.es/keel/datasets.php> [15]
4. <http://funapp.cs.bilkent.edu.tr/datasets> [10]

In order to reduce size of datasets, instance and feature selection was accomplished. Moreover, outliers were removed by means of three sigma method. Then the data was normalized using the min-max approach. Table 2 presents information about the datasets: code, name, content, number of instances and features, number of the link to site they come from (see the list above). 11 machine learning algorithms were run in KEEL individually for 29 datasets using 10-fold cross validation (10cv) and the prediction accuracy was measured with the mean square error (MSE).

**Table 2.** Datasets used in experiments

Code	Name	Content	Inst.	Feat.	Link
01	Abalone	predicting the age of abalone	4027	8	1,2
02	Ailerons	control problem (flying)	7154	8	2,3
03	Delta ailerons	controlling the ailerons (F16 aircraft)	6873	5	2
04	Stock	predicting daily stock prices	950	5	2,3
05	Bank8FM	predicting the fraction of bank customers who leave the bank because of full queues	4318	8	2
06	California Housing	predicting the median price of the house	7921	8	2,3
07	2Dplanes	obtain the value of the target variable Y	6560	7	2
08	House(8L)	predicting the median price of the house	7358	8	2
09	House(16H)	predicting the median price of the house	5626	8	2
10	Delta Elevators	predicting action taken on the elevators of the aircraft	4691	5	3
11	Elevators	predicting action taken on the elevators of the aircraft	7560	7	2
12	Friedman Example	obtain the value of the target variable Y	8217	5	1,2
13	Kinematics	predicting forward kinematics of robot arm	8190	8	2,4
14	ComputerActivity (1)	predicting usr, the portion of time that CPUs run in user mode from all attributes	6570	8	2
15	ComputerActivity (2)	predicting usr using a restricted number	6953	6	2
16	Boston Housing	predicting housing values in Boston	461	4	1,2
17	Diabetes	investigate the dependence of the level of serum C-peptide	43	2	2,4
18	Machine-CPU	predicting relative performance	188	6	1,4
19	Wisconsin Breast Cancer	predicting time to recur	152	6	1,2
20	Pumadyn (puma8NH)	predicting the angular acceleration of one of the robot arm's links	2984	8	2
21	Pumadyn (puma32H)	predicting the angular acceleration of one of the robot arm's links	1245	5	2
22	Baseball	predicting moving of player to other teams	337	6	4
23	Plastic	predicting the pressure	1650	2	3,4
24	Ele2-4 - Electrical-Length	estimate the minimum maintenance costs of the optimal electrical network	1056	4	3
25	Ele1-2 - Electrical-Length	predicting total length of low voltage line	495	2	3
26	Weather-Izmir	predicting the mean temperature	1461	7	3,4
27	Weather-Ankara	predicting the mean temperature	1609	8	3,4
28	Mortgage	predicting the 30-year mortgage rate	1049	6	3,4
29	Concrete Strength [29]	predicting concrete's compressive strength	72	5	1

## 4 Statistical Analysis of the Results of Experiments

Statistical analysis of the results of experiments was performed using a software available on the web page of Research Group "Soft Computing and Intelligent Information Systems" at the University of Granada (<http://sci2s.ugr.es/sicidm>). This open source JAVA program calculates multiple comparison procedures: Friedman, Iman-Davenport, Bonferroni-Dunn, Holm, Hochberg, Shaffer and Bergamnn-Hommel tests as well as adjusted p-values. An adjusted p-value can be directly taken as the p-value of a hypothesis belonging to a comparison of multiple algorithms. If the adjusted p-value for an individual null-hypothesis is less than the significance level, in our study  $\alpha=0.05$ , then this hypothesis is rejected [28]. For paired comparisons nonparametric Wilcoxon signed ranks tests were made using Statistica software.

### 4.1 Results for Deterministic Algorithms

MSE values obtained for 5 deterministic algorithms over 29 datasets are shown in Table 3. The lowest median and interquartile range (IQR) were obtained with SVM algorithm whereas the biggest values were produced by LRM algorithm. In turn M5T, M5R, and QRM provided similar results.

The Friedman and Iman-Davenport tests were performed in respect of average ranks, which use  $\chi^2$  and F statistics, respectively. The calculated values of these statistics

**Table 3.** MSE values for models built over 29 datasets using deterministic algorithms

Set	LRM	QRM	SVM	M5T	M5R
01	0.015687	0.014356	0.014255	0.015288	0.014774
02	0.005437	0.005162	0.005166	0.005157	0.005154
03	0.007314	0.007159	0.006858	0.006893	0.006888
04	0.008179	0.004200	0.002385	0.002473	0.001490
05	0.003341	0.002946	0.002427	0.002696	0.002606
06	0.019318	0.015852	0.014556	0.014951	0.015366
07	0.010351	0.001798	0.001987	0.001790	0.001790
08	0.013319	0.010826	0.010179	0.011371	0.010992
09	0.008385	0.007808	0.007052	0.006660	0.006662
10	0.011418	0.011342	0.010926	0.011123	0.011058
11	0.008379	0.007851	0.007230	0.007311	0.007332
12	0.007903	0.003263	0.001174	0.003760	0.003879
13	0.020297	0.015311	0.003158	0.013790	0.015198
14	0.003841	0.003662	0.003536	0.003691	0.003621
15	0.004251	0.003968	0.003732	0.004036	0.003958
16	0.012308	0.008282	0.007861	0.084322	0.009530
17	0.028618	0.023951	0.025154	0.033956	0.034820
18	0.008417	0.007597	0.007766	0.009801	0.008676
19	0.069762	0.084386	0.068615	0.069898	0.069898
20	0.035566	0.033178	0.020199	0.019260	0.019653
21	0.024852	0.025138	0.002690	0.002781	0.003227
22	0.024304	0.029267	0.025880	0.024009	0.024009
23	0.023453	0.023316	0.023264	0.023453	0.023453
24	0.000376	0.000354	0.000196	0.000105	0.000130
25	0.007285	0.006981	0.006352	0.008725	0.008725
26	0.000437	0.000350	0.000324	0.000397	0.000397
27	0.000454	0.000306	0.000273	0.000301	0.000300
28	0.000102	0.000092	0.000081	0.000098	0.000098
29	0.022580	0.015930	0.008509	0.019933	0.021959
<b>Med</b>	<b>0.008417</b>	<b>0.007808</b>	<b>0.006858</b>	<b>0.007311</b>	<b>0.007332</b>
<b>IQR</b>	<b>0.014860</b>	<b>0.012190</b>	<b>0.008499</b>	<b>0.012507</b>	<b>0.011971</b>

**Table 4.** Adjusted p-values for 1xn comparisons of deterministic algorithms over 29 datasets (SVM is the control algorithm)

<b>Alg</b>	<b>pUnadj</b>	<b>pBonf</b>	<b>pHolm</b>	<b>pHoch</b>	<b>pHomm</b>
LRM	<i>3.04E-12</i>	<i>1.22E-11</i>	<i>1.22E-11</i>	<i>1.22E-11</i>	<i>1.22E-11</i>
QRM	<i>0.000258</i>	<i>0.001033</i>	<i>0.000775</i>	<i>0.000775</i>	<i>0.000775</i>
M5T	<i>0.001200</i>	<i>0.004802</i>	<i>0.002401</i>	<i>0.002401</i>	<i>0.002401</i>
M5R	<i>0.006135</i>	<i>0.024538</i>	<i>0.006135</i>	<i>0.006135</i>	<i>0.006135</i>

**Table 5.** Adjusted p-values for nxn comparisons of deterministic algorithms over 29 datasets

<b>Alg vs Alg</b>	<b>pWilcox</b>	<b>pUnadj</b>	<b>pNeme</b>	<b>pHolm</b>	<b>pShaf</b>	<b>pBerg</b>
LRM vs SVM	<i>0.000016</i>	<i>3.04E-12</i>	<i>3.04E-11</i>	<i>3.04E-11</i>	<i>3.04E-11</i>	<i>3.04E-11</i>
LRM vs M5R	<i>0.000442</i>	<i>0.000023</i>	<i>0.000228</i>	<i>0.000205</i>	<i>0.000137</i>	<i>0.000137</i>
LRM vs M5T	<i>0.005835</i>	<i>0.000186</i>	<i>0.001862</i>	<i>0.001490</i>	<i>0.001117</i>	<i>0.000745</i>
QRM vs SVM	<i>0.000124</i>	<i>0.000258</i>	<i>0.002582</i>	<i>0.001807</i>	<i>0.001549</i>	<i>0.001549</i>
LRM vs QRM	<i>0.000975</i>	<i>0.000894</i>	<i>0.008943</i>	<i>0.005366</i>	<i>0.005366</i>	<i>0.003577</i>
SVM vs M5T	<i>0.002381</i>	<i>0.001200</i>	<i>0.012004</i>	<i>0.006002</i>	<i>0.005366</i>	<i>0.003601</i>
SVM vs M5R	<i>0.007101</i>	<i>0.006135</i>	<i>0.061346</i>	<i>0.024538</i>	<i>0.024538</i>	<i>0.012269</i>
QRM vs M5R	<i>0.335934</i>	<i>0.360979</i>	<i>3.609795</i>	<i>1.082938</i>	<i>1.082938</i>	<i>1.082938</i>
M5T vs M5R	<i>0.854173</i>	<i>0.618292</i>	<i>6.182917</i>	<i>1.236583</i>	<i>1.236583</i>	<i>1.082938</i>
QRM vs M5T	<i>0.611352</i>	<i>0.677975</i>	<i>6.779754</i>	<i>1.236583</i>	<i>1.236583</i>	<i>1.082938</i>

were 49.68 and 20.98, respectively, whereas the critical values at  $\alpha=0.05$  are  $\chi^2(4)=11.14$  and  $F(4,112)=2.45$ , so the null-hypothesis were rejected. Thus we were justified in proceeding to post-hoc procedures. Unadjusted and adjusted p-values for Bonferroni-Dunn, Holm, Hochberg, and Hommel tests for 1xn comparisons, where SVM was the control algorithm, are placed in Table 4. All adjusted p-values are less than 0.05 so that all hypothesis are rejected indicating that SVM revealed significantly better performance than any other algorithm.

In Table 5 p-values for Wilcoxon test, unadjusted values, and adjusted p-values for Nemenyi, Holm, Shaffer, and Bergmann-Hommel tests for nxn comparisons for all possible 10 pairs of algorithms are placed. The p-values below 0.05 indicate that respective algorithms differ significantly in prediction errors; they were marked with italic font. SVM algorithm revealed significantly better performance than others, in turn, LRM was significantly worse than any other algorithm.

## 4.2 Results for Neural Algorithms

The results obtained for 6 neural algorithms are shown in Tables 6, 7, and 8 which are constructed similarly to those for deterministic algorithms. In this group the lowest median and IQR were obtained with MLP and RBF algorithms whereas the biggest values were produced by RBD and SON algorithms.

For Friedman and Iman-Davenport tests, the calculated values of  $\chi^2$  and F statistics were 88.17 and 43.44, respectively, whereas the critical values at  $\alpha=0.05$  are  $\chi^2(5)=12.83$  and  $F(5,140)=2.28$ , so the null-hypothesis were rejected.

The p-values for 1xn comparisons, where MLP was the control algorithm, indicate that MLP revealed significantly better performance than any other algorithm except for RBF. Similar conclusion can be drawn on the basis of p-values for nxn comparisons for all possible 15 pairs of algorithms. It should be noted that with 15 hypothesis the differences between pairwise and multiple comparisons become apparent. Wilcoxon test allows for rejecting 13 hypotheses whereas Holm, Shaffer and Bergmann-Hommel ones discard only 10 and Nemenyi's method just 9.

**Table 6.** MSE values for models built over 29 datasets using neural algorithms

Set	MLP	RBF	RBI	RBD	IRP	SON
01	0.014617	0.015534	0.020329	0.021637	0.015411	0.015914
02	0.005285	0.005706	0.006174	0.222044	0.005961	0.009243
03	0.007186	0.008246	0.008836	0.028481	0.007723	0.008025
04	0.002406	0.002062	0.005222	0.005752	0.006480	0.023976
05	0.002408	0.002796	0.008383	0.012083	0.007221	0.023153
06	0.015729	0.017949	0.020691	0.055435	0.021870	0.022099
07	0.001848	0.001909	0.002569	0.013779	0.004350	0.030583
08	0.010652	0.011289	0.014580	0.021684	0.012842	0.019749
09	0.006988	0.007558	0.009174	0.009605	0.008432	0.009562
10	0.011375	0.011532	0.014138	0.041314	0.012275	0.017183
11	0.008034	0.007990	0.010917	0.012152	0.011719	0.010536
12	0.001314	0.001624	0.002626	0.009056	0.005608	0.018447
13	0.005366	0.007515	0.019045	0.016207	0.020944	0.024366
14	0.003768	0.003800	0.005609	0.135969	0.006132	0.006075
15	0.004069	0.003921	0.005481	0.298655	0.004270	0.009824
16	0.007809	0.008657	0.012165	0.012146	0.009109	0.015207
17	0.040079	0.043437	0.030105	0.051569	0.039685	0.035604
18	0.008000	0.007681	0.009722	0.017188	0.009781	0.017713
19	0.181193	0.070195	0.072858	0.087356	0.071879	0.072490
20	0.018491	0.030200	0.101113	0.035920	0.034526	0.041489
21	0.002079	0.003265	0.007112	0.010306	0.024989	0.024797
22	0.028246	0.028686	0.028494	0.033388	0.024977	0.026145
23	0.023220	0.027964	0.031375	0.034328	0.024938	0.030152
24	0.000347	0.000362	0.002157	0.037865	0.000741	0.002155
25	0.006417	0.007829	0.007208	0.009869	0.007049	0.016597
26	0.000441	0.000428	0.002141	0.003243	0.000705	0.003776
27	0.000389	0.000368	0.002667	0.003401	0.001415	0.022726
28	0.000144	0.000191	0.001723	0.002249	0.000366	0.001211
29	0.008368	0.013809	0.014812	0.015657	0.010894	0.064948
<b>Med</b>	<b>0.006988</b>	<b>0.007681</b>	<b>0.009174</b>	<b>0.017188</b>	<b>0.009109</b>	<b>0.018447</b>
<b>IQR</b>	<b>0.008969</b>	<b>0.011013</b>	<b>0.013564</b>	<b>0.027558</b>	<b>0.014983</b>	<b>0.014973</b>

**Table 7.** Adjusted p-values for 1×n comparisons of neural algorithms over 29 datasets (MLP is the control algorithm)

Alg	pUnadj	pBonf	pHolm	pHoch	pHomM
RBD	<i>2.01E-14</i>	<i>1.00E-13</i>	<i>1.00E-13</i>	<i>1.00E-13</i>	<i>1.00E-13</i>
SON	<i>4.18E-11</i>	<i>2.09E-10</i>	<i>1.67E-10</i>	<i>1.67E-10</i>	<i>1.67E-10</i>
RBI	<i>0.000002</i>	<i>0.000009</i>	<i>0.000005</i>	<i>0.000005</i>	<i>0.000005</i>
IRP	<i>0.000584</i>	<i>0.002918</i>	<i>0.001167</i>	<i>0.001167</i>	<i>0.001167</i>
RBF	0.182355	0.911776	0.182355	0.182355	0.182355

**Table 8.** Adjusted p-values for n×n comparisons of neural algorithms over 29 datasets

Alg vs Alg	pWilcox	pUnadj	pNeme	pHolm	pShaf	pBerg
MLP vs RBD	<i>0.000035</i>	<i>2.01E-14</i>	<i>3.01E-13</i>	<i>3.01E-13</i>	<i>3.01E-13</i>	<i>3.01E-13</i>
MLP vs SON	<i>0.000192</i>	<i>4.18E-11</i>	<i>6.27E-10</i>	<i>5.85E-10</i>	<i>4.18E-10</i>	<i>4.18E-10</i>
RBF vs RBD	<i>0.000003</i>	<i>2.67E-10</i>	<i>4.01E-09</i>	<i>3.47E-09</i>	<i>2.67E-09</i>	<i>2.67E-09</i>
RBF vs SON	<i>0.000038</i>	<i>1.41E-07</i>	<i>2.11E-06</i>	<i>1.69E-06</i>	<i>1.41E-06</i>	<i>8.46E-07</i>
MLP vs RBI	<i>0.000442</i>	<i>1.82E-06</i>	<i>2.73E-05</i>	<i>2.00E-05</i>	<i>1.82E-05</i>	<i>1.27E-05</i>
RBD vs IRP	<i>0.000073</i>	<i>0.000025</i>	<i>0.000381</i>	<i>0.000254</i>	<i>0.000254</i>	<i>0.000178</i>
MLP vs IRP	<i>0.000407</i>	<i>0.000584</i>	<i>0.008754</i>	<i>0.005252</i>	<i>0.004085</i>	<i>0.003502</i>
RBF vs RBI	<i>0.000066</i>	<i>0.000584</i>	<i>0.008754</i>	<i>0.005252</i>	<i>0.004085</i>	<i>0.003502</i>
IRP vs SON	<i>0.000042</i>	<i>0.001586</i>	<i>0.023797</i>	<i>0.011105</i>	<i>0.011105</i>	<i>0.006346</i>
RBI vs RBD	<i>0.000114</i>	<i>0.004007</i>	<i>0.060100</i>	<i>0.024040</i>	<i>0.024040</i>	<i>0.016027</i>
RBF vs IRP	<i>0.009767</i>	<i>0.035240</i>	<i>0.528603</i>	<i>0.176201</i>	<i>0.140961</i>	<i>0.070480</i>
RBI vs SON	<i>0.003634</i>	<i>0.068025</i>	<i>1.020373</i>	<i>0.272099</i>	<i>0.272099</i>	<i>0.136050</i>
MLP vs RBF	<i>0.001323</i>	<i>0.182355</i>	<i>2.735327</i>	<i>0.547065</i>	<i>0.547065</i>	<i>0.547065</i>
RBI vs IRP	<i>0.132888</i>	<i>0.182355</i>	<i>2.735327</i>	<i>0.547065</i>	<i>0.547065</i>	<i>0.547065</i>
RBD vs SON	<i>0.369526</i>	<i>0.292436</i>	<i>4.386534</i>	<i>0.547065</i>	<i>0.547065</i>	<i>0.547065</i>

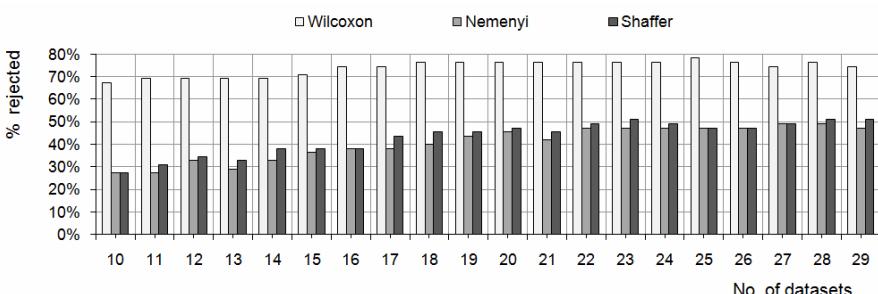
### 4.3 Results for All Algorithms

For 11 algorithms together, 55 pairs of algorithms are possible so that the presentation of all results obtained is not possible due to the limited size of the paper. When applying Friedman and Iman-Davenport tests, the null-hypothesis were rejected. Average ranks of individual algorithms are shown in Table 9. When analyzing adjusted p-values for Holm, Hochberg, and Hommel tests respectively for 1xn comparisons, where SVM was the control algorithm, it could be stated that SVM revealed significantly less prediction accuracy error than any other algorithm but one – M5R. In turn, when following Holm's and Shaffer's procedures for nxn comparisons, the null-hypotheses for the pairs of SVM with M5R, MLP, M5R and QRM algorithms could not be rejected.

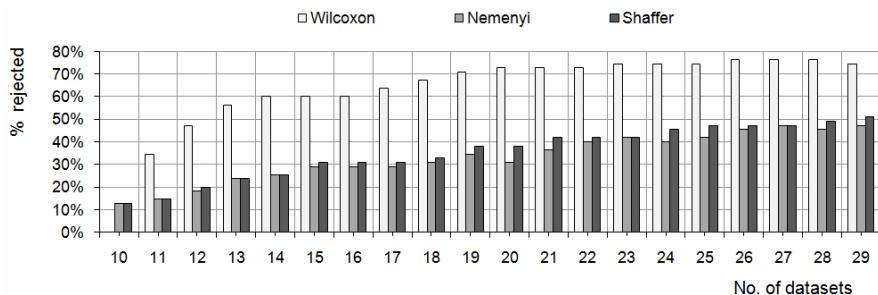
**Table 9.** Average rank positions of all deterministic and neural algorithms over 29 datasets

Rank	Algorithm	Rank	Algorithm	Rank	Algorithm	Rank	Algorithm
2.03	SVM	4.17	M5T	6.83	LRM	9.62	SON
3.69	M5R	4.52	QRM	7.52	IRP	9.97	RBD
4.07	MLP	5.48	RBF	8.10	RBI		

For all 11 algorithms the JAVA software we used to perform statistical tests did not produce any results for Bergmann-Hommel's method. Therefore, we compared the behaviour of three methods Wilcoxon test, Nemenyi's and Schaffer's procedures for nxn comparisons depending on the decreasing number of datasets. In Figures 1 and 2 the percentage of rejected null-hypotheses out of 55 possible pairs of algorithms over different number of datasets is shown. The decreasing number of datasets was obtained by stepwise elimination the datasets providing maximal average prediction error (Fig. 1) and minimal average MSE for all algorithms (Fig. 2). In Fig. 1 it can be observed that employing only pairwise Wilcoxon test would lead to overoptimistic conclusions because the number of rejected null-hypotheses by this test was from 13 to 22 greater than when following the Shaffer's scheme. In majority of instances the Shaffer's procedure turned out to be more powerful than Nemenyi's one. For multiple comparison procedures, the number of discarded null-hypotheses was larger for bigger number of datasets used. When comparing charts depicted in Fig. 1 and 2, it can be seen that the quality of datasets selected to experiments is also important. The bigger average prediction error provided by datasets the less number of null hypotheses rejected.



**Fig. 1.** Percentage of rejected null-hypotheses for Wilcoxon, Nemenyi, and Shaffer tests over different number of datasets (stepwise eliminating datasets providing maximal accuracy error)



**Fig. 2.** Percentage of rejected null-hypotheses for Wilcoxon, Nemenyi, and Shaffer tests over different number of datasets (stepwise eliminating datasets providing minimal accuracy error)

## 5 Conclusions and Future Work

We conducted a study on the application of statistical procedures designed especially for multiple  $1 \times n$  and  $n \times n$  comparisons to several regression algorithms implemented in KEEL. SVM among deterministic algorithms and MLP among neural algorithms revealed significantly better performance than any other technique. Nonparametric Wilcoxon test, when employed to multiple comparisons, would lead to overoptimistic conclusions. For multiple comparisons the more datasets used in tests the bigger the number of null-hypotheses rejected. The investigation proved the usefulness and strength of multiple comparison statistical procedures to analyse and select machine learning algorithms. Further research is planned to conduct similar experiments on evolutionary fuzzy systems for regression.

## References

1. Alcalá-Fdez, J., et al.: KEEL: A Software Tool to Assess Evolutionary Algorithms for Data Mining Problems. *Soft Computing* 13(3), 307–318 (2009)
2. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine (2007), <http://www.ics.uci.edu/~mlearn/MLRepository.html>
3. Bergmann, G., Hommel, G.: Improvements of general multiple test procedures for redundant systems of hypotheses. In: Bauer, P., Hommel, G., Sonnemann, E. (eds.) *Multiple Hypotheses Testing*, pp. 100–115. Springer, Berlin (1988)
4. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research* 7, 1–30 (2006)
5. Dunn, O.J.: Multiple comparisons among means. *Journal of the American Statistical Association* 56(238), 52–64 (1961)
6. Friedman, M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. of the American Statistical Assoc.* 32(200), 675–701 (1937)
7. García, S., Fernandez, A., Luengo, J., Herrera, F.: A Study of Statistical Techniques and Performance Measures for Genetics-Based Machine Learning: Accuracy and Interpretability. *Soft Computing* 13(10), 959–977 (2009)

8. García, S., Herrera, F.: An Extension on “Statistical Comparisons of Classifiers over Multiple Data Sets” for all Pairwise Comparisons. *Journal of Machine Learning Research* 9, 2677–2694 (2008)
9. Graczyk, M., Lasota, T., Trawiński, B.: Comparative Analysis of Premises Valuation Models Using KEEL, RapidMiner, and WEKA. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) *ICCCI 2009. LNCS (LNAI)*, vol. 5796, pp. 800–812. Springer, Heidelberg (2009)
10. Güvenir, H.A., Uysal, I.: Function Approximation Repository, Bilkent University (2000), <http://funapp.cs.bilkent.edu.tr>
11. Hochberg, Y.: A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75, 800–803 (1988)
12. Holm, S.: A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 65–70 (1979)
13. Hommel, G., Bernhard, G.: A rapid algorithm and a computer program for multiple test procedures using procedures using logical structures of hypotheses. *Computer Methods and Programs in Biomedicine* 43, 213–216 (1994)
14. Iman, R.L., Davenport, J.M.: Approximations of the critical region of the Friedman statistic. *Communications in Statistics* 18, 571–595 (1980)
15. KEEL (Knowledge Extraction based on Evolutionary Learning), KEEL-dataset, <http://www.keel.es>
16. Krzystanek, M., Lasota, T., Trawiński, B.: Comparative Analysis of Evolutionary Fuzzy Models for Premises Valuation Using KEEL. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) *ICCCI 2009. LNCS (LNAI)*, vol. 5796, pp. 838–849. Springer, Heidelberg (2009)
17. Lasota, T., Makos, M., Trawiński, B.: Comparative Analysis of Regression Tree Models for Premises Valuation Using Statistica Data Miner. In: Nguyen, N.T., et al. (eds.) *New Challenges in Computational Collective Intelligence. SCI*, vol. 244, pp. 337–348. Springer, Berlin (2009)
18. Lasota, T., Mazurkiewicz, J., Trawiński, B., Trawiński, K.: Comparison of Data Driven Models for the Validation of Residential Premises using KEEL. *International Journal of Hybrid Intelligent Systems* 7(1), 3–16 (2010)
19. Lasota, T., Sachnowski, P., Trawiński, B.: Comparative Analysis of Regression Tree Models for Premises Valuation Using Statistica Data Miner. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) *ICCCI 2009. LNCS*, vol. 5796, pp. 776–787. Springer, Heidelberg (2009)
20. Luengo, J., García, S., Herrera, F.: A Study on the Use of Statistical Tests for Experimentation with Neural Networks: Analysis of Parametric Test Conditions and Non-Parametric Tests. *Expert Systems with Applications* 36, 7798–7808 (2009)
21. Nemenyi, P.B.: Distribution-free Multiple comparisons. PhD thesis, Princeton University (1963)
22. Salzberg, S.L.: On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. *Data Mining and Knowledge Discovery* 1, 317–327 (1997)
23. Shaffer, J.P.: Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association* 81(395), 826–831 (1986)
24. Shaffer, J.P.: Multiple hypothesis testing. *Ann. Rev. of Psych.* 46, 561–584 (1995)
25. Sheskin, D.J.: *Handbook of Parametric and Nonparametric Statistical Procedures*, 4th edn. Chapman & Hall/CRC, Boca Raton (2007)
26. Torgo, L.: University of Porto (LIACC), Regression DataSets, <http://www.liaad.up.pt/~ltorgo/Regression/DataSets.html>
27. Wilcoxon, F.: Individual comparisons by ranking methods. *Biometrics* 1, 80–83 (1945)
28. Wright, S.P.: Adjusted p-values for simultaneous inference. *Biometrics* 48, 1005–1013 (1992)
29. Yeh, I.-C.: Modeling of strength of high performance concrete using artificial neural networks. *Cement and Concrete Research* 28(12), 1797–1808 (1998)
30. Zar, J.H.: *Biostatistical Analysis*, 5th edn. Prentice-Hall, Englewood Cliffs (2009)

# Adaptive Learning of Nominal Concepts for Supervised Classification

Nida Meddouri and Mondher Maddouri

Research Unit on Programming, Algorithmics and Heuristics - URPAH,  
Faculty of Science of Tunis - FST,  
Tunis - El Manar University,  
Campus Universitaire EL Manar, 1060, Tunis, Tunisia  
[nida.meddouri@gmail.com](mailto:nida.meddouri@gmail.com), [mondher.maddouri@fst.rnu.tn](mailto:mondher.maddouri@fst.rnu.tn)

**Abstract.** In recent decades, several machine learning methods based on *Formal Concept Analysis* have been proposed. The learning process is based on the construction of the mathematical structure of the *Galois* lattice. Two major limits characterize these methods. First, most of them are limited to the binary data processing. Second, the exponential complexity of a *Galois* lattice generation limits their fields of application. In this paper, we consider the *Boosting* of classifiers, which is an adaptive approach of classification. We propose the *Boosting* of classifiers based on *Nominal Concepts*. This method builds part of the lattice including the best concepts (pertinent concepts). It is distinguished from the other methods based on *Formal Concept Analysis* by its ability to handle nominal data. The discovered concepts are called *Nominal Concepts* and they are used as classification rules. The comparative studies and the experimental results carried out, prove the interest of this method compared to those existing in literature.

## 1 Introduction

Classification based on *Formal Concept Analysis* is an approach of *Data Mining* which makes it possible to extract correlations and rules, according to the generated concepts from the data. A great diversity of learning methods based on *Formal Concept Analysis* have been proposed [15]. A critical overflight of these methods, based on the extraction of the formal concepts, shows that the existing methods in literature are limited [15]. Several of these methods focus on the extraction of all the concepts from the *Galois* lattice, yielding exponential complexity [17] [19] [4] [8]. Other methods propose the partial construction of the *Galois* lattice [11] [16] [12] [5]. Thus generate less classification rules and less complexity. It should be noted that these methods can not reach the performance level of traditional methods like decision trees, neural networks, etc. We report also that several of these methods are limited since they can only handle binary data.

Recently, an important number of researches in machine learning have been concerned with the *Boosting* methods of classifiers that allow the improvement

of a single learner performances by the techniques of vote [3]. The *Boosting* is known for its ability to improve the performance of any learning algorithm, supposed nevertheless unstable and discovering weak classifiers (weak learner). Unfortunately, systems based on *Formal Concept Analysis* encountered some problems such as an exponential complexity (in the worse case), a high error rate and over-fitting. However, *Boosting* algorithms are known for improving the error rate of any single learner.

In this paper, we present the *Boosting* of *Nominal Concepts* (*BNC* method). This method, based on *Formal Concept Analysis*, improves the *Boosting* of *Formal Concepts* (*BFC* method) [13] [14]. It uses the basic algorithm of multi-class *Boosting*: *AdaBoost.M2* [7]. We suggest to improve the *Boosting of Formal Concepts*, by introducing a new notion: the *Nominal Concept*. This makes it possible to handle non-binary data (i.e nominal). We report that *BFC* calculates the *Informational Gain* of each modality of a binary attribute. So, we suggest to calculate the *Informational Gain* of the nominal attribute without resort to its modalities.

In section 2, we describe the proposed method: the *Boosting of Nominal Concepts*. Then, we present experimental results in section 3 to prove the validity of our proposed method.

## 2 Boosting of Nominal Concepts

*Boosting* is an adaptive approach, which makes it possible to correctly classify an object that can be badly classified by an ordinary classifier. The main idea of *Boosting* is to build many classifiers (experts) who complement each other, in order to build a more powerful classifier. It allows to decrease the error rates of any classifier [7]. So, the *Boosting* is a general method to convert a weak classifier into an effective classifier. There are two major interests for the combination of classifiers. First, a more reliable decision can be obtained by combining the opinion of several experts. Second, a complex problem can be decomposed into several sub problems which are easier to understand and to solve (divide and conquer).

The general idea of the algorithms which are based on the *Boosting* is iterative. At first, they select a subset of instances from the training data set (different subset from the training data set in each iteration). Then, they build a classifier using the selected subset of training instances. They evaluate the classifier on the training data set. And they start again  $T$  times.

Two frequent questions tackle the approach based on *Boosting*. How do we select the subsets? How do we combine the classifiers?

In fact, the diversity of responses leads to a diversity of *Boosting* algorithms. Among the most known ones, is the *Adaptive Boosting* (*AdaBoost*). Initially, *AdaBoost* assigns the training instances equal weights. It randomly selects a subset of the training instances. Then, it applies its learning algorithm on this subset to extract the resultant classifier. It calculates the error rate of the classifier over all the training set. So, if an instance is classified correctly by the

classifier, *AdaBoost* decreases its weight. Otherwise, it increases its weight. It standardizes the weights of the entire instances [6] and it starts all over again the procedure according to these conditions.

The first algorithm is called *AdaBoost.M1* [6],[7]. It repeats the previous process number of iterations fixed by the user in the beginning. If the error rate of one classifier becomes over 0.5, the current iteration is aborted and all the process is stopped. The second algorithm is called *AdaBoost.M2* [7]. It has the particularity to handle multi-class data and to operate whatever the error rate.

In [13] and [14], the authors have present the *BFC* (*Boosting of Formal Concepts*) method based on *Formal Concept Analysis* and exploiting the advantages of *Boosting* algorithms. This method handles only binary data and uses the basic algorithm of multi class *Boosting*: *AdaBoost.M2* [7]. Initially, the algorithm sets equal weights to the learning instances. It selects a subset of the training instances and extracts the pertinent concept within the binary data sets. Then, it uses the discovered formal concept to classify the learning instances and updates the weights of the learning instances by decreasing those of the well classified ones and by increasing the weights of the others (the bad instances). After that, it repeats the resampling based on the new weights, in order to discard the well classified instances and to consider only the bad ones.

The *BFC* method build adaptively a part of the concept lattice made up only by pertinent formal concepts. We report that *BFC* calculates the *Informational Gain* of each modality of a binary attribute. For example, to assess the nominal attribute *Color* from a binary context, *BFC* calculates the *Informational Gain* of each modality (binary attribute): *Color = Yellow*, *Color = Red* and so on. While, *BNC* calculates the *Informational Gain* of the nominal attribute without resort to its modalities.

With *BFC*, we are forced to make a transformation data of nominal into binary data. This transformation consumes additional memory and time resources compared to other methods like the *induction of decision trees*. We think that if we can handle directly nominal data, we will avoid the high cost of this transformation.

Also, a binary context calculated from a nominal context is described by a large number of binary attributes. Each nominal attribute will be replaced by many binary attributes (for each different value). Instead of calculating for **number of attributes** times the *Informational Gain*, we will calculate **it number of attributes \* number of values** times. We think that by handling the nominal data, we will avoid this lost of time.

Studying the rules generated by *BFC* and *ID3*, we noticed that *BFC* is unable to select the same attribute selected by *ID3* when compare the *Informational Gain* of the attributes. For illustration, with the *IRIS* data set, *BFC* chooses *petallength* as best attribute, while *ID3* chooses *petalwidth*. We think that this is due to the fact that *BFC* calculates the *Informational Gain* of an attribute value (i.e.: *petallength=(-inf-2.45])*). While *ID3* calculates the *Informational Gain* of the attribute (i.e.: *petalwidth*). We think that by handling the nominal data, we will be able to choose the same best attribute that is chosen by *ID3*.

## 2.1 Nominal Concepts

We consider that the whole of training instances  $\mathcal{O}$  is described by a whole of  $L'$  nominal attributes  $\mathcal{AN}$  (which are not necessary binary).

$$\mathcal{AN} = \{AN_l | l = \{1, \dots, L\}\}. \quad (1)$$

In the begining, it extracts the pertinent nominal concept within the training instances by selecting the nominal attribute which minimises the measure of *Informational Gain*. Once the nominal attribute  $AN^*$  is selected, we extract associated instances to each value  $v_j$  from this attribute:  $\delta(AN^* = v_j)$ .

**Proposition 1:** From a nominal context (multi-valued), the  $\delta$  operator is set by:

$$\delta(AN^* = v_j) = \{o \in O | AN^*(o) = v_j\}. \quad (2)$$

Then, we look for the other attributes describing all the extracted instances (using the closure operator  $\delta \circ \varphi$  ( $AN^* = v_j$ )). For this, we give the following proposition:

**Proposition 2:** From a nominal context (multi-valued), the  $\varphi$  operator is set by:

$$\varphi(B) = \{v_j | \forall o, o \in B \text{ and } \exists AN_l \in \mathcal{AN} | AN_l(o) = v_j\}. \quad (3)$$

## 2.2 Learning Concept Based Classifiers

Our approach is essentially based on *AdaBoost.M2* [7] (described with more details in section 2.3). We execute,  $T$  times, the learning algorithm on various distributions of the training instances. For each iteration, we do not obtain new training instances but we are satisfied to perturb the distribution by modifying the weights of the training instances.

Initially, the algorithm (the part based on *Adaboost.M2*) initialize the distribution of weights:

$$D_0(i) = (1/n) \text{ for } i = 1, \dots, n. \quad (4)$$

and sets weights to the training instances  $\mathcal{O}$ :

$$w_{i,y}^1 = D_0(i)/(k - 1) \text{ for } i = 1, \dots, n \text{ and each } y \neq y_i \quad (5)$$

Then, our proposed learning algorithm (Algorithm 1) starts by selecting another set:  $O_t$  (from the previous selected set) through resampling the learning nominal data by probabilistic drawing from  $\mathcal{O}$ :

$$O_t = \{(o_i, y_i) : i = \{1, \dots, n'\}, n' \leq n\} \quad (6)$$

Our proposed learning algorithm mines  $O_t$ . It extracts the pertinent nominal concept within the data sets  $O_t$  (described in the section 2.1).

We construct our pertinent concept associated to each value  $v_j$  of the best attribute  $AN^*$  ( $\delta(AN^* = v_j), \delta \circ \varphi(AN^* = v_j)$ ). A weak classifier is obtained

---

**Algorithm 1.** The learning algorithm of pertinent concept

---

INPUT: Sequence of  $n$  training instances  $\mathcal{O} = \{(o_1, y_1), \dots, (o_n, y_n)\}$   
with labels  $y_i \in \mathcal{Y}$ .

OUTPUT: Classifier rules:  $h_t$  (classifier).

BEGIN :

1. Select a subset of training instances by a probabilistic drawing from  $\mathcal{O}$  :  
 $O_t = \{(o_i, y_i) : i \in \{1, \dots, n'\}, n' \leq n\}$ .
  2. From  $O_t$ , find the attribute having the best Informational Gain value:  $AN^*$ .
  3. For each value  $v_j$  of  $AN^*$  do:
    - 3.1. Calculate the closure associated to this value  $v_j$  of the attribute  $AN^*$  in order to generate the pertinent concept :  $(\{\delta(AN^* = v_j)\}, \delta \circ \varphi(\{AN^* = v_j\}))$ .
    - 3.2. Determine the majority class associated to  $\delta(AN^* = v_j) : y^*$ .
    - 3.3. Induce the classification rule  $h_t$ : the conjunction of attributes from  $\delta \circ \varphi(AN^* = v_j)$  implies the membership to the same majority class  $y^*$ .
  - End of loop.
  4. Return  $h_t$  the set of obtained rules.
- 

END.

---

by seeking the majority class associated to the extent of the pertinent concept ( $\delta(AN^* = v_j)$ ). It induces a classification rule. The condition part of the rule is made up by the conjunction of the attributes included in the intent:  $\delta \circ \varphi(AN^* = v_j)$ . The conclusion part of the rule is made up by the majority class. At this level our proposed learning algorithm of nominal concept ends for this iteration.

### 2.3 Boosting Concept Based Classifiers

On each iteration of *Boosting*, we define:

$$W_i^t = \sum_{y \neq y_i} w_{i,j}^t \quad \text{and we set } q_t(i, y) = \frac{w_{i,j}^t}{W_i^t} \text{ for each } y \neq y_i \quad (7)$$

We calculate the distribution of weights:

$$D_t(i) = \frac{W_i^t}{\sum_{i=1}^n W_i^t} \quad (8)$$

Then, our learning algorithm generates a classifier  $h_t$  (a classification rule), in return to the algorithm of *AdaBoost.M2*.

This classifier gives an estimated probability to the class  $y_i$  from the entry  $o_i$ :  $h_t(o_i, y_i)$ . Three cases are presented:

- If  $h_t(o_i, y_i) = 1$  and  $h_t(o_i, y) = 0, \forall y \neq y_i$  then  $h_t$  has correctly predicted the class of  $o_i$ .
- If  $h_t(o_i, y_i) = 0$  and  $h_t(o_i, y) = 1, \exists y \neq y_i$  then  $h_t$  has opposing predicted the class of  $o_i$ .
- If  $h_t(o_i, y_i) = h_t(o_i, y), \forall y \neq y_i$  then the class of  $o_i$  is selected randomly between  $y$  and  $y_i$ .

From this interpretation, the pseudo-loss of the classifier  $h_t$  via the distribution  $W_t$ , defined by the algorithm (the part based on *AdaBoost.M2*) is:

$$\epsilon_t = 0.5 \times \sum_{i=1}^n D_t(i)(1 - h_t(o_i, y_i) + \sum_{y \neq y_i} q_t(i, y)h_t(o_i, y)) \quad (9)$$

So we calculate the error

$$\beta_t = \epsilon_t / (1 - \epsilon_t). \quad (10)$$

and we update the weights of all the training instances according to  $\beta_t$ . We set the new values of weights vector:

$$w_{i,y}^{t+1} = w_{i,y}^t \times \beta_t^{0.5 \times (1 + h_t(o_i, y_i) - h_t(o_i, y))} \text{ for } i = 1, \dots, n \text{ and } y \in Y - \{y_i\}. \quad (11)$$

We suppose that it updates the weights of all the training instances by decreasing those of the well classified ones and by increasing the weights of the others (the bad training instances). After  $T$  iterations, we obtain the final classifier via:

$$h_{fin}(o_i) = \arg \max_{y \in \mathcal{Y}} \sum_{t=1}^T \log(1/\beta_t) \times h_t(o_i, y_i). \quad (12)$$

In fact, the final classifier  $h_{fin}$  is a set of weak classifiers generated from the  $T$  iteration. Each iteration generates a weak classifier (a set of classification rules) or updates the weight of an existing weak classifier.

### 3 Comparative Study

In this section, we compare the proposed method *BNC* with existing ones based on *Formal Concept Analysis: IPR* [12], *CITREC* [5] and *BFC* [13][14]. Also, we compare the *BNC* method with existing ones in literature: *J48* and *AdaBoost.M1* using *J48* as weak learner. To compare the presented approaches, we focus on their classification rates and their numbers of generated concepts.

#### 3.1 Comparison of the Classification Rates

To compare the methods based on the classification rates, we used 19 known data sets from "UCI Machine Learning Repository" [1]. The chosen data sets were preprocessed with 2 filters under WEKA<sup>1</sup>. The first is an instance filter (weka.filters.supervised.attribute.Discretize-Rfirst-last) that converts a range of numeric attributes into nominal attributes (to evaluate *J48*, *AdaBoost.M1* and *BNC*). Then a second filter (weka.filters.supervised.attribute.Nominal-ToBinary) is applied to convert all nominal attributes into binary attributes (to evaluate *IPR*, *CITREC* and *BFC*). Table 1 presents the characteristics of these data sets.

---

<sup>1</sup> Available at <http://www.cs.waikato.ac.nz/ml/Weka>

**Table 1.** Data sets specification

Data Set	Instances	Attributes	Binary Attributes	Class
Balance-scale	625	4	4	3
Contact-lenses	24	4	6	3
Credit-g	1000	20	61	2
Diabets	768	8	13	2
Glass	214	9	19	6
Heart-statlog	270	13	13	2
Hypothyroid	3772	29	47	4
Ionosphere	351	34	144	2
Iris	150	4	12	3
Kr-vs-kp	3196	36	40	2
lymph	148	18	38	4
mushroom	8124	22	121	2
segment	2310	19	169	7
sick	3772	29	38	2
sonar	208	60	60	2
splice	3190	60	3189	3
vote	435	16	16	2
waveform-5000	5000	40	130	3
Weather	14	4	8	2

To calculate the classification rates generated by each method applied to each data set, we use the *WEKA* software. *IPR*, *CITREC*, *BFC* and *BNC* are integrated in a modified version of *WEKA* [2]. The experimentation is done by 10 cross-validation.

To decide the number of iterations for the suggested method *BNC*, we had resort to many experiments on all the mentioned data. These experiments are employed for various iteration counts (going from 1 up to 100 iterations). Finally, we reported that the classification rates calculated for these sets of data, stabilize starting from the 50th iteration.

As shown in Table 2, *BNC* has the specific ability to reduce the error rates compared to the methods based on *Formal Concept Analysis* (*IPR*, *CITREC* and *BFC*). We compare *BNC* with other methods known in the literature of supervised learning: *Induction of Decision tree J48* and *AdaBoost.M1* (with *J48* method for classifiers learning). These two methods are already present in *WEKA*.

We report according to Table 2, *BNC* provides the best results on 7 of 19 available data sets. Also on 9 data sets among 19, *AdaBoost.M1* gives the best results. And on 7 data sets among 19, *J48* gives the best classification rates. These results show that the proposed method reached a comparable level of precision with the known methods.

### 3.2 Comparison of the Number of Concepts

It's very important to determine the number of concepts obtained by each *Formal Concept Analysis* based method.

**Table 2.** Comparison of the success rates

<i>Data Sets</i>	<i>IPR</i>	<i>BFC</i>	<i>CITREC</i>	<i>J48</i>	<i>AdaBoost.M1</i>	<i>BNC</i>
<i>Balance-scale</i>	-	72.31	37.45	69.59	69.59	<b>72.80</b>
<i>Contact-lenses</i>	48.33	56.67	48.33	<b>81.67</b>	71.67	<b>81.67</b>
<i>Credit-g</i>	-	37.3	70.00	72.10	72.00	<b>72.30</b>
<i>Diabets</i>	-	-	65.11	<b>78.26</b>	76.57	74.74
<i>Glass</i>	-	49.94	54.11	<b>73.94</b>	<b>73.94</b>	52.88
<i>Heart-statlog</i>	-	74.44	58.52	81.85	83.33	<b>84.81</b>
<i>Hypothyroid</i>	-	-	92.55	<b>99.47</b>	99.36	94.49
<i>Ionosphere</i>	-	61.81	36.75	89.17	<b>92.03</b>	88.05
<i>Iris</i>	-	92.67	94.67	94.00	94.67	<b>95.33</b>
<i>Kr-vs-kp</i>	-	72.66	52.35	99.44	<b>99.50</b>	66.02
<i>Lymph</i>	-	20.24	1.33	78.33	<b>80.33</b>	74.14
<i>Mushroom</i>	-	52.58	51.80	<b>100.00</b>	<b>100.00</b>	78.38
<i>Segment</i>	-	32.94	14.29	95.32	<b>96.62</b>	47.4
<i>Sick</i>	-	-	90.83	<b>97.85</b>	97.59	93.74
<i>Sonar</i>	-	28.4	50.02	79.81	82.69	<b>82.71</b>
<i>Splice</i>	-	0	24.04	94.36	<b>94.42</b>	29.22
<i>Vote</i>	-	93.1	77.04	<b>96.33</b>	95.85	95.64
<i>Waveform-5000</i>	-	21.16	33.84	76.48	<b>79.68</b>	55.1
<i>Weather</i>	<b>75</b>	55	50	55	<b>75</b>	<b>75</b>

**Table 3.** Comparison of the numbers of concepts

<i>Data Set</i>	<i>Lattice</i>	<i>CITREC</i>	<i>IPR</i>	<i>BFC</i>	<i>BNC</i>
<i>Balance-scale</i>	16	7	4	4	8
<i>Contact-lenses</i>	33	<b>8</b>	14	<b>8</b>	12
<i>Credit-g</i>	-	<b>4</b>	-	10	11
<i>Diabets</i>	232	<b>4</b>	-	<b>4</b>	6
<i>Glass</i>	265	46	-	<b>8</b>	9
<i>Heart-statlog</i>	394	<b>4</b>	-	6	14
<i>Hypothyroid</i>	5743	<b>14</b>	-	-	19
<i>Ionosphere</i>	-	<b>4</b>	-	9	15
<i>Iris</i>	67	8	24	<b>5</b>	6
<i>Kr-vs-kp</i>	-	<b>4</b>	-	6	6
<i>Lymph</i>	7258	16	-	10	15
<i>Mushroom</i>	-	4	-	<b>2</b>	4
<i>Segment</i>	-	128	-	<b>3</b>	8
<i>Sick</i>	4973	<b>4</b>	-	-	27
<i>Sonar</i>	21630	4	-	<b>3</b>	25
<i>Splice</i>	-	8	-	9	<b>7</b>
<i>Vote</i>	7014	<b>4</b>	-	5	6
<i>Waveform-5000</i>	-	-	-	<b>4</b>	11
<i>Weather</i>	36	<b>4</b>	15	10	11

As shown in Table 3, the concept lattice contains a great number of concepts. *CITREC*, *BFC* and *BNC* induces a small part of the lattice. *BNC* reaches the best classification rates with a small number of concepts and gives a small number of concepts in less possible time compared to the other approaches.

## 4 Conclusion

**Formal Concept Analysis** is an interesting formalism to study machine learning and classification methods. **Formal Concept Analysis** allows a full construction of the concepts and the dependence relationships between concepts in order to build a lattice of *Formal Concepts*.

We report common limits between several supervised learning methods based on **Formal Concept Analysis**: handling only binary data and a high complexity. Also, the construction of the concepts is exhaustive or non-contextual.

We proposed an improvement of an adaptive classification method: **Boosting** classifiers based on **Formal Concepts** (*BFC*). The proposed method suggests to enlarge the application fields of **Formal Concept Analysis** focusing on a type of data rather than the binary one (i.e nominal). With each iteration, the learning algorithm selects the nominal attribute which maximizes the *Informational Gain* from the nominal data. The pertinent concepts are calculated from closer associated to this nominal attribute. A classification rule is obtained by associating a majority class to the extension of the pertinent concept.

We carried out an experimental study to show the importance of the proposed method by using known data sets. Our method reached good precision, small number of concepts (compared with the methods based on **Formal Concept Analysis**). We report that our method generated classification rates comparable with those generated by *J48* and *AdaBoost.M1*.

Other measures can be used to select the best attribute. We can also study the combination of measures to adopt an appropriate committee to our classifier. The concept of measures committee has shown its evidence in recent research [9], [10].

Many improvements of the **Boosting** can be realized [20] [13] [14] according to many recent methods of **Boosting** like *Globoost* [21].

## References

1. Asuncion, A., Newman, D.: UCI machine learning repository (2007)
2. Khalfi, B., Cherif, R., Meddouri, N., Maddouri, M.: Développement de méthodes de classification basées sur l'analyse de concepts formels sous la plateforme weka. In: 10èmes Journées Francophones en Extraction et Gestion des Connaissances, Hammamet, Tunisia (2010)
3. Breiman, L.: Bagging predictors. Machine Learning 24, 123–140 (1996)
4. Carpineto, C., Romano, G.: Galois: An order-theoretic approach to conceptual clustering. In: 10th International Conference on Machine Learning, Amherst, MA, USA (1993)

5. Douar, B., Latiri, C.C., Slimani, Y.: Approche hybride de classification supervisée à base de treillis de galois: application à la reconnaissance de visages. In: 8èmes Journées Francophones en Extraction et Gestion des Connaissances, Sophia Antipolis, France, pp. 309–320 (2008)
6. Freund, Y.: Boosting a weak learning algorithm by majority. *Information and Computation* 121, 256–285 (1995)
7. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: 13th International Conference on Machine Learning, Bari, Italy (1996)
8. Guillas, S., Bertet, K., Ogier, J.-M.: Reconnaissance de symboles bruités à l'aide d'un treillis de Galois. In: 9èmes Colloque International Francophone sur l'Ecrit et le Document, Fribourg, Switzerland (2006)
9. Guillet, F., Hamilton, H.J.: Quality Measures in Data Mining. Studies in Computational Intelligence, vol. 43. Springer, Heidelberg (2007)
10. Lenca, P., Meyer, P., Vaillant, B., Lallich, S.: On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. *European Journal of Operational Research* 184, 610–626 (2008)
11. Liquiere, M.: Mephuguiro, E.: Legal: learning with galois lattice. In: 5èmes Journées Françaises sur l'Apprentissage. Lannion, France (1990)
12. Maddouri, M.: Towards a machine learning approach based on incremental concept formation. *Intelligent Data Analysis* 8, 267–280 (2004)
13. Meddouri, N., Maddouri, M.: Générer des règles de classification par dopage de concepts formels. In: 9èmes Journées Francophones en Extraction et Gestion des Connaissances, Strasbourg, France (2009)
14. Meddouri, N., Maddouri, M.: Boosting formal concepts to discover classification rules. In: 22nd International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems, Tainan, Taiwan (2009)
15. Nguifo, E.M., Njiwoua, P.: Treillis de concepts et classification supervisée. *Technique et Science Informatiques* 24, 449–488 (2005)
16. Njiwoua, P., Mephuguiro, E.: Améliorer l'apprentissage partir d'instances grâce à l'induction de concepts: le système cible. *Revue d'Intelligence Artificielle* 13, 413–440 (1999)
17. Oosthuizen, D.: The use of a Lattice in Knowledge Processing. Thèse d'université, University of Strathclyde, Glasgow, UK (1988)
18. Oueslati, M., Zouari, H., Maddouri, M., Heutte, L.: Étude de la combinaison de classifieurs dans le cadre du boosting. In: 6ième édition des ateliers de travail sur le Traitement et Analyse de l'Information: Méthodes et Applications, Hammamet, Tunisia (2009)
19. Sahami, M.: Learning classification rules using lattices (extended abstract). In: 8th European Conference on Machine Learning, Heraklion, Crete, Greece (1995)
20. Sebban, M., Suchier, M.: On boosting improvement: Error reduction and convergence speed-up. In: 4th European Conference on Machine Learning, Cavtat-Dubrovnik, Croatia (2003)
21. Torre, F.: Globoost: Boosting de moindres généralisés. In: 6ième Conférence d'Apprentissage, Montpellier, France (2004)

# A Novel Approach of Process Mining with Event Graph

Hui Zhang<sup>1,2</sup>, Ying Liu<sup>2</sup>, Chunping Li<sup>1</sup>, and Roger Jiao<sup>3</sup>

<sup>1</sup> Tsinghua National Laboratory for Information Science and Technology,  
Key Laboratory for Information System Security, Ministry of Education,  
School of Software, Tsinghua University, Beijing, China, 100084

<sup>2</sup> Department of Industrial & Systems Engineering,  
Hong Kong Polytechnic University, Hong Kong SAR, China  
mfyliu@polyu.edu.hk

<sup>3</sup> The G.W. Woodruff School of Mechanical Engineering,  
Georgia Institute of Technology, USA

**Abstract.** Modern enterprises are increasingly moving towards the workflow paradigm in modeling their business process. One prevailing approach counts on process mining that aims to discover workflow models from log files which contain rich process information. The process models discovered are then used to model and design information systems intended for workflow management. Although workflow logs contain rich information, they have not been made full use in many existing modeling formalisms like Petri nets. In this paper, we propose a novel approach for process mining using event graph to integrate various process related information. Analysis is conducted to show the advantages of event graph based models compared to Petri nets. A case study is also reported to illustrate the entire mining process. Finally, a preliminary evaluation is conducted to show the merits of our method in terms of precision, generalization and robustness.

**Keywords:** process mining, workflow management, Petri nets, event graph.

## 1 Introduction

Because of globalization, modern enterprises are striving to reduce their operation costs and to develop new products and services rapidly. In order to address issues like enterprise performance improvement and business optimization, researches are looking into the possibility of reaching higher efficiency by centralizing the routine aspects of process activities. In manufacturing and administration, such process activities are often separated into well-defined process models incorporated with different kinds of enterprise information such as function roles, time, operation rules, which can be used to describe the process more completely [1].

Process mining, also known as workflow mining, aims to extract process information from workflow logs which are considered as valuable sources of information regarding the actual execution of business processes. The results of process mining are either business rules [2] or process models [3-5]. In this paper, we focus on the latter. With the help of such models, workflow management systems can improve

business processes by automating tasks, passing the right information to the right place for a specific job function, and achieving enterprise information integration [6]. So far, a few deployments of powerful workflow management systems have been reported, such as Staffware, IBM MQSeries and COSA [7].

For the majority of existing studies, Petri net is a prevailing formalism for process modeling in computing, engineering technology and automation. However, there are some notable difficulties in process mining using Petri nets. It is not a trivial task to mine the relations between activities, such as choice and parallelism. Although some tools can deal with loops, each of them imposes restrictions on the structure of these loops so that the integrity of the model discovered is not compromised. Although workflow logs contain rich information, e.g. timestamps, organizers and other process related data, they have not been made full use in the majority of existing modeling formalisms. In order to find out a better solution to tackle some of these difficulties, we present a novel approach for process mining using event graph where our proposed method is able to integrate various process related information.

The rest of this paper is organized as follows. Section 2 reviews related work on both process mining and event graph. Our approach for process mining with event graph is elaborated through a case study in Section 3. Section 4 reveals the evaluation of our method in terms of precision, generalization and robustness. Section 5 concludes.

## 2 Related Work

Process mining was first introduced by Agrawal et al. in 1998 [3]. Their work was built up based on the workflow graphs and it defined two problems, i.e. finding a conformal process model generating events appearing in logs given and finding the definition of edge conditions. A concrete algorithm was then given to tackle the first problem. Cook and Wolf investigated process discovery in the context of software engineering processes [8]. However, their tests were limited to sequential processes.

An important branch of studies in process mining is centered on the algorithms based on Petri nets. Weijters and van der Aalst constructed Petri nets through a heuristic approach, with an intermediate step to create “dependency/frequency tables” and “dependency/frequency graphs” using simple metrics [4; 9]. The classic algorithm named  $\alpha$ -algorithm [5] is able to construct a corresponding Petri net based on event logs. Later, various extensions based on  $\alpha$ -algorithm were reported. For instance, a algorithm was proposed based on Synchro-Net to deal with the invisible tasks and short-loops at ease [10], and a novel approach for process mining based on two event types, e.g. START and COMPLETE, to explicitly detect parallelism [11].

The focus of our approach lies in the way of model formation, using event graph, for process mining. Event graph was first proposed by Schruben to develop simulations of discrete-event systems [12]. Later in 1988, he presented simulation graphs as the mathematical formalization and extension of event graphs [13]. Researchers in the community actually did not make any distinction between event graph and simulation graph. The modeling power of event graphs was demonstrated by presenting a model simulating a Turing machine [14]. Because of its simplicity and modeling power, some researchers began to model and simulate processes with event graph. Robert had

modeled a simple manufacturing system via event graph to illustrate its modeling and analysis [15]. A simulation modeling methodology was developed and implemented combining discrete event simulation with qualitative simulation [16]. An extension to the classical event graphs was also presented towards the specification of component-based models. [17].

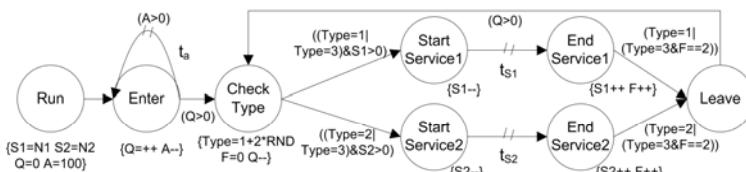
Through our literature review, we note that even though there have been some significant efforts in extending and refining event graph, it lacks of a clear and general way to deploy event graph for process modeling. At the same time, Petri nets are not making full use of the process information readily available. Therefore, it motivates us to explore the possibility of using event graph.

### 3 Approach for Process Mining with Event Graph

In this paper, we focus on constructing a model from event logs through process mining and presenting it in an event graph to integrate all aspects of enterprise process information stored in the logs. Five fundamental elements have played an important role in our model, event vertex, state variable, edge condition, transition function and time delay. Their functions are introduced in Section 3.2 respectively.

Fig.1. is an event graph model of a service system. Based on it, we summarize the merits of event graphs compared to Petri nets. Firstly, it is more flexible to express logical relations between activities because event graph possesses edge conditions for flow control. In this example, “CheckType” has both a XOR split and an AND split to “StartService1” and “StartService2”. Also “EndService1” and “EndService2” consist both a XOR join and an AND join to “Leave”. Such complex logical relations are very difficult to express in Petri nets. Secondly, event graphs have powerful expressibility by pulling together various aspects of a process and its different elements in a concise manner. This example contains rich information, such as the number of different idle servers as well as customers waiting in line and durations of important events, which can be used to track and analyze the process performance.

In order to demonstrate the entire process of our approach clearly, we use a case study through all the steps in this paper.

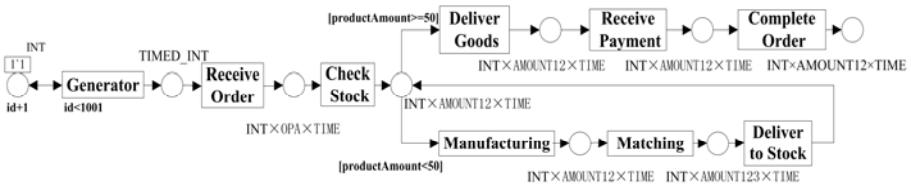


**Fig. 1.** An event graph model of a service system. The “Run” vertex simply initializes the model.  $S_1$  and  $S_2$  are the number of idle servers for two different kinds of services.  $Q$  is the number of customers waiting for service.  $A$  is a calling population of customers. There are three kinds of customer requirements denoted by Type. Customers with Type equaling 1 or 2 only require service1 or service2 respectively, while customers with Type equaling 3 require both of these two services.  $F$  is a state variable for controlling the flow.

### 3.1 Preparing Workflow Logs

Process mining aims at automatically generate process models from event logs. As a preliminary study, we start with simulation data using workflow logs generated from the predefined models. In this paper, we extend the colored Petri nets (CP-net) [18] as the predefined model and simulate them in CPN Tools to generate partial log files containing the various information we need. In order to make our case study more understandable, we align our simulation with a real-life manufacturing process. The extended CP-net in the case study is shown in Fig. 2. For the purpose of constructing the loop, we have an assumption that the production capacity here is limited, which is often true in reality. In addition, because the free variables in CPN Tools must be small color sets, containing less than 100 elements to enumerate, we hence specify the product amount in the range of 0 to 99 and the ordered product amount 40 to 50. Every cycle of production, it produces 20 products and the percentage of pass ranges from 80% to 95%. As a simulation study, the numbers here are relatively small. However, this won't affect the process modeling where in fact we can also consider the product yield to be one thousand units and more.

After the simulation of extended CP-net, it creates some partial MXML log files. We use the ProM<sub>import</sub> framework [19] to create the aggregated log file. Meanwhile, for comparison, we also add in noises in three manners, i.e. adding irrelevant events, deleting events and disordering events from the normal logs. The percentage of noise is 5%, 10% and 15% respectively in the new log files.



**Fig. 2.** The origin model of colored Petri net in CPN Tools

### 3.2 Mining Models from Logs

Besides the activity process, event graphs can combine other enterprise information so as to describe the process more completely. Our approach for process mining with event graph consists of five steps for different components as follows:

**Construct Activity Process.** Unlike the Petri net, our method does not concern about different types of logical relations between activities such as choice and parallelism. We only need to construct the causality relations because the control of flows relies on the edges conditions. Therefore, activity process with causality relations means the crucial architecture of an event graph. At first, we define causality relation.

#### Definition 1:

$A >_i B$  ( $0 \leq i \leq n_{max} - 2$ ): A is directly or indirectly followed by another event B but before the next appearance of A with intervals of  $i$  event(s).  $n_{max}$  is the maximum number for events in a trace in the logs and we assume that  $n_{max}$  is larger than 2.

**Definition 2:**

$V(A>_i B) = \delta^i$  ( $0 \leq i \leq 1$ ): When A is directly or indirectly followed by B with the inter-vallic event number  $i$ , the  $A \rightarrow B$  causality counter is increased with a factor  $\delta^i$ .  $\delta$  is a causality fall factor between 0 and 1.

**Definition 3:**

$F(A \rightarrow B) = (\sum V(A >_i B)) / (\#A)$ :  $F(A \rightarrow B)$  is to indicate the relation for A causing B. It equals the  $A \rightarrow B$  causality counter divided by the overall frequency of event A ( $\#A$ ).

With the workflow logs, we focus on the causality relation. For event A, we aim to calculate  $\sum V(A >_i B)$  for every other event B covering all the traces. For the parameter  $\delta$ , we specify it to be 0.8 here. We divide it by  $\#A$ , and then we get  $F(A \rightarrow B)$  from A to any other event B. After we shift other events for A and repeat this process, we can obtain the final causality matrix shown in Table 1.

**Table 1.** Causality value matrix.  $e_1$ -  $e_8$  are abbreviations of task names.  $e_1$ : Receive Order;  $e_2$ : Check Stock;  $e_3$ : Manufacturing;  $e_4$ : Matching;  $e_5$ : Deliver To Stock;  $e_6$ : Deliver Goods;  $e_7$ : Receive Payment;  $e_8$ : Complete Order.

	$e_1$	$e_2$	$e_3$	$e_4$	$e_5$	$e_6$	$e_7$	$e_8$
$e_1$		1.0	0.57	0.46	0.37	0.52	0.42	0.33
$e_2$	0.0		0.71	0.57	0.46	0.65	0.52	0.42
$e_3$	0.0	0.0		1.0	0.8	0.33	0.26	0.21
$e_4$	0.0	0.0	0.39		1.0	0.41	0.33	0.26
$e_5$	0.0	0.0	0.49	0.39		0.51	0.41	0.33
$e_6$	0.0	0.0	0.0	0.0	0.0		1.0	0.8
$e_7$	0.0	0.0	0.0	0.0	0.0	0.0		1.0
$e_8$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

For every row except the last one (because the last event has no subsequent event), we first choose the event with the maximum value  $f_{max}$ , which can surely fit the causality relation. Then we introduce a parameter  $\lambda$ . If other values in this row are larger than  $\lambda \cdot f_{max}$ , the corresponding events are also regarded as subsequent events. In this experiment, we set  $\lambda$  to be 0.85. In this way, we obtain the causality relations of  $e_1 \rightarrow e_2$ ,  $e_2 \rightarrow e_3$ ,  $e_2 \rightarrow e_6$ ,  $e_3 \rightarrow e_4$ ,  $e_4 \rightarrow e_5$ ,  $e_5 \rightarrow e_3$ ,  $e_5 \rightarrow e_6$ ,  $e_6 \rightarrow e_7$  and  $e_7 \rightarrow e_8$ .

**Extract State Variables.** In this step, we need to extract all the information in the workflow logs and determine how they can be mapped into state variables. After obtaining the raw data from logs, we first carry out in-depth statistic analysis with the help of tools like Weka [20]. Based on the statistical results, we determine which data features should be mapped into the state variables and what their types and ranges are. Finally, we obtain numeric state variables of product amount (PA), ordered product amount (OPA) and newly produced product amount (NNPA). Their ranges are [0, 99], [40, 50] and [16, 19] respectively.

**Mine Edge Conditions.** With the causality relations between activities captured and state variables identified, we then aim to mine the edge conditions with the help of classification in data mining. We extract the events with two or more subsequent events, their subsequent events and all the interrelated state variables. The state variables are kept as attributes and the names of subsequent events are kept as class labels.

By this means, all the data are extracted and transformed into a .arff file for classification in Weka and the classification rules are generated. We take the classification rules as edge conditions and they can be verified by human experts. In this way, condition “PA $\leq$ 49” is generated for edges of e<sub>2</sub> to e<sub>3</sub> and e<sub>5</sub> to e<sub>3</sub> and condition “PA $>$ 49” is generated for edges of e<sub>2</sub> to e<sub>6</sub> and e<sub>5</sub> to e<sub>6</sub>.

**Establish Transition Functions.** Transition functions represent how an event vertex affects state variables. Here we consider two types of situations. One is that a state variable shows up for the first time at an event vertex. In this case, we consider that the event vertex will alter the state variable from its default value (for example, zero for numeric) to a random value in the corresponding range. The other is for state variables not showing up for the first time but being changed. In this case, we have made a simple assumption that all the state variables, before and after a certain event, are basically located in the same semantic layer and are in the same order so that the transition function can be realized through some simple mathematic formation like addition and subtraction. Therefore, we can deal with the transition functions for the latter situation with multilinear fitting, where the coefficients in the multilinear equations are 0, 1 or -1. By this means, we obtain the transition functions in Table 2.

**Table 2.** Transition functions table. FSV is state variables first appeared. CSV is changed state variables. e<sub>1</sub>- e<sub>8</sub> denote for the events in Table 1. PA, OPA and NPPA are state variables.

event	FSV	CSV	Transition Functions
e <sub>1</sub>	OPA	N/A	OPA =[40,50]
e <sub>2</sub>	PA	N/A	PA =[0,99]
e <sub>3</sub>	N/A	N/A	N/A
e <sub>4</sub>	NPPA	N/A	NPPA =[16,19]
e <sub>5</sub>	N/A	PA	PA $\sqcup$ = NPPA
e <sub>6</sub>	N/A	PA	PA $\sqcup$ = OPA
e <sub>7</sub>	N/A	N/A	N/A
e <sub>8</sub>	N/A	N/A	N/A

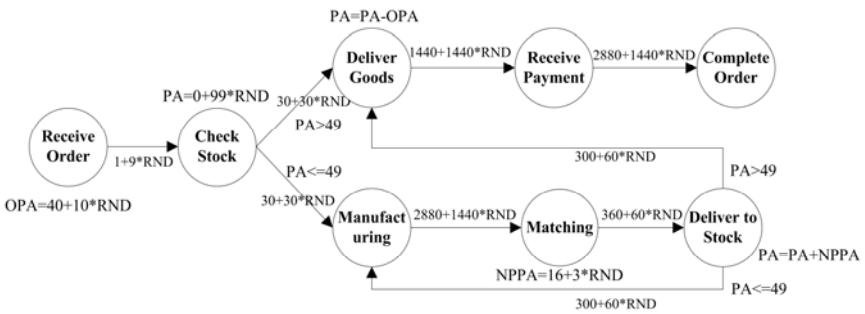
**Deal with Time Delays.** Temporal information is also a significant factor in process. Event graph contains time delays on edges and needs a global clock in simulation. In event graph, every event vertex is regarded as instantaneous and time delays can be expressed on edges. The time delay between event i and j denotes for the interval from the start of event i to the start of event j. However, if there is no time gap, it can also be considered as the duration of event i. In this case study, we extract timestamps of two events and calculate their time gap. Based on the same method used in the previous step of mining state variables, we obtain the statistical results of time delay shown in Table 3.

### 3.3 Models of Event Graph

After the five steps above, we have identified activity process, state variables, edge conditions, transition functions and time delays. Now we are able to construct the

**Table 3.** Time delay table,  $e_1$ -  $e_8$  denote for the same events in Table 1

Start	End	Statistics (min)		Range (min)
		Min	Max	
$e_1$	$e_2$	1	10	[1,10]
$e_2$	$e_3$	30	60	[30,60]
$e_2$	$e_5$	30	60	[30,60]
$e_3$	$e_4$	2880	4320	[2880, 4320]
$e_4$	$e_5$	360	420	[360,420]
$e_5$	$e_3$	300	360	[300,360]
$e_5$	$e_6$	300	360	[300,360]
$e_6$	$e_7$	1440	2880	[1440, 2880]
$e_7$	$e_8$	2880	4320	[2880, 4320]



**Fig. 3.** Constructed model of event graph. PA, OPA and NPPA are the state variables. The notations with state variables along with the edges denote for edges conditions. The notations near the vertexes denote for transition functions. The notations with RND along with the edges denote for time delays. RND is a Sigma function for random numbers. For example,  $1+9*RND$  denotes for the random number between 1 and 10.

event graph model. Here we use Sigma, a modeling and simulation tool for event graph [21]. Note that only activity process can be shown in the main window in Sigma and other elements are embedded in other windows behind. Here we give a full picture of the constructed event graph model for better visualization in Fig. 3.

## 4 Evaluation

It has been pointed out that although there has been a lot of progresses in developing progress mining algorithms in recent years, not much efforts have been invested in developing a common means to assess the quality of the models discovered [22]. Combing existing evaluation strategies with our own ideas, we aim to evaluate our approach by measuring precision, generalization and robustness.

Firstly, we compare our model with the benchmark baseline, i.e. the predefined CPN model in CPN Tools. After our careful and comprehensive examination, we find out that the basic activity process in our event graph is well aligned with the origin model. It counts for the correctness of the crucial architecture of our model. Next, we aim to measure how well the event logs and constructed process models

match. One approach is to generate all execution sequences allowed by the model and then compare them with the log traces [23]. In this experiment we simulate the event graph model in Sigma and obtain 1000 epoches (LogB) and compare it with the origin logs generated by CPN Tools (LogA). In this case, we are able to measure the precision and generalization by calculating how many new traces are included in the origin traces and by calculating how many origin log traces are included in the new traces generated by Sigma respectively. The results of comparison in Table 4 show that our approach is able to achieve good precision and generalization.

**Table 4.** Comparison of LogA and LogB. Name is the name of logs. NoC is the number of traces. NoT is the number of task names. NoE is the number of total events in the log. NoIT is the number of traces of one log included in the other log.

Name	NoC	NoT	NoE	NoIT	Ratio
LogA	1000	8	8039	1000	100%
LogB	1000	8	7865	999	99%

**Table 5.** Test result for constructing activities process with different noisy logs. AC is the couples of activities with causality relation. CV is the causality value for the corresponding couples of activities, kept with four decimal places. CD indicates whether the causality relation is correctly detected.  $e_1$ -  $e_8$  denote for the events in Table 1.

AC	CV				CD
	Log1 (no noise)	Log2 (5% noise)	Log3 (10% noise)	Log4 (15% noise)	
$e_1 \rightarrow e_2$	1.0	0.9971	0.9867	0.9791	yes
$e_2 \rightarrow e_3$	0.7131	0.7235	0.7173	0.7092	yes
$e_2 \rightarrow e_6$	0.6520	0.6356	0.6323	0.6307	yes
$e_3 \rightarrow e_4$	1.0	0.9950	0.9921	0.9896	yes
$e_4 \rightarrow e_5$	1.0	0.9983	0.9916	0.9873	yes
$e_5 \rightarrow e_3$	0.4888	0.4855	0.4814	0.4763	yes
$e_5 \rightarrow e_6$	0.5112	0.5088	0.5077	0.5092	yes
$e_6 \rightarrow e_7$	1.0	0.9923	0.9747	0.9649	yes
$e_7 \rightarrow e_8$	1.0	0.9938	0.9806	0.9681	yes

Besides precision and generalization, we also attempt to evaluate the robustness of our approach. In our study, the crucial part is the step to construct activity process. Therefore, we assess the performance of this step with different datasets which contain noisy logs being purposely introduced. The goal is to see whether our approach can still obtain the correct causality relations in the presence of noises. Through Table 5 we can see that our approach can identify correct causality relations under different noisy conditions and the performance compromise is acceptable. Even with 15% noise added, the maximum reduction counts for less than 4% in  $e_6 \rightarrow e_7$  and  $e_7 \rightarrow e_8$  only.

## 5 Conclusion and Future Work

In this paper, we have proposed a novel approach for process mining with event graph which is able to combine both data structure and activity process in a single entity.

Through this means, various aspects of enterprise information can be integrated so that the business process is described more completely. Analysis is conducted to demonstrate the advantages of event graph models compared to Petri nets. A case study is reported to reveal the entire mining process. Through evaluation, it shows that our method can deliver good performance in terms of both precision and generalization. Moreover, we have also tested that it is robust to deal with noisy data. In the future, efforts in evaluating and polishing our approach using real world data will be constantly pursued. Further applications based on the constructed event graph models, e.g. process tracking and analyzing change impact, will also be studied.

## Acknowledgement

The work described in this paper was supported by a research grant from the Hong Kong Polytechnic University, Hong Kong SAR, CHINA (Grant No: A-PD0M).

## References

1. Georgakopoulos, D., Hornick, M., Sheth, A.: An overview of workflow management: from process modeling to workflow automation infrastructure. *Journal of Distributed and Parallel Databases* 3(2), 119–153 (1995)
2. Crerie, R., Baião, F.A., Santoro, F.M.: Discovering business rules through process mining. *Lecture Notes in Business Information Processing* 29, 136–148 (2009)
3. Agrawal, R., Gunopulos, D., Leymann, F.: Mining process models from workflow logs. In: *Proceedings of the 6th International Conference on Extending Database Technology: Advances in database Technology*, Valencia, Spain (1998)
4. Weijters, A.J.M.M., van der Aalst, W.M.P.: Process mining: discovering workflow models from event-based data. In: *Proceedings of the ECAI Workshop on Knowledge Discovery and Spatial Data*, Sydney, Australia (2001)
5. van der Aalst, W.M.P., Weijters, A.J.M.M., Maruster, L.: Workflow mining: discovering process models from event logs. *Journal of IEEE Transactions on Knowledge and Data Engineering* 16(9), 1128–1142 (2004)
6. Eder, J., Olivotto, G., Gruber, W.: A data warehouse for workflow logs. In: *Proceedings of the First International Conference on Engineering and Deployment of Cooperative Information Systems*, Beijing, China (2002)
7. van der Aalst, W.M.P., Hee, K.v.: *Workflow management: Models, methods, and systems*. MIT Press, Cambridge (2002)
8. Cook, J.E., Wolf, A.L.: Discovering models of software processes from event-based data. *Journal ACM Transactions on Software Engineering and Methodology* 7(3), 215–249 (1998)
9. Weijters, A.J.M.M., van der Aalst, W.M.P.: Workflow mining: discovering workflow models from event-based data. In: *Proceedings of the ECAI Workshop on Knowledge Discovery and Spatial Data*, Lyon, France (2002)
10. Xing, Q., Fuwang, L., Zhao, W., Kunzhang, S., Yiyuan, C.: A workflow process mining algorithm based on Synchro-net. *Journal of Computer Science and Technology* 21(1), 66–72 (2006)

11. Wen, L., Wang, J., van der Aalst, W.M.P., Huang, B., Sun, J.: A novel approach for process mining based on event types. *Journal of Intelligent Information Systems* 32(2), 163–190 (2009)
12. Schruben, L.W.: Simulation modeling with event graphs. *Journal of Communications of ACM* 26(11), 957–963 (1983)
13. Schruben, L.W., Yucesan, E.: Simulation graphs. In: Proceedings of the 20th conference on Winter simulation, San Diego, California, USA (1988)
14. Savage, E.L., Schruben, L.W., Yucesan, E.: On the generality of event-graph models. *Informs Journal on Computing* 17(1), 3–9 (2005)
15. Sargent, R.G.: Event graph modelling for simulation with an application to flexible manufacturing systems. *Journal of Management Science* 34(10), 1231–1251 (1988)
16. Ingalls, R.G., Morrice, D.J., Whinston, A.B.: The implementation of temporal intervals in qualitative simulation graphs. *Journal of ACM Transactions on Modeling and Computer Simulation* 10(3), 215–240 (2000)
17. Lara, J.d.: Distributed event graphs: formalizing component-based modelling and simulation. *Journal of Electronic Notes in Theoretical Computer Science* 127(4), 145–162 (2005)
18. Jensen, K.: Coloured Petri Nets: Basic Concepts, Analysis Methods and Practical Use. Springer, Berlin (1997)
19. Medeiros, A.K.A.d., Gunther, C.W.: Process mining: using CPN Tools to create test logs for mining algorithms. In: Proceedings of the Sixth Workshop and Tutorial on Practical Use of Coloured Petri Nets and the CPN Tools, Aarhus, Denmark (2005)
20. Holmes, G., Donkin, A., Witten, I.H.: WEKA: a machine learning workbench. In: Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia (1994)
21. Schruben, L.W.: Graphical simulation modeling and analysis: using SIGMA for windows. Boyd & Fraser, Danvers (1995)
22. Rozinat, A., Medeiros, A.K.A.d., Günther, C.W., Weijters, A.J.M.M., van der Aalst, W.M.P.: Towards an evaluation framework for process mining algorithms. BPM Center Report BPM-07-06, BPMcenter.org (2007)
23. Rozinat, A.: Conformance testing: measuring the fit and appropriateness of event logs and process models. In: Bussler, C.J., Haller, A. (eds.) BPM 2005. LNCS, vol. 3812, pp. 163–176. Springer, Heidelberg (2006)

# A Classification Algorithm for Process Sequences Based on Markov Chains and Bayesian Networks

Katharina Tschumitschew<sup>1</sup>, Detlef Nauck<sup>2</sup>, and Frank Klawonn<sup>1</sup>

<sup>1</sup> Department of Computer Science Ostfalia University of Applied Sciences Salzdahlumer Str.  
46/48, D-38302 Wolfenbuettel, Germany

<sup>2</sup> BT Group, Chief Technology Office, Research and Venturing Intelligent Systems Research  
Centre Adastral Park, Orion Building pp1/12, Ipswich IP5 3RE, UK

**Abstract.** Companies in the customer-centric service sector deal with thousands of business processes on a daily basis. As different tasks progress along the process sequence, the process owners are interested in three key questions: which task could be the next in the remaining process path, how and when will the process finish? In this paper, we focus mainly on the first two question, the prediction of the length of the process path is part of further research. We propose a classification method based on Markov chains and Bayesian networks for predicting such properties like the remaining process flow, especially the next task in the process path, and the class of the process. This approach is applied and tested on real-world data, showing some interesting results and hints for further research.

## 1 Introduction

Process mining could be understood as a part of the data mining [1]. General approaches to business process mining and workflow management are provided for instance in [2], [3], [4] and [5]. However the majority of the above cited works focus on the designing and modelling of workflow processes, for instance using Petri-net formalization [2].

Here, process mining is applied to business processes within British Telecom. In this context, a process is a sequence of work steps to resolve errors or faults that have occurred. Such a process starts when a problem has been detected and terminates when the problem is resolved successfully. Once the cause for the problem is discovered, it can usually be resolved in a short time. Therefore, it is crucial to be able to predict as early as possible which kind of problem or fault needs to be resolved. After a problem has been resolved, its cause is assigned to the corresponding process. The causes are considered as classes assigned to processes.

The workflow of a process contains at least one, but usually a sequence of actions called tasks which are carried out one after the other. At each point in time, only one task can be active. The processes can vary in the number of tasks as well as in the tasks themselves. Table I shows the structure of a process.

The processes can be considered as time series with a discrete and finite domain.

As mentioned before, it is important to find the cause or the diagnosis to finish the process as early as possible. In this way, unnecessary steps or tasks can be avoided and the time to resolve problems can be shortened significantly. It also important to know the next step, i.e. the next task in the sequence of the process making the planning of

**Table 1.** Process

$$\boxed{Task_1 \rightarrow Task_2 \rightarrow Task_3 \rightarrow \dots \rightarrow Task_n : clearcode}$$

work tasks and resources easier. If the length of the process is known, the customer who is affected by the problem can be informed about how long it will take to resolve the problem. The following aspects are therefore of interest:

- Prediction of the class (clearcode) for each process.
- Prediction of the next state (task).
- Estimation of the length of the process sequence.

First, we consider the prediction of the class and of the next task. The following aspects should be taken into account.

- The processes correspond to events in time which should be modelled accordingly.
- The length in terms of the number of tasks of the processes can vary.

Therefore, standard classification algorithms cannot be applied directly to this problem, since there is no fixed set of predictor attributes. The problem of varying lengths of the processes could be amended by extending each process to a maximum number of tasks by filling it up with empty tasks or one could restrict the considerations to the last  $k$  tasks. The first case will lead to unnecessary increase of the memory and a non-canonical representation. The second case leads to a loss of information. The time-component is also not taken into account by standard classifiers.

Interpreting the processes as Markov processes, the temporal aspect as well as the varying length can be integrated easily in the model. Since a class will be assigned to each process, the tasks in the Markov processes should not only depend on the previous tasks, but also on the assigned class.

## 2 Markov Chain and Bayesian Network

Markov chains [6] are special cases of stochastic processes [7]. A Markov chain can be described as a sequence of random variables  $\xi(t)$  ( $t = 0, 1, \dots$  discrete time) having the Markov property, namely that, given the present the future is conditionally independent of the past.

$$p(\xi(t+1) = s_i^{(t+1)} | \xi(t) = s_i^{(t)}, \dots, \xi(0) = s_i^{(0)}) = p(\xi(t+1) = s_i^{(t+1)} | \xi(t) = s_i^{(t)}) \quad (1)$$

Therefore, the state  $\xi(t+1)$  is conditionally independent of  $\xi(t-1), \dots, \xi(0)$  given  $\xi(t)$ .

$S$  is a finite set and is called the state-space.

$$\xi(t) \in S = \{s_1, s_2, \dots, s_m\}, t = 0, 1, 2, \dots$$

The Markov process starts ( $t = 0$ ) in one of these states and runs consecutively from one state to another. If the process is currently in state  $s_i$  then the probability to move at the next step to the state  $s_j$  is  $p_{ij}$ .  $p_{ij}$  is called transition probability and does not depend upon previous states in the chain.

The transition probabilities are furthermore time independent (time-homogeneous Markov chains [6]):

$$p(\xi(t+1) = s_i | \xi(t) = s_j) = p(\xi(t) = s_i | \xi(t-1) = s_j) \quad (2)$$

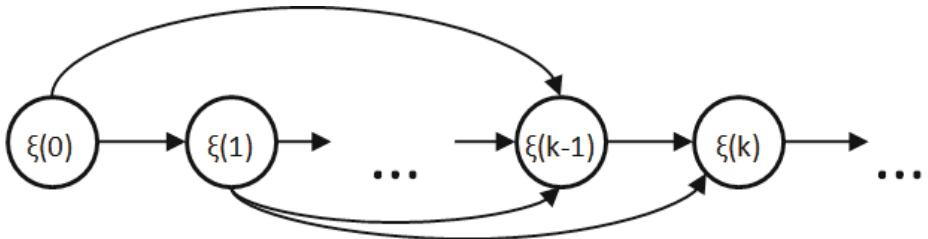
For the purpose of improving readability, we write furthermore

$$p(\xi(t+1) | \xi(t)) \text{ instead of } p(\xi(t+1) = s_i | \xi(t) = s_j).$$

A Markov chain of order  $k$  is a stochastic process with the following property:

$$p(\xi(t+1) | \xi(t), \dots, \xi(0)) = p(\xi(t+1) | \xi(t), \dots, \xi(t-k+1)) \quad (3)$$

Hence, the state  $\xi(t+1)$  is conditionally independent of  $\xi(t-k), \dots, \xi(0)$  given  $\xi(t), \dots, \xi(t-k+1)$ . Figure 1 shows a Markov chain of order  $k$ .



**Fig. 1.** Markov Model  $k$ -the order

In order to model the dependencies between states and the assigned class of the process, we extend the Markov Chain with a class node ( $C = \{c_1, c_2, \dots, c_\nu\}$ ). Therefore, equation (3) transforms to (4).

$$p(\xi(t+1) | \xi(t), \dots, \xi(0), C) = p(\xi(t+1) | \xi(t), \dots, \xi(t-k+1), C) \quad (4)$$

Figure 2 illustrates this model. Hence, the model is a hybrid approach based on Markov chains and Bayesian networks [8].

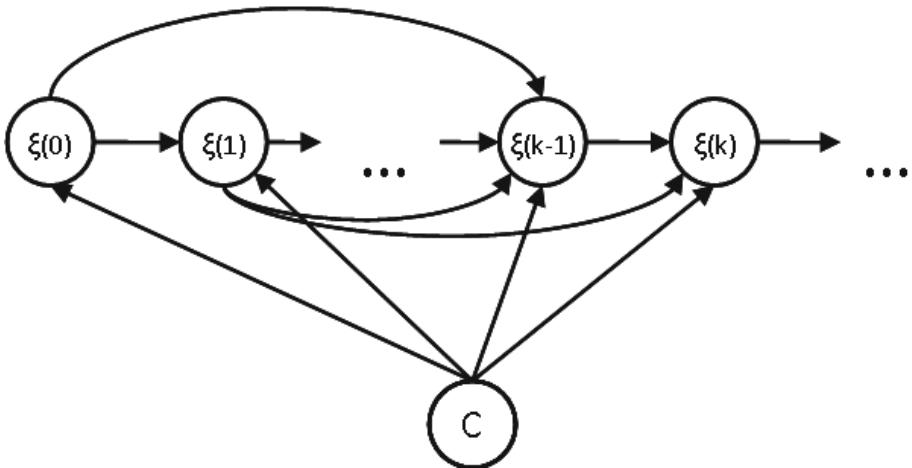
The joint probability for all nodes in this hybrid model is calculated according to equation (5):

$$p(C, \xi(0), \xi(1), \dots, \xi(l)) = p(C) \prod_{i=0}^l p(\xi(i) | \text{par}(\xi(i))) \quad (5)$$

where  $\text{par}(\xi(i))$  defines the set of parents nodes of  $\xi(i)$ .

$$p(C = c_j | \xi(0), \xi(1), \dots, \xi(l)) = \frac{p(C = c_j, \xi(0), \xi(1), \dots, \xi(l))}{p(\xi(0), \xi(1), \dots, \xi(l))} \quad (6)$$

We are only interested in the numerator of the fraction (6), since the denominator does not depend on the class and the values of the features  $\xi(0), \dots, \xi(l)$  are given, so that the denominator is constant for all  $c_j, j = 1, \dots, \nu$ . Therefore, the fraction (6) simplifies to the following:

**Fig. 2.** Bayesian network with markov model properties

$$L(C = c_j | \xi(0), \xi(1), \dots, \xi(l)) = p(C = c_j, \xi(0), \xi(1), \dots, \xi(l)) \quad (7)$$

Taking (5) into account we modify (7) to:

$$L(C = c_j | \xi(0), \xi(1), \dots, \xi(l)) = p(C = c_j) \prod_{i=0}^l p(\xi(i) | \text{par}(\xi(i))) \quad (8)$$

The values provided by the formula (8) should be normalized.

The equation for the probability for the next state is:

$$\frac{p(\xi(t+1) = s_i^{(t+1)} | \xi(t) = s_i^{(t)}, \dots, \xi(t-k+1) = s_i^{(t-k+1)})}{\sum_{\{C\}} p(\xi(t+1) = s_i^{(t+1)}, \xi(t) = s_i^{(t)}, \dots, \xi(t-k+1) = s_i^{(t-k+1)}, C)} = \frac{\sum_{\{C,S\}} p(S, \xi(t) = s_i^{(t)}, \dots, \xi(t-k+1) = s_i^{(t-k+1)}, C)}{\sum_{\{C,S\}} p(S, \xi(t+1) = s_i^{(t+1)}, \dots, \xi(t-k+1) = s_i^{(t-k+1)}, C)} \quad (9)$$

The best  $k$  can be selected by various heuristic techniques, for example, cross-validation.

For the estimation of the length of the process we need a different approach. The length should be estimated as a weighted mean from the length of similar processes. For this approach a similarity measure should be defined. The weights could be calculated as a combination of the similarity and the probability of the process in the data. Several similarity measures for sequential data have been proposed for instance in [9], [10], [11].

### 3 Results

The model can learn probabilities for the class and the next tasks from data. For prediction, the class (task) with the highest probability is chosen. The model has been implemented in Java and has been tested with artificial as well as with real-world data.

The real data consist of 108358 instances assigned to 30 different classes. Random guessing of the classes leads to a rate of roughly 3% correctly classified instances. The proposed model can improve this rate to 36.15%. The file structure: each row represent one process as a tasks sequence. Table 2 represents the example of such sequences.

**Table 2.** Processes

1489793,1489672,1489961,1489717,C1
1489672,R6
1489769,1489717,1489672,R3
1489902,1489769,1489717,1489672,F3
1489902,1489793,1489717,1489793,1489717,1489793,1489672,C2
1489672,1489961,1489717,1489961,1489717,1489672,H4
1489769,1489717,1492609,1489717,1489672,H4
1489672,1489961,1489717,1489672,1489786,1489717,1489672,1492209,1489672,C8

A naive Bayes classifier has also been tested. For this purpose, for each process the number of occurrences for each task were computed. Therefore, the data are stored in an  $m \times n$  matrix.  $m$  is the number of different tasks. The classification accuracy of the naive Bayes classifier is 32%. Other standard classifiers have yielded similar results. The classification results provided by different approaches are listed in table 3.

**Table 3.** Comparison of different classification methods

classification method	correctly classified instances
Naive Bayes	32.11%
Bayesian network	35.02%
Bayesian network with markov model properties	36.15%

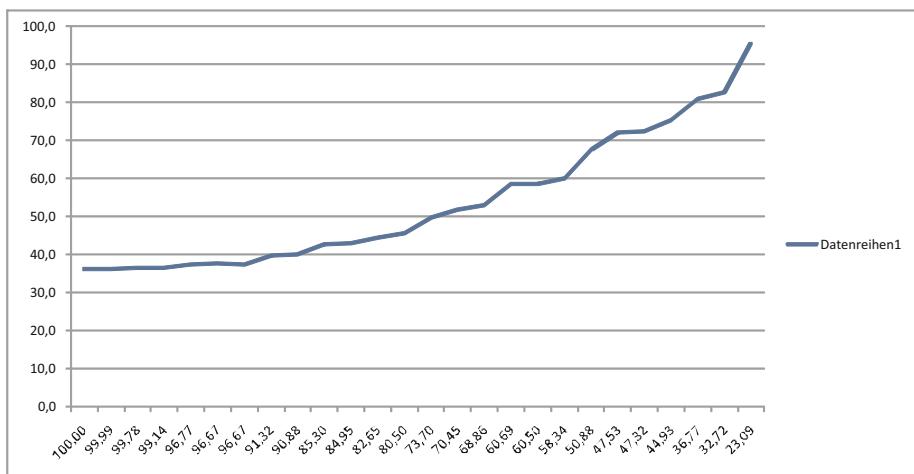
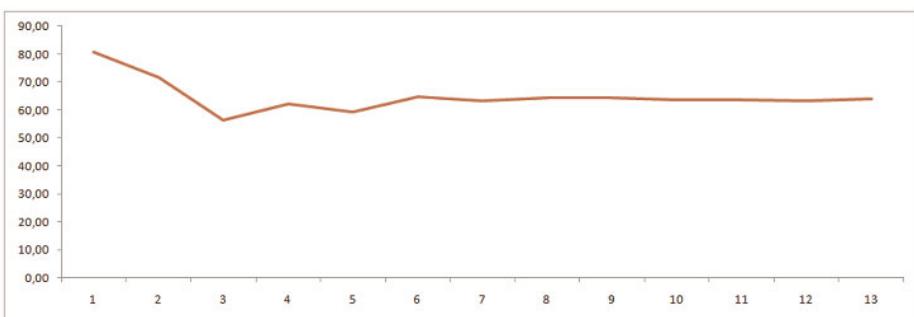
For some of the classes, the misclassification rate is very high, so that they might be strongly overlapping. Restricting, for instance, the data set to the 11 best classes (best in the sense of classification)), the classification accuracy can be increased to 53%.

Figure 3 shows the rate of correct classifications. The classes are ordered with respect to the classification accuracy (for all data), starting with the worst class which is then removed from the data set. The last  $Y/X$ -value represents the classification accuracy for the two best classes.

It was also tested to predict each classes versus all other classes together. This means, the classifier must only decide, whether an instance belongs to one specific class or not. This leads to an average of 75% correctly classified instances.

For the prediction of the next task, it must be taken into account, how many tasks precede the task to be predicted. When all tasks except the last one are already known, the last task can be predicted with an accuracy of 80.7%.

Figure 4 shows the prediction accuracy for tasks depending on the number of tasks that have already been carried out within a process. The number on the  $X$ -axis refers to the position of the task to be predicted counting from the end of the process. Therefore,

**Fig. 3.** Classification accuracy**Fig. 4.** Prediction of the next task

1 on the  $X$ -axis means that the last task in a sequence was predicted, 2 refers to the next to last task and so on. The  $Y$ -axis is for the prediction accuracy.

## 4 Conclusions

In this paper, we have proposed a classification method to solve the problem of process mining based on Markov chains and Bayesian networks. The proposed approach has shown interesting results and justifies further research. So far we have only considered processes that are given as a sequence of tasks. However, in order to improve the misclassification rate, it could be useful to take additional information about the process and the tasks into account. Future work will focus on improving the so far promising results and on predicting the length of the process. This prediction could be based on

prototypes – processes that occur quite frequently. The length of a process whose starting sequence of tasks is known can then be estimated as a weighted mean of the lengths of the processes with a similar starting sequence. This requires a suitable similarity measure on the set of task sequences.

**Acknowledgements.** This work was partially supported by the European Science Foundation through COST Action IC0702.

## References

1. Bishop, C.M.: *Pattern Recognition and Machine Learning*. Springer, Berlin (2007)
2. Van der Aalst, W., van Hee, K.: *Workflow Management: Models, Methods, and Systems*. MIT Press, Cambridge (2002)
3. Van der Aalst, W., van Dongen, B., Herbst, J.: Workflow mining: a survey of issues and approaches. *SData and Knowledge Engineering* 2(47), 237–267 (2003)
4. Bonifati, A., Casati, F., Dayal, U., Shan, M.: Warehousing workflow data: Challenges and opportunities. In: Procs. of VLDB 2001, Rome, Italy (September 2001)
5. Panagos, E., Rabinovich, M.: Escalations in workflow management systems. In: Procs. of DART 1997, Rockville, Maryland (November 1997)
6. Norris, J.: *Markov Chains* (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge University Press, Cambridge (1998)
7. Lawler, G.F.: *Introduction to Stochastic Processes*. Crc Pr Inc., Boca Raton (2006)
8. Jensen, F.V., Nielsen, T.: *Bayesian Networks and Decision Graphs*. Springer, Berlin (2007)
9. Tajima, F., Nei, M.: Estimation of evolutionary distance between nucleotide sequences. *Molecular Biology and Evolution* 1, 269–285 (1984)
10. Miller, W., Myers, E.W.: Sequence comparison with concave weighting functions. *Bulletin of Mathematical Biology* 50(2), 97–120 (1988)
11. Gotoh, O.: An improved algorithm for matching biological sequences. *Journal of Molecular Biology* 162, 705–708 (1982)

# Coaching to Enhance the Online Behavior Learning of a Robotic Agent

Masakazu Hirokawa and Kenji Suzuki

University of Tsukuba, Japan  
[hirokawa@ai.iit.tsukuba.ac.jp](mailto:hirokawa@ai.iit.tsukuba.ac.jp), [kenji@ieee.org](mailto:kenji@ieee.org)

**Abstract.** This paper describes a novel methodology for behavior learning of an agent, called “Coaching”. This is an interactive learning method which allows a human trainer to give a subjective evaluation to the robotic agent in real time, and the agent can update the reward function dynamically based on the evaluation. We demonstrated that the agent is capable of learning the desired behavior by being given simple and subjective instruction such as “good and bad”. The proposed approach is also effective when it is difficult to determine a suitable reward function for the learning situation in advance.

**Keywords:** Behavior Learning, Reinforcement Learning, Coaching.

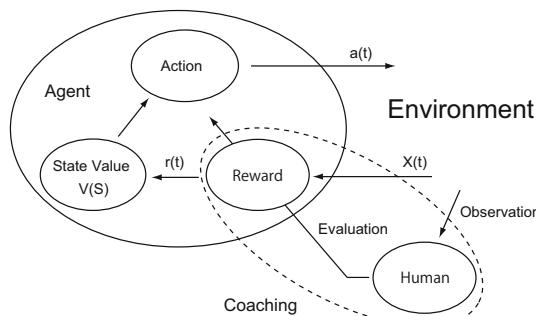
## 1 Introduction

In general, most of the methodologies for behavior learning consist of the following steps : Performing an action, Evaluating the result of the action and Modifying the parameters accordingly. The evaluation step in particular is important in order to modify each of the parameters. Although the best way to achieve learning is to design an evaluation function which gives the correct evaluation for every one of the agent’s actions immediately, it is often not practical to design a specific evaluation function in advance due to the necessity of surveying the whole state space. Autonomous machine learning methods, for example reinforcement learning (RL), have attracted attention for many years [1]. In the RL, the agent updates the expected reward, called the state-value, according to a simple evaluation function, called the reward function and acquires the action rules to maximize the summation of rewards. Therefore, RL can be applied to an environment even with delayed rewards. For this reason, RL is a method which can help to acquire a certain behavior in a real environment. Most conventional methodologies based on RL are based on the assumption that the agent can get its first reward in a reasonable time. At the beginning of learning, the agent explores the state space randomly until it discovers the first reward, and then starts learning based on that reward. Therefore, if the possibility of discovering a reward within the explorable area with respect to the initial state is low, the agent cannot learn for a considerable amount of time. Moreover, this problem cannot be solved by the learning methodologies which focus on how to improve

the learning efficiency based on rewards. To avoid this problem, we should design the reward function carefully. Although the learning efficiency is strongly affected by the reward function, methodologies to design the reward function are still not available. Consequently, we design the reward function depending on the task and the environment according to our own experience. As mentioned above, although there are many kinds of techniques available to improve the learning efficiency based on the obtained reward, we still need a human's knowledge to design the reward function. Especially when the agent tries to learn a complicated task in a real environment, the suitable reward function cannot be created definitely in advance. In such cases, it is necessary to find out the optimal reward function for the learning algorithm used through trial and error, and this is usually very hard.

To reduce the load on humans, several intuitive ways to acquire complex and diverse behavior through the interaction with the trainer, such as Imitation learning and Programming by demonstration (PbD), are attracting attention in the last several years [2][3]. Although several studies have shown that those techniques are effective for real environments, there are some problems. For instance, in the case of Imitation learning, it is assumed that the agent and its trainer have similar body structures and degrees of freedom. Thus, its application range is limited. Further, PbD, it requires demonstration by a trainer for learning. Therefore, PbD is not suitable for tasks which a human cannot demonstrate. Moreover, it is also necessary to prepare different interfaces between the trainer and the agent for demonstrations in different environments. These issues are further hampered by the necessity of detailed information, for example the trajectory of the joint movement, from the trainer regarding the desired behavior, because Imitation learning and PbD focuses on how to imitate the trainer's behavior correctly. On the other hand, the only thing RL requires for learning is the reward function which can be defined in every task as a specific function. This is the reason why the RL can be applied to a variety of situations.

We assume that the conventional RL is useful for behavior learning but, the question of designing the reward function remains unanswered. On the other



**Fig. 1.** The learning model of proposal method. Human trainer can intervene to the learning process of the agent by the updating of the reward function based on Coaching as a learning assistance.

hand, learning through the interaction with a human trainer can realize the action acquisition without a reward function but, giving too much concrete information regarding each task and its environment decreases its general versatility. Thus, the teaching signals from the trainer to the agent should be as simple and intuitive as possible. Therefore, we propose a new learning framework which allows human trainers to give a subjective and fuzzy evaluation to the agent asynchronously and in real time, to lead the agent towards learning the desired behavior. This framework is intuitive for human trainers and we define this framework which is inspired by the human-to-human skill transfer process, as “Coaching”, as opposed to “Teaching” a robot. In this paper, we introduce how to achieve the implementation of the Coaching framework on a typical RL agent. In particular, as indicated in the Fig. 1, a human trainer can intervene in the learning process of the agent and “coach” it, in order to update its reward function according to the situation. In the typical RL method, modifying the reward function and running learning algorithms are individual processes but, the proposed method aims to achieve both processes in parallel. By using this method, the human trainer can support the agent’s learning and reduce the load of designing a reward function. Moreover, we use abstract and primitive binary values as the evaluation quantity of Coaching, namely “good or bad”. Thus, the purpose of this study is to realize interactive and intuitive behavior learning without any technical knowledge about machine learning or use of special interfaces.

## 2 Methodology

### 2.1 Coaching

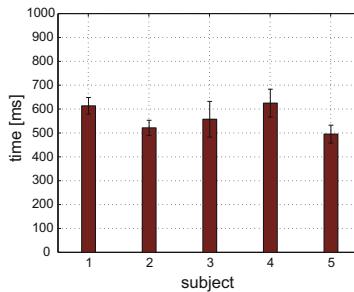
In the conventional method for controlling a robotic system, it is essential to build a model of the controlled object that includes its dynamic specifications. Generally speaking, that process is hard and time-consuming in a machine with the complicated structure. Machine learning methodologies have been proposed to reduce the load of such processes. However, as we described in the previous chapter, they still need human effort to design the evaluation function. On the other hand, Coaching aims to achieve the behavior acquisition without a mathematical model or prior design of the evaluation function, by implementing a mechanism to infer the subjective evaluation given by the trainer interactively and learning the behavior based on that evaluation. In addition, Coaching allows the emergence of difference of the results of behavior learning. It means, even for the same task, there is a possibility of the different result emerging depending on the trainer. We have proposed the original idea of Coaching [4] and have demonstrated that the biped robot can adjust its own parameters for balancing and walking based on the human’s subjective evaluation. Riley et al. have developed a system which can refine the motion of the humanoid by the subjective evaluation from a human [5]. In above studies, the agent can modify its own parameters based on the evaluation given by the trainer but, the agent does not have the ability to learn automatically by the internal value like RL.

On the contrary, the proposed method aims to implement Coaching using an agent that has a learning ability based on its own internal values. Moreover, this method can be applied to other autonomous machine learning methods, for example RL, as an extrapolating algorithm.

Coaching is an interactive learning methodology in which a human trainer can assist the agent in an intuitive manner by giving subjective evaluation in real time. To do this, we must consider the characteristics of human evaluation, especially the time delay and consistency of human's evaluation.

## 2.2 Characteristics of Human Evaluation

When a human trainer gives an evaluation in real time, a time delay occurs between the evaluation timing and the target behavior. Thus, to determine the target behavior of that evaluation, we must measure the time delay as a characteristic of human evaluation. The result of an experiment to measure the time delay by several subjects is shown in Fig. 2.



**Fig. 2.** Result of the experiment to measure the time delay of human evaluation. The procedure is that measurement of the reaction time where the subject did 2 kinds of reaction 10 times against 2 kinds of stimulation selectively, and calculation of the average and the standard deviation for 5 subjects.

We can see from Fig 2 that the difference of the average between subjects is only about 100[ms], and the deviation within each subject is also small. It means that it is possible to narrow down the time delay to a constant range of time. For this reason, by considering the mean and the standard deviation, we defined two kinds of time constants ( $T_1, T_2$ ) which corresponds to the minimum and maximum delay time as below.

$$T_1 = 300 \text{ [ms]}, \quad T_2 = 800 \text{ [ms]} \quad (1)$$

By introducing these time constants we can avoid searching all the action history in an episode for coaching feedback. In this study, the time delay is obtained in advance through a preliminary experiment. However, it is possible to embed it in the process in order to obtain the time delay at the initial stage of learning.

### 2.3 Reinforcement Learning in Continuous State-Action Space

This study attempts to achieve behavior learning in a real environment. Thus, we must deal with agents in a continuous state-action space. In this study, we use the approximated expression of continuous state-action space by using a Radial Basis Function (RBF) network [6][7],

$$\sum_k w_k b_k(\mathbf{x}_t) \quad (2)$$

$\mathbf{x}_t$  is the observed state variable at time  $t$ ,  $b_k(\mathbf{x}_t)$  is the base function unit number  $k$ , and  $w_k$  is a weight variable for  $k$ . The base function is given by the Gaussian function,

$$b_k(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x} - \mu_k\|}{\rho\sigma_k^2}\right) \quad (3)$$

where  $\mu_k$  and  $\sigma_k$  correspond to the center and standard deviation of the Gaussian function. The state value  $V(\mathbf{x}_t)$  and the action output  $u(\mathbf{x}_t)$  can be expressed as:

$$V(\mathbf{x}_t) = \sum_k w_k b_k(\mathbf{x}_t) \quad (4)$$

$$u(\mathbf{x}_t) = \sum_k v_k b_k(\mathbf{x}_t) + n_t \quad (5)$$

$w_k$  and  $v_k$  are weight variables, and  $n_t$  is a random number to explore the state space. The agent updates its own state value and action output by repeating the following steps.

$$\delta = r(\mathbf{x}_t) + \gamma V(\mathbf{x}_t) - V(\mathbf{x}_{t-1}) \quad (6)$$

$$w_k \leftarrow w_k + \alpha \delta b_k(\mathbf{x}) \quad (7)$$

$$v_k \leftarrow v_k + \beta \delta u(\mathbf{x}) b_k(\mathbf{x}) \quad (8)$$

$\delta$  is a TD error,  $\gamma$  is the discount rate and  $\alpha$  and  $\beta$  are learning coefficients.

### 2.4 Implementation of Coaching

To implement Coaching on the above mentioned RL agent, the reward function should also be expressed as a continuous formula. In the same way, using an RBF network, we defined the reward function as,

$$r(\mathbf{x}_t) = r_{init}(\mathbf{x}_t) + \sum_k w_k b_k(\mathbf{x}_t) \quad (9)$$

The first item,  $r_{init}$ , refers to the initial reward function which defines the goal state of the task given at the beginning of the learning algorithm, and the second one means the dynamic reward which has been updated by Coaching during learning.

When the human trainer notices the target behavior, he gives the evaluation at time  $t$ . Using  $T_1$  and  $T_2$ , the state variable  $X$  is defined as:

$$X = \{\mathbf{x}_{t-T_2}, \mathbf{x}_{t-T_2+1}, \dots, \mathbf{x}_{t-T_1}\} \quad (10)$$

$$\equiv \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\} \quad (11)$$

We assume that the target behavior happens during  $X$ . If the evaluation is “good”, it is appropriate to increase the reward around the state set  $X$ . Thus, the weight parameters of the RBF network are updated in a similar way to RL,

$$\delta_i = \lambda^{n-i-1} R + \gamma r(\mathbf{x}_i) - r(\mathbf{x}_{i-1}) \quad (12)$$

$$w_k \leftarrow w_k + \alpha \delta_i b_k(\mathbf{x}_i) \quad (13)$$

Where  $R$  is the immediate reward by Coaching and  $\lambda$  ( $0 < \lambda \leq 1$ ) is a discount rate. By iterating  $i$  from  $n$  to 1, a gradually decreasing gradient from the state  $\mathbf{x}(n)$  to  $\mathbf{x}(1)$  is generated on the reward function.

On the other hand, if the evaluation is “bad”, by applying the following process to  $V(X), u(X)$  and  $r(X)$ ,

$$if \parallel \mathbf{x}_i - \boldsymbol{\mu}_k \parallel \leq d \quad (14)$$

$$w_k = 0 \quad (15)$$

the agent can quickly forget its mistaking knowledge about that area of the state space around  $X$ , and restart learning.

## 2.5 The Basic Problem and the Solution Approach

In this section, we define the basic problem which should be solved by the proposed method (Fig. 3). The upper row of Fig. 3 shows a situation in which a simple reward function can be defined only around the target state in advance, due to the lack of information about the environment or difficulty of the task. As stated in the introduction, from a reward accessibility point of view, it is clear that this situation is difficult for conventional reward-based learning methods.

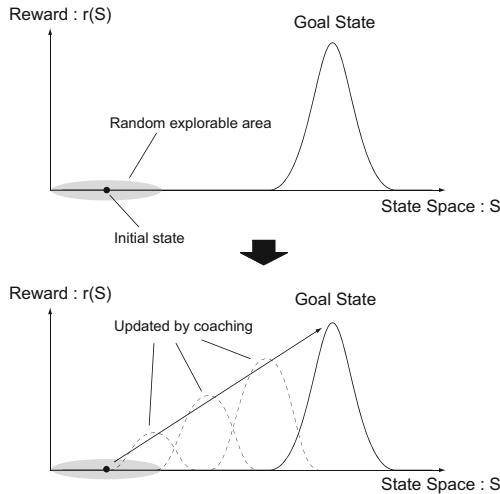
By using the proposed method, we consider that updating the reward function according to the progress of the agent in incremental steps is effective in improving the learning efficiency, as shown in the lower row of Fig. 3. Furthermore, this approach is an attempt to integrate the prior design of the reward function into the learning mechanism.

## 3 Evaluation of Behavior Learning

### 3.1 Definition of the Task

Here we evaluate the performance of the proposed method through a learning experiment that consists of keeping the balance of an inverted pendulum both in a simulation environment and in a real robot. In the experimental procedure, the state space is defined as a 2 dimensional continuous space in terms of pendulum angle  $\theta$  and angular velocity  $\omega$ , and the agent can add the bidirectional force to the root of the pendulum as an action output  $u(t)$ . It is as follows:

1. Set the initial posture of the pendulum as vertically downward.
2. Start the action and measure the total time which the inverted pendulum spends within the target posture range of  $\pi \pm \epsilon$  [rad], where  $\epsilon$  is a small value.
3. Repeat above steps 100 times.



**Fig. 3.** The basic problem of this research. The upper stand of this figure shows difficult situation for conventional reward-based learning method, and the lower stand shows the solution approach by Coaching against this problem.

The reward function is given below.

$$r(\theta) = \begin{cases} 1 & \text{where } |\theta| = \pi \\ 0 & \text{else} \end{cases} \quad (16)$$

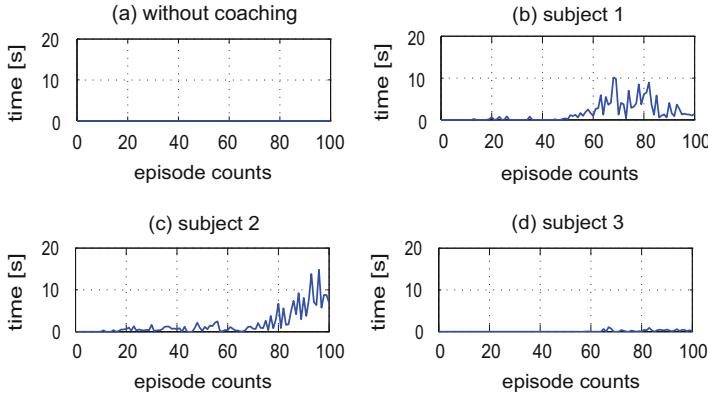
This reward function allows the agent to obtain the reward only at the target state. Moreover, it is necessary to move the control right and left synchronizing it with the oscillation of the tip of the pendulum to explore this state space widely. The probability that such a movement is generated from a random output is very low. Thus, this learning environment is similar to the basic problem which we described in the previous section.

With the proposed method, we conducted an experiment by using three subjects as trainers and a conventional RL method as the control experiment. The instruction for subjects was as follows:

- The purpose of this task is to make the agent learn to keep the balance of the inverted pendulum.
- Give a “good or bad” evaluation if you think that the agent did or didn’t do an action that brings it closer to achieving this task.

### 3.2 Results

The Fig. 4 shows the results of the experiment. The graph, Fig. 4(a), corresponds to the case of learning without Coaching. Although in this case the agent could not learn the desired behavior, in two other cases which used Coaching the agent

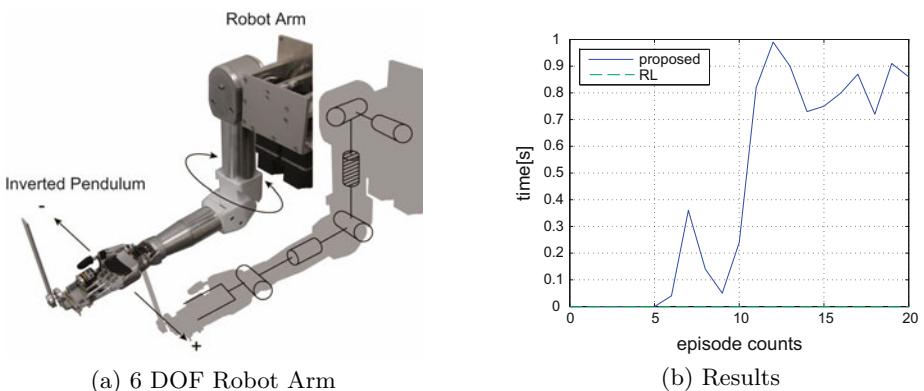


**Fig. 4.** Results of simulation experiment. The horizontal axis is number of episodes and the vertical axis is the time the agent could keep balancing.

acquired the knowledge of how to keep balancing the pendulum when the trainer was subject #1 and #2. The reason why the subject #3 could not make the agent acquire the behavior was that the subject had been giving too many “bad” evaluation compare to “good”. In the proposed algorithm, the “bad” evaluation resets the knowledge, hence does not give any information that can carry learning forward. This problem is one of the considerations for the future works. However, overall, the proposed method successfully improved the learning ability even when the reward function can be defined only in out of the explorable area of the agent.

### 3.3 Experiment by Using a Real Robotic Agent

Finally, we conducted an experiment on the same task by using a real robot arm. We used the robot arm with 6 degrees of freedom shown in Fig. 5(a) as the robotic

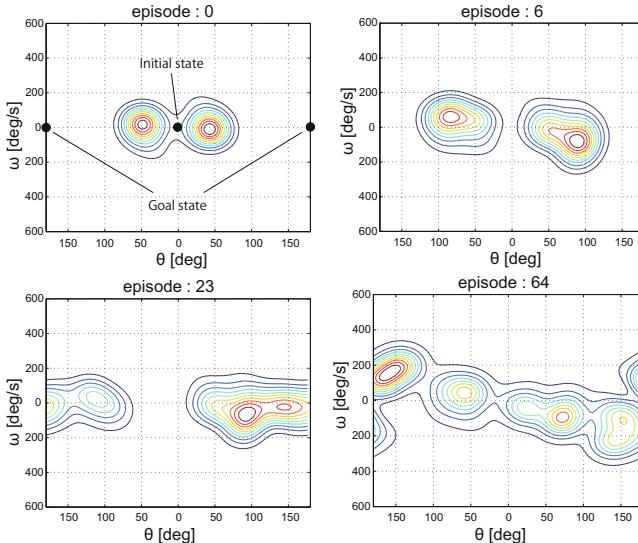


**Fig. 5.** Results of the experiment by the real robotic agent

agent which attempts to learn the behavior. In this study, it controlled only the upper arm joint, which is shown shaded in the figure, to swing the inverted pendulum. The result of this experiment is shown in Fig. 5(b). The proposed method has shown significant improvement compared to the conventional RL with the same reward function, as the latter, completely failed to achieve the task.

### 3.4 Discussion

To evaluate the feasibility of our approach in the simulation experiment, we verified the transition of the reward function modified by subject #2 whose performance was best. The Fig. 6 shows the reward function at the end of episodes 0, 6, 23 and 64. We can see from this figure that the peak of the reward function has been moving from the initial state towards the goal state. Consequently, the solution approach we described shown in Fig. 3 was achieved. However, in some cases Coaching may not help for the robot as shown by the results of subject #3. It can be said that this is because the robot reflected the difference of the Coaching strategies of trainers. On the other hand, in the real environment, due to the limitation of hardware the total time the agent could keep the balance was less than 1 second, however, this could be improved by fine-tuning of the reward function. From this, regarding the reward function and the state value, it can be said that the proposed method is effective behavior acquisition in situations in which conventional methodologies are not very effective.



**Fig. 6.** The transition of the reward function updated by subject #2

## 4 Conclusions and Future-Works

In this paper, we proposed a novel methodology for behavior learning, called “Coaching”. This method allows a human trainer to intervene the learning process of the agent by giving subjective evaluations such as “good and bad”. The agent updates its own reward function based on the evaluation which is enhanced by considering the characteristics of human evaluation. Then, we confirmed the effectiveness of the proposed method using both simulations and a real robot. The learning task which we used for the experiment, inverted pendulum, had been studied well, and there are various methodologies to achieve the task. However, generally these methodologies includes an element of heuristics for customizing to each task. In particular, our method has an advantage in the designing of the reward function, which allows users to modify the function during the learning phase.

Currently, we are working on the development of a universal interface device for Coaching towards the agent having real body in the real environment, and refinement of the learning algorithm which can utilize negative evaluations.

## References

1. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction, A Bradford Book. MIT Press, Cambridge (1998)
2. Inamura, T., Tanie, H., Nakamura, Y.: From Stochastic Motion Generation and Recognition to Geometric Symbol Development and Manipulation. In: Proc. of IEEE-RAS Int'l. Conf. on Humanoid Robots (2003)
3. Young, J.E., Igarashi, T., Sharlin., E.: Puppet Master: Designing Reactive Character Behavior by Demonstration. In: Proc. SCA 2008, pp. 183–191 (2008)
4. Nakatani, M., Suzuki, K., Hashimoto, S.: Subjective-Evaluation Oriented Teaching Scheme for a Biped Humanoid Robot. In: Proc. of IEEE-RAS Intl. Conf. on Humanoid Robots (2003)
5. Riley, M., Ude, A., Atkeson, C., Chang, G.: Coaching: An Approach to Efficiently and Intuitively Create Humanoid Robot Behaviors. In: Proc. of IEEE-RAS Intl. Conf. on Humanoid Robots, pp. 567–574 (2006)
6. Doya, K.: Reinforcement learning in continuous time and space. Neural Computation 12, 219–245 (2000)
7. Kamatani, H., Kitayama, K., Fujimura, A., Abe, K.: Reinforcement Learning in Continuous State Space. In: SICE Tohoku Chapter Workshop, pp. 229–11 (2006) (in Japanese)

# Cooperation of AGVs' Head-on Collision Avoidance by Knowledge Exchange in Autonomous Decentralized FMS

Hidehiko Yamamoto<sup>1</sup> and Takayoshi Yamada<sup>2</sup>

Department of Human and Information System , Gifu University, Japan  
`{yam-h, yamat}@gifu-u.ac.jp`

**Abstract.** This paper describes the method of cooperation by knowledge exchange in automated guided vehicles (AGVs) moving autonomously in autonomous decentralized flexible manufacturing systems (AD-FMSs). The method gives the AGV an individual knowledge called AGV-knowledge, and by the exchange of which, each AGV can avoid collisions. This method does not use the conventional control by a host computer but applies communication among AGVs. Head-on collisions were prevented by applying this method to 9 types of FMSs constructed in a computer.

**Keywords:** Autonomous decentralized system, FMS, AGV, Head-on collision avoidance.

## 1 Introduction

Autonomous decentralized flexible manufacturing systems (AD-FMSs), one of the next generation intelligent production systems have been studied previously [1]-[3]. These studies on production planning include agent systems where agents act independently in order to realize an AD-FMS and achieve high production efficiency [4],[5]. However, for each agent to be able to act independently, the agents require not only the objective of production but also a method to move independently. This means that automated guided vehicles (AGVs) in an AD-FMS must move independently but in coordination with other AGVs. The conventional FMS operates according to the host computers' orders. AGVs also move according to the prescheduled plans made by the host computers. When an FMS becomes more complicated, such that the number of AGVs increase and have two-way routes, AGVs will have a high possibility of colliding if they move according to prescheduled orders. This is because each machining center (MC) of an FMS in a real factory does not always finish its jobs on schedule and AGVs sometimes break down. That is, a few AGVs may close in on the same point at the same time because very few AGVs' actual locations match the prescheduled locations.

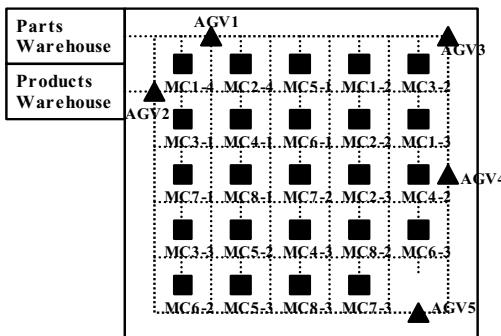
This paper proposes a method to solve the problem of possible collisions within an AD-FMS by enabling AGVs to exchange knowledge and cooperate with each other. The AD-FMS adopts the condition of AGVs' moving in two directions on a guide line. This paper describes the knowledge that AGVs hold, the knowledge that is exchanged between AGVs to cooperate and the method of exchanging that knowledge.

Various methods for controlling AGVs motion have been studied [6],[7]. These include a priori path optimization of right-of-the-way determination and rules control.

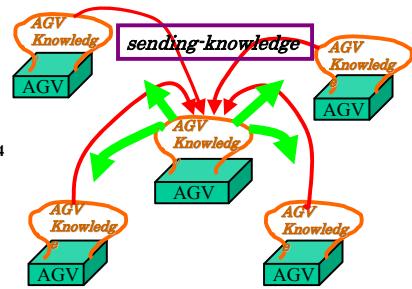
These are different, however, from our research: controlling AGV motion by knowledge exchange.

## 2 Factory Infrastructure Conditions

AD-FMS research considers the following infrastructure conditions. AGVs can move at a uniform speed on a lattice road as shown in Fig. 1. When an AGV passes an intersection, it can recognize the intersection coordinates. At the same time, the AGV sends some of the information that it holds to other AGVs. The other AGVs also receive this information.



**Fig. 1.** Autonomous decentralized FMS layout



**Fig. 2.** AGVs' communication by knowledge

## 3 Collision Avoidance by Knowledge Exchange and Cooperation

In present study, each AGV in the AD-FMS works as an agent of the autonomous decentralized system, and each agent cooperates by knowledge exchange, as shown in Fig. 2. The cooperative actions that comprise the following three elements are proposed: AGV-knowledge (knowledge that each AGV holds), sending-knowledge (knowledge that an AGV sends to other AGVs), and each AGV's consultations about the route and avoid collisions by exchanging sending-knowledge .

### 3.1 AGV Knowledge

AGV-knowledge is a set of knowledge to move independently and has four components, as shown in Fig. 3: (1) path-knowledge, (2) self-knowledge, (3) neighbor-knowledge, and (4) emergency-knowledge. Each AGV holds its AGV-knowledge. The four types of knowledge are defined as follows.

[Definition] path-knowledge: path-knowledge expresses a set of elements indicating sequential intersection coordinates along which the AGV moves. For example, when an

AGV receives a part, it memorizes the sequential intersection coordinates indicating the shortest route to take the part to the MC.

[Definition] self-knowledge: self-knowledge is knowledge about the AGV itself and consists of four elements: self-name, last-crossing, next-crossing, and following-crossing. Self-name expresses the AGV's own name, last-crossing expresses the coordinates of the intersection that the AGV passed through last, next-crossing expresses the coordinates of the intersection that the AGV moves through next, and following-crossing expresses the coordinates of the intersection to which the AGV moves two points ahead. Self-knowledge is expressed as a list below.

$$\text{self-knowledge} = ((\text{self-name}) (\text{last-crossing}) (\text{next-crossing}) \\ (\text{following-crossing})) \quad (1)$$

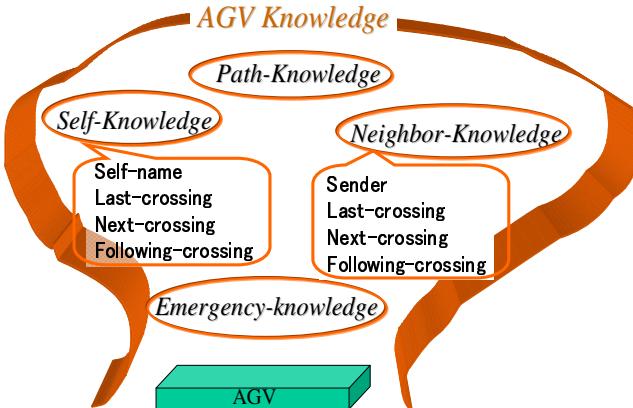
Each AGV recognizes its current intersection (or path) by checking the second and third elements. Each AGV also recognizes the following-path along which it moves by checking the third and fourth elements. AGVs generate and update their self-knowledge when they pass through an intersection.

[Definition] neighbor-knowledge: It is difficult for AGVs to move independently if they do not hold other AGVs' information. However, it is not feasible for each AGV to hold all information about every other AGV. To solve this problem, AGVs are given the minimum knowledge on the current and subsequent paths of the AGV's. This knowledge is called neighbor-knowledge and consists of four elements. The first element is sender (the name of other AGVs). The second and subsequent elements are expressed as the same intersection coordinates as self-knowledge elements. That is, the four elements for one AGV, as described in Equation (2), are considered as a subset of AGV-knowledge, and each AGV holds as neighbor-knowledge all subsets of other AGVs' knowledge for AGVs moving in the AD-FMS. Neighbor-knowledge is created by the received knowledge as described below.

$$\text{neighbor-knowledge} = (((\text{sender-1}) (\text{last-crossing1}) (\text{next-crossing1}) (\text{following-crossing1})) \\ ((\text{sender-2}) ((\text{last-crossing2}) (\text{next-crossing2}) \\ (\text{following-crossing2}))) \dots ) \quad (2)$$

[Definition] emergency-knowledge: emergency-knowledge is a set of knowledge that is memorized when emergency sending-knowledge (hereafter, E-sending-knowledge) is received. Each AGV holds the number of received E-sending-knowledge. Except for the empty element of emergency-knowledge, AGVs recognize that a collision avoidance action is being executed somewhere.

As described above, for the AGVs to move independently—in other words, to recognize which passage it (AGV) itself is now on and which paths other AGVs are now on—they hold three types of knowledge: knowledge about whole path from the start to the goal, path-knowledge, self-knowledge, and neighbor-knowledge, and when needed, emergency-knowledge to avoid collisions. By means of these knowledges, AGVs can autonomously move, judge, and avoid collisions when needed.



**Fig. 3.** AGV-knowledge

### 3.2 Sending-Knowledge and Message-Knowledge

AGVs can move autonomously by sending some of their AGV-knowledge to others. This information is called sending-knowledge and is sent from an AGV to other AGVs in the form of a radio broadcast, as shown in Fig. 2. Most sending-knowledge is intersection coordinates; however, just sending intersection coordinates does not always correspond to agent cooperation. A certain will that an AGV holds has to be sent if we want to consider it agent cooperation. One characteristic of our research is that sending-knowledge includes exchanging wills.

Sending-knowledge is created and sent when one of the following two conditions is satisfied.

Condition 1: AGV passes through the intersection.

Condition 2: AGV receives sending-knowledge from another AGV.

Sending-knowledge created in condition 1 corresponds to informing the AGV of its current location and is particularly called ordinal sending-knowledge (O-sending-knowledge). Sending-knowledge created in condition 2 is formed when a collision is about to occur and is called emergency sending-knowledge (E-sending-knowledge).

Ordinal sending-knowledge consists of four elements. The first element (sender) is the AGV corresponding to the sending AGV. The second and later correspond to the intersection coordinates. That is, the second indicates the intersection coordinates that the sender has just transmitted, and the intersection is called last-crossing. The third indicates the intersection coordinates that the sender moves to next, and the intersection is called next-crossing. The fourth indicates the intersection coordinates on which the sender will move two intersections later and is called following-crossing. Equation (3) shows the ordinal sending-knowledge. When an AGV receives this ordinal sending-knowledge, the AGV memorizes it as an element of neighbor-knowledge in AGV-knowledge.

$$\text{O-sending-knowledge} = ((\text{sender}) \ (\text{last-crossing}) \ (\text{next-crossing}) \ (\text{following-crossing})) \quad (3)$$

E-sending-knowledge does not hold information for intersection coordinates but holds certain messages expressing the wills of AGVs, which are as indispensable as the coordinates between AGVs. The first element of the knowledge is sender name, the second is the content to inform (message), the third is the name of the AGV that receives (addressee), and the fourth and the subsequent are the intersection coordinates. Based on the type of message, the coordinates described as the fourth element differ. The second element, message, expresses the will of AGVs, and the following eight terms are the corresponding possible messages: (1) Ask, (2) Answer, (3) Ask-again, (4) Answer-again, (5) Reverse, (6) Clear, (7) After-you, (8) Thanks.

Emergency-knowledge describes one of the eight messages that can be sent as emergency-knowledge. Equation (4) expresses emergency-knowledge. Intersection coordinates are described in the blanks of the parentheses and the number of the parentheses depend on the types of messages.

$$\text{E-sending-knowledge} = ((\text{sender}) \ (\text{message}) \ (\text{address}) ( \ ) .....( \ )) \quad (4)$$

Each message for emergency-knowledge is defined below.

[Definition] Ask: Ask is the message, “What are the three intersections ahead of you?” Usually, AGV store neighbor-knowledge, which includes the last intersection coordinates that other AGVs have passed through, the coordinates of the intersection immediately ahead, and those of the intersections two points ahead. As a result of this memory, AGVs can recognize the current paths of other AGVs and the next paths that other AGVs will move on. In addition, when needed, the message Ask can express the will to know the following path. For example, E-sending-knowledge = ((AGV-1) (Ask) (AGV-2)) is sent and the meaning is AGV-1 asks AGV-2 which intersection AGV-2 passes three intersections ahead.

[Definition] Answer: Answer holds the message that expresses the coordinates of the third intersection ahead. When an AGV receives the message whose addressee is the AGV itself, the AGV returns the message as Answer. For example, when E-sending-knowledge = ((AGV-2) (Answer) (AVG-1) (1, 3)) is sent, it means that AGV-2 informs AGV-1 that the coordinates of the third intersection ahead of AGV-2 are (1, 3).

[Definition] Ask-again: Ask-again is the message to ask for the coordinates of an intersection one intersection ahead of the intersection asked for with Ask.

[Definition] Answer-again: When an AGV receives the message with Ask-again and the addressee that the message includes is itself, the AGV returns the message as Answer-again. The message Ask-again gives the coordinates the intersection one point ahead of the intersection at which the Ask message is sent.

[Definition] Reverse: Reverse is the message to inform others, “I’m moving backward to avoid a collision.” The message Reverse is the most urgent message, and if an AGV receives the message, other AGVs must not interfere with its movement.

[Definition] Clear: Clear is the message to inform others, “Please move backward because there is no problem.”

[Definition] After-you: After-you is the message to inform others, “You can move first.”

[Definition] Thanks: Thanks is the message to inform others that “The situation is over” as the reverse movement is finished. This means that when Thanks is received, all AGVs return to their normal situation.

### 3.3 Cooperation by Exchanging Sending-Knowledge

AGVs always repeat their basic and fixed processes to execute cooperation actions. These processes are called routine AGV processes (RAP) and the following are the details.

#### [RAP]

Step1-1: When an AGV updates its self-knowledge, go to Step1-2, and when it receives emergency-knowledge, go to Step1-3.

Step1-2: The AGV compares its self-knowledge with its neighbor-knowledge and judges the possibility of collision in the following manner. Judge whether the third and the fourth elements of self-knowledge are the same as the fourth and third elements of each neighbor-knowledge, and if the same is found, go to Step 1-2-2; if not, go to Step 1-2-1.

Step1-2-1: The AGV finishes RAP.

Step1-2-2: The AGV, to judge the possibility for a collision, stops at the middle of the current path, creates emergency-knowledge including the message Ask, sends it, and finishes RAP.

Step 1-3: The AGV creates emergency-knowledge according to the received message, sends it, and finishes RAP.

In this manner, AGVs check the possibility for a collision by performing RAP whenever they update self-knowledge and receive sending-knowledge. Foreseeing a possible collision in Step 1-2, AGVs start to exchange emergency-knowledge. Here, we use the term foreseeing since a collision does not always occur at this step.

### 3.4 Message Exchange for Head-On Collisions

AGVs foresee the possibility for a collision by RAP described in Section 3.3. Several patterns for collisions are considered, such as head-on collisions and flank collisions. This section deals with a head-on collision, the most popular collision pattern, and explains how messages are exchanged to avoid collisions. The strategy for collision avoidance involves two processes: [a] an AGV remains stationary there if the other AGVs get into its current path, and [b] in the case where the other AGVs get into its current path, the AGV moves backward and waits till the other AGVs pass through. The strategy carries out the two processes in the sequence [a] → [b].

In Step1-2-2 of RAP, when an AGV foresees the possibility of collision, it starts to exchange emergency-knowledge based on the following rules of exchange (ROE) for avoiding head-on collisions. Priority AGV and Concessive AGV are determined. The former means the AGV that is given the priority to move in order to avoid collisions,

and the latter means the AGV that makes way for the Priority AGV to avoid collisions. To be specific, the AGV that has updated self-knowledge in Step 1-2-2 of RAP will be changed to the Concessive AGV. Then, part of ROE adopts certain terms as well as the eight messages described in Section 3.2. The terms are defined as below.

[Definition] description style {[S] → [V] [O]}: The letter in the parenthesis before “→” corresponds to subject and the letters in the parenthesis after “→” correspond to verb and object. The composition SVO has several permutations of verb and object. Further, the symbol “+” expresses additional actions. For example, {[A] → ([message-1] [B] + [Act-1])} means A sends message-1 to B and executes the action, Act-1.

[Definition] Stay-1: Move onto the middle of the current path and wait.

[Definition] Check-1: Check whether the two intersections ahead of the current path of the Concessive AGV and the path of the Priority AGV are the same or not. To be specific, check whether the third and the fourth elements of the Concessive AGV’s self-knowledge and the coordinates of the Priority AGV’s two intersections ahead and three intersections ahead, which are the intersection coordinates included in the message Answer, are the same.

[Definition] Find-1: Find the path that is different from the path three intersections ahead of the Priority AGV. The path found corresponds to an evacuated path.

[Definition] Check-2: Check whether the evacuated path and the one intersection ahead path are the same or not. If same, wait on the middle of the current path.

[Definition] Move-1: Move on the middle of an evacuated path and stay there.

[Definition] Move-2: Move out as scheduled.

[ROE]

Rule-1: *if* {an AGV foresees the possibility of a collision}, *then* {[Concessive AGV] → [Ask] [Priority AGV] + [Stay-1]}

Rule-2: *if* {an AGV receives Ask}, *then* {[Priority AGV] → [Answer] [Concessive AGV] + [Stay-1]}

Rule-3: *if* {an AGV receives Answer}, *then* {[Concessive AGV] → [Check-1] + [Ask-again] [Priority AGV]} **or**  
{[Concessive AGV] → [Check-1] + [After-you] [Priority AGV]}

Rule-4: *if* {an AGV receives Ask-again}, *then* {[Priority AGV] → [Answer-again] [Concessive AGV]}

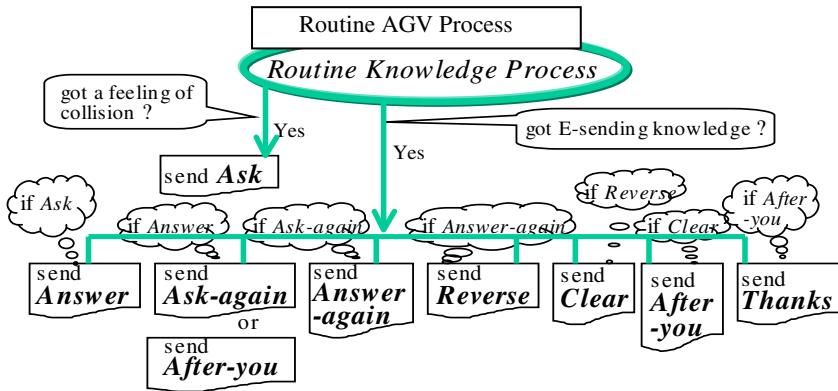
Rule-5: *if* {an AGV receives Answer-again}, *then* {[Concessive AGV] → [Find-1] + [Reverse] [all AGVs]}

Rule-6: *if* {an AGV receives Reverse}, *then* {[all AGVs] → [Check-2] + [Clear] [Concessive AGV]}

Rule-7: *if* {an AGV receives Clear}, *then* {[Concessive AGV] → [Move-1] + [After-you] [Priority AGV]}

Rule-8: *if* {an AGV receives After-you}, *then* {[Priority AGV] → [Move-3] + [Thanks] [all AGVs]}

Fig.4 shows the relationship between RAP and ROE. The detailed procedures of ROE are described below.



**Fig. 4.** Routine AGV process and message exchange rules

#### 4 Simulation Application Examples for Head-on Collisions

The idea of cooperation by exchanging knowledge proposed in Section 3 was programmed and applied for an operating AD-FMS. Nine types of AD-FMSs from Style-1 to Style-9 were adopted and the differences are shown in Table 1. The table lists (from left to right) AD-FMS types (Styles), part types, number of MCs (<sub>g</sub>MC), and number of AGVs. The number outside the parenthesis in the MC column is the type number of MC and the number inside the parenthesis is the MC number for each type. For example, the part type of Style-9 is 9, MC numbers are 8 and three MCs for each MC type are located in the AD-FMS. Fig. 1 shows the layout of Style-1. The layouts of other FMSs have a smaller number of MCs than Style-9. Each part has a different machining time for each MC. Table 2 shows the machining times and process sequences for each part of Style-4 through Style-9. The sequence from the top to bottom of Table 2 indicates the process sequence of the part being machined. The examples adopted unpredicted troubles in operating the AD-FMS, which were not dealt with by preplanned schedules. The adopted troubles have two conditions. One is that each AGV breaks down randomly three times in 24 hours and the AGV stops for five minutes each time. The other is that each MC randomly finishes its machining after a delay of 10%.

The AGVs' collision times were examined after operating each AD-FMS for 24 hours. As Due to space limitations, the simulation results of Style-3, Style-6, and Style-9 are shown in Tables 3, 4, and 5. The results correspond to the values of developed system in Tables 3, 4, and 5. The tables also include each part output and the

occurrences of collision avoidance. The times indicate how many AGVs change paths and wait to avoid a collision. The results of the conventional system that does not adopt knowledge-exchange cooperation are also described in the right-most column of the tables. For example, the outputs of the developed system of Table 3 are 200 for part\_1, 242 for part\_2, and 80 for part\_3. The collision avoidance times are 1,879, which indicate that a large number of avoidances occurred for 24 hours. Collisions occurred 984 times. On the contrary, the conventional system's outputs are 221 for part\_1, 266 for part\_2, and 80 for part\_3, and collisions occurred 1,488 times.

**Table 1.**

Styles	Parts kinds	gMC(MC)	AGV
1	3	3(1,1,1)	3
2	3	3(1,2,1)	3
3	3	3(2,2,2)	3
4	6	6(1,1,1,1,1,1)	5
5	6	6(2,2,2,2,2,2)	5
6	6	6(3,3,3,3,3,3)	5
7	9	8(1,1,1,1,1,1,1)	5
8	9	8(2,2,2,2,2,2,2)	5
9	9	8(3,3,3,3,3,3,3)	5

decreased from 1,488 to 984 owing to the knowledge-exchange cooperation. As for Style-6 and Style-9, the number of collisions reduced from 3,792 to 1,776 and from 3,336 to 1,896, respectively. The other styles that are not shown in the tables had the same reduced results, and all simulation results indicate that the collision times were reduced. The results signify that ROE and the research to adopt knowledge-exchange cooperation are effective.

Although many collisions occurred, after investigations it was found that they did not correspond to head-on collision but to other types of collisions such as flank collisions and rear-end collisions.

## 5 Conclusions

The research dealt with the knowledge-exchange cooperation for autonomously moving AGVs, which is essential for operating an AD-FMS. The proposed method avoids collisions by adopting three steps: (1) each AGV is given knowledge, (2) by exchanging knowledge, some AGVs change to AGVs that have the priority to move and some that make concession moves,

which indicate that a large number of avoidances occurred for 24 hours. Collisions occurred 984 times. On the contrary, the conventional system's outputs are 221 for part\_1, 266 for part\_2, and 80 for part\_3, and collisions occurred 1,488 times. The number of collisions in the developed system, when compared to the conventional system that does not adopt the collision avoidance measures,

**Table 2.** Simulation results of Style -9

	Deve loped system	C on ventional system
Product output	Part_1	50
	Part_2	71
	Part_3	36
	Part_4	35
	Part_5	23
	Part_6	12
	Part_7	46
	Part_8	60
	Part_9	23
Number to avoid collisions	<b>2,491</b>	-----
Number of collisions	<b>1,896</b>	<b>3,336</b>

and (3) between these AGVs, collision avoidance is executed by exchanging knowledge.

This method involves each AGV holding four types of AGV knowledge: path-knowledge, self-knowledge, neighbor-knowledge, and emergency-knowledge. Furthermore, AGV collisions are autonomously avoided by exchanging the sending-knowledge, including the messages of wills of AGVs. The method was applied to nine types of AD-FMSs built in a computer and complete preventions of all head-on collisions were confirmed.

## References

- [1] Sugimura, N., Shrestha, R., Inoue, J.: Integrated process planning and scheduling in holonic manufacturing systems -Optimization based on shop time and machining cost-. In: Proc. of the 2003 IEEE International symposium on Assembly and task planning (ISATP2003), pp. 36–41 (2003)
- [2] Kouiss, K., Pierreval, H., Mebarki, N.: Using multi-agent architecture in FMS for dynamic scheduling. *Journal of Intelligent Manufacturing* 8(1), 41–47 (1997)
- [3] Moriwaki, T., Hino, R.: Decentralized Job Shop Scheduling by Recursive Propagation Method. *International Journal of JSME, Series C* 45(2), 551–557 (2002)
- [4] Yamamoto, H., Ramli, R.B.: Real-time Decision Making of Agents to Realize Decentralized Autonomous FMS by Anticipation. *International Journal of Computer Science and Network Security* 6(12), 7–17 (2006)
- [5] Yamamoto, H., Ramli, R.B.: Real-time control of decentralized autonomous flexible manufacturing systems by using memory and oblivion. *International Journal of Intelligent Information and Database Systems* 1(3/4), 346–355 (2007)
- [6] Berman, S., Edan, Y.: Decentralized autonomous AGV system for material handling. *International Journal of Production Research* 40(15)
- [7] Wallace, A.: Application of AI to AGV control. *International Journal of Production Research* 39(4), 709–726 (2001), 3995–4006 (October 2002)

# A Log Analyzer Agent for Intrusion Detection in a Multi-Agent System<sup>\*</sup>

Iago Porto-Díaz, Óscar Fontenla-Romero, and Amparo Alonso-Betanzos

Department of Computer Science, University of A Coruña, Spain  
`{iporto,ofontenla,ciamparo}@udc.es`

**Abstract.** In this work, the design and implementation of a log analyzer agent is described. This agent is conceived to act as a part of a multi-agent Intrusion Detection System. The agent analyzes log files of services, applications or operating systems contrasting every log line with a set of security rules defined by experts. These rules can be created using a new easy to use XML-based format founded on an object-oriented model. Whenever a security match is found, the agent sends a security report to the next level of the multi-agent system using the IDMEF (Intrusion Detection Message Exchange Format) and the IDXP (Intrusion Detection Exchange Protocol).

## 1 Introduction

The development of Intrusion Detection Systems (IDS) has flourished along with the spreading of computer networks and the Internet. An increasing number of services and a growing number of businesses are available every day with plenty of potential. As a consequence, intrusion detection has become increasingly difficult because of the complexity of the domain and the great heterogeneity of computer networks. According to these circumstances, one can infer some desirable features in IDSs [1]: high adaptability and configurability, fault tolerance and resistance to attacks. These features are hardly fulfilled by traditional IDSs, which present certain weaknesses [2]: the central analyzer is a single point of failure; scalability is limited, it is difficult to reconfigure or add capabilities to the IDS and the analysis of network data can be flawed. Moreover, dealing with fragmentation or denial of service attacks and real time processing have proved to be tenacious challenges.

The idea of intrusion detection was first introduced by J.P. Anderson [3] in 1980. He defined an intrusion as a deliberate unauthorized attempt to access information, manipulate information or make a system unreliable or unusable. Since then, research products such as EMERALD (Event Monitoring Enabling

---

\* This work was supported in part by Spanish Ministerio de Ciencia e Innovación under Project Code TIN 2006-02402 and by Xunta de Galicia under Project Code PGIDIT06PXIB105205PR and under the program “Axudas para a consolidación e a estruturación de unidades de investigación competitivas” (code 2007/134), all of them partially supported by the European Union ERDF.

Responses to Anomalous Live Disturbances) [4], NetSTAT [5] and Bro [6] have been developed, as well as commercial products such as CMDS (Computer Misuse Detection System), NetProwler, NetRanger, Centrax and Real Secure [7].

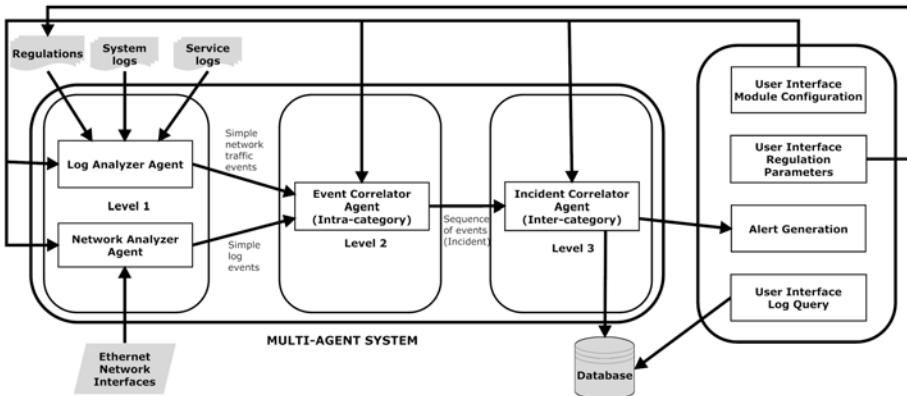
A multi-agent system is composed of multiple interacting intelligent agents, which are at least partially autonomous, local and decentralized [8]. Using agents in IDSs may help to reduce the network load, providing higher scalability, autonomy and dynamic adaptation. One of the most relevant agent-based research products is AAFID (Autonomous Agents For Intrusion Detection) [9]. In this architecture agents are arranged in a hierarchical tree-based structure. The model is composed of agents, filters, transceivers and monitors. New models have been proposed based on this architecture, e.g. [9] and [10].

When it comes to ensure security in a system, log files are one of the most important sources of information. Applications and operating systems use log files not only for security matters but also for recovery and performance studies. By using log services, an application can record a detailed relation of its events of execution, both errors and messages of normal operation. In this manner, log files can be examined to detect unauthorized or suspicious patterns that may represent intrusions. The main drawback with log files is that they hastily become crowded with irrelevant messages, which makes the intrusion detection process very difficult. Therefore, a log analyzer agent is one of the crucial parts of an agent-based IDS.

The objective of this work is the design and implementation of a log analyzer agent as a part of a multi-agent system for intrusion detection. This paper is divided as follows. In Sect. 2 the whole multi-agent system is described. Section 3 focuses on the design of the log analyzer agent. Finally, in Sect. 4 the conclusions obtained are exposed.

## 2 Description of the Multi-Agent System

The agent described in this work is a part of a multi-agent system under development which main schema is shown in Fig. 1. The system works hierarchically in three levels. Level one contains the agents which receive inputs from the outside. These are the log analyzer agent, which is the focus of this work and will be seen with more details later on, and the network analyzer agent. Their job is performing a preliminary detection upon the external inputs. Specifically, the network agent analyzes network packages acquired through a sniffer and confronts them with known signatures of attacks. The log analyzer agent monitors service logs, application logs and operating system logs. Level two consists of the event correlator agent, which receives output events from level one and tries to correlate them from a semantic point of view, generating incidents, which are delivered to level three. This level comprises the incident correlator agent, that is in charge of looking for relationships between incidents and generating user friendly alerts according to these. The internal architecture of levels two and three has not been decided yet, but could be e.g. rule-based, as level one, probabilistic, as in [12] or state-based, as in [13]. Using multiple agents in levels two and three by means of a voting scheme, as in [14], could also be taken into consideration.



**Fig. 1.** Multi-agent system architecture

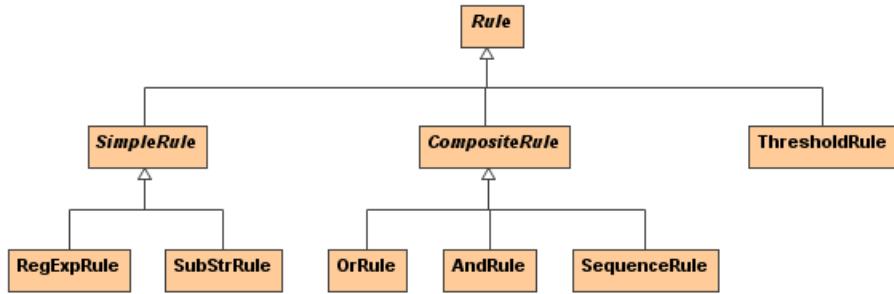
In accordance with the presented design, a new need arises: defining a common language for the communication between agents. The sources of information are very heterogeneous, as they comprehend several different log files and network data. Because of this, getting the information integrated is far from a trivial task. The proposed multi-agent system utilizes the data format IDMEF (Intrusion Detection Message Exchange Format), defined by the IETF (Internet Engineering Task Force) [1], described in the RFC (Request For Comments) 4765 [15] combined with the application-layer protocol IDXp (Intrusion Detection Exchange Protocol), described in the RFC 4767 [16]. This communication schema uses an object-oriented data model and is implemented in XML (Extensible Markup Language) [17].

### 3 Design of the Log Analyzer Agent

From this point on, this work will be focused on describing the design and implementation of the log analyzer agent. It is intended to function either as stand-alone software or as a component of a multi-agent system for intrusion detection. Its purpose is to contrast – in real time – a log file with a series of rules previously defined by security experts.

Several instances of the log analyzer can be found running at once in a single system. Each of them will look after the specific logs of each sensitive application of the system. For this reason, the design of the agent presented in this paper will be carried out from a general perspective, in a way that permits an easy application to any log file format of any application or operating system by the user. The only specific part to each concrete application is the description of security rules. Therefore, the examples provided in this work will be focused on a commonly used web server, such as the *Apache HTTP Server* and the *Common Log Format*.

As stated before, the behavior of the log analyzer agent is based on contrasting log entries against rules defined by experts, which are stored in a rule-base.



**Fig. 2.** Rules diagram

To define the format of the rules, an object-oriented approach has been followed and a subsequent XML implementation has been carried out.

Figure 2 shows a diagram of the defined hierarchy of rules. Three main types of rules are considered:

- Simple rules affect just one line of the log file and can be defined by means of a regular expression or a substring. It is this regular expression or substring which describes the information the rule aims to detect in the log file.
- Composite rules are needed to model relationships between rules, e.g. temporal relationships. Thus, they affect more than just one log line and they are made up of one or more rules, which in turn can be simple or composite. Thus, a composite rule is a composition tree in itself. Three types of composite rules have been defined: AND rule, OR rule and SEQUENCE rule.
  - A composite rule AND will be triggered when all of its components are satisfied.
  - An OR rule will be triggered when at least one of its components is satisfied.
  - A SEQUENCE rule will be triggered when all of its components are satisfied, in a determined temporal order.
- Threshold rules detect several occurrences of the same rule and get triggered when its component rule is satisfied a specific number of times.

To identify component rules that can be related with others, a session identifier is used. For instance, a composite rule can be made up of three simple rules, but they may be wanted to be referred to the same client IP address. In this case, the session identifier would be the client IP address, and the composite rule would be triggered when all three simple rules matched, but with the same session identifier (same client IP address).

Algorithm 1 shows how the agent works. Upon the arrival of a new log line, all simple rules and component rules in the rule base are checked for matches. All matching rules are stored in the agenda. In a second step, the algorithm iterates over the agenda looking for match completions on composite rules. Every time a rule is triggered, a security report is emitted.

---

**Algorithm 1.** Log analyzer algorithm

---

Inputs: rule base, log line.

1. foreach rule R in rule base,
    - (a) if R is a substring rule,
      - i. Check whether the log line contains the substring or not.
    - (b) else if R is a regular expression (regexp) rule,
      - i. Match the log line against the regular expression.
    - (c) if match found,
      - i. Add R to agenda
  2. foreach rule S in agenda,
    - (a) if S is not a component of a composite rule,
      - i. Report security match.
    - (b) if S is a component of a composite rule,
      - i. Mark the component rule S as active.
      - ii. Recursively check if the root composite rule is active.
      - iii. If the root composite rule is active,
        - A. Report security match.
- 

Rules are implemented using a XML-based format. Using XML is advantageous because it is very portable and the parser is a standard component, which prevents bugs and accelerates the development. Moreover, it is easily extendable just by adding new tags and its structure is easy to be learned and processed by a third party. In the following subsections, the XML tags of the rule definition format are described thoroughly.

### 3.1 Element <rules>

The element <rules> defines the root element in a document of rule definitions. It contains the definitions of all the rules defined in the system. Its syntax is presented below and table 1 shows the description of its attributes.

```
<rules xmlns=anyURI xmlns:xsi=anyURI xsi:schemaLocation=string>
  (simpleRule|compositeRule|thresholdRule)*
</rules>
```

The example below shows how the rules element would look like using a fictional namespace and location of the XML schema.

```
<rules xmlns=http://www.dc.fi.udc.es/lidia/xml
  xmlns:xsi=http://www.w3.org/2001/XMLSchema-instance
  xsi:schemaLocation="http://www.dc.fi.udc.es/lidia/xml ruleset.xsd">
  ...
</rules>
```

**Table 1.** Attributes of the element <rules>

Attribute	Description
xmns	Mandatory. Declaration of the default namespace.
xmns:xsi	Mandatory. Declaration of the namespace <i>xsi</i> , needed to declare the attribute <i>xsi:schemaLocation</i> .
xsi:schemaLocation	Mandatory. Association of a <i>XML Schema</i> with a namespace.

### 3.2 Element <simpleRule>

The element <simpleRule> defines a simple rule (*SimpleRule*) by indicating its name, description, type – substring or regular expression – and the pattern to be detected. Its syntax is presented below and table 2 shows the description of its attributes.

```
<simpleRule name=token desc=string
ptype=SubStr|RegExp pattern=string>
</simpleRule>
```

**Table 2.** Attributes of the element <simpleRule>

Attribute	Description
name	Mandatory. Descriptor name of the rule.
desc	Facultative. Informal description of the rule.
ptype	Mandatory. Type of simple rule. It can take as value <i>SubStr</i> – the rule tries to detect a substring – or <i>RegExp</i> – the rule is defined by means of a regular expression.
pattern	Mandatory. The pattern that defines the rule. According to the value of <i>ptype</i> , <i>pattern</i> may contain a regular expression or a text substring.

The example below shows a rule that detects intrusions of the Nimda computer worm in the Apache error.log file.

```
<simpleRule name="nimda"
desc="Detects NIMDA attempts" ptype="RegExp"
pattern="File does not exist: .*scripts/root\.exe" />
```

### 3.3 Element <compositeRule>

The element <compositeRule> defines a composite rule (*CompositeRule*). A composite rule can contain other rules, both simple, threshold and composite rules. Besides the name and the description, the operation – AND, OR or SEQ –, the session identifier and the maximum time to live (ttl) are provided. The ttl field indicates the maximum time allowed for the rule to get triggered. Its syntax is presented below and table 3 shows the description of its attributes.

```

<compositeRule name=token desc=string
op=AND|OR|SEQ sessionID=string ttl=int>

(simpleRule|compositeRule|thresholdRule)*

</compositeRule>

```

**Table 3.** Attributes of the element <compositeRule>

Attribute	Description
name	Mandatory. Descriptor name of the rule.
desc	Facultative. Informal description of the rule.
op	Mandatory. Logical operation that relates the subrules of a composite rule. It may take values <i>AND</i> , <i>OR</i> or <i>SEQ</i> , meaning, respectively, an <i>AndRule</i> , an <i>OrRule</i> or a <i>SequenceRule</i> .
sessionID	Facultative. Session identifier used for relating the subrules of a composite rule. This field is dependent on the examined application. For the case of the Combined Log Format, it may take values %h, %r, %t or %{User-agent}i, meaning, respectively, the source IP address, the text of the <i>request</i> , the date/time or the <i>User-Agent</i> field that indicates the browser.
ttl	Facultative. Time to live for the rule in minutes. Indicates the time window for all component rules to be triggered. Otherwise partial occurrences are discarded.

The example below detects typical attacks to Mambo content management system in the Apache error\_log file.

```

<compositeRule name=Mambo desc="Detects Mambo attacks"
op="OR" ttl="60">

<simpleRule name="mambo1" ptype="RegExp"
pattern="&mosConfig_absolute_path=http\:*\\?\&cmd=cd%20/tmp/;
wget%20http\:*mambo.txt;rm%20-rf" />

<simpleRule name="mambo2" ptype="RegExp"
pattern="&mosConfig_absolute_path=http\:*\\?\&cmd=cd%20/tmp/;
wget%20http\:*perl\%20mambo.txt;rm%20-rf" />

</compositeRule>

```

### 3.4 Element <thresholdRule>

The element <thresholdRule> defines a threshold rule (*ThresholdRule*). The attributes required by a threshold rule are name, description, session identifier, time to live, and threshold. The latter indicates the number of times the

component rule has to be activated. A threshold rule must contain a simple or composite rule. Its syntax is presented below and table 4 shows the description of its attributes.

```
<thresholdRule name=token desc=string
sessionID=string ttl=int threshold=int>

(simpleRule|compositeRule)

</thresholdRule>
```

**Table 4.** Attributes of the element <thresholdRule>

Attribute	Description
name	Mandatory. Descriptor name of the rule.
desc	Facultative. Informal description of the rule.
sessionID	Facultative. Session identifier used for relating the occurrences of the component rule. This field is dependent on the examined application. For the case of the Combined Log Format, it may take values %h, %r, %t or %{User-agent}i, meaning, respectively, the source IP address, the text of the <i>request</i> , the date/time or the <i>User-Agent</i> field that indicates the browser.
ttl	Facultative. Time to live for the rule in minutes.
threshold	Mandatory. Number of times the rule has to be activated.

The example below shows a rule that detects five failed login attempts within fifteen minutes in the Apache error.log file.

```
<thresholdRule name="5loginFailures"
desc="Detects 5 login failures in 15 minutes"
threshold="5" ttl="15">

  <simpleRule name="loginFailure" desc="Detects login failures"
  ptype="RegExp"
  pattern="authentication failure for ".*";
  : password mismatch" />

</thresholdRule>
```

## 4 Conclusions

The log analyzer agent described with details in this work is a part of an IDS based on agent technologies. It has been designed to ease configurability and extensibility. In this manner, it is very simple for an expert to create various rule sets for several services, applications and operating systems. Moreover, the

use of a XML-based format for describing the rules grants universality and standardization, in such a way that the created rules are easy to read, debug and modify. Examples of rules related to the Apache web server are provided.

Besides the log analyzer agent, the multi-agent system consists of some other agents: a network analyzer agent, an event correlator agent and an incident correlator agent, all of which communicate with each other using the IDMEF format and the IDXp protocol.

As future work, the network analyzer agent and the agents of levels two and three will be developed, and the whole system will be validated by injecting a number of intrusions into the system and finding out its success rate. A study of the performance will be carried out, comparing the proposed method with other state-of-the-art techniques.

## References

1. Spafford, E., Zamboni, D.: Intrusion Detection Using Autonomous Agents. *Computer Networks* 34(4), 547–570 (2000)
2. Balasubramaniyan, J.S., Garcia-Fernandez, J.O., Isacoff, D., Spafford, E., Zamboni, D.: An Architecture for Intrusion Detection Using Autonomous Agents. CERIAS Technical Report 98/05 42(7) (July 1999)
3. Anderson, J.P.: Computer Security Threat Monitoring and Surveillance. Technical Report. James P. Anderson Co., Fort Washington PA (April 1980)
4. Porras, P.A., Neumann, P.G.: EMERALD: Event Monitoring Enabling Responses to Anomalous Live Disturbances. In: Proceedings of the 20th National Information Systems Security Conference, pp. 353–365 (1997)
5. Vigna, G., Kemmerer, R.A.: NetSTAT: A Network-based Intrusion Detection System. *Journal of Computer Security* 7(1), 37–71 (1999)
6. Paxson, V.: Bro: A System for Detecting Network Intruders in Real-time. *Computer Networks* 31(23), 2435–2463 (1999)
7. Allen, J., McHugh, J., Fithen, W., Christie, A., Pickel, J.: State of the Practice of Intrusion Detection Technologies, Software Engineering Institute. Carnegie Mellon University, Pittsburgh (2000)
8. Wooldridge, M.: An Introduction to MultiAgent Systems. John Wiley and Sons Ltd., Chichester (2002)
9. Frincke, D.: A Framework for Cooperative Intrusion Detection. In: Proceedings of the 21st National Information Systems Security Conference (1998)
10. Lee, W.: Data Mining and CIDF Based Approach for Detecting Novel and Distributed Intrusions. In: Debar, H., Mé, L., Wu, S.F. (eds.) RAID 2000. LNCS, vol. 1907, pp. 49–65. Springer, Heidelberg (2000)
11. IETF (Internet Engineering Task Force), <http://www.ietf.org> (cited February 2010)
12. Gowadia, V., Farkas, C., Valtorta, M.: PAID: A Probabilistic Agent-Based Intrusion Detection System. *Computers & Security* 24(7), 529–545 (2005)
13. Do-hyeon, L., Doo-young, K., Jae-il, J.: Mobile Agent Based Intrusion Detection System Adopting Hidden Markov Model. In: Gervasi, O., Gavrilova, M.L. (eds.) ICCSA 2007, Part II. LNCS, vol. 4706, pp. 122–130. Springer, Heidelberg (2007)

14. Rehak, M., Pěchouček, M., Bartoš, K., Grill, M., Čeleda, P., Krmíček, V.: CAM-NEP: An Intrusion Detection System for High-Speed Networks. *Progress in Informatics* (5), 65–74 (2008)
15. Debar, H., Curry, D., Feinstein, B.: The Intrusion Detection Message Exchange Format (IDMEF). RFC 4765 (March 2007)
16. Feinstein, B., Matthews, G.: The Intrusion Detection Exchange Protocol (IDX). RFC 4767 (March 2007)
17. XML (Extensible Markup Language), <http://www.w3.org/XML/> (cited February 2010)

# A Proof System for Time-Dependent Multi-agents

Norihiro Kamide

Waseda Institute for Advanced Study,  
Waseda University,  
1-6-1 Nishi Waseda, Shinjuku-ku, Tokyo 169-8050, Japan  
[logician-kamide@aoni.waseda.jp](mailto:logician-kamide@aoni.waseda.jp)

**Abstract.** An extension of linear-time temporal logic (LTL), called an agents-indexed linear-time temporal logic (ALTL), is introduced as a Gentzen-type sequent calculus. ALTL is intended to appropriately express reasoning about time-dependent multi-agents within a proof system. The cut-elimination and completeness theorems for ALTL are shown.

**Keywords:** Linear-time temporal logic, sequent calculus, multi-agents, cut-elimination theorem, completeness theorem.

## 1 Introduction

Verifying and specifying time-dependent multi-agent systems are growing importance in Computer Science, since computer systems are generally used by or composed of time-dependent multi-agents. It is known that *linear-time temporal logic* (LTL) [110] is one of the most useful temporal logics for verifying and specifying time-dependent and concurrent systems. In this paper, an extension of LTL, called an *agents-indexed linear-time temporal logic* (ALTL), is introduced as a *Gentzen-type sequent calculus*, which is known to be a useful proof system for automated theorem proving.

ALTL has some temporal, agent and fixpoint operators. By using these operators, reasoning about time-dependent multi-agents can appropriately be expressed. The proposed fixpoint operator in ALTL can express common knowledge (or information) of multi-agents. ALTL is also regarded as a combination of LTL and a *fixpoint logic*. The cut-elimination and completeness theorems for ALTL are shown as the main results of this paper. These theorems show that ALTL is attractive as a theoretical basis for automated theorem proving about time-dependent multi-agents.

Fixpoint logics (or fixed point logics) are regarded as logics with a *fixpoint* (or *fixed point*) operator. Typical examples of fixpoint logics are *propositional  $\mu$ -calculus* [9], which is more expressive than temporal logics, and *common knowledge logic* [2], which is an extension of multi-agent epistemic logic. It is known that fixpoint logics are useful for representing temporal and knowledge-based reasoning. Combining LTL and a fixpoint logic is thus an attractive issue for

representing time-dependent multi-agent systems. Indeed, a model checking (i.e. model-theoretic) approach to combine LTL and a common knowledge operator has been studied successfully [11].

A cut-free and complete Gentzen-type sequent calculus for combining LTL and a fixpoint logic has been required for providing a theoretical basis for automated theorem proving. However, a Gentzen-type sequent calculus for such a temporal fixpoint logic has not yet been studied. A reason may be that proving the cut-elimination and completeness theorems for such a combined logic is difficult since the traditional formulations of fixpoint operators are rather complex. This paper tries to overcome such a difficulty by introducing a new simple formulation of a fixpoint operator and by using an embedding-based proof method.

In the following, we roughly explain the proposed formulation of the fixpoint operator. The symbol  $\omega$  is used to represent the set of natural numbers. The symbol  $K$  is used to represent the set  $\{\heartsuit_i \mid i \in \omega\}$  of agent modal operators, and the symbol  $K^*$  is used to represent the set of all words of finite length of the alphabet  $K$ . Greek lower-case letters  $\iota$  and  $\kappa$  are used to represent any members of  $K^*$ . The characteristic inference rules for a fixpoint operator  $\heartsuit_c$  are as follows:

$$\frac{\iota\kappa\alpha, \Gamma \Rightarrow \Delta}{\iota\heartsuit_c\alpha, \Gamma \Rightarrow \Delta} (\heartsuit_c\text{left}) \quad \frac{\{\Gamma \Rightarrow \Delta, \iota\kappa\alpha \mid \kappa \in K^*\}}{\Gamma \Rightarrow \Delta, \iota\heartsuit_c\alpha} (\heartsuit_c\text{right}).$$

These inference rules are intended to imply the following axiom scheme:  $\heartsuit_c\alpha \leftrightarrow \bigwedge\{\iota\alpha \mid \iota \in K^*\}$ . This axiom scheme corresponds to the so-called iterative interpretation of common knowledge. Indeed, if we can read  $\heartsuit_i\alpha$  as “agent  $i$  knows  $\alpha$ ,” then we can understand  $\heartsuit_c\alpha$  as “ $\alpha$  is common knowledge of agents.” Suppose that for any formula  $\alpha$ ,  $f_\alpha$  is a mapping on the set of formulas such that  $f_\alpha(x) := \bigwedge\{\heartsuit_i(x \wedge \alpha) \mid i \in \omega\}$ . Then,  $\heartsuit_c\alpha$  becomes a fixpoint of  $f_\alpha$ .

Based upon the interpretation explained above, ALTL has some useful descriptions. An example of such descriptions is:  $G(\heartsuit_c \text{ password} \rightarrow \heartsuit_i F \text{ login})$  which means:

“If the login password of a computer is regarded as common information in the group  $A := \{1, 2, \dots, n\}$  of agents, then an agent  $i$  in  $A$  will eventually be able to login the computer.”

The contents of this paper are then summarized as follows. In Section 2, ALTL is introduced as a Gentzen-type sequent calculus. The cut-elimination theorem for ALTL is proved using a theorem for syntactically embedding ALTL into a sequent calculus  $LK_\omega$  for infinitary logic. In Section 3, an *agent-time indexed semantics*, which is an extension of a semantics for LTL, is introduced for ALTL, and the completeness theorem with respect to this semantics is proved by combining two theorems for syntactically and semantically embedding ALTL into  $LK_\omega$ . In Section 4, this paper is concluded and some related works are addressed.

## 2 Sequent Calculus and Cut-Elimination

Let  $n$  be a fixed positive integer. Then, the symbol  $N$  is used to represent the set  $\{1, 2, \dots, n\}$  of agents. The following list is adopted for the language  $\mathcal{L}$

of the underlying logic: (countable) propositional variables,  $\rightarrow$  (implication),  $\neg$  (negation),  $\wedge$  (conjunction),  $\vee$  (disjunction),  $\heartsuit_i$  ( $i \in N$ ) (agent  $i$  knows),  $\heartsuit_c$  (least fixpoint or common knowledge),  $\heartsuit_d$  (greatest fixpoint),  $X$  (next-time),  $G$  (globally in the future) and  $F$  (eventually in the future). Small letters  $p, q, \dots$  are used to denote propositional variables, Greek lower-case letters  $\alpha, \beta, \dots$  are used to denote formulas, and Greek capital letters  $\Gamma, \Delta, \dots$  are used to represent finite (possibly empty) sets of formulas. An expression  $\circ\Gamma$  where  $\circ \in \{\heartsuit_i \mid i \in \omega, \heartsuit_c, \heartsuit_d, X, G, F\}$  is used to denote the set  $\{\circ\gamma \mid \gamma \in \Gamma\}$ . An expression  $A \equiv B$  denotes the syntactical identity between  $A$  and  $B$ . The symbol  $\omega$  is used to represent the set of natural numbers. The symbol  $K$  is used to represent the set  $\{\heartsuit_i \mid i \in N\}$ , and the symbol  $K^*$  is used to represent the set of all words of finite length of the alphabet  $K$ . Remark that  $K^*$  includes  $\emptyset$  and hence  $\{\iota\alpha \mid \iota \in K^*\}$  includes  $\alpha$ . Greek lower-case letters  $\iota$  and  $\kappa$  are used to denote any members of  $K^*$ . Lower-case letters  $i, j$  and  $k$  are sometimes used to denote any natural numbers. An expression  $X^i\alpha$  for any  $i \in \omega$  is defined inductively by  $X^0\alpha \equiv \alpha$  and  $X^{n+1}\alpha \equiv XX^n\alpha$ . Let  $T$  be  $K \cup \{X\}$ . Then,  $T^*$  is used to represent the set of all words of finite length of the alphabet  $T$ . For example,  $X^i\iota X^j\kappa$  is in  $T^*$ . An expression  $\sharp$  (or  $\sharp_i$ ) is used to represent an arbitrary member of  $T^*$ .  $\sharp_1$  ( $\in T^*$ ) is called a *permutation* of  $\sharp$  ( $\in T^*$ ) if  $\sharp_1$  is obtained from  $\sharp$  by shifting the places of the occurrences of  $X$ . For example,  $X^i\iota X^j\kappa$  is a permutation of  $X^{i+j}\iota\kappa$ . Remark that  $\sharp$  itself is a permutation of  $\sharp$ . An expression of the form  $\Gamma \Rightarrow \Delta$  is called a *sequent*. An expression  $L \vdash S$  or  $\vdash S$  is used to denote the fact that a sequent  $S$  is provable in a sequent calculus  $L$ . A rule  $R$  of inference is said to be *admissible* in a sequent calculus  $L$  if the following condition is satisfied: for any instance

$$\frac{S_1 \cdots S_n}{S}$$

of  $R$ , if  $L \vdash S_i$  for all  $i$ , then  $L \vdash S$ .

A sequent calculus ALTL is introduced below.

**Definition 1 (ALTL).** The initial sequents of ALTL are of the form: for any propositional variable  $p$ ,

$$\sharp p \Rightarrow \sharp p.$$

The structural inference rules of ALTL are of the form:

$$\frac{\Gamma \Rightarrow \Delta, \alpha \quad \alpha, \Sigma \Rightarrow \Pi}{\Gamma, \Sigma \Rightarrow \Delta, \Pi} \text{ (cut)} \quad \frac{\Gamma \Rightarrow \Delta}{\Sigma, \Gamma \Rightarrow \Delta, \Pi} \text{ (we).}$$

The logical inference rules of ALTL are of the form: for any  $m \in N$ ,

$$\begin{array}{c} \frac{\Gamma \Rightarrow \Delta, \sharp\alpha \quad \sharp\beta, \Sigma \Rightarrow \Pi}{\sharp(\alpha \rightarrow \beta), \Gamma, \Sigma \Rightarrow \Delta, \Pi} \text{ ($\rightarrow$left)} \quad \frac{\sharp\alpha, \Gamma \Rightarrow \Delta, \sharp\beta}{\Gamma \Rightarrow \Delta, \sharp(\alpha \rightarrow \beta)} \text{ ($\rightarrow$right)} \\ \\ \frac{\sharp\alpha, \sharp\beta, \Gamma \Rightarrow \Delta}{\sharp(\alpha \wedge \beta), \Gamma \Rightarrow \Delta} \text{ ($\wedge$left)} \quad \frac{\Gamma \Rightarrow \Delta, \sharp\alpha \quad \Gamma \Rightarrow \Delta, \sharp\beta}{\Gamma \Rightarrow \Delta, \sharp(\alpha \wedge \beta)} \text{ ($\wedge$right)} \end{array}$$

$$\begin{array}{c}
\frac{\# \alpha, \Gamma \Rightarrow \Delta \quad \# \beta, \Gamma \Rightarrow \Delta}{\#(\alpha \vee \beta), \Gamma \Rightarrow \Delta} (\vee\text{left}) \quad \frac{\Gamma \Rightarrow \Delta, \# \alpha, \# \beta}{\Gamma \Rightarrow \Delta, \#(\alpha \vee \beta)} (\vee\text{right}) \\
\frac{\Gamma \Rightarrow \Delta, \# \alpha}{\# \neg \alpha, \Gamma \Rightarrow \Delta} (\neg\text{left}) \quad \frac{\# \alpha, \Gamma \Rightarrow \Delta}{\Gamma \Rightarrow \Delta, \# \neg \alpha} (\neg\text{right}) \\
\frac{\# \heartsuit_m X \alpha, \Gamma \Rightarrow \Delta}{\# X \heartsuit_m \alpha, \Gamma \Rightarrow \Delta} (\text{X}\heartsuit\text{left}) \quad \frac{\Gamma \Rightarrow \Delta, \# \heartsuit_m X \alpha}{\Gamma \Rightarrow \Delta, \# X \heartsuit_m \alpha} (\text{X}\heartsuit\text{right}) \\
\frac{\# X^k \alpha, \Gamma \Rightarrow \Delta}{\# G \alpha, \Gamma \Rightarrow \Delta} (\text{Gleft}) \quad \frac{\{ \Gamma \Rightarrow \Delta, \# X^j \alpha \mid j \in \omega \}}{\Gamma \Rightarrow \Delta, \# G \alpha} (\text{Gright}) \\
\frac{\{ \# X^j \alpha, \Gamma \Rightarrow \Delta \mid j \in \omega \}}{\# F \alpha, \Gamma \Rightarrow \Delta} (\text{Fleft}) \quad \frac{\Gamma \Rightarrow \Delta, \# X^k \alpha}{\Gamma \Rightarrow \Delta, \# F \alpha} (\text{Fright}) \\
\frac{\# \kappa \alpha, \Gamma \Rightarrow \Delta \quad (\kappa \in K^*)}{\# \heartsuit_c \alpha, \Gamma \Rightarrow \Delta} (\heartsuit_c\text{left}) \quad \frac{\{ \Gamma \Rightarrow \Delta, \# \kappa \alpha \mid \kappa \in K^* \}}{\Gamma \Rightarrow \Delta, \# \heartsuit_c \alpha} (\heartsuit_c\text{right}) \\
\frac{\{ \# \kappa \alpha, \Gamma \Rightarrow \Delta \mid \kappa \in K^* \}}{\# \heartsuit_d \alpha, \Gamma \Rightarrow \Delta} (\heartsuit_d\text{left}) \quad \frac{\Gamma \Rightarrow \Delta, \# \kappa \alpha \quad (\kappa \in K^*)}{\Gamma \Rightarrow \Delta, \# \heartsuit_d \alpha} (\heartsuit_d\text{right}).
\end{array}$$

Note that (Gright), (Fleft), ( $\heartsuit_c$ right) and ( $\heartsuit_d$ left) have infinite premises. Remark that Gentzen's sequent calculus LK for classical logic and Kawai's sequent calculus  $\text{LT}_\omega$  [8] for LTL are subsystems of ALTL.

Remark that the sequents of the form  $\# \alpha \Rightarrow \# \alpha$  for any formula  $\alpha$  are provable in cut-free ALTL. This fact can be proved by induction on  $\alpha$ .

An expression  $\alpha \Leftrightarrow \beta$  represents two sequents  $\alpha \Rightarrow \beta$  and  $\beta \Rightarrow \alpha$ .

**Proposition 2.** *The following sequents concerning  $\heartsuit_i$  and  $\heartsuit_c$  are provable in cut-free ALTL: for any  $i \in N$ ,*

1.  $\heartsuit_i(\alpha \rightarrow \beta) \Rightarrow \heartsuit_i \alpha \rightarrow \heartsuit_i \beta$ ,
2.  $\heartsuit_c \alpha \Rightarrow \heartsuit_i \alpha$ ,
3.  $\heartsuit_c \alpha \Rightarrow \kappa \alpha$  for any  $\kappa \in K^*$ ,
4.  $\heartsuit_c \alpha \Rightarrow \heartsuit_i \heartsuit_c \alpha$ ,
5.  $\heartsuit_i X \alpha \Leftrightarrow X \heartsuit_i \alpha$ .

**Proposition 3.** *The rules of the form: for any  $i, m \in N$ ,*

$$\frac{\Gamma \Rightarrow \Delta}{\heartsuit_i \Gamma \Rightarrow \heartsuit_i \Delta} (\heartsuit\text{regu}) \quad \frac{\# X \heartsuit_m \alpha, \Gamma \Rightarrow \Delta}{\# X \heartsuit_m \alpha, \Gamma \Rightarrow \Delta} (\text{X}\heartsuit\text{left}^{-1}) \quad \frac{\Gamma \Rightarrow \Delta, \# X \heartsuit_m \alpha}{\Gamma \Rightarrow \Delta, \# X \heartsuit_m \alpha} (\text{X}\heartsuit\text{right}^{-1})$$

*are admissible in cut-free ALTL.*

Remark that the rule ( $\heartsuit$ regu) is more expressive than the following standard inference rules for the normal modal logics K and KD:

$$\frac{\Gamma \Rightarrow \alpha}{\heartsuit_i \Gamma \Rightarrow \heartsuit_i \alpha} \quad \frac{\Gamma \Rightarrow \gamma}{\heartsuit_i \Gamma \Rightarrow \heartsuit_i \gamma}$$

where  $\gamma$  can be empty. Thus, the operator  $\heartsuit_i$  in ALTL is stronger than those in K and KD.  $\heartsuit_i$  may thus be used for an alternative to the operators in K and KD.

Next, we consider a sequent calculus  $\text{LK}_\omega$  for infinitary logic in order to show the syntactical embedding theorem of ALTL into  $\text{LK}_\omega$ . A language of  $\text{LK}_\omega$  is obtained from the language  $\mathcal{L}$  of ALTL by deleting  $X, G, F, \heartsuit_i$  ( $i \in N$ ),  $\heartsuit_c$  and  $\heartsuit_d$  and adding  $\wedge$  (infinitary conjunction) and  $\vee$  (infinitary disjunction). For  $\wedge$  and  $\vee$ , if  $\Theta$  is a countable nonempty set of formulas, then  $\bigwedge \Theta$  and  $\bigvee \Theta$  are also formulas. Expressions  $\wedge\{\alpha\}$  and  $\vee\{\alpha\}$  are equivalent to  $\alpha$ . The standard binary connectives  $\wedge$  and  $\vee$  are regarded as special cases of  $\wedge$  and  $\vee$ , respectively.

A sequent calculus  $\text{LK}_\omega$  for infinitary logic is then presented below.

**Definition 4 ( $\text{LK}_\omega$ ).** *The initial sequents of  $\text{LK}_\omega$  are of the form: for any propositional variable  $p$ ,*

$$p \Rightarrow p.$$

*The structural rules of  $\text{LK}_\omega$  are (cut) and (we) presented in Definition 1.*

*The logical inference rules of  $\text{LK}_\omega$  are of the form:*

$$\begin{array}{c} \frac{\Gamma \Rightarrow \Sigma, \alpha \quad \beta, \Delta \Rightarrow \Pi}{\alpha \rightarrow \beta, \Gamma, \Delta \Rightarrow \Sigma, \Pi} (\rightarrow \text{left}^\emptyset) \quad \frac{\alpha, \Gamma \Rightarrow \Delta, \beta}{\Gamma \Rightarrow \Delta, \alpha \rightarrow \beta} (\rightarrow \text{right}^\emptyset) \\ \\ \frac{\Gamma \Rightarrow \Delta, \alpha}{\neg \alpha, \Gamma \Rightarrow \Delta} (\neg \text{left}^\emptyset) \quad \frac{\alpha, \Gamma \Rightarrow \Delta}{\Gamma \Rightarrow \Delta, \neg \alpha} (\neg \text{right}^\emptyset) \\ \\ \frac{\alpha, \Gamma \Rightarrow \Delta \quad (\alpha \in \Theta)}{\bigwedge \Theta, \Gamma \Rightarrow \Delta} (\wedge \text{left}) \quad \frac{\{\Gamma \Rightarrow \Delta, \alpha \mid \alpha \in \Theta\}}{\Gamma \Rightarrow \Delta, \bigwedge \Theta} (\wedge \text{right}) \\ \\ \frac{\{\alpha, \Gamma \Rightarrow \Delta \mid \alpha \in \Theta\}}{\bigvee \Theta, \Gamma \Rightarrow \Delta} (\vee \text{left}) \quad \frac{\Gamma \Rightarrow \Delta, \alpha \quad (\alpha \in \Theta)}{\Gamma \Rightarrow \Delta, \bigvee \Theta} (\vee \text{right}) \end{array}$$

where  $\Theta$  is a countable nonempty set of formulas.

The superscript “ $\emptyset$ ” in the rule names in  $\text{LK}_\omega$  means that these rules are the special cases of the corresponding rules of ALTL, i.e., the case that  $\sharp$  is  $\emptyset$ . The sequents of the form  $\alpha \Rightarrow \alpha$  for any formula  $\alpha$  are provable in cut-free  $\text{LK}_\omega$ . As well-known,  $\text{LK}_\omega$  enjoys cut-elimination.

The following definition of an embedding function is a modified extension of the embedding function of LTL into infinitary logic, which was proposed in [6] and was further studied in [7].

**Definition 5.** *Fix a countable non-empty set  $\Phi$  of propositional variables, and define the sets  $\Phi_\sharp := \{p_\sharp \mid p \in \Phi\}$  ( $\sharp \in T^*$ ) of propositional variables with  $p_\emptyset := p$  (i.e.,  $\Phi_\emptyset := \Phi$ ). The language  $\mathcal{L}^t$  (or the set of formulas) of ALTL is defined using  $\Phi$ ,  $\rightarrow, \neg, \wedge, \vee, \heartsuit_i$  ( $i \in N$ ),  $\heartsuit_c$ ,  $\heartsuit_d$ ,  $X$ ,  $G$  and  $F$ . The language  $\mathcal{L}^i$  of  $\text{LK}_\omega$  is defined using  $\bigcup_{\sharp \in T^*} \Phi_\sharp, \rightarrow, \neg, \wedge$  and  $\vee$ . The binary versions of  $\wedge$  and  $\vee$  are also denoted as  $\wedge$  and  $\vee$ , respectively, and these binary symbols are assumed to be included in  $\mathcal{L}^i$ . Moreover, we assume the following condition: for any permutations  $\sharp_1$  and  $\sharp_2$  of  $\sharp$  ( $\in T^*$ ) and any  $p \in \Phi$ ,  $p_{\sharp_1} = p_{\sharp_2}$ , i.e.,  $\Phi_{\sharp_1} = \Phi_{\sharp_2}$ .*

*A mapping  $f$  from  $\mathcal{L}^t$  to  $\mathcal{L}^i$  is defined as follows.*

1. for any  $p \in \Phi$ ,  $f(\sharp p) := p_\sharp \in \Phi_\sharp$  ( $\sharp \in T^*$ ), esp.,  $f(p) := p \in \Phi_\emptyset$ ,
2.  $f(\sharp(\alpha \circ \beta)) := f(\sharp \alpha) \circ f(\sharp \beta)$  where  $\circ \in \{\rightarrow, \wedge, \vee\}$ ,

3.  $f(\sharp\neg\alpha) := \neg f(\sharp\alpha)$ ,
4.  $f(\sharp\heartsuit_c\alpha) := \bigwedge\{f(\sharp\kappa\alpha) \mid \kappa \in K^*\}$ ,
5.  $f(\sharp\heartsuit_d\alpha) := \bigvee\{f(\sharp\kappa\alpha) \mid \kappa \in K^*\}$ ,
6.  $f(\sharp G\alpha) := \bigwedge\{f(\sharp X^j\alpha) \mid j \in \omega\}$ ,
7.  $f(\sharp F\alpha) := \bigvee\{f(\sharp X^j\alpha) \mid j \in \omega\}$ ,
8.  $f(\sharp\heartsuit_i X\alpha) := f(\sharp X\heartsuit_i\alpha)$ .

An expression  $f(\Gamma)$  denotes the result of replacing every occurrence of a formula  $\alpha$  in  $\Gamma$  by an occurrence of  $f(\alpha)$ .

**Theorem 6 (Syntactical embedding).** *Let  $\Gamma$  and  $\Delta$  be sets of formulas in  $\mathcal{L}^t$ , and  $f$  be the mapping defined in Definition 5. Then:*

1. if  $\text{ALTL} \vdash \Gamma \Rightarrow \Delta$ , then  $\text{LK}_\omega \vdash f(\Gamma) \Rightarrow f(\Delta)$ .
2. if  $\text{LK}_\omega - (\text{cut}) \vdash f(\Gamma) \Rightarrow f(\Delta)$ , then  $\text{ALTL} - (\text{cut}) \vdash \Gamma \Rightarrow \Delta$ .

**Proof.** • (1) : By induction on the proofs  $P$  of  $\Gamma \Rightarrow \Delta$  in ALTL. We distinguish the cases according to the last inference of  $P$ . We show only the following cases.

Case  $(\sharp p \Rightarrow \sharp p)$ : The last inference of  $P$  is of the form:  $\sharp p \Rightarrow \sharp p$ . In this case, we obtain  $\text{LK}_\omega \vdash f(\sharp p) \Rightarrow f(\sharp p)$ , i.e.,  $\text{LK}_\omega \vdash p_\sharp \Rightarrow p_\sharp$  ( $p_\sharp \in \Phi_\sharp$ ) by the definition of  $f$ .

Case  $(\heartsuit_c \text{right})$ : The last inference of  $P$  is of the form:

$$\frac{\{ \Gamma \Rightarrow \Delta, \sharp\kappa\alpha \mid \kappa \in K^* \}}{\Gamma \Rightarrow \Delta, \sharp\heartsuit_c\alpha} (\heartsuit_c \text{right}).$$

By induction hypothesis, we have  $\text{LK}_\omega \vdash f(\Gamma) \Rightarrow f(\Delta), f(\sharp\kappa\alpha)$  for all  $\kappa \in K^*$ . Let  $\Phi$  be  $\{f(\sharp\kappa\alpha) \mid \kappa \in K^*\}$ . We obtain the required fact:

$$\frac{\vdots}{\frac{\{ f(\Gamma) \Rightarrow f(\Delta), f(\sharp\kappa\alpha) \mid f(\sharp\kappa\alpha) \in \Phi \}}{f(\Gamma) \Rightarrow f(\Delta), \bigwedge \Phi}} (\wedge \text{right})$$

where  $\bigwedge \Phi$  coincides with  $f(\sharp\heartsuit_c\alpha)$  by the definition of  $f$ .

Case  $(X\heartsuit \text{left})$ : The last inference of  $P$  is of the form:

$$\frac{\sharp\heartsuit_m X\alpha, \Gamma \Rightarrow \Delta}{\sharp X\heartsuit_m\alpha, \Gamma \Rightarrow \Delta} (X\heartsuit \text{left}).$$

By induction hypothesis, we have  $\text{LK}_\omega \vdash f(\sharp\heartsuit_m X\alpha), f(\Gamma) \Rightarrow f(\Delta)$  and hence obtain the required fact  $\text{LK}_\omega \vdash f(\sharp X\heartsuit_m\alpha), f(\Gamma) \Rightarrow f(\Delta)$ , since  $f(\sharp X\heartsuit_m\alpha)$  coincides with  $f(\sharp\heartsuit_m X\alpha)$  by the definition of  $f$ .

• (2) : By induction on the proofs  $Q$  of  $f(\Gamma) \Rightarrow f(\Delta)$  in  $\text{LK}_\omega$ . We distinguish the cases according to the last inference of  $Q$ . We show only the following case. The last inference of  $Q$  is of the form:

$$\frac{\{ f(\Gamma) \Rightarrow f(\Delta), f(\sharp\kappa\alpha) \mid f(\sharp\kappa\alpha) \in \Phi \}}{f(\Gamma) \Rightarrow f(\Delta), \bigwedge \Phi} (\wedge \text{right})$$

where  $\Phi = \{f(\sharp\kappa\alpha) \mid \kappa \in K^*\}$ , and  $\bigwedge \Phi$  coincides with  $f(\sharp\heartsuit_c\alpha)$  by the definition of  $f$ . By induction hypothesis, we have  $\text{ALTL} \vdash \Gamma \Rightarrow \Delta, \sharp\kappa\alpha$  for all  $\kappa \in K^*$ . We thus obtain the required fact:

$$\frac{\vdots}{\{ \Gamma \Rightarrow \Delta, \sharp\kappa\alpha \mid \kappa \in K^* \}} \quad (\heartsuit_c\text{right}). \quad \blacksquare$$

**Theorem 7 (Cut-elimination).** *The rule (cut) is admissible in cut-free ALTL.*

**Proof.** Suppose  $\text{ALTL} \vdash \Gamma \Rightarrow \Delta$ . Then, we have  $\text{LK}_\omega \vdash f(\Gamma) \Rightarrow f(\Delta)$  by Theorem 6 (1), and hence  $\text{LK}_\omega - (\text{cut}) \vdash f(\Gamma) \Rightarrow f(\Delta)$  by the cut-elimination theorem for  $\text{LK}_\omega$ . By Theorem 6 (2), we obtain  $\text{ALTL} - (\text{cut}) \vdash \Gamma \Rightarrow \Delta$ .  $\blacksquare$

Remark that by Theorem 7, we can strengthen the statements of Theorem 6 by replacing “if then” with “iff”. This fact will be used to prove the completeness theorem for ALTL.

### 3 Semantics and Completeness

Let  $\Gamma$  be a set  $\{\alpha_1, \dots, \alpha_m\}$  ( $m \geq 0$ ) of formulas. Then,  $\Gamma^*$  represents  $\alpha_1 \vee \dots \vee \alpha_m$  if  $m \geq 1$ , and otherwise  $\neg(p \rightarrow p)$  where  $p$  is a fixed propositional variable. Also  $\Gamma_*$  represents  $\alpha_1 \wedge \dots \wedge \alpha_m$  if  $m \geq 1$ , and otherwise  $p \rightarrow p$  where  $p$  is a fixed propositional variable. The symbol  $\geq$  or  $\leq$  is used to represent a linear order on  $\omega$ .

A semantics for ALTL is defined below.

**Definition 8.** *Agent-time indexed valuations  $I^{\iota:i}$  ( $\iota \in K^*, i \in \omega$ ) are mappings from the set of all propositional variables to the set  $\{t, f\}$  of truth values. Then, agent-time indexed satisfaction relations  $\models_{\iota;i} \alpha$  ( $\iota \in K^*, i \in \omega$ ) for any formula  $\alpha$  are defined inductively by:*

1. *for any propositional variable  $p$ ,  $\models_{\iota;i} p$  iff  $I^{\iota:i}(p) = t$ ,*
2.  *$\models_{\iota;i} \alpha \wedge \beta$  iff  $\models_{\iota;i} \alpha$  and  $\models_{\iota;i} \beta$ ,*
3.  *$\models_{\iota;i} \alpha \vee \beta$  iff  $\models_{\iota;i} \alpha$  or  $\models_{\iota;i} \beta$ ,*
4.  *$\models_{\iota;i} \alpha \rightarrow \beta$  iff  $\text{not}-(\models_{\iota;i} \alpha)$  or  $\models_{\iota;i} \beta$ ,*
5.  *$\models_{\iota;i} \neg \alpha$  iff  $\text{not}-(\models_{\iota;i} \alpha)$ ,*
6. *for any  $k \in N$ ,  $\models_{\iota;i} \heartsuit_k \alpha$  iff  $\models_{\iota;\varphi_k;i} \alpha$ ,*
7.  *$\models_{\iota;i} \heartsuit_c \alpha$  iff  $\models_{\iota;\kappa;i} \alpha$  for all  $\kappa \in K^*$ ,*
8.  *$\models_{\iota;i} \heartsuit_d \alpha$  iff  $\models_{\iota;\kappa;i} \alpha$  for some  $\kappa \in K^*$ ,*
9.  *$\models_{\iota;i} X\alpha$  iff  $\models_{\iota;i+1} \alpha$ ,*
10.  *$\models_{\iota;i} G\alpha$  iff  $\models_{\iota;j} \alpha$  for any  $j \geq i$ ,*
11.  *$\models_{\iota;i} F\alpha$  iff  $\models_{\iota;j} \alpha$  for some  $j \geq i$ .*

A formula  $\alpha$  is called ALTL-valid if  $\models_{\emptyset;0} \alpha$  holds for any agent-time indexed satisfaction relations  $\models_{\iota;i}$  ( $\iota \in K^*, i \in \omega$ ). A sequent  $\Gamma \Rightarrow \Delta$  is called ALTL-valid if so is the formula  $\Gamma_* \rightarrow \Delta^*$ .

Remark that the following clause holds for any agent-time indexed satisfaction relations  $\models_{\iota;i}$ , any formula  $\alpha$ , any  $i \in \omega$  and any  $\kappa \in K^*$ ,  $\models_{\iota;i} \kappa\alpha$  iff  $\models_{\iota\kappa;i} \alpha$ .

Next, a semantics for  $LK_\omega$  is defined below.

**Definition 9.** Let  $\Theta$  be a countable nonempty set of formulas. A valuation  $I$  is a mapping from the set of all propositional variables to the set  $\{t, f\}$  of truth values. A satisfaction relation  $\models \alpha$  for any formula  $\alpha$  is defined inductively by:

1.  $\models p$  iff  $I(p) = t$  for any propositional variable  $p$ ,
2.  $\models \bigwedge \Theta$  iff  $\models \alpha$  for any  $\alpha \in \Theta$ ,
3.  $\models \bigvee \Theta$  iff  $\models \alpha$  for some  $\alpha \in \Theta$ ,
4.  $\models \alpha \rightarrow \beta$  iff  $\text{not-}(\models \alpha)$  or  $\models \beta$ ,
5.  $\models \neg \alpha$  iff  $\text{not-}(\models \alpha)$ .

A formula  $\alpha$  is called  $LK_\omega$ -valid if  $\models \alpha$  holds for any satisfaction relation  $\models$ . A sequent  $\Gamma \Rightarrow \Delta$  is called  $LK_\omega$ -valid if so is the formula  $\Gamma_* \rightarrow \Delta^*$ .

As well known, the following completeness theorem holds for  $LK_\omega$ : For any sequent  $S$ ,  $LK_\omega \vdash S$  iff  $S$  is  $LK_\omega$ -valid.

**Lemma 10.** Let  $f$  be the mapping defined in Definition 5. For any agent-time indexed satisfaction relation  $\models_{\iota;i}$  ( $\iota \in K^*, i \in \omega$ ), we can construct a satisfaction relation  $\models$  such that for any formula  $\alpha$  in  $\mathcal{L}^t$ ,  $\models_{\iota;i} \alpha$  iff  $\models f(\iota X^i \alpha)$ .

**Proof.** Let  $\Phi$  be a set of propositional variables and  $\Phi_\sharp$  be the set  $\{p_\sharp \mid p \in \Phi\}$  of propositional variables with  $p_\emptyset := p$ . Suppose that  $I^{\iota;i}$  ( $\iota \in K^*, i \in \omega$ ) are mappings from  $\Phi$  to  $\{t, f\}$ . Suppose that  $I$  is a mapping from  $\bigcup_{\sharp \in T^*} \Phi_\sharp$  to  $\{t, f\}$ .

Suppose moreover that  $I^{\iota;i}(p) = t$  iff  $I(p_{\iota X^i}) = t$ . Then the lemma is proved by induction on the complexity of  $\alpha$ .

- Base step:

Case ( $\alpha \equiv p \in \Phi$ ):  $\models_{\iota;i} p$  iff  $I^{\iota;i}(p) = t$  iff  $I(p_{\iota X^i}) = t$  iff  $\models p_{\iota X^i}$  iff  $\models f(\iota X^i p)$  (by the definition of  $f$ ).

- Induction step:

Case ( $\alpha \equiv \alpha_1 \rightarrow \alpha_2$ ):  $\models_{\iota;i} \alpha_1 \rightarrow \alpha_2$  iff  $\text{not-}(\models_{\iota;i} \alpha_1)$  or  $\models_{\iota;i} \alpha_2$  iff  $\text{not-}(\models f(\iota X^i \alpha_1))$  or  $\models f(\iota X^i \alpha_2)$  (by induction hypothesis) iff  $\models f(\iota X^i (\alpha_1 \rightarrow \alpha_2))$  (by the definition of  $f$ ).

Cases ( $\alpha \equiv \alpha_1 \wedge \alpha_2$  and  $\alpha \equiv \alpha_1 \vee \alpha_2$ ): Similar to Case ( $\alpha \equiv \alpha_1 \rightarrow \alpha_2$ ).

Case ( $\alpha \equiv \neg \beta$ ):  $\models_{\iota;i} \neg \beta$  iff  $\text{not-}(\models_{\iota;i} \beta)$  iff  $\text{not-}(\models f(\iota X^i \beta))$  (by induction hypothesis) iff  $\models \neg f(\iota X^i \beta)$  iff  $\models f(\iota X^i \neg \beta)$  (by the definition of  $f$ ).

Case ( $\alpha \equiv \heartsuit_k \beta$ ):  $\models_{\iota;i} \heartsuit_k \beta$  iff  $\models_{\iota \heartsuit_k; i} \beta$  iff  $\models f(\iota \heartsuit_k X^i \beta)$  (by induction hypothesis) iff  $\models f(\iota X^i \heartsuit_k \beta)$  (by the definition of  $f$ ).

Case ( $\alpha \equiv \heartsuit_c \beta$ ):  $\models_{\iota;i} \heartsuit_c \beta$  iff  $\models_{\iota \kappa; i} \beta$  for any  $\kappa \in K^*$  iff  $\models f(\iota \kappa X^i \beta)$  for any  $\kappa \in K^*$  (by induction hypothesis) iff  $\models f(\iota X^i \kappa \beta)$  for any  $\kappa \in K^*$  (by the definition of  $f$ ) iff  $\models \bigwedge \{f(\iota X^i \kappa \beta) \mid \kappa \in K^*\}$  iff  $\models f(\iota X^i \heartsuit_c \beta)$  (by the definition of  $f$ ).

Case ( $\alpha \equiv \heartsuit_d \beta$ ): Similar to Case ( $\alpha \equiv \heartsuit_c \beta$ ).

Case ( $\alpha \equiv X\beta$ ):  $\models_{\iota; i} X\beta$  iff  $\models_{\iota; i+1} \beta$  iff  $\models f(\iota X^{i+1}\beta)$  (by induction hypothesis) iff  $\models f(\iota X^i X\beta)$ .

Case ( $\alpha \equiv G\beta$ ):  $\models_{\iota; i} G\beta$  iff  $\models_{\iota; j} \beta$  for any  $j \geq i$  iff  $\models f(\iota X^j \beta)$  for any  $j \geq i$  (by induction hypothesis) iff  $\forall k \in \omega [\models f(\iota X^{i+k}\beta)]$  iff  $\models \gamma$  for any  $\gamma \in \{f(\iota X^{i+k}\beta) \mid k \in \omega\}$  iff  $\models \bigwedge \{f(\iota X^{i+k}\beta) \mid k \in \omega\}$  iff  $\models f(\iota X^i G\beta)$  (by the definition of  $f$ ).

Case ( $\alpha \equiv F\beta$ ): Similar to Case ( $\alpha \equiv G\beta$ ). ■

**Lemma 11.** *Let  $f$  be the mapping defined in Definition 5. For any satisfaction relation  $\models$ , we can construct a agent-time indexed satisfaction relation  $\models_{\iota; i}$  such that for any formula  $\alpha$  in  $\mathcal{L}^t$ ,  $\models f(\iota X^i \alpha)$  iff  $\models_{\iota; i} \alpha$ .*

**Proof.** Similar to the proof of Lemma 10. ■

**Theorem 12 (Semantical embedding).** *Let  $f$  be the mapping defined in Definition 5. For any formula  $\alpha$  in  $\mathcal{L}^t$ ,  $\alpha$  is ALTL-valid iff  $f(\alpha)$  is  $LK_\omega$ -valid.*

**Proof.** By Lemmas 10 and 11. We take 0 for  $i$  and take  $\emptyset$  for  $\iota$ . ■

**Theorem 13 (Completeness).** *For any sequent  $S$ ,  $ALTL \vdash S$  iff  $S$  is ALTL-valid.*

**Proof.** Let  $\Gamma \Rightarrow \Delta$  be  $S$  and  $\alpha$  be  $\Gamma_* \rightarrow \Delta^*$ . It is sufficient to show that  $ALTL \vdash \alpha$  iff  $\alpha$  is ALTL-valid. We show this as follows.  $ALTL \vdash \alpha$  iff  $LK_\omega \vdash \Rightarrow f(\alpha)$  (by Theorem 6 and Theorem 7) iff  $f(\alpha)$  is  $LK_\omega$ -valid (by the completeness theorem for  $LK_\omega$ ) iff  $\alpha$  is ALTL-valid (by Theorem 12). ■

## 4 Concluding Remarks

In this paper, the logic ALTL (agents-indexed linear-time temporal logic) was introduced as a Gentzen-type sequent calculus. ALTL was intended to appropriately represent reasoning about time-dependent multi-agents within a proof system. As the main results of this paper, the cut-elimination and completeness theorems for ALTL were shown by combining two theorems for syntactically and semantically embedding ALTL into infinitary logic. It was thus shown in this paper that ALTL is attractive as a theoretical basis for automated theorem proving about time-dependent multi-agents.

The rest of this paper is devoted to address some related works. Some agents-based or knowledge-based model checkers have successfully been developed by many researchers (e.g., [11, 3, 5, 4]). For example, an approach to model checking for the *modal logic of knowledge and linear-time in distributed systems with perfect recall* was established by van der Meyden and Shilov [11]. They showed that some model checking problems with or without a common knowledge operator are undecidable or PSPACE-complete. A model checker for real-time and multi-agent systems, called *VerICS*, has been developed by Kacprzak et al. [4]. They focused on SAT-based model checking for multi-agent systems and several extensions and implementations to real-time systems' verification. Although ALTL is

intended to provide a good proof theory, we believe that the proposed semantics for ALTL may also be applicable to model checking, satisfiability checking and validity checking for time-dependent multi-agent systems.

**Acknowledgments.** This work was partially supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Young Scientists (B) 20700015.

## References

1. Emerson, E.A.: Temporal and modal logic. In: van Leeuwen, J. (ed.) Formal Models and Semantics (B). Handbook of Theoretical Computer Science, pp. 995–1072. Elsevier and MIT Press (1990)
2. Fagin, R., Halpern, J.Y., Moses, Y., Vardi, M.Y.: Reasoning about knowledge. MIT Press, Cambridge (1995)
3. Gammie, P., van der Meyden, R.: MCK: Model checking the logic of knowledge. In: Alur, R., Peled, D.A. (eds.) CAV 2004. LNCS, vol. 3114, pp. 479–483. Springer, Heidelberg (2004)
4. Kacprzak, M., Nabialek, W., Niewiadomski, A., Penczek, W., Polroa, A., Szczerter, M., Wozawa, B., Zbrzezny, A.: VerIcs 2007: A model checker for real-time and multi-agent systems. *Fundamenta Informaticae* 85(1–4), 313–328 (2008)
5. Lomuscio, A., Raimondi, F.: A model checker for multi-agent systems. In: Hermanns, H., Palsberg, J. (eds.) TACAS 2006. LNCS, vol. 3920, pp. 450–454. Springer, Heidelberg (2006)
6. Kamide, N.: Embedding linear-time temporal logic into infinitary logic: Application to cut-elimination for multi-agent infinitary epistemic linear-time temporal logic. In: Fisher, M., Sadri, F., Thielscher, M. (eds.) CLIMA IX. LNCS (LNAI), vol. 5405, pp. 57–76. Springer, Heidelberg (2009)
7. Kamide, N., Wansing, H.: Combining linear-time temporal logic with constructiveness and paraconsistency. *Journal of Applied Logic* 8, 33–61 (2010)
8. Kawai, H.: Sequential calculus for a first order infinitary temporal logic. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik* 33, 423–432 (1987)
9. Kozen, D.: Results on the propositional mu-calculus. *Theoretical Computer Science* 27, 333–354 (1983)
10. Pnueli, A.: The temporal logic of programs. In: Proceedings of the 18th IEEE Symposium on Foundations of Computer Science, pp. 46–57 (1977)
11. van der Meyden, R., Shilov, N.V.: Model checking knowledge and time in systems with perfect recall (extended abstract). In: Pandu Rangan, C., Raman, V., Sarukkai, S. (eds.) FST TCS 1999. LNCS, vol. 1738, pp. 432–445. Springer, Heidelberg (1999)

# Toward Emotional E-Commerce: Formalizing Agents for a Simple Negotiation Protocol

Veronica Jascanu, Nicolae Jascanu, and Severin Bumbaru

Department of Computer Science, University „Dunarea de Jos” of Galati, Romania  
`{veronica.jascanu,nicolae.jascanu,severin.bumbaru}@ugal.ro`

**Abstract.** Electronic commerce has become a central pillar of the Internet. Easy access, mobile devices with permanent connection, social networks and the real-time conversation streams have a big influence over B2C and C2C commerce. Currently, e-commerce becomes a social commerce, much closer to the traditional paradigm. The inclusion of emotional components in the act of trading complies with the current social trends and further approaches the electronic commerce to the traditional one. This paper continues the work on an emotional e-commerce platform by formalizing the customer, supplier and community agents. We present a simple negotiation protocol as a proof of concept.

**Keywords:** Multi-agent systems, negotiation, e-commerce, affective computing.

## 1 Introduction

In recent years, e-commerce has gained a key role in modern society. Currently, online e-commerce includes many directions like advertising and marketing, payment mechanisms, security and privacy, reputation and trust, contracting and economic legislation, business management, distribution, sale and purchase of goods and services [1]. Huge amount of products and services offered online and the scenarios in which trading occur electronically, requires the development of automatic tools. The goal is to understand the user and to give what he wants or what he needs at the right time. Using a multi-agent system to represent the various entities participating at the act of commerce is a proven method for addressing the complexity of the system. In recent years, research on electronic commerce shaped around intelligent agents [2].

The goal for service providers is to understand the customer and give appropriate products and services. All major service providers have specific methods for monitoring and identification of consumer preferences. Amazon is the representative service that, based on the history of interactions between products and customers, is able to recommend similar products that may be useful in the given context. Over 60% of customers of the Netflix movie rental service are using automatic recommendations for choosing the movies. The emergence of social communities has a major impact on e-commerce. Many online commerce services have begun to include social elements to attract a greater number of clients. Opportunity to express your opinion, to talk about a product or service, to influence the others view represents a natural evolution of the online trading. There are studies about the dynamics of information flow in social groups and the influence of the online interactions over the real life. As electronic commerce becomes a

permanent part of our contemporary society, the need to use intelligent and automated systems to facilitate various operations becomes more pressing. In terms of consumer behavior have been identified six stages in which intelligent agents may have significant contributions: necessity identification, products and suppliers brokering, social interactions, negotiations, payment, delivery, and after sales services.

Emotion is a fundamental aspect of life. Extensive research in psychology shows that even a random emotion, triggered by unrelated events can have a major influence over the decision. Incorporating emotions in decision-making system is necessary for solving complex problems and better understanding the decisions. Today, emotional theories are a multi-disciplinary research area, which includes cognitive psychology, neurology, genetics etc. One of the leaders in emotional research is the European project FP6 HUMAINE (Human-Machine Interaction Network on Emotion) [3], which bring together over 33 partners from 14 European countries. Emotional research has taken such a magnitude that the W3C consortium is seeking to define a markup emotional language EmotionML that standardizes the description of emotional knowledge [4].

We should treat emotions as knowledge in order to integrate them in a system. Various emotional models such as discrete models, in which each response to an action is associated to a distinct emotion, could represent the emotional knowledge. We could also represent the emotional knowledge by using dimensional models. The circumplex model is a powerful theoretical tool, which describes the relations between emotions, and predicts the effects on behavior and knowledge. The structural model assumes that emotional states, depending on intensity are positive, zero or negative correlated. Russell's circumplex is a two-dimensional model with the following axes: pleasant-unpleasant or valence and aroused-relaxed axis or excitation [5]. Russell's circumplex model proved over decades that it could represent an impressive number of distinct emotional terms. The model is currently being used in a variety of areas, from customer satisfaction analysis and extraction of qualitative knowledge related to products or services, to mobile applications and interactive games [6], [7].

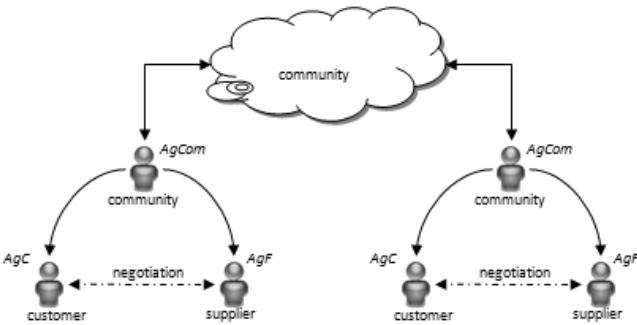
Since the early 90's, emotional theories began to be used in the field of intelligent agents. Picard [8] separates the human emotion from the one of a software agent. For the agents, emotion is just a label that describes a certain state and the corresponding action. Many psychologists have developed emotional theories in such a way that researchers in artificial intelligence can easily assimilate them [9].

Electronic commerce should finally meet the client and his style of doing trades. Traditional trade has a history of thousands of years and the online version must take into account the many subtleties of human nature.

In this paper, we continue our work on a multi-agent system for electronic commerce that integrates emotional models for each one of the three agents: the customer, the supplier and the community [10]. Using the formalism proposed by Parsons and Sabater [11], [12], we formalize each agent for a simple negotiation protocol.

## 2 A Brief Presentation of the Emotional E-Commerce Platform

The platform has three agents: the customer, the supplier and the community. The community agent has a supportive role during negotiation for both customer and supplier



**Fig. 1.** The three agents of the multi-agent platform: AgC - customer agent, AgF - supplier agent and AgCom - community agent

agents (Figure 1). At its turn, each agent is specified as a multi-agent system using the formalism of Parsons and Sabater.

Each agent has an emotional component and a specific knowledge acquisition method. Each agent uses the Russell's circumplex model of emotion in a slightly different way. The knowledge is directly acquired from the human partner whenever the system does not know details on the subject of negotiation. We are trying to develop learning mechanisms from past configurations and from the community. An agent should be able to learn that some characteristics are desirable no matter what products are negotiated. The learning mechanisms that we are developing are not explained in this paper.

## 2.1 The Customer Agent

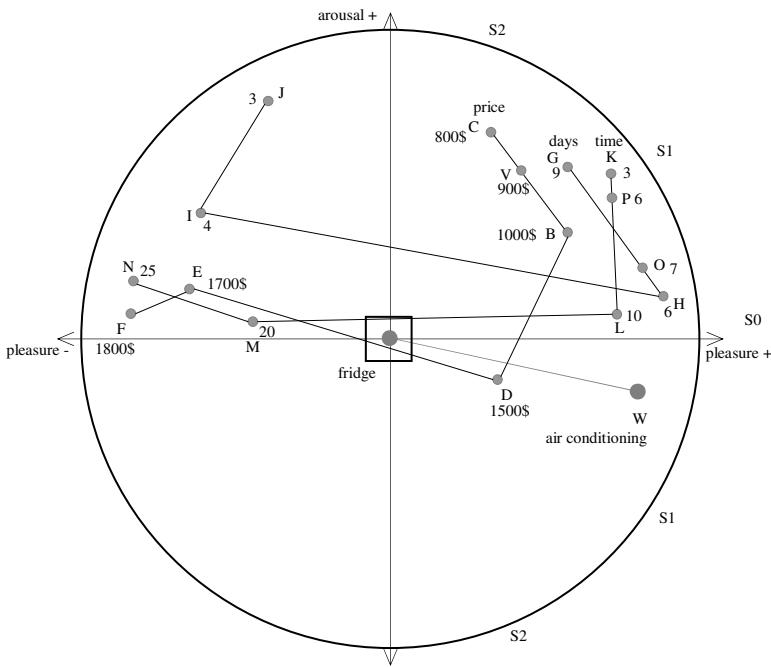
We are using the circumplex to gather the emotional aspects of the negotiation act. The system is able to capture the fine aspects of emotional-rational conflict inherent in any negotiation process. For example, when negotiating a holiday destination, we have to make a rational decision over some emotional issues: the destination is not so expensive and the location is superb. In this situation, it is likely that the emotional aspect will prevail and the rational will switch to second place. However, if we find that the services are of poor quality during that period, the weather is bad, and there are not many tourists, it is possible for the rational to win.

Figure 2 shows a negotiation configuration for a trip to the sea, with the following characteristics: price, number of days, time from the hotel to the beach. The system is able to decide between any sets of values for characteristics:

[B = 1000 \$, G = 9 days, L = 10 min] and [B = 1000 \$, O = 7 days, P = 6 min]

where B, G, L, O and P are represented on figure 2.

The *cost* value for the [B, G, L] configuration is 1.33 and for [B, O, P] of 1.36. In conclusion, the system will choose the configuration with a lower cost, namely [B, G, L]. To select between various configurations of negotiation, the system performs a quantitative analysis that provides an order for the emotional references on a characteristic (e.g. the ordering for the characteristic *price* is C, V, B, D, E and F) and a



**Fig. 2.** The customer agent configuration for negotiating a sea trip. There are five characteristics: price, number of days, time to reach the beach, air conditioning unit and fridge.

qualitative analysis that sets a *cost* for each segment separately. The order of the emotional moments is deduced by using a geometric surface and the cost is determined through a fuzzy inference system. With a sort and a cost per segment, the system can decide and choose between any configurations simultaneously. Explanations that are more detailed are presented in Jascanu's PhD Summary [13].

## 2.2 The Supplier Agent

The supplier's agent specification can be easily adapted to any system already in production. In essence, the supplier sets a list of products and services, each entry having associated an order. The order represents the supplier's preferences for selling those products. Only the supplier knows the list in its entirety. The list is minimally exposed during negotiations.

The emotional component of the supplier agent is a step forward for the B2C and C2C commerce. For instance, during off-season it is more beneficial to negotiate the price for an accommodation than to impose a fixed one. Whenever it is hard to decide over rental details, it is better to use an emotional configuration and let the system to negotiate. The supplier uses the same model described for the customer agent, such as the system is consistent in representation. Table 1 shows a fragment from the supplier negotiation configuration. The last entry in table has two emotional characteristics defined using the circumplex model. The formalism used in table 1 was developed in

order to be able to represent domain values for characteristics. A number of 10 to 12 persons is represented as  $\llbracket\{10\dots12\}\rrbracket$  and a number of more than 15 persons as  $\llbracket15\mapsto\rrbracket$ . If there is a non-smoker room, we use the  $\llbracket\!\!\mid\!\!\rrbracket$  symbol. If the room has a fridge or any other feature, we use the  $\llbracket\{\}\rrbracket$  symbol. The formalism is detailed in [13].

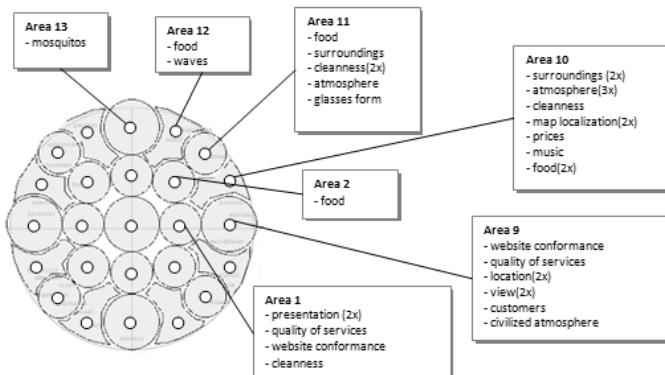
**Table 1.** The negotiation configuration for the supplier agent. The formalism is similar with the one at the customer agent.

!	* no. pers.	* price	* days	time	* smoker	A/C	fridge
$\llbracket\{9\}\rrbracket$	$\llbracket\{10\dots12\}\rrbracket$	$\llbracket\{350\}\rrbracket$	$\llbracket\{5\}\rrbracket$	$\llbracket\{4\dots6\}\rrbracket$	$\llbracket\!\!\mid\!\!\rrbracket$	$\llbracket\{\}\rrbracket$	$\llbracket\{\}\rrbracket$
$\llbracket\{10\}\rrbracket$	$\llbracket15\mapsto\rrbracket$	$\llbracket\{280\}\rrbracket$	$\llbracket\{5\dots9\}\rrbracket$	$\llbracket\{4\dots6\}\rrbracket$	$\llbracket\!\!\mid\!\!\rrbracket$	$\llbracket\{\}\rrbracket$	$\llbracket\{\}\rrbracket$
$\llbracket\{15\}\rrbracket$	<i>cplx</i>	<i>cplx</i>	$\llbracket\{3\}\mapsto\rrbracket$	$\llbracket\{4\dots6\}\rrbracket$	$\llbracket\!\!\mid\!\!\rrbracket$	$\llbracket\{\}\rrbracket$	$\llbracket\{\}\rrbracket$

## 2.3 The Community Agent

The community has a consultative role for both customer and supplier agents. E-commerce systems have solved the problem of facilitating search and selection of products and services, but they created another problem, namely: information overloading. There are so many e-commerce stores, that the user hardly finds what he needs. The selection of a product even by a single parameter, the *price*, proves to be a laborious job. Furthermore, if you do not know exactly what or where to look it is even more difficult to choose from the huge available supplies.

The knowledge base of the community agent is an emotional one. Instead of writing a textual description with your feelings about a place or event, you could place an emotional reference on the circumplex. We describe the emotional references using a set of keywords as in the figure 3. Therefore, the keywords are associated with a feeling. Therefore, you are able to express your feelings at that moment. During a vacation, you will easily express tenths of emotional references. All this references could



**Fig. 3.** The emotional knowledge acquired during a practical experiment. The subjects expressed feelings about a touristic location. The figure shows information only for the first quadrant.

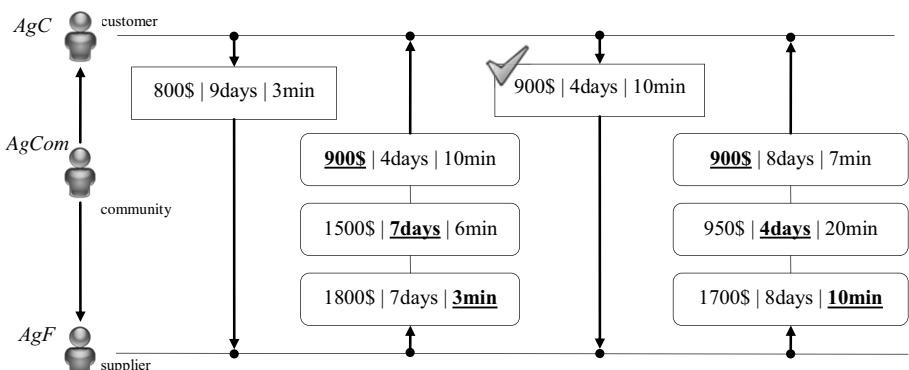
easily replace the textual description. Moreover, the references are already rated in a very intuitive and flexible system. The system will aggregate all the information even there are many opinions. An overview for a vacation to the Caribbean islands will show what is exciting, boring or relaxing.

The circumplex model is used to acquire emotional impressions on any topic or experience. The emotional opinions are indexed and can be interrogated in various ways. The emotional opinions are qualitative parameters of products and services. Thus, a customer agent that has no previous experience in a specific negotiation could use the information provided by the community. Such qualitative knowledge is extremely useful during negotiation, and can significantly influence the choices of both agents. In addition, for the supplier agent, the community plays a major role. The agent will better understand the product or service and will change the negotiation parameters appropriately.

### 3 A Simple Negotiation Protocol

The customer agent is able to analyze and sort out any number of negotiation instances received from the supplier. The negotiation starts with customer agent sending the most favorable configuration: the one with a zero cost (figure 4). In our case, the best configuration is [800\$, 9 days, 3 min]. The supplier agent receives this configuration, and for each characteristic tries to find the best offer around the specified value. Therefore, the supplier will generate three offers. The customer agent calculates the offers and selects the one with the minimal cost. In this case, the selected offer is [900\$, 4 days, 10 min]. The agent sends this negotiation instance to the supplier agent. The costs of the newly generated offers are bigger so that the negotiation is over.

We keep the protocol very simple as a proof of concept. The ability of customer agent to sort out any number of negotiation instances is essential for our algorithm. Therefore, we cut down many negotiation steps between the customer and supplier agents. The community agent has a consultative role for both agents. The information from community acts as a weight over the negotiation characteristics.

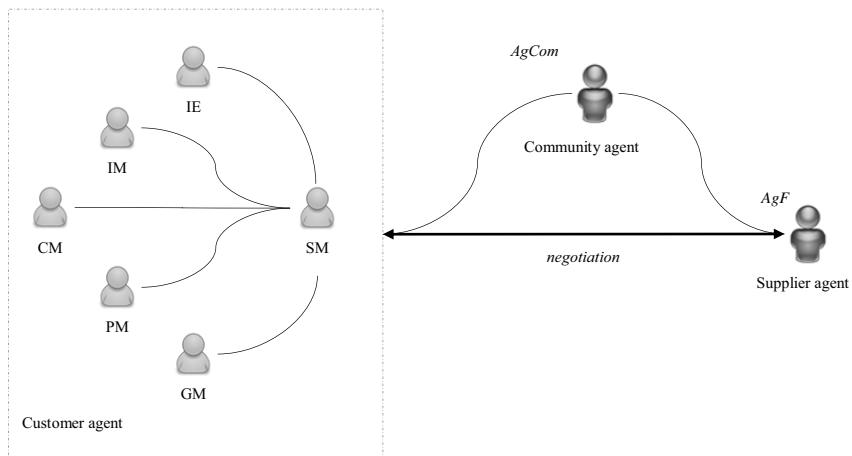


**Fig. 4.** The sequence diagram of a simple negotiation protocol between customer and supplier agents

## 4 Using a Multi-context System to Formalize the Agents

In order to implement a practical platform, we have formalized the customer, supplier and community agents using the multi-context approach of Parsons and Sabater.

The customer agent is a multi-agent system with the following entities (figure 5): GM (Goal Manager) - generates the necessary goals to solve a situation and monitor their status; PM (Plan Manager) - it is a repository of plans to accomplish each goal; CM (Configuration Manager) - represents the active negotiation configuration; IM (Instance Manager) - it is a store for the negotiation instances received from the supplier; IE (Inference Engine) - it calculates the cost of each received instance and selects the one with the smallest cost; SM (Social Manager) - it is the communication node between internal and external entities.



**Fig. 5.** The functional schema of the customer agent

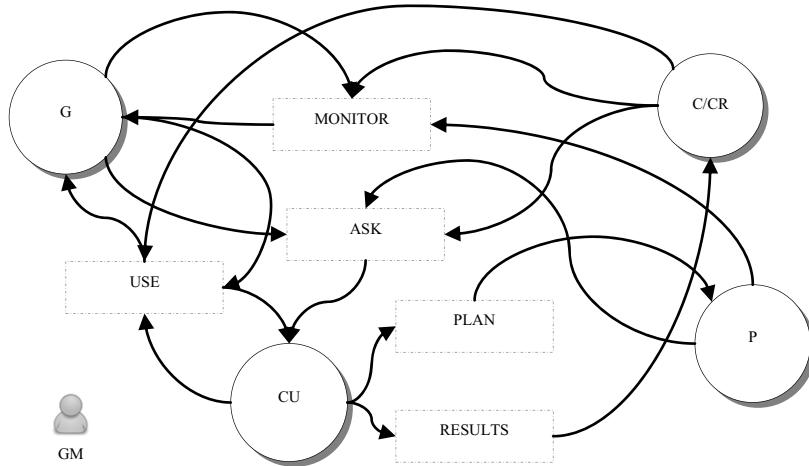
The Goal Manager has the following internal units (figure 6): CU (Communication Unit) - receives and sends messages; G (Goals) - it is a repository of goals; C/CR (Community opinions and costs storage) - it is a storage unit for the community opinions and the calculated costs of negotiation instances; P - it is a storage for plans.

We define a *negotiation offer* as the sets of values sent by the supplier agent. The *negotiation instance* represents the set of values received by the supplier agent.

$$\begin{aligned} O_F &= \{ [id_{ch_i}, (ch_1, val), (ch_2, val), \dots] \dots \} \\ I_C &= \{ [(ch_1, val), (ch_2, val), \dots] \} \end{aligned} \quad (1)$$

where *ch* is the characteristic and *val* is the value.

For each entity with have bridge rules that relates formulae in different units. When the communication unit CU receives an *ask* message from the social manager agent SM to analyze the supplier's *offer*, the USE bridge generates the analyze goal G.



**Fig. 6.** GM (Goal Manger) - generates the necessary goals to solve a situation

$$USE = \frac{CU > ask(AgC/SM, AgC/GM, offer(AgC, AgF, O_F), \{\})}{G : goal(analyze(O_F))} . \quad (2)$$

If the functional unit G has fulfilled its goals (*done*) and the computation was not yet flagged as finished, and we have the *result* at the C/CR unit, the CU unit may send the *answer* to the IM agent.

$$G : done(community(O_F)) \\ G : done(analyzeEach(O_F)) \\ G : not(done(analyze(O_F))) \\ USE = \frac{C / CR : result(analyzeEach(O_F), \{I_C\})}{CU : answer(AgC/GM, AgC/IM, result(O_F, I_C))} . \quad (3)$$

*G : done(analyze(O\_F))*

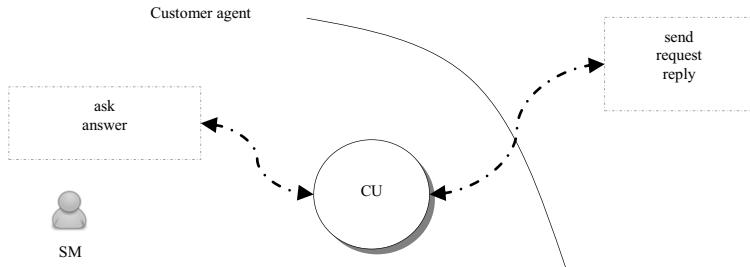
The IM agent *updates* its knowledge base and sends the instance to the SM agent.

$$USE = \frac{CU > answer(AgC/GM, AgC/IM, result(O_F, I_C))}{H : update(result(O_F, I_C))} . \quad (4)$$

$$USE = \frac{H : update(result(O_F, I_C))}{CU : answer(AgC/IM, AgC/SM, check(AgF, AgC, I_C), \{\})}$$

The SM agent (figure 7) is a communication router that translates messages from and to the supplier agent and routes the messages internally.

The *check* type messages received from the IM agent are routed to the supplier agent.



**Fig. 7.** SM (Social Manger) - routes the messages internally and externally

$$\begin{aligned}
 & \text{answer}(\text{AgC}/\text{IM}, \text{AgC}/\text{SM}, \text{check}(\text{AgF}, \text{AgC}, I_c), \{\}) \\
 \rightarrow & \text{send}(\text{AgC}, \text{AgF}/\text{SM}, \text{check}(\text{AgF}, \text{AgC}, I_c))
 \end{aligned} \tag{5}$$

## 5 Conclusion

In this paper, we briefly presented a simple negotiation protocol and the formalism for agents to accomplish it. Using the formalism of Parsons and Sabater, we have formalized completely each agent. At present time, we have implemented a prototype using the JADE multi-agent platform. Regarding the community agent, we have an Android mobile application that gathers emotional knowledge in different experiments. Further research will augment the negotiation protocol with an argumentation framework that deals with the defined emotional aspects. We think that this is a natural step forward for a more human-like experience for the retail area of electronic commerce.

## References

1. Sierra, C., Dignum, F.: Agent-mediated Electronic Commerce: Scientific and Technological Roadmap. In: Sierra, C., Dignum, F.P.M. (eds.) *AgentLink 2000*. LNCS (LNAI), vol. 1991, pp. 1–18. Springer, Heidelberg (2001)
2. Fatima, S.S., Wooldridge, M., Jennings, N.R.: On Efficient Procedures for Multi-issue Negotiation. In: Fasli, M., Shehory, O. (eds.) *TADA/AMEC 2006*. LNCS (LNAI), vol. 4452, pp. 31–45. Springer, Heidelberg (2007)
3. Humaine Network of Excellence, IST FP6 (2004), <http://www.emotion-research.net>
4. Emotion Markup Language EmotionML, World Wide Web Consortium W3C (2008) <http://www.w3.org/2005/Incubator/emotion/>  
XGR-emotionml-20081120/
5. Russell, J.: A circumplex model of affect. In: Judd, C., Simpson, J., King, L. (eds.) *Journal of Personality and Social Psychology*, vol. 39(6), pp. 1161–1178 (1980)
6. Stahl, A.: Designing for Emotional Expressivity. Licentiate Thesis. Institute of Design, Umeå University. Sweden (2006)

7. Adam, C.: The Emotions: From Psychological Theories to Logical Formalization and Implementation in a BDI agent. PhD Thesis, Institut de Recherche en Informatique de Toulouse - IRIT, France (2007)
8. Picard, R.W.: Affective Computing. MIT Press, Cambridge (1997), ISBN 0-262-16170-2
9. Ortony, A., Clore, G.L., Collins, A.: The cognitive structure of emotions. Cambridge University Press, Trumpington Street (1988)
10. Jascanu, N., Jascanu, V., Bumbaru, S.: Toward Emotional E-Commerce: The customer agent. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part I. LNCS (LNAI), vol. 5177, pp. 202–209. Springer, Heidelberg (2008)
11. Parsons, S., Jennings, N.R., Sabater, J., Sierra, C.: Agent Specifications Using Multi-context Systems. In: d'Inverno, M., Luck, M., Fisher, M., Preist, C. (eds.) UKMAS Workshops 1996-2000. LNCS (LNAI), vol. 2403, pp. 205–226. Springer, Heidelberg (2002)
12. Sabater, J., Sierra, C., Parsons, S., Jennings, N.R.: Using Multi-context Systems to Engineer Executable Agents. In: Jennings, N.R. (ed.) ATAL 1999. LNCS, vol. 1757, pp. 260–276. Springer, Heidelberg (2000)
13. Jascanu, N.: Intelligent software agents for electronic commerce. Contributions. PhD Summary, supervisor Prof.dr.ing. Severin Bumbaru, Department of Computer Science, University “Dunarea de Jos” of Galati (2009),  
[http://www.ugal.ro/doc/Rezumat\\_teza\\_doctorat\\_Jascanu\\_Nicola\\_e.pdf](http://www.ugal.ro/doc/Rezumat_teza_doctorat_Jascanu_Nicola_e.pdf) (Available in Romanian)

# Distributed Ant Colony Clustering Using Mobile Agents and Its Effects

Ryotaro Oikawa<sup>1</sup>, Masashi Mizutani<sup>1</sup>, Munehiro Takimoto<sup>1</sup>,  
and Yasushi Kambayashi<sup>2</sup>

<sup>1</sup> Department of Information Sciences, Tokyo University of Science, Japan

<sup>2</sup> Department of Computer and Information Engineering,  
Nippon Institute of Technology, Japan

**Abstract.** This paper presents a new approach for controlling mobile multiple robots connected by communication networks. The control mechanism is based on a specific Ant Colony Clustering (ACC) algorithm. In traditional ACC, an ant convey an object, but in our approach, the ant is implemented as a mobile software agent that controls the robot which is corresponding to an object, so that the object moves to the direction ordered by the ant agent. In this time, the process in which an ant searches an object corresponds to a sequence of migrations of the ant agent, which is much more efficient than the search by a mobile robot. In our approach, not only the ant but also the pheromone is implemented as a mobile software agent. The mobile software agents can migrate from one robot to another, so that they can diffuse over robots within their scopes. In addition, since they have their strengths as vector values, they can represent mutual intensification as synthesis of vectors. We have been developing elemental techniques for controlling multiple robots using mobile software agents, and showed effectiveness of applying them to the previous ACC approach which requires a host computer that centrally controls mobile robots. The new ACC approach decentralizes the mobile robot system, and makes the system free from special devices for checking locations.

**Keywords:** Mobile agent, Ant Colony Clustering, Intelligent robot control.

## 1 Introduction

Ant colony clustering is one of the clustering methods that model the behaviors of social insects such as ants. The ants collect objects that are scattered in a field. In ant colony clustering, artificial ants imitate the real ants and gradually form several clusters. The application we have in our mind is a kind of intelligent carts.

When we pass through terminals of the airport, we often see carts scattered in the walkway and laborers manually collecting them one by one. It is a laborious task and not a fascinating job. It would be much easier if carts were roughly gathered in any way before the laborers begin to collect them.

For example, in order to achieve such clustering, we can take advantage of Ant Colony Clustering (ACC) algorithm which is an Ant Colony Optimization (ACO) specialized for clustering objects. ACO is a swarm intelligence-based method and a multi-agent system that exploits artificial stigmergy for the solution of combinatorial optimization problems. ACC is inspired by the collective behaviors of ants, and Deneubourg formulated an algorithm that simulates the ant corps gathering and brood sorting behaviors [2].

We previously proposed an ACC approach using mobile software agents [2]. In the approach, a mobile software agents traverses all the robots that are corresponding to objects, collecting the information of their locations. The mobile software agent conveys the location information to a host computer, where the ACC algorithm is performed to determine the locations of clusters to which robots are supposed to move. Each robot has certain pheromone value so that if the strength of the pheromone is strong, it means a substantial number of robots form a cluster and the robots nearby are supposed to move toward the cluster. Also the robots in the cluster are *locked* to prevent being moved. This mechanism contributes to stabilizing the clusters and giving them relatively monotonic growth.

Although the previous approach yielded favorable results in preliminary experiments, it required a host computer for centrally managing locations of robots and executing ACC algorithm. Therefore, there is some time lag until the robots were clustered, and it could prevent the robots from dynamically reflecting changes of the circumstance.

In this paper, we propose a new pheromone base ACC approach using mobile software agents. In our new approach, called distributed ACC, the host for centrally controlling robots is not necessary. Some *Ant* agents, which is mobile software agents corresponding to ants, iteratively traverse robots (intelligent carts). Furthermore, the pheromone is also implemented as a collection of mobile software agents. We call them *Pheromone* agent. Each Pheromone agent is created on a robot included in a cluster. Once it is created on the robot, it migrates to other robots within the scope. It has a datum representing strength and direction, which is used for guiding Ant agents. Multiple Pheromone agents reaching the same robot are combined into one single agent with the synthesized strength and direction.

These features of our distributed ACC algorithm provide the following contributions:

1. dynamically reflecting circumstances, and
2. saving energy consumption because robots without Ant agent consume very little energy.

We also show that the number of Ant agents can be decreased without sacrificing efficiency of clustering in our experimental results.

The structure of the balance of this paper is as follows. In the second section, we describe the background. The third section briefly describes the traditional ACC algorithms. The fourth section describes the mobile software agent system that performs the arrangement of the multiple robots. The agent system consists of

several Ant and Pheromone agents. In this section, we show how our distributed ACC algorithm performs the quasi optimal clustering of the mobile robots. The fifth section describes some numerical experiments based on a simulator. Finally, we conclude in the sixth section and discuss future research directions.

## 2 Background

Kambayashi and Takimoto have proposed a framework for controlling intelligent multiple robots using higher-order mobile agents [34]. The framework helps users to construct intelligent robot control software by migration of mobile agents. Since the migrating agents are higher-order, the control software can be hierarchically assembled while they are running. Dynamically extending control software by the migration of mobile agents enables them to make base control software relatively simple, and to add functionalities one by one as they know the working environment. Thus they do not have to make the intelligent robot smart from the beginning or make the robot learn by itself. They can send intelligence later as new agents. Even though they demonstrate the usefulness of the dynamic extension of the robot control software by using the higher-order mobile agents, such higher-order property is not necessary in our setting. We have employed a simple, non higher-order mobile agent system for our framework. They have implemented a team of cooperative search robots to show the effectiveness of their framework, and demonstrated that their framework contributes to energy saving of multiple robots [45]. They have achieved significant saving of energy. Our simple agent system should achieve similar performance on the task of clustering mobile robots.

On the other hand, algorithms that are inspired by behaviors of social insects such as ants to communicate to each other by an indirect communication called stigmergy are becoming popular [67]. Upon observing real ants' behaviors, Dorigo et al. found that ants exchanged information by laying down a trail of a chemical substance (called pheromone) that is followed by other ants. They adopted this ant strategy, known as ant colony optimization (ACO), to solve various optimization problems such as the traveling salesman problem (TSP) [7]. Deneubourg has originally formulated the biology inspired behavioral algorithm that simulates the ant corps gathering and brood sorting behaviors [1]. Wang and Zhang proposed an ant inspired approach along this line of research that sorts objects with multiple robots [8]. Lumer has improved Deneubourg's model and proposed a new simulation model that is called Ant Colony Clustering [9]. His method could cluster similar objects into a few groups.

## 3 The Ant Colony Clustering

The coordination of an ant colony is composed by the indirect communication through pheromones. In traditional ACO system, artificial ants leave pheromone signals so that other artificial ant can trace the same path [67]. Kambayashi et al. have developed an ACC system based on pheromone signals [2]. Randomly

walking artificial ants have high probability to pick up an object with weak pheromone, and to put the object where it senses strong pheromone. They are not supposed to walk long distance so that the artificial ants tend to pick up a scattered object and produce many small clusters of objects. When a few clusters are generated, they tend to grow.

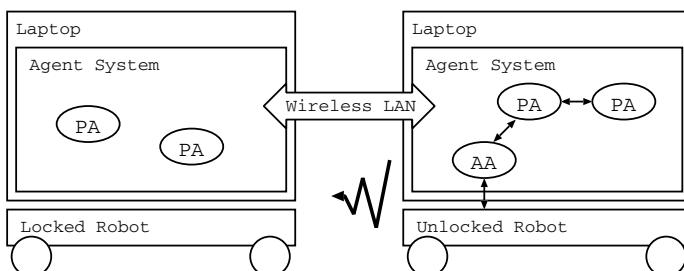
Since the purpose of the traditional ACC is clustering or grouping objects into several different classes based on some properties; it is desirable that the generated chunks of clusters grow into one big cluster so that each group has distinct characteristic. In our system, however, we want to produce several roughly clustered groups of the same type, and make each robot have minimum movement. (We assume we have one kind of cart robots, and we do not want robots move long distance.)

In the implementation of our ACC algorithm, when the artificial ants are generated, they have randomly supplied initial positions and walking directions. While an artificial ant performs random walk, when it finds an isolated object, it picks up the object, and continues random walk. While the artificial ant performs random walk, when it senses strong pheromone, it put the conveying object. The artificial ants repeat this simple procedure until the terminate condition is satisfied. These behaviors in the ACC are achieved using mobile software agents below.

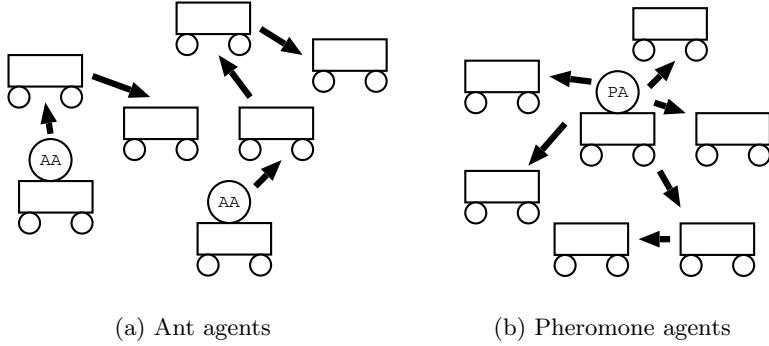
## 4 The Mobile Agents

Our system model consists of robots and two kinds of mobile software agents as shown by Fig. 1. The robots have simple capabilities of movement such as driving wheels, detecting objects through a camera, and checking obstacles through supersonic sensors, and they can communicate each other through a communication network such as a wireless LAN. Any other noble capability like determining absolute locations and directions through GPS, RFIDs, or other devices are not required.

All the controls for the mobile robots are achieved through the mobile agents. They are: 1) Ant agents (AA), and 2) Pheromone agents (PA). Some mobile agents (AA) traverse robots scattered in the field one by one to search isolated



**Fig. 1.** The relation between agents and robots

**Fig. 2.** Traversals of mobile agents

robots as shown by Fig. 2(a). In this traversal, migrating to a robot imitates a behavior of an ant that finds an object, and picks it up. Upon arriving on a robot, AA controls the robot and drives it. AA is the agent that drives a robot. Without AA, robots are just sitting. But AA has no knowledge of direction it should lead the robot so that it just give the robot a random walk. PA is the agent that guides AA to which direction it should drive the robot. PA is created on one of the robots that form a cluster. Randomly walking robots happen to bump each other and to create a cluster by chance. Robots that form a certain size of cluster are locked so that they become a nucleus of a growing cluster. A robot situated in a center location of such a growing cluster creates PA. In order to imitate disseminating pheromone, PA migrates to other robots as shown by Fig. 2(b). Once PA reaches the robot where AA exists, the PA guides the AA to the locked robot where the PA originated.

In the following sections, we describe the details of Ant agents and Pheromone agents in our distributed ACC algorithm.

#### 4.1 Ant Agents

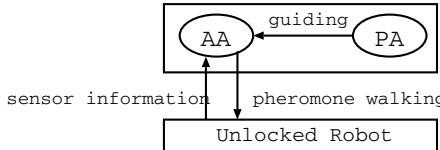
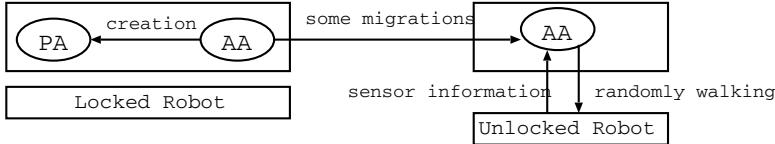
AA has IP list of all the robots in order to traverse them one by one. If it has visited all the robots, it goes back to the home host for administration of the robot system to check the number of robots, and updates its IP list in the following cases:

1. some new robots have been added, and
2. some robots have been broken.

However, those cases are so rare that the home host almost never interferes.

In addition, AA can observe the states of robots as follows:

1. it is being used by a customer,
2. it is locked, and
3. it is free, i.e., unlocked and not used by any customer.

**Fig. 3.** Pheromone walking**Fig. 4.** The behavior of picking up an object and randomly walking

When AA visits a robot that is used by customers or that is locked, it immediately leaves it without doing anything. If it visits a free robot i.e. not used and isolated, it begins to control it. At this time, if there is PA on that robot, AA makes the robot move following the guidance of the PA as shown by Fig. 3, otherwise AA drives the robot randomly as shown by Fig. 4.

Once the robot controlled by AA is locked because of reaching a suitable cluster, AA has to leave the robot and start migration again to find another free robot. Before that, AA creates PA if there is no PA on the robot, as shown by Fig. 4. As a result, PA starts behaving as a pheromone as shown in the next section.

## 4.2 Pheromone Agents

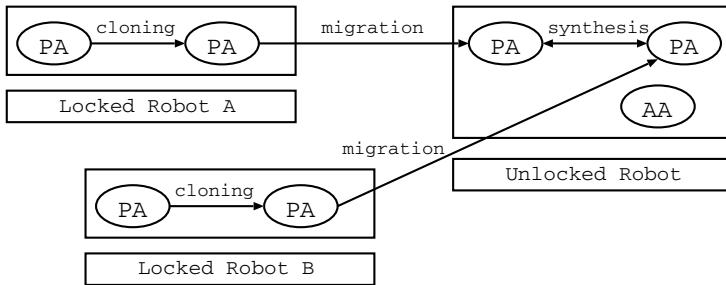
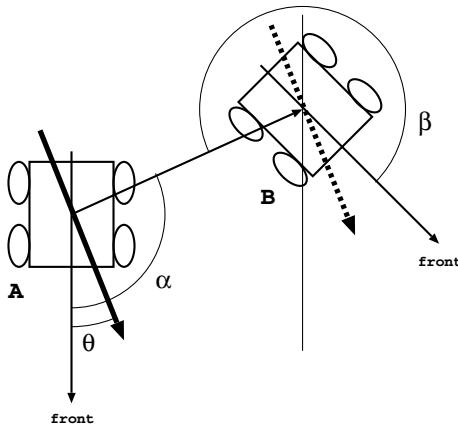
Since the purpose of ACC is to nurture clusters, the number of objects in a cluster affects the probability of picking up and putting down an object. Such a property of a cluster is modeled by a pheromone attracting ants.

The pheromone intrinsically has the following properties:

1. the strength increases in proportional to the number of objects,
2. the strength decreases in proportional to the distance,
3. it has a scope and does not affect out of the scope, and
4. the strength decreases as time elapses.

The state of a PA depends on the number of objects, the distance between the objects, and elapsing time. Robots acquire those data through the camera and the timer on them. The data are represented as a vector value inside PA.

**The Migration of PA.** PA is created by AA on a locked robot as shown by Fig. 4. Once PA is created, it clones itself, and the newly created PA migrate to other robot as shown by Fig. 5. PA has a vector value in it. The length and the direction of

**Fig. 5.** Synthesis of pheromones**Fig. 6.** Migration of pheromones

the vector value means strength of attracting and guiding direction relative to the front of a robot respectively. In Fig. 6, thick arrows represent vector value, where the direction of vector value on robot A is represented as the angle  $\theta$ .

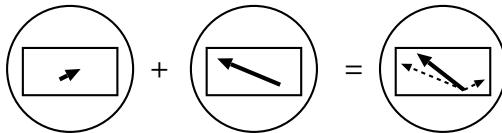
Considering scope size  $S$  and hold time  $T$ , the absolute value of vector  $v$  is represented as follows, where  $distance$  is the distance that the PA moved,  $time$  is elapsing time since the PA was born,  $K$  is the maximal value that  $v$  can take, and  $C$  is a suitable coefficient:

$$|v| = \min(C * (S - distance) * (T - time), K)$$

$C$  is just used to adjust the equation for each circumstance, and  $K$  is used to prevent the synthesized value from being too big.

Vector value  $v$  is initially computed by PA itself after the migration as follows:

1. PA migrates from robot A to robot B,
2. PA on B observes A where PA was born, gets the distance between A and B, and the direction ( $\beta$  in Fig. 6) from B to A, and
3. PA sets these values as an initial vector value.



**Fig. 7.** Synthesizing Pheromone agents

On the other hand, the direction after some migrations is not so simple. If PA migrates from robot A to robot B as shown in Fig. 6, direction  $\theta'$  relative to the front of robot B is computed by the following formula:

$$\theta' = \pi - \alpha + \beta + \theta$$

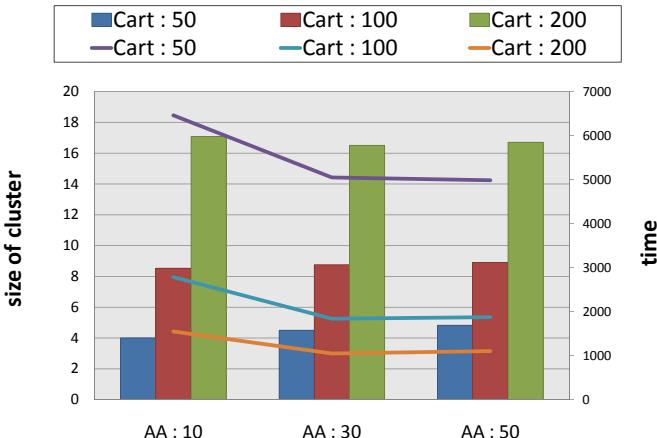
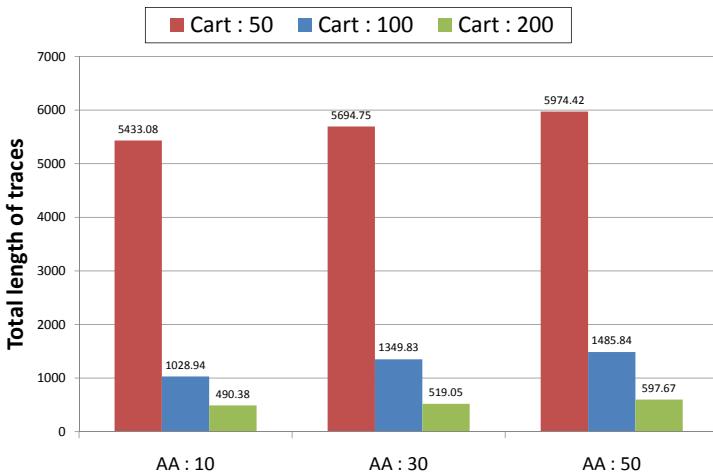
**The Fusion of PAs.** The moving robot controlled by AA can receive several PAs. In such a case, AA needs the consistent guidance of the PAs. Several PAs are fused into one PA as shown by Fig. 5. Since the data with PAs are vector values, they can be easily synthesized as shown by Fig. 7.

## 5 Experimental Results

In order to demonstrate the effectiveness of our distributed ACC algorithm, we have built a simulator for clustering robots (intelligent carts) and conducted experiments on it. On the simulator, moving and rotating speed of robots, and lags required in agent migration and object recognition are based on real values in the previous experiments [5]. In the experiments, we employed three wheeled mobile



**Fig. 8.** Simulated robots

**Fig. 9.** The number of clusters and the time for convergence**Fig. 10.** The total length of traces

robots, which are called PIONEER 3-DX (Fig. 8), as the platform for our prototype system. Each robot has two servo-motors with tires, one camera and sixteen sonic sensors. The power is supplied by rechargeable battery. A PIONEER 3-DX has one servo-motor and sensor controller board that sends/receives data to/from a host computer through a USB cable. The camera is directly operated by the host with the middleware which is called ERSP. Each robot holds one notebook computer as its host computer. Our control agents migrate to these host computers by wireless LAN.

We have conducted several experiments with different number of robots and AAs, and compared their results. Fig. 9 shows the average size of created clusters (rectangles) and the average time taken till convergence (lines). As shown by the figure, the size of cluster seem to be around 8.5 % of the number of robots for any number of AAs. Also, the time till convergence for 50 AAs is equal to the time for 30 AAs though it is less than the time for 10 AAs. Fig. 10 shows the total length of traces of all robots. As shown in the figure, the less the number of AAs is, the shorter the length of traces is. Notice that the shorter trace means less energy consumption. These results demonstrate the beneficial features of our ACC, in which the energy consumption can be decreased on some levels without sacrificing efficiency.

## 6 Conclusions

We have proposed a framework for controlling mobile multiple robots connected by communication networks. In this framework, scattered mobile multiple robots autonomously form into several clusters based on the ant colony clustering (ACC) algorithm. The ACC algorithm finds quasi-optimal positions for the mobile multiple robots to form the clusters.

In our distributed ACC algorithm, we introduced two kinds of mobile software agents:, i.e. ant agents and pheromone agents. The ant agents represent the artificial ants. They see the mobile robots as objects and drive them to the quasi-optimal positions. The pheromone agents represent pheromone and diffuse the effects by migrations. In general, making mobile multiple robots perform the ant colony optimization is impossible due to enormous inefficiency. Our approach does not need the ant-like robots and other special devices. The preliminary experiments shows that our approach is efficient enough and it enables suppressing energy consumption.

So far we are not aware of any multi-robot system that integrates pheromone as a control means as Deneubourg envisaged in his monumental paper [1]. The preliminary experiments suggest favorable results. We will show the feasibility of our multi-robot system using Ant and Pheromone agent base ACC by further numerical experiments.

## Acknowledgements

This work is supported in part by Japan Society for Promotion of Science (JSPS), with the basic research program (C) (No. 20510141), Grant-in-Aid for Scientific Research.

## References

1. Deneubourg, J., Goss, S., Franks, N.R., Sendova-Franks, A.B., Detrain, C., Chreien, L.: The dynamics of collective sorting: Robot-like ant and ant-like robot. In: Proceedings of the First Conference on Simulation of Adaptive Behavior: From Animals to Animats, pp. 356–363. MIT Press, Cambridge (1991)

2. Kambayashi, Y., Ugajin, M., Sato, O., Tsujimura, Y., Yamachi, H., Takimoto, M., Yamamoto, H.: Integrating ant colony clustering to a multi-robot system using mobile agents. *Industrial Engineering and Management Systems* 8(3), 181–193 (2009)
3. Kambayashi, Y., Takimoto, M.: Higher-order mobile agents for controlling intelligent robots. *International Journal of Intelligent Information Technologies* 1(2), 28–42 (2005)
4. Takimoto, M., Mizuno, M., Kurio, M., Kambayashi, Y.: Saving energy consumption of multi-robots using higher-order mobile agents. In: Nguyen, N.T., Grzech, A., Howlett, R.J., Jain, L.C. (eds.) *KES-AMSTA 2007. LNCS (LNAI)*, vol. 4496, pp. 549–558. Springer, Heidelberg (2007)
5. Nagata, T., Takimoto, M., Kambayashi, Y.: Suppressing the total costs of executing tasks using mobile agents. In: Proceedings of the 42nd Hawaii International Conference on System Sciences. IEEE Computer Society CD-ROM (2009)
6. Dorigo, M., Birattari, M., Stützle, T.: Ant colony optimization—artificial ants as a computational intelligence technique. *IEEE Computational Intelligence Magazine* 1(4), 28–39 (2006)
7. Dorigo, M., Gambardella, L.M.: Ant colony system: a cooperative learning approach to the traveling salesman. *IEEE Transaction on Evolutionary Computation* 1(1), 53–66 (1996)
8. Wand, T., Zhang, H.: Collective sorting with multi-robot. In: Proceedings of the First IEEE International Conference on Robotics and Biomimetics, pp. 716–720 (2004)
9. Lumer, E.D., Faiesta, B.: Diversity and adaptation in populations of clustering ants, from animals to animats 3. In: Proceedings of the 3rd International Conference on the Simulation of Adaptive Behavior, pp. 501–508. MIT Press, Cambridge (1994)

# Monitoring a Multi-Agent System Evolution through Iterative Development

Yves Wautelet<sup>1</sup> and Manuel Kolp<sup>2</sup>

<sup>1</sup> Hogeschool-Universiteit Brussel, Belgium

[yves.wautelet@hubrussel.be](mailto:yves.wautelet@hubrussel.be)

<sup>2</sup> Université catholique de Louvain, Belgium

[manuel.kolp@uclouvain.be](mailto:manuel.kolp@uclouvain.be)

**Abstract.** Iterative development is a well known project management technique which has notably been popularized in object-oriented development through the use of the Rational Unified Process. Such an approach is nevertheless always driven by milestones rules on requirements or design models while it could be applied onto the evolution of a Multi-Agent System (MAS) among the software project. We propose in this paper to define a MAS acting as a wrapper for COTS-based software development and to monitor its evolution through an iterative template. Rules are evaluated by project managers so that the MAS evolves on the bases of agents decisions with human monitoring. The paper formalizes some relevant concepts in the perspective of a component-based MAS development, it defines what happens at transaction level as well as a series of constraints to properly monitor the project evolution.

## 1 Introduction

Multi-Agent Systems (MAS) have driven a lot of research over the last 20 years. Even if MAS development methodologies as Tropos [1] define adequate models for describing organizations and designing active software, the processes remain basic with no or poor focus on project management.

In this paper we firstly define the concepts required to build a MAS acting as a wrapper for COTS components-based software development. We then define specific rules (i.e. objectives) to monitor its evolution over the 4 milestones of an iterative project management template. At runtime, the MAS agents ask for the resolution of defined procedures to a mediator agent delegating their functional execution to the best possible component offer. The MAS is interfaced to the components using an API or any other standardized technical mean. The MAS is subject to evolution in performance since it exploits in the form of low-level functions (i.e. capabilities) various components that can be integrated or not and customized on the basis of managerial decisions. Evolution is monitored iteratively by project managers ensuring that phase objectives are met at milestones.

Contributions of this paper are multiple:

- Since the goal is to propose a framework for monitoring a MAS evolution in terms of component selection we need a formal model to build it up.

This formal model is the first contribution of this paper. It goes beyond a pure definition of agent concepts by incorporating access, dialogue and delegation to (third party) components. A transaction procedure is also defined. The MAS represents an abstraction layer between the modelled business processes and the available components so that it constitutes an integration framework for COTS component-based software development;

- Then, we need a template and the definition of milestone rules to monitor the project evolution. To that extend, we tailor the I-Tropos [2] iterative template to the evaluation of the MAS performance at runtime. The latter is made of 4 phases separated by milestones where the project is subject to evaluation. The formal definition of the conditions for evaluating the project tailored to the MAS runtime reports is the second contribution of the paper;
- Finally, the framework as a whole constitutes an attempt to develop a methodology where the MAS contextual issues (here in terms of component usage) are considered in parallel with the project management ones. This reinforces the interest of adopting iterative templates as the Rational Unified Process [3] in agent-oriented and component-based software development.

The paper is structured as follows. Section 2 overviews the most important software engineering (SE) trends and frameworks in the context of this paper. Section 3 introduces the MAS conceptual model while section 4 the rules for monitoring components integration using an iterative template. Section 5 illustrates the concepts on the development of an OL collaborative platform. Section 6 overviews related work and finally section 7 concludes the paper.

## 2 Problem Statement

This section introduces relevant SE concepts.

### 2.1 Component-Based Software Development

Multiple definitions of COTS-components have been proposed in literature. More than a variety of definitions, Vidger and Dean [4], Carney and Long [5], Basili and Boehm [6] and the Software Engineering Institute in [7] have a variety of views, approaches and conceptualizations of what COTS-components are. These approaches are notably compared in [8] and the key factors differentiating these definitions we want to focus on here are the **accessibility to source code and its modifiability** and **technical specifications**.

While [4], [6] and [7] tend to consider COTS-components as "black boxes", [5] defines a typology for studying the COTS origin and openness (through modification levels). This view is by nature more flexible and in line with the definition proposed by [9]; namely *A COTS product is a commercially available or open source piece of software that other software projects can reuse and integrate into their own products*. The customization dimension is very important in our approach and we only consider customizable components. We moreover define customizable components as COTS products where low level functionalities

following a defined specification can be modified or added into the component with defined cost and quality of service (QoS).

*Component-Based Software Development* (CBSD) is based on the idea to develop software systems by selecting appropriate commercial off-the-shelf (COTS) components and assembling them with well-defined software architecture [10].

## 2.2 The I-Tropos Process

Iterative development is commonly used to get an early user feedback from rapidly sketched software prototypes so that the analysis disciplines can be re-evaluated and corrected to take risks early on into the software project and better meet stakeholders' expectations (see [11][3]). To that extend, iterative templates as I-Tropos [2] have been developed. The latter is tailored on monitoring the MAS evolution within component selection. Iterative refinements are provided by agent delegation choices but also by managerial decisions to develop missing (or existing under performing) capabilities. In an I-Tropos development, each iteration belongs to one of the four phases; a summary of each phase objective is depicted into section 4. These phases are achieved sequentially and have different goals evaluated at milestones through metrics tailored here to MAS component selection at runtime. I-Tropos has notably been adopted here because it uses the i\* (i-star, [12]) framework at analysis stage. In the framework defined here, analysis is performed using the *Strategic Dependency (SD)* and *Strategic Rationale (SR)* diagrams from which tasks and goals are forward engineered through a MAS for resolution (see section 3).

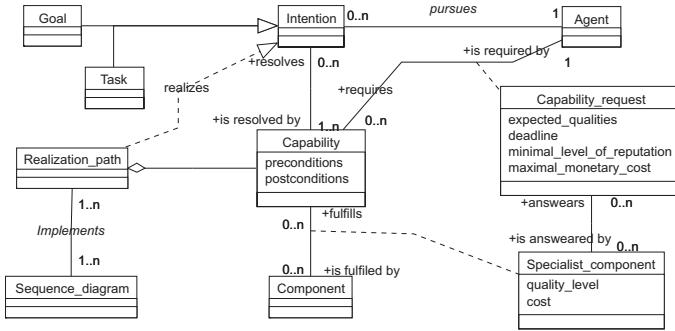
## 3 MAS Conceptual Foundations

This section defines a conceptual model for building a MAS.

### 3.1 MAS Definitions

Figure 1 depicts the relevant MAS concepts and their dependencies using a UML class diagram. The model is structured as follows: the agent pursues a series of intentions which can be tasks or goals in the sense defined by i\* and are then resolved through a series of capabilities by software components under the responsibility of agents in the form of a realization path. A UML sequence diagram can be used to document those realization paths. The MAS architecture assumes capabilities' realization are delegated by agents to software components by the use of capability requests. Components are source code packages able to achieve a capability that is part of its offer.

**Definition 1.** A tuple  $\langle \{(cp_i, q_{cp_i}^c), \dots, (cp_{i+m}, q_{cp_{i+m}}^c)\}, Comp^c \rangle$  is called an component  $c$ , where  $cp_i$  is a capability. The component advertises to the mediator agent its ability to resolve a capability at defined QoS level and cost  $q_{cp_i}^a$ .  $Comp^c$  is assumed to contain all additional properties of the component irrelevant for the present discussion, yet necessary when integrating the component.

**Fig. 1.** The MAS Meta-Model

Format and content of  $Comp^c$  will depend on the component technologies being used. A *Capability* is part of a *Goal* or *Task* realization (those are generalized as *Intentions* into the UML class diagram). Non-available capabilities must be customly developed inside the scope of a component.

**Definition 2.**  $\langle c, cp_i, q_{cp_i}^c, ct_{cp_i}^c \rangle$  associating a capability  $cp_i$  to a quality level  $q_{cp_i}^c$  at cost  $ct_{cp_i}^c$  advertised by component  $c$  is a specialist component  $c_i^{SC}$ . The component must be capable of performing the capability:  $\forall c_i^{SC} = \langle c, cp_i, q_{cp_i}^c, ct_{cp_i}^c \rangle, (cp_i, q_{cp_i}^c, ct_{cp_i}^c) \in c$ .  $ct_{cp_i}^c$  specifies the cost at which the specialist component  $c_i^{SC}$  can realize the capability. Its definition follows a specific cost ontology.

Any component  $c$  that can accomplish  $m > 1$  capabilities can also be seen as a set of specialized components:  $\{c_i^{SC}, \dots, c_{i+m}^{SC}\}$ .

**Definition 3.** A capability  $cp_i$  is  $\langle cp_i^{pre}, \tau_i, cp_i^{post} \rangle$ , where  $cp_i^{pre}$  describes the capability preconditions,  $\tau_i$  is a specification (in some language) of how the component is to execute the capability and  $cp_i^{post}$  describes the conditions true after the capability is executed. Capabilities belong to the set  $\mathbb{CP}$ .

The capability is realized by a component under the responsibility of an agent. Agents are issued of the domain model on the exception of a special agent called the *mediator agent*. The latter is in charge of managing the transaction between the domain specific agents and the integrated COTS-components.

**Definition 4.** A service mediator agent  $a_c^{SM}$  in the multi-agent system is an agent that can execute the transaction procedure:  $t_A \in a_c^{SM}$ .

The transaction procedure  $t_A$  is documented into section [\[3.2\]](#).

**Definition 5.**  $\langle rp_j^t, rp_{cp_j}^c, rp_{cp_j}^a, rpTransit_j, rpState_j \rangle$  is a realization path  $rp_j$ , where  $rp_j^t$  provides the details of the functional specification of the realized intention,  $(rp_{cp_j}^c, rp_{cp_j}^a)$  defines a sequence diagram where  $rp_{cp_j}^c$  represents the

series of capabilities required to realize the intention and  $rp_{cp_j}^a$  the agents responsible for their realization. The two functions label swimlanes and messages with capability information:  $rpTransit_j : rp_{cp_j}^c \mapsto \mathbb{CP}$  is a partial function returning the capability for a given message in the sequence diagram, while  $rpState_j : rp_{cp_j}^a \mapsto \{cp_i^{pre}\}_{cp_i \in \mathbb{CP}} \cup \{cp_i^{post}\}_{cp_i \in \mathbb{CP}}$  maps each message to a condition from the set of all capability preconditions (i.e.,  $\{cp_i^{pre}\}_{cp_i \in \mathbb{CP}}$ ) and postconditions (i.e.,  $\{cp_i^{post}\}_{cp_i \in \mathbb{CP}}$ ). The capability specified on a message must have the precondition and postcondition corresponding to conditions given, respectively, on its origin and its destination swimlane. Realization paths belong to the set  $\mathbb{RP}$ .

Capabilities can thus be understood as a functional decomposition of a "higher level" intention (a task or goal) with the realization path as a success scenario. However in the context of this paper we are mostly focusing on the lowest level functional decomposition.

**Definition 6.** A capability request  $\hat{cp}_j$  is  $\langle Ag^a, cp_j, cp_j^{QoS}, cp_j^D, cp_j^R, cp_j^{cost} \rangle$ , where:

- $Ag^a$  is the Agent requesting the capability realization by software components i.e. the one responsible for the capability realization.
- $cp_j$  is the capability to provide.
- $cp_j^{QoS}$  specifies expected qualities and their required level. Its definition follows a QoS ontology. Whatever the specific QoS ontology, expected qualities are likely to be specified as (at least)  $cp_j^{QoS} = \langle (p_1, d_1, v_1, u_1), \dots, (p_r, d_r, v_r, u_r) \rangle$ , where:
  - $p_k$  is the name of the QoS parameter (e.g., connection delay, standards compliance, and so on).
  - $d_k$  gives the type of the parameter (e.g., nominal, ordinal, interval, ratio).
  - $v_k$  is the set of desired values of the parameter, or the constraint  $<, \leq, =, \geq, >$  on the value of the parameter.
  - $u_k$  is the unit of the property value.
- $cp_j^D$  is a deadline, specified as a natural.
- $cp_j^R$  specifies minimal levels of reputation over task quality parameters that any component participating in the provision of the given capability must satisfy.
- $cp_j^{cost}$  is the maximal monetary cost the agent requesting the capability is ready to pay to obtain the service.

Capability requests belong to the set  $\mathbb{REQ}$ .

### 3.2 The Transaction

When a capability request  $\hat{cp}_i$  is submitted to the mediator agent with  $cp_i$  as capability to provide. Let  $W[]$  be the ordered vector containing a weighting

factor for all of the aspects of the QoS ontology for the particular business logic,  $\forall cp_j \in \mathbb{CP}$  where  $cp_j$  meets the functional specification of  $cp_i$ ,  $cp_j^{QoS} = \langle (p_1, d_1, v_1, u_1), \dots, (p_r, d_r, v_r, u_r) \rangle$ , with  $v_k$  being the value of the quality parameter:  $cp_{best}$  is the capability with  $\max(W[1]*v_1 + W[2]*v_2 + \dots + W[r]*v_r)$  value. The mediator agent  $a_c^{SM}$  will delegate the functional execution of the capability request to the component advertizing  $cp_{best}$ .

## 4 Monitoring the MAS' Components with Iteration Milestones

Iterative MASs is not a genuine subject. Nevertheless this paper focuses on monitoring the MAS at runtime using the I-Tropos project management template (largely inspired by the Rational Unified Process) traditionnaly used to monitor requirements elicitation and software design. This is original in two ways. First such a template does traditionnaly not evaluate an application at runtime and second this allows integrating agent-oriented development within project management. Iterative refinements are provided by agent delegation choices but also by managerial decisions to develop missing (or existing under performing) capabilities. In an I-Tropos development, each iteration belongs to one of the four phases; a summary of each phase objective is depicted into this section as well as, for each of them, rules applied to capabilities at runtime.

**Setting.** The setting phase is concerned with the first sketch of the MAS through identifying the main agents and the capabilities they are responsible for. Each of the capability requests must be able to find at least one specialized component without focus on cost and QoS.

**Condition 1.** A condition for evaluating the MAS state at setting phase milestone is:  $\forall \hat{cp}_j \in \mathbb{REQ}$  where  $cp_j$  is the capability to provide,  $\exists c_i^{SC} \mid cp_j \in c_i^{SC}$ .

**Blueprinting.** The blueprinting phase is concerned with a first executable version of the MAS with each of the capability requests adressable by at least one of the specialist components over minimal QoS.

**Condition 2.** A condition for evaluating the MAS state at blueprinting phase milestone is:  $\forall \hat{cp}_j \in \mathbb{REQ}$  where  $cp_j$  is the capability to provide,  $\exists c_i^{SC} \mid cp_j \in c_i^{SC} \wedge q_{cp_i}^c \geq cp_j^{QoS}(v_k)$ .

**Building.** The building phase is concerned with a first executable version of the MAS with each of the capability requests adressable by at least one of the specialist components over minimal QoS and under maximal cost.

**Condition 3.** A condition for evaluating the MAS state at building phase milestone is:  $\forall \hat{cp}_j \in \mathbb{REQ}$  where  $cp_j$  is the capability to provide,  $\exists c_i^{SC} \mid cp_j \in c_i^{SC} \wedge q_{cp_i}^c \geq cp_j^{QoS}(v_k) \wedge ct_{cp_i}^c \leq cp_j^{cost}$ .

**Setuping.** Even if the MAS is running "at equilibrium" (which means it uses the best possible specialist component available for each of the capability requests) new components can be available onto the market with better performance. This last phase monitors the inclusion of such components and has no formal exit milestone since it virtually never ends.

## 5 Motivating Example

This section introduces the application of the framework onto the development of an e-collaboration platform for outbound logistics (OL).

### 5.1 Application Domain

OL is the process related to the movement and storage of products from the supplier to the end user. In the context of this paper we mostly focus on transportation decisions, which will additionally provide information for better internal storage. The actors of the supply chain play different roles in the OL flow. The producer will be a logistic client in its relationship with the raw material supplier, which will be considered as the shipper. The carrier will receive transportation orders from the shipper and deliver goods to the client, while relying on the infrastructure holder and manager. In its relation with the intermediary wholesaler, the producer will then play the role of the shipper and the wholesaler will be the client.

One of the goals of the *TransLogistTIC* project ([www.translogistic.be](http://www.translogistic.be)) is to develop of an online collaborative platform allowing chargers, carriers, infrastructure managers and final clients (the major OL actors) to share information for a better optimization of the logistic chain. Such a platform is developed on the basis of the business logic of each actor and requires the development of a custom applicative package having access to third party components delivering

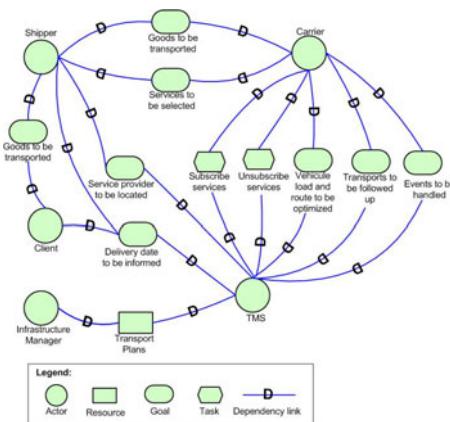


Fig. 2. Strategic Dependency Diagram of Track Transports

functionalities and information for specific issues. More particularly, reuse can be made from *Fleet Management Systems*, *Warehouse Management Systems*, *Enterprise Resource Planning* systems and *Transportation Management Systems (TMS)*. TMSs are computer software designed to manage transportation operations, they notably aid in determining the most efficient and most cost-effective way to execute the movement of product(s). They are the ones useful in the context of this illustrative example. Figure 2 depicts the i\* SD diagram of the *Track Transports* service which is (notably) provided by the platform and where TMS-components are involved in.

## 5.2 Platform Development

Due to a lack of space, we only focus here on the *Select Most Adequate Transport* task. The aim of this task is, for a carrier, to select the offer the most adequate transport (if he has one) to transport the shipper's demand (for a defined amount of merchandise to be transported from a source to a destination under defined constraints as dates, mode of transport, etc).

To select the most adequate transport the MAS involves three main actors: the *Shipper*, *Carrier* and *Infrastructure Manager* (agents are the instance of these actors). Capabilities that agents are responsible for and modelled following the model of section 3 are extensively described in Figure 3. Capabilities realization are delegated to the best possible specialist component using the procedure

No	Capabilities	Description	Responsible Agent	Ressource
1	createLogisticRequest()	When the Shipper requires to transport merchandise from a source to a destination (generally to fulfill an order or for internal reasons) it creates a logistic request (which remains an internal concept)	Shipper	logistic_request
2	handleTransportationCall(logistic_request)	The Shipper transmits a Logistic_Request in the form of a Transportation_Call to a Carrier which is in charge of proposing a planning to realize them (note that the Logistic_Request is an abstract requirement, the Transportation_Call is a Logistic_Request with requested departures and arrival dates/times addressed to a Carrier)	Carrier	logistic_request
3	evaluateTransportationServiceOffer(transportation_call)	The Carrier disposes of a transportation offer in the form of Transportation Services. So it is in charge of evaluating the possible matches between the demand (the transportation call) and its own offer.	Carrier	service
4	relaxConstraints(transportation_call)	The Carrier eventually needs relaxing the constraints to match transportation offer and demand.	Carrier	transportation_proposal
5	computeRequiredOperations(transportation_proposal)	The Carrier computes the required logistic operations needed to sucessfully achieve the possible transports	Carrier	transportation_proposal
6	evaluateInfrastructureAvailability(transportation_proposal)	The Infrastructure Manager manages a series of ressources to make transportsations possible. Ressources are limited and should be booked; on request it answers on disposals.	Infrastructure_Manager	transportation_proposal
7	acceptResolutionSequence(transportation_call_resolution_sequence)	The Shipper is responsible for accepting the transportation call resolution sequence proposal submitted by the Carrier.	Shipper	transportation_call_resolution_sequence
8	handleProposalAcceptance(transportation_call_resolution_sequence)	When the Shipper has accepted the transportation call resolution sequence proposal, the Carrier is in charge of achieving the procedures required to adequately fulfill it.	Carrier	transportation_call_resolution_sequence
9	createTransport(transport_proposal)	The Carrier is in charge of creating the transport so that it will be part of the Carrier and other involved agents future plannings.	Carrier	transport
10	bookRequiredRessources(operation_list)	Required resources to adequately fulfill the transport are booked at the infrastructure manager	Infrastructure_Manager	operation_list

**Fig. 3.** The *Select Most Adequate Transport* Task's Capabilities

	Capab. 1	Capab. 2	Capab. 3	Capab. 4	Capab. 5	Capab. 6	Capab. 7	Capab. 8	Capab. 9	Capab. 10
TMS1	best	A	N/A	A	best	A	A	A	N/A	best
TMS2	A	best	best	best	A	N/A	A	best	best	N/A
TMS3	N/A	N/A	A	N/A	A	best	best	N/A	N/A	A

Fig. 4. A MAS Runtime Report at Setuping Phase

described in section 3.2, execution sequence could be represented graphically in a AUML sequence diagram (due to a lack of space we do not include it here).

Figure 4 resumes a MAS runtime report of the Setuping phase. It documents the capabilities that are available (A) or non-available (N/A) in each of the integrated TMS components as well as if they are selected for capability execution at runtime (best). Since we are at Setuping phase, each of the capabilities is fulfilled under QoS and cost constraints. As long as there is no new component integrated into the system, the MAS is at equilibrium.

## 6 Related Work

Series of CBSD methods have been proposed in literature. Two of them have been selected because they focus on the use of multiple components in a single project, on functional aspects and use an iterative development life cycle.

PORE (Procurement-Oriented Requirements) [13] is a template-based approach for evaluating COTS-components in an iterative manner. The study of the existing COTS' functional aspects as well as the "traditional" requirements engineering constitute parallel activities with an iterative process for acquiring requirements and evaluating available components. The idea is to progressively reject non suitable products. This approach is also present here with components progressively adopted or rejected on the basis of their performance. PORE however considers component selection and removal onto a higher level basis. Such a selection is recommended in our process and documented in [4] but this remains a primary facultative step since all available components can be evaluated and selected on a low functional level basis at runtime. This has the advantage of always selecting the best possible offer both for integration and customization. The main drawback of avoiding an high level analysis is that every available component (even the poorest ones) should be interfaced with the MAS which leads to a waste of resources during the selection process.

CARE (COTS-Aware Requirements Engineering) [15] defines a goal-oriented process for describing and selecting COTS-components from a technical view. For any identified component, they capture its functional specification in the form of hard and softgoals and further define their specification. Then, they are stored and maintained in a knowledge base called *Digital Library System*. On the basis of the COTS descriptions, a research can determine which product(s) can be useful. We consider here functional specifications on the lowest level; we consequently do not point to a global library but rather a selection in the development process of potential components candidates competing at runtime

within the developed MAS. The idea of a lower level library could nevertheless be interesting but since we focus on applications with a higher degree of integration thus requiring more customization than in a purely "service-oriented" approach.

## 7 Conclusion

Iterative development is often used in an "ad hoc" manner without formal rules to guide and monitor the software process evolution. Indeed, such frameworks often lack of formality. Rules to evaluate proper advancement or guiding lines are seldom well defined and, when they are, they apply to analysis, design or test artifacts but not to runtime reports. That is why we have proposed in this paper to define overall rules in the context of iterative development to monitor the evolution of a MAS acting as a wrapper over COTS-components.

The development of the collaborative platform is in an advanced state. Prototypes are working and subject to detailed tests before they can fully support an industrial load. More measurements are being taken to further demonstrate the benefits of the proposed approach. Finally, the framework can be further refined. Indeed, it has been targeted for monitoring the evolution of a MAS dealing with COTS components. Other aspects can be included for monitoring a MAS evolution through the iterative template defined by I-Tropos.

## References

1. Castro, J., Kolp, M., Mylopoulos, J.: Towards requirements-driven information systems engineering: the tropos project. *Inf. Syst.* 27(6), 365–389 (2002)
2. Wautelet, Y.: A goal-driven project management framework for multi-agent software development: The case of i-tropos. PhD thesis, Université catholique de Louvain, Louvain-La-Neuve, Belgium (2008)
3. Kruchten, P.: The rational unified process: An introduction. Longman (Wokingham). Addison-Wesley, Reading (December 2003)
4. Vigder, M.R., Dean, J.C.: An architectural approach to building systems from cots software components. In: Johnson, J.H. (ed.) CASCON. IBM, p. 22 (1997)
5. Carney, D.J., Long, F.: What do you mean by cots? finally, a useful answer. *IEEE Software* 17(2) (2000)
6. Basili, V.R., Boehm, B.W.: Cots-based systems top 10 list. *IEEE Computer* 34(5), 91–93 (2001)
7. Brownsword, L., Oberndorf, T., Sledge, C.A.: Developing new processes for cots-based systems. *IEEE Software* 17(4), 48–55 (2000)
8. Ayala, C.: Systematic construction of goal-oriented cots taxonomies. PhD Thesis, Technical University of Catalunya (2008)
9. Torchiano, M., Morisio, M.: Overlooked aspects of cots-based development. *IEEE Software* 21(2), 88–93 (2004)
10. Pour, G.: Component-based software development approach: New opportunities and challenge. In: Proceedings Technology of Object-Oriented Languages, TOOLS 26, pp. 375–383 (1998)
11. Boehm, B.W.: A spiral model of software development and enhancement. *IEEE Computer* 21(5), 61–72 (1988)

12. Yu, E.: Modeling strategic relationships for process reengineering. PhD thesis, University of Toronto, Department of Computer Science, Canada (1995)
13. Maiden, N.A.M., Kim, H., Ncube, C.: Rethinking process guidance for selecting software components. In: Dean, J., Gravel, A. (eds.) ICCBSS 2002. LNCS, vol. 2255, pp. 151–164. Springer, Heidelberg (2002)
14. Wautelet, Y., Achbany, Y., Kiv, S., Kolp, M.: A service-oriented framework for component-based software development: An  $i^*$  driven approach. In: proceedings of the 11th International Conference on Enterprise Information Systems (ICEIS 2009). LNBIP, pp. 551–563 (2009)
15. Chung, L., Cooper, K.: Defining goals in a cots-aware requirements engineering approach. System Engineering 7(1), 61–83 (2002)

# An Agent for Ecological Deliberation

John Debenham<sup>1</sup> and Carles Sierra<sup>2</sup>

<sup>1</sup> University of Technology, Sydney, Australia  
debenham@it.uts.edu.au

<sup>2</sup> Institut d'Investigació en Intel·ligència Artificial - IIIA,  
Spanish Scientific Research Council, CSIC  
08193 Bellaterra, Catalonia, Spain  
sierra@iiia.csic.es

**Abstract.** An agent architecture supports the two forms of deliberation used by human agents. Cartesian, constructivist rationalism leads to game theory, decision theory and logical models. Ecological rationalism leads to deliberative actions that are derived from agents' prior interactions and are *not* designed; i.e., they are strictly *emergent*. This paper aims to address the scant attention paid by the agent community to the predominant form of deliberation used by mankind.

## 1 Introduction

This paper describes a form of agency that enables rational agents to move beyond Cartesian rationalism. The work is founded on the two forms of rationality described by the two Nobel Laureates Friedrich Hayek [1] and Vernon Smith [2] as being within 'two worlds'. Hayek and Smith identify; *constructivist rationality* that underpins rational predictive models of decision making; and, *ecological rationality* that refers to social institutions and practices that *emerge* from the history of an agent's interactions and are *not* pre-designed.

For intelligent agency we interpret Hayek and Smith's two rationalities as:

- Constructivist. An agent's actions are determined by a theory that may be independent of the particular environment in which the agent is situated, and typically requires access to data.
- Ecological. An agent's actions are the product of prior agents' actions only — deliberation that uses past experience and contextual triggers to build action sequences from experiential memory.

This paper is concerned with the issue generally known as *bounded rationality* that dates back to David Hume [3] and more recently to the early work of Herbert Simon. Bounded rationality refers to systems that are not founded on Cartesian rationalism; it has been widely addressed in economics [4], and is discussed in all good books on artificial intelligence, e.g. [5].

For over fifty years artificial intelligence research has spawned countless theories and systems that are *not* founded on Cartesian rationalism; one classic contribution being Rodney Brooks' work reported in his 'Computers and Thought' award-winning paper [6]. Despite these advances, work in multiagent systems has been heavily influenced by

game theory, decision theory and logic [7]; this is in contrast to an original motivation for investigating ‘distributed artificial intelligence’ in the mid 1970s where intelligence emerged from the interactions between systems.

Why would an agent be motivated to deliberate in a non-constructivist way? First, it may not be aware of a constructivist theory that addresses its goals. Second, it may have difficulty articulating its needs and its context completely and accurately in the theory. Third, the data required by the theory to determine its actions may not be readily available. Fourth, it may not have sufficient time for all this to happen. Fifth, it may favour ecological deliberation simply because it leads to a superior outcome. For example, when selecting a bottle of wine, some human agents refer to books of ratings and prices and make a constructivist choice, whereas others rely on their merchant to make a choice for them — this choice is purely ecological, its ‘rationality’ is in the trust that has been built through repeated interaction.

The main contribution of this paper is to describe a single agent that exhibits ecological deliberation, we show how it evolves as its experience grows. Various preliminaries are described in Section 2. Section 3 introduces the essential features of the agent architecture including the world model, and a ‘social model’ that is essential to ecological deliberation. Section 4 describes expectations of the effect of actions in the experiential memory— these expectations include measures of trust. Section 5 describes the ecological deliberative process, and Section 6 concludes.

## 2 Preliminaries

A multiagent system  $\{\alpha, \beta_1, \dots, \beta_o, \xi, \theta_1, \dots, \theta_t\}$ , contains an agent  $\alpha$  that interacts with negotiating agents,  $\beta_i$ , and information providing agents,  $\theta_j$ . We assume that each dialogical interaction takes place within a particular institution that is represented by an *institutional agent*,  $\xi$ , [8]. Institutions, or normative systems, play a central role in this work. We will describe an *ontology* that will permit us both to structure the dialogues and to structure the processing of the information gathered by agents. Our agent  $\alpha$  has two languages:  $\mathcal{C}$  is an illocutionary-based language for communication, and  $\mathcal{L}$  is a probabilistic first-order language for internal representation including the representation of its *world model*  $\mathcal{M}^t$ .  $\mathcal{C}$  is described in [9].

An agent’s *in-coming messages* and *observations* of the effect of its own actions are tagged with the identity of the sending agent and the time received, and are stored in a *repository*. A *world model* contains beliefs of the state of the other agents and the environment, and a *social model* contains beliefs of the state of the agent’s *relationships* with the other agents. The agent’s *experiential memory* contains complete historic information concerning prior actions and sequences of actions — this is detailed in Section 3.

Some messages trigger the agent’s *reactive logic* that overrides other activities. The agent aims to satisfy its *needs* using one of two forms of *deliberation*: *constructivist* (described in [10]) that is based on theories that call on *plans*, and *ecological* that uses past experience and contextual triggers to retrieve or build action sequences from experiential memory.

An *ecologically rational* agent  $\alpha$  with need  $\nu$  in context  $\Theta^t$  will act using the lottery  $E_\alpha(\mathcal{H}_\alpha^t, \nu, \Theta^t) \in \Delta(\mathcal{M} \times \mathcal{B})$  where:

$$E_\alpha : \mathcal{H} \times \mathcal{N} \times \mathcal{I} \rightarrow \Delta(\mathcal{M} \times \mathcal{B}) \quad (1)$$

where  $E_\alpha$  is a function that is *not* founded on an abstraction or theory that models, explains, constrains, describes or prescribes the behaviour of agents or the environment; it encapsulates  $\alpha$ 's particular ecological rationality. As above, if  $\alpha$  has an open dialogue then Equation 1 determines (non-deterministically) the next utterance that  $\alpha$  makes in that dialogue. Ecologically rational agents are non-deterministic. The action performed by a *deterministic ecologically rational agent* is determined by:  $E_\alpha : \mathcal{H} \times \mathcal{N} \times \mathcal{I} \rightarrow \mathcal{M} \times \mathcal{B}$ .

The “ecological rationality” in  $E_\alpha$  is based on the belief that the wisdom in  $\mathcal{H}_\alpha^t$  can somehow tell  $\alpha$  how to act rationally. This belief is not an abstraction or “theory” (as described above); it is simply a belief that the wisdom embedded in prior observations are a basis for rational action. In a simple form, this belief may be that prior agent behaviour reliably indicates future behaviour<sup>1</sup>. That is, ecological rationality may be founded on a *sense of trust* that  $\alpha$  has in agents, i.e. that they will continue to behave with no less integrity than that which they have displayed in the past. Ecological rationality may also be founded on the reputation that another agent has, or on trust in the institution (i.e. a normative multiagent system to which all the agents belong) to ensure that agents behave in some way [10]. In addition, Ecological rationality may be founded on subtle observations mined from  $\mathcal{H}_\alpha^t$ . As a simple example, “Whenever John is running a marketing promotion Carles invariably gives excellent service”. In general, ecologically rational behaviour will be derived from  $\mathcal{H}_\alpha^t$  using data mining techniques. Ecological rationality, like trust, is an experience-based phenomenon — it can not exist without experience. At the “top end”, ecological rationality embodies all the models that have been described from information-based agents including the important notion of integrity<sup>2</sup>.

### 3 Agent Architecture

$\alpha$  acts to satisfy a *need*. Needs either trigger  $\alpha$ 's constructivist, goal/plan deliberative reasoning described in [10], or ecological deliberation described in Section 5.

$\alpha$ 's *experiential memory* contains a history of what happened when any goal-directed sequence of actions was triggered or when any individual action was observed. First an individual *action experience*,  $a$ , consists of: (i) the *action*,  $a_{act}$ , i.e. the utterance, the sending and receiving agents, and the time at which the action was taken, (ii) the *trigger*, or precondition, that signalled when the action was to be performed,  $a_{trig}$ , and (iii) any observed *effect(s)*,  $a_{effect}$ , i.e. any identifiable responses that are an effect of  $a_{act}$  — see Section 4.

<sup>1</sup> Such a belief may incorporate context. For example, “John is most reliable *except* on Mondays”.

<sup>2</sup> The extent to which a partner agent will take advantage of his private information when enacting his commitments. E.g. “I haven't got the strawberries you ordered because they were all rain damaged.”

Then a *sequence experience*,  $s$ , consists of: (i) the *goal* of the sequence,  $s_{\text{goal}}$ , that may have been to satisfy a need, (ii) a sequence of *action experiences*,  $s_a = (a_i)_{i=1}^n$ , where each action experience  $a_i$  is described as above, (iii) beliefs of the prevailing *environment*,  $s_{\text{env}}$ , that includes: the institutional norms that apply at the time,  $s_{\text{norm}}$ , the agents involved in the interaction,  $s_{\text{agents}}$ , and the state of the *social model* (see Section 3.2) between the agents,  $s_{\text{social}}$ , i.e.  $s_{\text{env}} = \{s_{\text{norm}}, s_{\text{agents}}, s_{\text{social}}\}$ , (iv) a *rating* of the outcome of the action sequence,  $s_{\text{rate}}$ , that enables an ecologically rational agent to develop its repertoire of actions.

This rating is not simply in terms of the extent to which the sequence outcome met the original need, but in a sense that includes the possibility that the other agents involved may have adapted their actions to take account of changes in circumstance that occur during the sequence itself, or even that they went “over the odds” and gave more than was expected of them in some sense. These ratings are on a fuzzy scale from  $-5$  to  $+5$  where  $0$  means “is perfectly acceptable”,  $-5$  means “ghastly, completely unacceptable” and  $+5$  means “better than I could have dreamed of”. Ratings are not a ‘utility function’ in any sense — they are a subjective assessment of outcomes that is totally dependent on the prevailing state of the environment.

$\alpha$  uses the contents of its experiential memory to: reuse successful action sequences, build new sequences from individual actions, and improve prior sequences by using its knowledge of individual action experiences.

The integrity of beliefs derived from observations decreases in time.  $\alpha$  may have background knowledge concerning the expected integrity of a belief as  $t \rightarrow \infty$ . Such background knowledge is represented as a *decay limit distribution*. If the background knowledge is incomplete then one possibility is for  $\alpha$  to assume that the decay limit distribution has maximum entropy whilst being consistent with the data. Given an uncertain belief represented as the distribution,  $\mathbb{P}(X_i)$ , and a decay limit distribution  $\mathbb{D}(X_i)$ ,  $\mathbb{P}(X_i)$  decays by:

$$\mathbb{P}^{t+1}(X_i) = \Delta_i(\mathbb{D}(X_i), \mathbb{P}^t(X_i)) \quad (2)$$

where  $\Delta_i$  is the *decay function* for the  $X_i$  satisfying the property that  $\lim_{t \rightarrow \infty} \mathbb{P}^t(X_i) = \mathbb{D}(X_i)$ . For example,  $\Delta_i$  could be linear:  $\mathbb{P}^{t+1}(X_i) = (1 - \nu_i) \times \mathbb{D}(X_i) + \nu_i \times \mathbb{P}^t(X_i)$ , where  $\nu_i < 1$  is the decay rate for the  $i$ 'th distribution. Either the decay function or the decay limit distribution could also be a function of time:  $\Delta_i^t$  and  $\mathbb{D}^t(X_i)$ .

### 3.1 World Model

In the absence of in-coming messages the integrity of  $\mathcal{M}^t$  decays by Equation 2. The following procedure updates  $\mathcal{M}^t$  for all utterances expressed in  $C$ . Suppose that  $\alpha$  receives a message  $\mu$  from agent  $\beta$  at time  $t$ . Suppose that this message states that something is so with probability  $z$ , and suppose that  $\alpha$  attaches an epistemic belief  $\mathbb{R}^t(\alpha, \beta, \mu)$  to  $\mu$  — this probability reflects  $\alpha$ 's level of personal *caution*. Each of  $\alpha$ 's active plans,  $s$ , contains constructors for a set of distributions  $\{X_i\} \in \mathcal{M}^t$  together with associated *update functions*,  $J_s(\cdot)$ , such that  $J_s^{X_i}(\mu)$  is a set of linear constraints on the posterior distribution for  $X_i$ . Examples of these update functions are given in [12]. Denote the prior distribution  $\mathbb{P}^t(X_i)$  by  $p$ , and let  $p_{(\mu)}$  be the distribution with

minimum relative entropy<sup>3</sup> with respect to  $\mathbf{p}$ :  $\mathbf{p}_{(\mu)} = \arg \min_{\mathbf{p}} \sum_j r_j \log \frac{r_j}{p_j}$  that satisfies the constraints  $J_s^{X_i}(\mu)$ . Then let  $\mathbf{q}_{(\mu)}$  be the distribution:

$$\mathbf{q}_{(\mu)} = \mathbb{R}^t(\alpha, \beta, \mu) \times \mathbf{p}_{(\mu)} + (1 - \mathbb{R}^t(\alpha, \beta, \mu)) \times \mathbf{p} \quad (3)$$

and then let:

$$\mathbb{P}^t(X_{i(\mu)}) = \begin{cases} \mathbf{q}_{(\mu)} & \mathbf{q}_{(\mu)} \text{ is more interesting than } \mathbf{p} \\ \mathbf{p} & \text{otherwise} \end{cases} \quad (4)$$

A general measure of whether  $\mathbf{q}_{(\mu)}$  is ‘more interesting than’  $\mathbf{p}$  is:  $\mathbb{K}(\mathbf{q}_{(\mu)} \parallel \mathbb{D}(X_i)) > \mathbb{K}(\mathbf{p} \parallel \mathbb{D}(X_i))$ , where  $\mathbb{K}(\mathbf{x} \parallel \mathbf{y}) = \sum_j x_j \ln \frac{x_j}{y_j}$  is the Kullback-Leibler distance between two probability distributions  $\mathbf{x}$  and  $\mathbf{y}$ .

Finally merging Equations 4 and 2 we obtain the method for updating a distribution  $X_i$  on receipt of a message  $\mu$ :

$$\mathbb{P}^{t+1}(X_i) = \Delta_i(\mathbb{D}(X_i), \mathbb{P}^t(X_{i(\mu)})) \quad (5)$$

This procedure deals with integrity decay, and with two probabilities: first, any probability  $z$  in the message  $\mu$ , and second the belief  $\mathbb{R}^t(\alpha, \beta, \mu)$  that  $\alpha$  attached to  $\mu$ .

$\mathbb{R}^t(\alpha, \beta, \mu)$  is estimated by measuring the ‘difference’ between  $\mu$  and its subsequent verification. Suppose that  $\mu$  is received from agent  $\beta$  at time  $u$  and is verified by  $\xi$  as  $\mu'$  at some later time  $t$ . Denote the prior  $\mathbb{P}^u(X_i)$  by  $\mathbf{p}$ . Let  $\mathbf{p}_{(\mu)}$  be the posterior minimum relative entropy distribution subject to the constraints  $J_s^{X_i}(\mu)$ , and let  $\mathbf{p}_{(\mu')}$  be that distribution subject to  $J_s^{X_i}(\mu')$ . We now estimate what  $\mathbb{R}^u(\alpha, \beta, \mu)$  should have been in the light of knowing *now*, at time  $t$ , that  $\mu$  should have been  $\mu'$ .

The idea of Equation 3, is that  $\mathbb{R}^t(\alpha, \beta, \mu)$  should be such that, *on average* across  $\mathcal{M}^t$ ,  $\mathbf{q}_{(\mu)}$  will predict  $\mathbf{p}_{(\mu')}$  — no matter whether or not  $\mu$  was used to update the distribution for  $X_i$ , as determined by the condition in Equation 4 at time  $u$ . The *observed reliability* for  $\mu$  and distribution  $X_i$ ,  $\mathbb{R}_{X_i}^t(\alpha, \beta, \mu) | \mu'$ , on the basis of the verification of  $\mu$  with  $\mu'$ , is the value of  $k$  that minimises the Kullback-Leibler distance:

$$\mathbb{R}_{X_i}^t(\alpha, \beta, \mu) | \mu' = \arg \min_k \mathbb{K}(k \cdot \mathbf{p}_{(\mu)} + (1 - k) \cdot \mathbf{p} \parallel \mathbf{p}_{(\mu')})$$

### 3.2 Social Model

The *social model* contains beliefs of the state of  $\alpha$ ’s relationships with other agents — it consists of two components. First, an *intimacy model* that for each agent  $\beta$  consists of  $\alpha$ ’s model of  $\beta$ ’s private information, and,  $\alpha$ ’s model of the private information that  $\beta$  has about  $\alpha$ . Second, a *balance model* of the extent of reciprocity between pairs of agents.

Intimacy and balance were first reported in [9] to support argumentative negotiation where they were based on five illocutionary categories. Our requirements here are more general, and the models are quite different but we retain the same names. The spirit

<sup>3</sup> Entropy-based inference is a form of Bayesian inference that is convenient when the data is sparse [13] and encapsulates common-sense reasoning.

of them remains the same: *intimacy* — degree of closeness, and *balance* — degree of fairness. Intimacy is defined in terms of information gain, and balance in terms of ratings.

Private information is categorised first by the type of statement, using a set of illocutionary particles  $\mathcal{F}$ , and second by the contents of the statement, using the ontology  $\mathcal{O}$ . A categorising function  $\kappa : U \rightarrow \mathcal{P}(\mathcal{F})$ , where  $U$  is the set of utterances, allocates utterances to one or more category in the framework. The power set,  $\mathcal{P}(\mathcal{F})$ , is required as some utterances belong to multiple categories.

$I_{\alpha/\beta}^t$  is  $\alpha$ 's model of  $\beta$ 's private information; it is represented as real numeric values over  $\mathcal{F} \times \mathcal{O}$ . Suppose  $\alpha$  receives utterance  $u$  from  $\beta$  and that category  $f \in \kappa(u)$  then:  $I_{\alpha/\beta(f,c)}^t = I_{\alpha/\beta(f,c)}^{t-1} + \lambda \times \mathbb{I}(u) \times \text{Sim}(u, c)$  for any  $c \in \mathcal{O}$ , where  $\text{Sim}(\cdot)$  is a semantic similarity function [4],  $\lambda$  is the learning rate,  $I_{\alpha/\beta(f,c)}^t$  is the intimacy value in the  $(f, c)$  position in  $\mathcal{F} \times \mathcal{O}$ ,  $\mathbb{I}(u)$  is the Shannon information gain in  $\mathcal{M}^t$  due to receiving  $u$  using Equation 5, and  $\text{Sim}$  is as above. Additionally, the intimacy model decays in time in any case by  $I_{\alpha/\beta}^t = \delta \times I_{\alpha/\beta}^{t-1}$  where  $\delta < 1$  and very close to 1 is the decay rate.

$I_{\alpha \setminus \beta}^t$  is  $\alpha$ 's model of the private information that  $\beta$  has about  $\alpha$ . Assuming that confidential information is treated in confidence<sup>4</sup>  $\alpha$  will know what  $\beta$  knows about  $\alpha$ . This means that the same method can be used to model  $I_{\alpha \setminus \beta}^t$  as  $I_{\alpha/\beta}^t$  with the exception of estimating  $\mathbb{I}(u)$  as it is most unlikely that  $\alpha$  will know the precise state of  $\beta$ 's world model — for this we resort to the assumption that  $\beta$ 's world model mirrors  $\alpha$ 's and 'estimate' the information gain. Then the *intimacy model* is  $I_{\alpha \beta}^t = (I_{\alpha/\beta}^t, I_{\alpha \setminus \beta}^t)$ . In [9] balance was defined as the element by element numeric difference of  $I_{\alpha/\beta}^t$  and  $I_{\alpha \setminus \beta}^t$ . That definition is not suitable here.

$R_{\alpha/\beta}^t$  is a model of  $\alpha$ 's aggregated rating of  $\beta$ 's actions in assisting  $\alpha$  to achieve her goals and satisfy her needs.  $\alpha$  will have a variety of goals including the acquisition of goods, information, offering and receiving advice, gossip, and so on. These goals are categorised using a set of illocutionary particles  $\mathcal{G}$  and the ontology  $\mathcal{O}$ . Suppose  $\alpha$  triggers an action sequence  $s$  with goal  $g = (k, d)$  when the state of the environment is  $e$  and on completion of the sequence rates the outcome as  $\rho(\alpha, s, e)$  then:

$$R_{\alpha/\beta(k,c)}^t = R_{\alpha/\beta(k,c)}^{t-1} + \lambda \times \rho(\alpha, s, e) \times \text{Sim}(d, c)$$

for any  $c \in \mathcal{O}$ , where  $\rho(\alpha, s, e)$  is the fuzzy rating of the outcome of  $s$  as an integer in the range  $[-5, +5]$ ,  $\lambda$  is the learning rate,  $R_{\alpha/\beta(k,c)}^t$  is the aggregated rating in the  $(k, c)$  position in  $\mathcal{G} \times \mathcal{O}$ , and  $\text{Sim}$  is as above. Additionally, the model decays in time in any case by  $R_{\alpha/\beta}^t = \delta \times R_{\alpha/\beta}^{t-1}$  where  $\delta < 1$  and very close to 1 is the decay rate. This form of decay means that in the limit all values in the model decay to 0 meaning "is perfectly acceptable". This may appear to be odd, but the model is used only to gauge divergence from the norm; it is *not* used to select a trading partner — that is a job for the trust model.

$\alpha$  should have "a pretty good idea" of how  $\beta$  rates  $\alpha$ 's actions in assisting  $\beta$  to achieve her goals, and  $R_{\alpha \setminus \beta}^t$  models  $\alpha$ 's estimates of  $\beta$ 's rating of  $\alpha$ 's performance. Then the *balance model* is the pair  $R_{\alpha \beta}^t = (R_{\alpha/\beta}^t, R_{\alpha \setminus \beta}^t)$ . This structure is a historical summary

<sup>4</sup> See [10] for a discussion on measuring *confidentiality* i.e. 'information leakage'.

of how  $\alpha$  believes it has “done the right thing”, or otherwise, by other agents. It also exposes social debts, obligations and opportunities.

## 4 Expectations

An ecologically rational agent’s rationality lies only in its past experience. To behave rationally it will require some expectation, based on that experience, of what other agents will do. Experiential memory records each of the agent’s individual experiences; it does not address expectation. We now derive expectations from this historic data. Expectations are considered for the two classes of experience in experiential memory.

We consider expectations concerning the effect of triggering an action sequence. Suppose that  $\alpha$  triggers an action sequence,  $s$  with goal  $g$  where the state of the environment is  $e$  then we are interested in the rating of the outcome  $r$ . Given the rich meaning of the environment, as described in Section 3, it is reasonable to consider:

$$\mathbb{P}(\text{Observe}^{t'}(r) \mid \text{Enact}^t(s), e) \quad (6)$$

If  $\Omega \in e$  is the set of agents in  $e$ , then the aggregated rating of their responsive actions leading to the sequence outcome is a subjective measure of their collective *trust*, *honour* or *reliability* — a fuller account of these estimates is given in [12].

We first consider a special case of the expected rating of a diminutive action sequence consisting of a single agent,  $\Omega = \{\beta\}$ , and a single action — as is observed in the case of “commitment followed by subsequent enactment”. In this case if we estimate  $\mathbb{P}_\beta^t(v|u)$  where  $u$  is the commitment and  $v$  the enactment then:

$$T_\alpha(\beta, u, e) = \sum_v \rho(\alpha, v, e) \times \mathbb{P}_\beta^t(v|u)$$

Then  $\alpha$ ’s estimate of the *trust*, *honour* or *reliability* of  $\beta$  with respect to a class of utterances  $U$  will be:

$$T_\alpha(\beta, U, e) = \sum_{u \in U} T_\alpha(\beta, u, e) \times \mathbb{P}_\alpha^t(u)$$

where  $\mathbb{P}_\alpha^t(u)$  is as above.

For action sequences in general we abbreviate the expectation of Equation 6 to  $\mathbb{P}^t(r|s, e)$  that we may estimate directly using the same reasoning for estimating  $\mathbb{P}_\beta^t(v|u)$  because  $r$  is over a discrete space. Then  $T_\alpha(\Omega, s, e) = \mathbb{E}_\Omega^t(r|s, e)$  and  $T_\alpha(\Omega, S, e) = \sum_{s \in S} T_\alpha(\Omega, s, e) \times \mathbb{P}_\alpha^t(s)$ .

## 5 Ecological Deliberation

Human agents employ ecological deliberation for all but a very small proportion of the decisions that they make [15]. The neurological processes that enable human non-Cartesian deliberative processes are not well understood. It appears that given a need, contextual triggers somehow retrieve appropriate action sequences from experiential

memory. The retrieval process does not require a complete match and operates tentatively when the perceived environment is new, possibly by adapting the action sequence. This is reminiscent of the work of Roger Schank on dynamic memory.

When an agent is ‘born’ it will have no experiential memory, and its only rational basis for deliberative action will be either through pre-programmed constructivist deliberation or by imitating a ‘parent’ or ‘teacher’. Thereafter, whenever it acts it will observe the effects of its actions and its experiential memory will expand.

$\alpha$  has the following assets at its disposal to support ecological deliberation:

- an *experiential memory* — Section 3
- *expectations* — Section 4
- a *world model* — Section 3.1
- a *social model* — Section 3.2

Together experiential memory and expectations make a potent pair. Experiential memory contains details of action sequences, and expectations tell us what to expect if those sequences are reused. The world and social models describe the states of affairs that  $\alpha$  desires to change.

An agent acts to satisfy its needs. An ecological agent’s rationality lies in the trust that it has of others. This means that an ecological agent’s desires should address its social needs as well as its material needs — these may not be compatible. And this means that the actions that an ecological agent takes should attempt to shape its social model as well as its world model.<sup>5</sup> An agent’s social structures, and the structures of the institutions that it inhabits, are its means to transcend its individual ability.

Rather than give a tedious description of how each of the above operations may be performed we simply assume that they all have been, and that we are confronted with an enormous selection of previous, improved, adapted, simplified and created action sequences.

Our problem then is: given a current need, the current norm state, and the current states of the world and social models, to select one sequence. We deal with the complexity of matching the current goal and environment to those of previously observed sequences with a ‘super-Sim’ function that moderates the expected rating (Section 4) of each previously recorded sequence,  $s$ , to give expectations of the rating,  $r(s) \in [0, 1]$ , of how that sequence would perform if it was reused now for the current need.

Given that we now face the problem of devising a method that selects an action sequence it is worth considering first what we expect of that method. What it should *not* do is to say “That one is the best choice” that is pure constructivism — it says “Carles and John have greater knowledge than can ever emerge from the environment”. Worse still it would mean that by determining the agent’s actions it would then pervert the agent’s experiential memory for ever more.<sup>6</sup>

What is needed is an evolutionary method of some sort — that may well be how humans operate. A problem with evolutionary methods is that we may not be prepared to accept poor performance while the method evolves, although permitting a method to explore and make mistakes may also enable it to discover.

---

<sup>5</sup> In future work we propose to address how it should also attempt to shape the norms of the institutions that it inhabits.

<sup>6</sup> Unfortunately this complication also applies to the definition of ‘super-Sim’.

The point of this digression is to excuse ourselves for presenting only a ‘quasi-ecological’ method that permits the agent to explore whilst guiding it in an apparently sensible direction. A strategy is reported in [16] on how to place all of one’s wealth as win-bets indefinitely on successive horse races so as to maximise the rate of growth; this is achieved by proportional gambling, i.e. by betting a proportion of one’s wealth on each horse equal to the probability that that horse will win. This result is interesting as the strategy is independent of the betting odds. The situation that we have is not equivalent to the horse race, but it is tempting to consider the strategy that selects sequence  $s_i$  with probability  $q_i$ :

$$q_i = \frac{r(s_i)^c}{\sum_k r(s_k)^c} \quad (7)$$

where  $c > 0$  is a real constant that moderates the degree of exploration. This strategy will favour those sequences whose expected performance and moderated by ‘super-Sim’ is greater.

Finally we consider how an agent combines constructivist and ecological deliberation.

Ecological deliberation is by no means the poor relation of its Cartesian brother. Referring back to the ‘wine merchant’ example in Section 1, it may simply be that the recommendations of the wine merchant are better in all respects than those that the agent could derive from the data available. If this is so then a rational agent should surely prefer ecological deliberation.

A committed constructivist might respond by saying that clearly the agent should learn as much about wine as the merchant and then everything becomes Cartesian again. Creating a “Mr Know-It-All” agent is dangerous if it means that the agent believes his knowledge will remain superior in a competitive world to that of other agents, he may then live and decay in a state of sublime ignorance.

A rational agent builds an experiential memory and maintains an open mind on whether to choose constructivist or ecological deliberation. It reinforces the choices it makes by forming a view on which performs better by using its subjective ability to evaluate outcomes.

## 6 Discussion

The full realisation of the Hayekian vision of self-evolving agents situated in a world of self-evolving institutions is an extensive research agenda that is the subject of on-going research. For example, there is no clear means of achieving an orderly self-evolution of normative systems in a multi-system context. The contribution of this paper is to describe how a single agent can engage in ecological deliberation in addition to well-understood constructivist deliberation. This enables agents to evolve and adapt their deliberative processes as their environment and their fellow agents evolve.

The social model contains beliefs of the strength of agents’ relationships, enables agents to form desires of how those relationships could be, and to form intentions of how to make them so. A possible next step is to experiment with a norm model in a similar fashion. If this can be achieved through ecological deliberation then we will be close to understanding self-evolving electronic institutions that will take multiagent systems technology to a new level.

## References

1. Hayek, F.A.: *The Fatal Conceit: The Errors of Socialism*. University Of Chicago Press (1991)
2. Smith, V.L.: *Rationality in Economics: Constructivist and Ecological Forms*. Cambridge University Press, Cambridge (2007)
3. Hume, D.: *An Enquiry concerning Human Understanding*, 3rd edn, 1975 edn. Clarendon Press, Oxford (1977)
4. Rubinstein, A.: *Modeling Bounded Rationality*. MIT Press, Cambridge (1998)
5. Russell, S., Norvig, P.: *Artificial Intelligence: A Modern Approach*, 2nd edn. Prentice-Hall, Englewood Cliffs (2002)
6. Brooks, R.A.: Intelligence without reason. In: Myopoulos, R., Reiter, J. (eds.) *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, Sydney, Australia, pp. 569–595. Morgan Kaufmann, San Francisco (1991)
7. Russell, S.: Rationality and intelligence. *Artificial Intelligence* 94, 57–77 (1997)
8. Arcos, J.L., Esteva, M., Noriega, P., Rodríguez, J.A., Sierra, C.: Environment engineering for multiagent systems. *Journal on Engineering Applications of Artificial Intelligence* 18 (2005)
9. Sierra, C., Debenham, J.: The LOGIC Negotiation Model. In: *Proceedings Sixth International Conference on Autonomous Agents and Multi Agent Systems AAMAS-2007*, Honolulu, Hawai'i, pp. 1026–1033 (2007)
10. Sierra, C., Debenham, J.: Information-based deliberation. In: Padgham, L., Parkes, D., Müller, J., Parsons, S. (eds.) *Proceedings Seventh International Conference on Autonomous Agents and Multi Agent Systems AAMAS-2008*, Estoril, Portugal. ACM Press, New York (2008)
11. Sierra, C., Debenham, J.: Information-based reputation. In: Paolucci, M. (ed.) *First International Conference on Reputation: Theory and Technology (ICORE 2009)*, Gargonzia, Italy, pp. 5–19 (2009)
12. Sierra, C., Debenham, J.: Trust and honour in information-based agency. In: Stone, P., Weiss, G. (eds.) *Proceedings Fifth International Conference on Autonomous Agents and Multi Agent Systems AAMAS-2006*. pp. 1225–1232. ACM Press, New York (2006)
13. Cheeseman, P., Stutz, J.: On The Relationship between Bayesian and Maximum Entropy Inference. In: *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, pp. 445–461. American Institute of Physics, Melville (2004)
14. Li, Y., Bandar, Z.A., McLean, D.: An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering* 15, 871–882 (2003)
15. Levine, D.S.: Brain mechanisms for making, breaking, and changing rules. In: Huang II, D.S., Levine, D.C.W., Levine, D.S., Jo, K.H. (eds.) *Advanced Intelligent Computing Theories and Applications. With Aspects of Contemporary Intelligent Computing Techniques*, pp. 345–355. Springer, Heidelberg (2008)
16. Kelly, J.J.: A new interpretation of information rate. *IEEE Transactions on Information Theory* 2, 185–189 (1956)

# A Framework to Compute Inference Rules Valid in Agents' Temporal Logics

Sergey Babenyshev and Vladimir Rybakov

Department of Computing and Mathematics, Manchester Metropolitan University,  
Manchester, U.K.

Siberian Federal University, Krasnoyarsk, Russia  
[sergei.babyonyshев@gmail.com](mailto:sergei.babyonyshев@gmail.com), [V.Rybakov@mmu.ac.uk](mailto:V.Rybakov@mmu.ac.uk)

**Abstract.** Our paper<sup>1</sup> suggests a computational framework for verification valid inference in agents' temporal logics. As a tool, describing human reasoning procedure, we suggest valid inference rules (valid semantically - in Kripke-like frames generating logic). We investigate valid inference rules in agents' temporal logics with linear and branching intransitive time. Main results of our paper are suggested algorithms which allow to compute valid inference rules in agents' liner time logics  $\mathcal{LTL}_K$  and  $\mathcal{LTL}_K(Z)$ , agents' logic with branching intransitive time  $\mathcal{L}_{TA_i}$ , and the logic with branching transitive time  $\mathcal{L}_{TA_t}$ .

## 1 Introduction

Prime objective of AI is: to describe the property of humans, intelligence - the sapience of Homo sapiens - so precisely that it can be simulated by a machine. Evidently, this been standing a set of issues about the nature of the mind and limits of scientific hubris, and one of most intriguing questions was what is human reasoning procedure. In turn, this generated study and research of mathematical symbolic logic describing reasoning and true statements. Today, it has become an essential part of AI, providing the heavy lifting for many difficult problems in computer science. Usage of symbolic mathematical logic, and especially various systems of non-classical logic in AI is very popular (cf. van Benthem [12], - temporal and modal logic, Gabbay (et.al) [345] - modal, multi-modal, temporal logics, inference, Hodkinson (at. al) [6789] - temporal, temporal linear and real time logic, first-order branching time logic, time and automata). A corner part of representation human reasoning is inference: the process of drawing a conclusion by applying rules to observations or hypotheses. Human inference (i.e. how humans draw conclusions) is traditionally studied within the field of cognitive psychology. Mathematical logic studies the laws of valid inference, and we would like work out a logical framework for computation valid inference in multi-agents' logics, which are combinations (or hybrids) of multi-modal and temporal logics.

---

<sup>1</sup> This research is supported by Engineering and Physical Sciences Research Council (EPSRC), U.K., grant EP/F014406/1.

Various applications of multi-modal logic to AI and CS is a popular area. In particular, they are based on usage of multi-agent logic, as a part (or implementation) of epistemic logic. Among epistemic logics to model knowledge a range from  $S4$  over the intermediate systems  $S4.2 - S4.4$  to  $S5$  has been investigated (Hintikka –  $S4$  (1962), Kutschera argued for  $S4.4$  (1976), Lenzen suggested  $S4.2$  (1978), van der Hoek has proposed to strengthen knowledge according to system  $S4.3$  (1996), van Ditmarsch, van der Hoek and Kooi together with Fagin, Halpern, Moses and Vardi (Fagin et al. 1995 [10], cf. also [11]) and others assume knowledge to be  $S5$  valid). The developed approach to model multi-agent environment in AI often combines, not only modal operations for agents' knowledge and Boolean logical operations, but also other ones - e.g. operations for time - temporal operations, dynamic logic operations (cf. R.Schmidt et.al [12]). Through multi-agent approach we may view the logic of discovery, which has a solid prehistory, maybe, starting from the monograph Logic of Discovery and Logic of Discourse by Jaakko Hintikka and Fernand Vandamme [13]. The mentioned above logical systems are, as a rule, some hybrids of individual ones. They are usually introduced by combination (fusion) of several non-classical logics. Among these logics, the linear temporal logic LTL was very popular.

Temporal logics and, in particular, LTL, is currently the most widely used specification formalism for reactive systems. They were first suggested to be used for specifying properties of programs in late 1970's (cf. Pnueli [14]). The most used temporal framework is the linear-time propositional temporal logic LTL, which has been studied from various viewpoints (cf. e.g. Manna and Pnueli [15][16], Clark E. et al. [17]). Temporal logics have numerous applications to safety, liveness and fairness, to various problems arising in computing.

Model checking for LTL formed a noticeable trend in applications of logic in Computer Science, which uses, in particular, applications of automata theory (cf. Vardi [18][19]). Temporal logics themselves can be considered as special cases of hybrid logics, e.g. as a bimodal logic with some additional laws imposed on interaction of modalities to emulate the flow of time. The mathematical theory dedicated to study of various aspects of interaction of temporal operations (e.g. axiomatizations of temporal logics) and to construction of effective semantic theory based on Kripke/Hintikka-like models and temporal Boolean algebras, formed a highly technical branch in non-classical logics (cf. e.g. van Benthem [20][21], Goldblatt [22], Gabbay and Hodkinson [23], Hodkinson [24]). Admissible inference rules is some temporal logics where investigated in Rybakov [25][26][27].

Often another components of hybrids' logics are multi-agents' knowledge logics, which form active area in AI and Knowledge Representation (cf. Bordini et al. [28], Dix et al. [29], Hoek et al. [30], Fisher [31], Hendler [32], Kacprzak [33], Wooldridge [34]). In a sense, multi-agent logics came from a particular application of multi-modal logics to reasoning about knowledge (cf. Fagin et al. [10][35], Halpern and Shore [11]), where modal-like operations  $K_i$  (responsible for knowledge of individual agents) were embedded in the language of propositional logic. These operations are intended to model effects and properties of agents' knowledge in changing environment. When we make combination a background logic

(say for time), a reasonable question which logic should be used as the basic logic. For instance, if the logic used as the basis is too expressive — the undecidability phenomenon can occur (cf. Kacprzak [33] with reduction of the decidability to the domino problem). If the basic logic is just the boolean logic, and the agents are autonomous, decidability for the standard systems usually can be obtained by techniques from modal logic, e.g. filtration, (cf. [10,35]).

In this paper we intend to suggest algorithms to compute valid inference rules in agents' temporal logics, therefore we have to care about correct choice of background temporal logics. We will consider some extensions of the linear temporal logic *LTL* (which is simplest case) and branching temporal logic with potentially intransitive time accessibility relation (though the technique work for transitive time as well). Here we adopt and extend our approach applied previously in Rybakov et al. [36,37,38]. The main result of our paper is suggested algorithms, which reduce computation of validity for inference rules to model checking on (effectively finite) models. In the final part of the paper we briefly discuss extension of our research to verification of admissible rules. Some necessary conditions for admissibility immediately follow from obtained results.

## 2 Preliminaries, Case of Linear Agent's Temporal Logics

Here we firstly recall necessary definitions and notation to follow the paper. Though some preliminary acquaintance with non-classical propositional logics (symbolic/mathematical approach) and their algebraic and Kripke-Hintikka semantics is advisable. Our main aim is to study a framework which would allow correctly compute valid inference rules, valid logical inference. To proceed, consider any propositional logic (modal, temporal, intuitionistic etc.)  $\Lambda$ , whose set of all formulas is denoted by  $Fm_\Lambda$ .

A rule is an expression  $\alpha_1, \dots, \alpha_n / \beta$ , where  $\alpha_i, \beta \in Fm_\Lambda$ . An informal meaning of this inference rule is: (i)  $\beta$  follows from premisses (assumptions, hypothesis)  $\alpha_1, \dots, \alpha_n$ , or (ii) this rule infers  $\beta$  from  $\alpha_1, \dots, \alpha_n$ . As an informal example, consider: "When it rains, the grass gets wet. It rains. Therefore, the grass is wet." For the formalization the notions of intelligence and argumentation in all forms, logical inference is of fundamental importance. Which inference rules may be taken as correct formalism is non-trivial question. In this paper we will study so called valid inference rules.

Given a logic  $\Lambda$  defined by a class  $K$  of Kripke-like models, based on a class  $F$  of Kripke-like frames. A rule  $r = \alpha_1 \dots \alpha_m / \beta$  is *valid in a model*  $\mathcal{M} = \langle \mathcal{F}, Val \rangle \in M$ , where  $\mathcal{F} \in F$ , (symbolically  $\mathcal{M} \Vdash r$ ), if

$$[\forall_i (\mathcal{M} \Vdash_{Val} \alpha_i)] \implies \mathcal{M} \Vdash_{Val} \beta.$$

So, the meaning of the rule is  $r$  is valid in  $\mathcal{M}$  if as soon as all premisses of  $r$  are valid the conclusion is valid as well.

A rule  $r$  is *valid in a frame*  $\mathcal{F} \in F$ , if, for every valuation  $Val$  of variables from  $r$ ,  $\mathcal{F} \Vdash_{Val} r$ . Finally,  $r$  is valid in logic  $\Lambda$  generated by a class of frames  $F$ , if for any frame  $\mathcal{F} \in F$ ,  $r$  is valid in  $\mathcal{F}$ . This definition looks quite correct from general

viewpoint: the rule is valid if it is valid everywhere - in all models generating  $\Lambda$ . Basic question which we are dealing with is how to compute valid inference (how verify if a given rule is valid). Below we start from illustration on how to compute valid inference in agents' logics based on linear time background.

The logic  $\mathcal{LTL}_K$  introduced in [39] is the agents' logic based on linear time. We recall basis definition and facts concerning this logic. The logic  $\mathcal{LTL}_K$  is generated by frames  $\mathcal{N}_C := \langle \bigcup_{i \in N} C(i), R, R_1, \dots, R_m, \text{Next} \rangle$  ( $m$  is the number of accessibility relations), which are tuples, where  $N$  is the set of natural numbers,  $C(i), i \in N$  are some pairwise disjoint nonempty sets,  $R, R_1, \dots, R_m$  are binary relations, emulating agents' accessibility. For all elements  $a$  and  $b$  from  $\bigcup_{i \in N} C(i)$ ,  $aRb \iff [a \in C(i) \text{ and } b \in C(j) \text{ and } i < j] \text{ or } [a, b \in C(i) \text{ for some } i]$ ; any  $R_j$  is a reflexive, transitive and symmetric relation (i.e. an equivalence relation), and for all  $j$ ,  $\forall a, b \in \bigcup_{i \in N} C(i)$ ,  $aR_jb \implies [a, b \in C(i) \text{ for some } i]$ ;  $a \text{ Next } b \iff \exists i[a \in C(i) \& b \in C(i+1)]$ . These frames are intended to model the reasoning (or computation) in discrete time, so each  $i \in N$  (any natural number  $i$ ) is the time index for a cluster of states arising at a step in current computation. Any  $C(i)$  is a set of all possible states in the time point  $i$ , and the relation  $R$  models discrete current of time. Relations  $R_j$  are intended to model knowledge-accessibility relations of agents in any cluster of states  $C(i)$  at the time point  $i$ . Thus any  $R_j$  is supposed to be  $S5$ -like relation, i.e. an equivalence relation. The  $\text{Next}$  relation is the standard one – it describes all states available in the next time point cluster.

The language of  $\mathcal{LTL}_K$  uses the language of Boolean logic, operations  $\mathbf{N}$  (next),  $\mathbf{U}$  (until),  $\mathbf{U}_w$  (weak until),  $\mathbf{U}_s$  (strong until), agents' knowledge (unary) operations  $K_j$ ,  $1 \leq j \leq m$  and *knowledge via interaction operation*  $\mathbf{IntK}$ . Formation rules for formulas are as usual. For any collection of propositional letters  $Prop$  and any frame  $\mathcal{N}_C$ , a valuation  $V$  in  $\mathcal{N}_C$  is a mapping which assigns truth values to elements of  $Prop$  in  $\mathcal{N}_C$ . So, for any  $p \in Prop$ ,  $V(p) \subseteq \mathcal{N}_C$ . We will call  $\langle \mathcal{N}_C, V \rangle$  a model (a Kripke/Hintikka model). For any model  $\mathcal{M}$ , the truth values are extended from propositions (in  $Prop$ ) to arbitrary formulas (built over  $Prop$ ) as follows (for  $a \in \mathcal{N}_C$ , notation  $(\mathcal{N}_C, a) \Vdash_V \varphi$  says that the formula  $\varphi$  is true at  $a$  in  $\mathcal{N}_C$  w.r.t.  $V$ ). The rules are:

$$\forall p \in Prop, \quad (\mathcal{M}, a) \Vdash_V p \Leftrightarrow a \in V(p);$$

$$(\mathcal{M}, a) \Vdash_V \varphi \wedge \psi \Leftrightarrow (\mathcal{M}, a) \Vdash_V \varphi \& (\mathcal{M}, a) \Vdash_V \psi;$$

$$(\mathcal{M}, a) \Vdash_V \neg \varphi \Leftrightarrow \text{not}[(\mathcal{M}, a) \Vdash_V \varphi];$$

$$(\mathcal{M}, a) \Vdash_V \mathbf{N} \varphi \Leftrightarrow \forall b[(a \text{ Next } b) \Rightarrow (\mathcal{M}, b) \Vdash_V \varphi];$$

$$(\mathcal{M}, a) \Vdash_V \varphi \mathbf{U} \psi \Leftrightarrow \exists b[(aRb) \&$$

$$((\mathcal{M}, b) \Vdash_V \psi) \& \forall c[(aRc \& cRb) \& \neg(bRc) \Rightarrow (\mathcal{M}, c) \Vdash_V \varphi]].$$

Similar rules (to the above case  $\mathbf{U}$ ) work for  $\mathbf{U}_w$  and  $\mathbf{U}_s$ . Further,

$$(\mathcal{M}, a) \Vdash_V K_j \varphi \Leftrightarrow \forall b[(a R_j b) \Rightarrow (\mathcal{M}, b) \Vdash_V \varphi], \text{ and } (\mathcal{M}, a) \Vdash_V \mathbf{IntK} \varphi \Leftrightarrow$$

$$\exists a_{i_1} \dots \exists a_{i_s} \in \mathcal{M}[a R_{i_1} a_{i_1} R_{i_2} a_{i_2} \dots R_{i_s} a_{i_s} \ \& \ (\mathcal{M}, a_{i_s}) \Vdash_V \varphi].$$

The logic  $\mathcal{LTL}_K$  is the set of all formulas valid in all models. The paper [39] proves that this logic is decidable: there is an algorithm which for any formula  $\alpha$  computes if  $\alpha \in \mathcal{LTL}_K$ . If we are interested to compute valid for logic  $\mathcal{LTL}_K$  inference rules, it is very easy to observe that, for any rule  $r = \alpha_1 \dots \alpha_m / \beta$ ,  $r$  is valid in  $\mathcal{LTL}_K$  iff

$$[\bigwedge_{1 \leq i \leq m} (\square \alpha_i) \rightarrow \beta] \in \mathcal{LTL}_K.$$

Therefore, from decidability of  $\mathcal{LTL}_K$  we immediately obtain

**Proposition 1.** *There is an algorithm computing inference rules valid in  $\mathcal{LTL}_K$ .*

Similar statement holds for linear temporal logic with past. To be precise we give definition of this logic and the detailed statement. First we introduce (cf. [39]) the logic  $\mathcal{LTL}_K(Z)$ . The logic  $\mathcal{LTL}_K$  is based on a flow of time modeled by natural numbers, which matches well with human intuition. If we intend to model past, first candidate for time indexes is the set  $Z$  of all integer numbers with standard ordering. Semantic definition of the logic is as follows. The frame  $\mathcal{Z}_C := \langle \bigcup_{i \in Z} C(i), R, R_1, \dots, R_m, Next, Prev \rangle$  is a tuple, where  $Z$  is the set of all integer numbers,  $C(i)$  are some nonempty (pairwise disjoint) sets,  $R$  is a binary linear relation for time,  $R_1, \dots, R_m$  are binary accessibility relations imitating possible agents' transitions.  $\forall a, b \in \bigcup_{i \in Z} C(i) (a R b) \Leftrightarrow [a \in C(i) \& b \in C(j) \& i \leq j]$ . As before,  $R_j$  are reflexive, transitive and symmetric relations, and  $\forall a, b \in \bigcup_{i \in Z} C(i), a R_j b \Rightarrow \exists i \in Z[a, b \in C(i)]$ . Any  $R_j$  is an equivalence relation, at clusters  $C(i)$ . Further,  $a Next b \Leftrightarrow [\exists i((a \in C(i)) \& (b \in C(i+1)))]$ ;  $a Prev b \Leftrightarrow [\exists i((a \in C(i)) \& (b \in C(i-1)))]$ . The language of  $\mathcal{LTL}_K(Z)$  extends the language of  $\mathcal{LTL}_K$  by four more logical operations:  $\mathbf{S}$  (since),  $\mathbf{S}_w$  (weak since),  $\mathbf{S}_s$  (strong since),  $\mathbf{N}^{-1}$  (previous). For a frame  $\mathcal{Z}_C$  with a valuation  $V$ , the rules of computation for truth values of formulas in the model  $\mathcal{M} := \langle \mathcal{Z}_C, V \rangle$  are similar to given above ones for the case of logic  $\mathcal{LTL}_K$ , only we compute operations  $\mathbf{S}$ ,  $\mathbf{S}_w$ ,  $\mathbf{S}_s$  and  $\mathbf{N}^{-1}$  towards past.

Again, in the paper [39] it was proved that  $\mathcal{LTL}_K(Z)$  is decidable, and an algorithm for computation of true and satisfiable formulas was found. And if we want to compute valid in  $\mathcal{LTL}_K(Z)$  inferences, we may use the following simple observation: for any rule  $r = \alpha_1 \dots \alpha_m / \beta$ ,  $r$  is valid in  $\mathcal{LTL}_K(Z)$  iff

$$[\bigwedge_{1 \leq i \leq m} (\square^+ \alpha_i \wedge \square^- \alpha_i) \rightarrow \beta] \in \mathcal{LTL}_K(Z),$$

where  $\square^+$  is operation *necessary in future*, and  $\square^-$ , respectively, is operation *was necessary in past*. Hence, using decidability of  $\mathcal{LTL}_K(Z)$  we derive.

**Proposition 2.** *The logic  $\mathcal{LTL}_K(Z)$  is decidable w.r.t. valid inference rules. There is an algorithm computing inference rules valid in  $\mathcal{LTL}_K(Z)$ .*

If we will take to consideration agents' logics with non-linear, or not transitive time then the question is already not so evident. We describe a suitable technique to handle this case in next section.

### 3 Logics with Non-linear and Intransitive Time

To consider agents' logics with non-linear and intransitive time, we start from a semantical definition of such logics. The models for these logics are based on the Kripke/Hintikka frames

$$C_Y := \langle \bigcup_{i \in Y} C(i), R, R_1, \dots, R_m \rangle,$$

where  $Y = \langle Y, R_t \rangle$  can be any set with a binary relation  $R_t$ , each  $i \in Y$  is the time index for a set  $C(i)$  of possible states (so, any  $C(i)$  is just a non-empty set of worlds – states).  $Y := \langle Y, R_t \rangle$  is the time frame (viewed as steps in a computation or stages of evolving a system). Any  $i \in Y$  is a time point, and  $C(i)$  is a set of states/elements (*worlds* or *states of affairs* in terms of Kripke/Hintikka semantics) in the current time point  $i$ .

The branching-time flow is modeled by the binary accessibility relation  $R$  in  $C_Y := \langle \bigcup_{i \in Y} C(i), R, R_1, \dots, R_m \rangle$  where for all elements  $a$  and  $b$  from  $\bigcup_{i \in Y} C(i)$ ,  $aRb \iff \exists i, j \in Y : iR_t j \& a \in C(i) \& b \in C(j)$ . Less formally,  $R$  imitates the flow of time connecting the states, so,  $aRb$  means that  $a$  the state  $b$  is situated in a time point, which is accessible from the time point where  $a$  is situated.

We do not assume now that the accessibility relation  $R$  from frames  $C_Y$  is compulsory transitive, because it would not correspond to all possible interpretations. For instance, if  $R$  is imitated by possible links within WEB, it could be not transitive (if we see a web page  $wp_a$  and yet some else web page  $wp_b$  is visible from the intermediate  $wp_a$ , this does not mean that we can see  $wp_b$  from the starting web page). Relations  $R_1, \dots, R_m$  are binary relations on  $\bigcup_{i \in Y} C(i)$  imitating agents' accessibility within time clusters  $C(i)$ . So, any  $R_j$  is a reflexive, transitive and symmetric relation, and  $\forall a, b \in \bigcup_{i \in Y} C(i), aR_j b \Rightarrow [a, b \in C(i) \text{ for some } i]$ .

The language of the corresponding logic includes propositional letters, Boolean logical operations, operations  $K_i$ ,  $1 \leq i \leq m$  for agents' knowledge, and temporal operations:  $\diamond^+$  (will be in future),  $\diamond^-$  (was in past). (Since relations  $R$  is not obligatory linear, no very natural way to input operations *until* and *next*, as well as similar ones for past)). Formation rules for formulas are as usual.

A model  $M$  on a frame  $C_Y$  is this frame together with a valuation  $V$  of propositional letters, i.e.  $V$  maps letters in subsets of the base set of  $C_Y$ . As in previous section this valuation may be extended from propositional letters to arbitrary formulas. The rules are as follows,  $\forall a \in \bigcup_{i \in Y} C(i), \forall p \in Prop, (\mathcal{M}, a) \Vdash_V p \Leftrightarrow a \in V(p)$ , then the similar, as earlier, steps for Boolean operations and agents' knowledge operations, and for temporal operations the rules are

$$\begin{aligned} \forall a \in \bigcup_{i \in Y} C(i), \forall p \in Prop(C_Y, a) \Vdash_V \diamond^+ \varphi \iff \\ \exists b \in C_Y[(aRb) \text{ and } (C_Y, b) \Vdash_V \varphi]; \\ \forall a \in \bigcup_{i \in Y} C(i), \forall p \in Prop(C_Y, a) \Vdash_V \diamond^- \varphi \iff \\ \exists b \in C_Y[(bRa) \text{ and } (C_Y, b) \Vdash_V \varphi]. \end{aligned}$$

**Definition 1.** The logic  $\mathcal{L}_{TA_i}$  is the set of all formulas valid in all worlds of all models based on all possible frames  $C_Y$ .

If we wish to describe inference rules valid in  $\mathcal{L}_{TA_i}$ , approach from previous section does not work because we cannot say that a rule  $r = \alpha_1 \dots \alpha_m / \beta$ ,  $r$  is valid in  $\mathcal{L}_{TA_i}$  iff  $[\bigwedge_{1 \leq i \leq m} (\square^+ \alpha_i \wedge \square^- \alpha_i) \rightarrow \beta] \in \mathcal{L}_{TA_i}$  (because both possible intransitivity of accessibility relations and full temporal language for future and past - possible time zigzags). So, even decidability of  $\mathcal{L}_{TA_i}$  itself cannot be used directly.

But using some translation inference rules to reduced forms and some technique based on advanced filtration may help. First we define reduced forms for inference rules. A rule  $\mathbf{r}$  is said to have the *reduced normal form* if  $\mathbf{r} = \varepsilon/x_1$  where

$$\begin{aligned} \varepsilon := \bigvee_{1 \leq j \leq n_1} \left( \bigwedge_{1 \leq i \leq n} [x_i^{t(j,i,0)} \wedge (\diamond^+ x_i)^{t(j,i,1)} \wedge (\diamond^- x_i)^{t(j,i,2)} \wedge \right. \\ \left. \bigwedge_{1 \leq s \leq m} (\neg \mathbf{K}_s \neg x_i)^{t(j,i,s,3)} \right), \end{aligned}$$

and all  $x_k$  are certain letters (variables),  $t(j, i, z), t(j, i, p, z) \in \{0, 1\}$  and, for any formula  $\alpha$  above, we define  $\alpha^0 := \alpha$ ,  $\alpha^1 := \neg \alpha$ .

Similar to Lemma 3.1.3 and Theorem 3.1.11 from [40], it follows.

**Theorem 1.** There exists an algorithm running in (single) exponential time, which, for any given rule  $\mathbf{r}$ , constructs its equivalent normal reduced form  $\mathbf{r}_{\text{nf}}$ .

Most important for us is

**Theorem 2.** If a rule in reduced form  $\mathbf{r}_{\text{nf}}$  is invalid in a model based at some  $C_Y$  then  $\mathbf{r}_{\text{nf}}$  is invalid in some such finite model, size of which is computable from  $\mathbf{r}_{\text{nf}}$ .

And combining these two theorems we obtain

**Theorem 3.** The logic  $\mathcal{L}_{TA_i}$  is decidable w.r.t. valid inference rules. There is an algorithm which by any rule  $\mathbf{r}$  checks if  $\mathbf{r}$  is valid in  $\mathcal{L}_{TA_i}$ .

It is relevant to say that this theorem just reduce validity problem for inference rules to model checking which itself is a difficult computational task.

Approach applied in this section is rather flexible, in particular, we may to impose some restrictions on possible time admissibility relations, and yet the approach can work. For instance if possible time accessibility relations  $R$  at frames  $C_Y$  may be only transitive, we obtain the logic  $\mathcal{L}_{TA_r}$ . Using exactly the same approach and scheme of proofs we obtain

**Theorem 4.** *The logic  $\mathcal{L}_{TA_r}$  is decidable w.r.t. valid inference rules. There is an algorithm which by any rule  $r$  checks if  $r$  is valid in  $\mathcal{L}_{TA_r}$ .*

Valid rules are a subclass of more general class of admissible rules, which were introduced to consideration by P.Lorentzen (1955). A rule  $r$  is admissible in a propositional logic  $\mathcal{L}$  if  $\mathcal{L}$ , as the set of formulas, is closed w.r.t. applications of  $r$ . That is, for any substitution  $\sigma$ , if  $\sigma$  turns all premisses of  $r$  to theorems of  $\mathcal{L}$ , then the conclusion  $\psi$  turns in a theorem after application of  $\sigma$ . It is very interesting and challenging to develop a technique to determine inference rules admissible in described logics, but up to day we see no closure solution. Only a point is: it is immediate to see that if a rule  $r$  is not admissible then  $r$  is not valid. Therefore theorems of this paper concerning decidability w.r.t. valid rules give a necessary condition for rules to be admissible.

## 4 Conclusion, Future Work

Our paper suggests methods to check valid inference rules for several agents' temporal logic. Some merely reduce the problem to model checking via decidability of logics (for linear time), others (for more complex cases) are using more intelligent technique involving construction reduced form or rules and an advanced filtration technique. In sum, in mathematical terms, we show that considered agents' temporal logics are decidable w.r.t. valid inference rules.

There are many open venues for the future research, including (i) extensions of results to other logics from AI area, in particular with more complicated restriction on structures of time frames, (ii) construction of more efficient algorithms checking validity of inference rules in considered logics. The mentioned above area of questions concerning admissibility of inference rules in agents' temporal logics is very interesting and actual as well. The results of our paper might be useful for researchers from AI community interested in logical inference in agents' logics, knowledge representation, and human(automated) reasoning.

## References

1. van Benthem, J.: The Logic of Time.- A Model-Theoretic Investigation into the Varieties Temporal Ontology and Temporal Discourse. Kluwer, Dordrecht (1991)
2. van Benthem, J.: Modality, bisimulation and interpolation in infinitary logic. Ann. Pure Appl. Logic 96 (1999)
3. Gabbay, D.M., Schlechta, K.: A theory of hierarchical consequence and conditionals. Journal of Logic, Language and Information 19(1), 3–32 (2010)
4. Gabbay, D.M., Rodrigues, O., Pigozzi, G.: Connections between belief revision, belief merging and social choice. J. Log. Comput. 19(3), 445–446 (2009)

5. Gabbay, D., Kurucz, A., Wolter, F., Zakharyaschev, M.: *Stud. Logic Found. Math.* Elsevier Sci. Publ., Noth-Holland (2003)
6. Ian Hodkinson, A.M., Sciavicco, G.: Non-finite axiomatisability and undecidability of interval temporal logics with c, d, and t. In: Kaminski, M., Martini, S. (eds.) CSL 2008. LNCS, vol. 5213, pp. 308–322. Springer, Heidelberg (2008)
7. Hodkinson, I.M.: Complexity of monodic guarded fragments over linear and real time. *Ann. Pure Appl. Logic* 138(1–3), 94–125 (2006)
8. Hodkinson, I., Woter, F., Zakharyaschev, M.: Undecidable fragments of first-order branching time logic. In: LICS 2002, pp. 393–402 (2002)
9. Hodkinson, I.: Temporal logic and automata. In: Temporal Logic. Math. Found. and Comp. Asp., ch. 2, vol. 2, pp. 30–72. Clarendon Press (2000)
10. Fagin, R., Halpern, J., Moses, Y., Vardi, M.: *Reasoning About Knowledge*. The MIT Press, Cambridge (1995)
11. Halpern, J., Shore, R.: Reasoning about common knowledge with infinitely many agents. *Information and Computation* 191(1), 1–40 (2004)
12. Schmidt, R., Tishkovsky, D.: Multi-agent dynamic logics with informational test. *Annals of Mathematics and Artificial Intelligence* 42(1–3), 5–36 (September 2004)
13. Hintikka, J., Vandamme, F.: *Logic of Discovery and Logic of Discourse*. Springer, Heidelberg (1986)
14. Pnueli, A.: The temporal logic of programs. In: Proc. of the 18th Annual Symp. on Foundations of Computer Science, pp. 46–57. IEEE, Los Alamitos (1977)
15. Manna, Z., Pnueli, A.: *Temporal Verification of Reactive Systems: Safety*. Springer, Heidelberg (1995)
16. Manna, Z., Pnueli, A.: *The Temporal Logic of Reactive and Concurrent Systems: Specification*. Springer, Heidelberg (1992)
17. Clarke, E., Grumberg, O., Hamaguchi, K.P.: Another look at ltl model checking. In: Dill, D.L. (ed.) CAV 1994. LNCS, vol. 818. Springer, Heidelberg (1994)
18. Daniele, M., Giunchiglia, F., Vardi, M.: Improved automata generation for linear temporal logic. In: Halbwachs, N., Peled, D.A. (eds.) CAV 1999. LNCS, vol. 1633, pp. 249–260. Springer, Heidelberg (1999)
19. Vardi, M.: An automata-theoretic approach to linear temporal logic. In: Proceedings of the Banff Workshop on Knowledge Acquisition, Banff 1994 (1994)
20. van Benthem, J.: *The Logic of Time*. Kluwer, Dordrecht (1991)
21. van Benthem, J., Bergstra, J.: Logic of transition systems. *Journal of Logic, Language and Information* 3(4), 247–283 (1994)
22. Goldblatt, R.: *Logics of Time and Computation*. CSLI Lecture Notes, vol. 7 (1992)
23. Gabbay, D., Hodkinson, I.: An axiomatisation of the temporal logic with until and since over the real numbers. *Journal of Logic and Computation* 1(2), 229–260 (1990)
24. Hodkinson, I.: *Temporal Logic and Automata*, Chapter II of *Temporal Logic*, vol. 2, pp. 30–72. Clarendon Press, Oxford (2000)
25. Rybakov, V.: Logical consecutions in discrete linear temporal logic. *Journal of Symbolic Logic* 70(4), 1137–1149 (2005)
26. Rybakov, V.: Logical consecutions in intransitive temporal linear logic of finite intervals. *Journal of Logic Computation* 15(5), 633–657 (2005)
27. Rybakov, V.: Until-Since Temporal Logic Based on Parallel Time with Common Past. Deciding Algorithms. In: Artemov, S., Nerode, A. (eds.) LFCS 2007. LNCS, vol. 4514, pp. 487–497. Springer, Heidelberg (2007)
28. Bordini, R.H., Fisher, M., Visser, W., Wooldridge, M.: Model checking rational agents. *IEEE Intelligent Systems* 19, 46–52 (September/October 2004)

29. Dix, J., Fisher, M., Levesque, H., Sterling, L.: Editorial. *Annals of Mathematics and Artificial Intelligence* 41(2–4), 131–133 (2004)
30. van der Hoek, W., Wooldridge, M.: Towards a logic of rational agency. *Logic Journal of the IGPL* 11(2), 133–157 (2003)
31. Fisher, M.: Temporal development methods for agent-based systems. *Journal of Autonomous Agents and Multi-Agent Systems* 10(1), 41–66 (2005)
32. Hendler, J.: Agents and the semantic web. *IEEE Intelligent Systems* 16(2), 30–37 (2001)
33. Kacprzak, M.: Undecidability of a multi-agent logic. *Fundamenta Informaticae* 45(2–3), 213–220 (2003)
34. Wooldridge, M., Weiss, G., Ciancarini, P. (eds.): AOSE 2001. LNCS, vol. 2222. Springer, Heidelberg (2002)
35. Fagin, R., Geanakoplos, J., Halpern, J., Vardi, M.: The hierarchical approach to modeling knowledge and common knowledge. *International Journal of Game Theory* 28(3), 331–365 (1999)
36. Rybakov, V.: Logic of discovery in uncertain situations – deciding algorithms. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 950–958. Springer, Heidelberg (2007)
37. Babenyshev, S., Rybakov, V.V.: Decidability of hybrid logic with local common knowledge based on linear temporal logic *LTL*. In: Beckmann, A., Dimitracopoulos, C., Löwe, B. (eds.) CiE 2008. LNCS, vol. 5028, pp. 32–41. Springer, Heidelberg (2008)
38. Babenyshev, S., Rybakov, V.V.: Describing evolutions of multi-agent systems. In: Velásquez, J.D., Ríos, S.A., Howlett, R.J., Jain, L.C. (eds.) Knowledge-Based and Intelligent Information and Engineering Systems. LNCS (LNAI), vol. 5711, pp. 38–45. Springer, Heidelberg (2009)
39. Rybakov, V.V.: Linear temporal logic *K* extended by multi-agent logic *K<sub>n</sub>* with interacting agents. *J. of Logic Computation* 19, 989–1017 (2009)
40. Rybakov, V.: Admissible Logical Inference Rules. Studies in Logic and the Foundations of Mathematics, vol. 136. Elsevier Sci. Publ., North-Holland (1997)

# Statecharts-Based JADE Agents and Tools for Engineering Multi-Agent Systems

Giancarlo Fortino, Francesco Rango, and Wilma Russo

Dept. of Electronics, Informatics and Systems (DEIS) – University of Calabria,  
Via P. Bucci, cubo 41C, 87036 Rende (CS), Italy  
`{g.fortino,w.russo}@unical.it, frango@si.deis.unical.it`

**Abstract.** In this paper we propose frameworks and tools supporting a Statecharts-based development of JADE-based multi-agent systems (MASs) with the purpose of providing an effective approach for engineering MASs and leveraging agent-oriented development methodologies and processes adopting JADE as target agent platform. In particular, a framework for programming JADE behaviors through a variant of the Statecharts, named Distilled StateCharts (DSCs), has been first developed by enhancing the JADE add-on HSMBehaviour. Then, to enable rapid prototyping of JADE agents, a visual tool for DSCs has been extended with translation rules based on the developed framework which allows to automatically translate DSC specifications into DSC-based JADE behaviors. The proposed approach is exemplified through a case study concerning an agent-based meeting organization system.

**Keywords:** Statecharts, Software agents, JADE, Visual programming, Automatic code generation, CASE tool.

## 1 Introduction

In the last decade the agent oriented software engineering (AOSE) research area has produced a rich set of methodologies and tools which can be actually exploited for the development of complex software systems in terms of multi-agent systems (MASs) [1]. In parallel with AOSE, the mainstream software engineering area has driven UML 2.0 [2] along with related methodologies and tools to become the *de facto* standard for the development of software systems. In particular, the UML state machines, derived from the Harel's Statecharts [3], are an effective and widely adopted formalism for the specification of active component behavior and protocols in general-purpose and real-time systems. It is widely recognized that the benefits provided by Statecharts for engineering complex software systems are mainly visual programming, executable specifications, protocol-oriented specifications, and a set of CASE tools facilitating software development. In this context, to effectively develop multi-agent systems (MAS), models, frameworks and tools are needed to support flexible and rigorous specification and subsequent implementation of agent behaviors and agent-to-agent interaction protocols [4]. Thus the use of Statecharts-based models, frameworks and tools for the development of MASs could provide the same benefits in the AOSE research area as those provided in the context of traditional software

engineering. However, in the AOSE research area, Statecharts are still under-used to specify agent behaviors and protocols even though some proposed agent models founded on different types of state machines are available. Among such proposals, the most known and interesting ones are the JADE FSMBehaviour [5], the SmartAgent framework [6], the ELDA agent model [7], and the Bond agent framework [8]. In particular, The JADE framework [5], one of the most used agent-oriented framework in academy and industry, provides the FSMBehaviour [9] for the modeling of agent behaviors based on finite state machines (FSMs). However their programming is not flexible as they are not based on ECA (Event-Condition-Action)-rule based transitions, and do not provide important mechanisms for reducing behavior complexity such as well structured OR-decomposition and history entrances. In fact, although states of the FSMBehaviour can be FSMBehaviors or other behaviors, mechanisms for handling this induced state hierarchy are not provided. The SmartAgent model [10, 6] extends the JADE CompositeBehaviour and provides a behavior based on hierarchical state machine driven by ECA rules, named HSMBehaviour. However, it does not support shallow and deep history entrance mechanisms, useful for reducing behavior complexity even further and for transparently archiving agent states. Although visual modeling and emulation of HSMBehaviour agents can be done with the provided HSMEditor [11], automatic translation of modeled agents into JADE code is not yet available. The ELDA (Event-driven Lightweight Distilled Statecharts-based Agents) agent model [7] is based on a Statecharts-like machine, providing or-decomposition and history entrance mechanisms, named Distilled StateCharts [12] suitable for the modeling of lightweight agents for distributed computing. Moreover, they can be effectively modeled through the ELDATool, a graphical tool for visual specification, automatic code translation and simulation of ELDA-based systems [13]. However, an ELDA-based execution platform is not yet available so confining the use of ELDA agents in the MAS simulation domain. The behavior of the Bond agents [8] is based on a multi-plane state machine where each plane is modeled as a finite state machine (FSM). However, the Bond agent model does not offer the state hierarchy and history mechanisms, which are important for coping with complexity, and tools for automating agent prototyping. Finally other previous agent frameworks are ZEUS [14], which provides a more complete non-hierarchical state machine execution subsystem but the API is not well-developed, and the JACKAL conversation engine which also uses a state machine model [15].

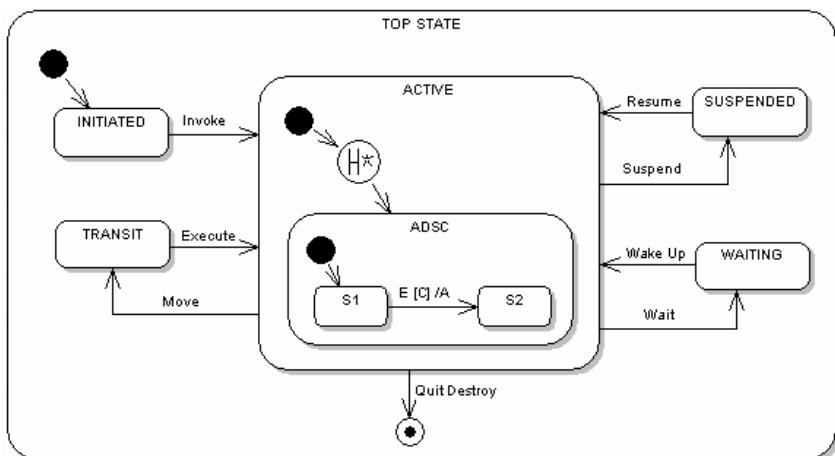
In this paper we propose programming frameworks and techniques supporting a Statecharts-based development of JADE multi-agent systems. This research contribution is important as it aims at integrating Statecharts and MASs to deliver the same important benefits provided by Statecharts for the engineering of traditional software systems. Moreover, the proposed approach can be fruitfully exploited to leverage already existing agent-oriented development methodologies and processes adopting JADE as target agent platform (e.g. INGENIAS [16], PASSI [17], MESSAGE [18]). In particular, a framework for programming JADE behaviors through the Distilled StateCharts (DSCs) formalism, named DistilledStateChartBehaviour, has been developed by enhancing the JADE HSMBehaviour. To enable rapid prototyping of JADE agents, a CASE tool obtained by enhancing the ELDATool with a new component based on the DistilledStateChartBehaviour for automatic code generation of DSC-based behaviors

into JADE code, is made available. The proposed approach is exemplified with reference to an agent-based meeting organization system.

The rest of this paper is organized as follows. In section 2, after an introduction of the basic concepts of the Distilled StateCharts formalism, the JADE DistilledStateChartBehaviour is described and a case study exemplifying the proposed Statecharts-based approach is proposed. In section 3 a CASE tool-driven approach for engineering JADE-based MAS from modeling to implementation, is presented. Finally, conclusions are drawn and on-going work delineated.

## 2 Statecharts-Based JADE Agents

In this section, the DSC formalism, which provides a powerful and rich set of modeling concepts enabling an effective specification of agent behavior, is overviewed. Then, the proposed programming framework for DSC-based JADE agents is described so enhancing JADE with the benefits deriving from Statecharts. Finally an example of an agent-based system modeled through DSC-based JADE agents is presented.



**Fig. 1.** A FIPA compliant DSC-based agent behavior

### 2.1 The Distilled StateCharts Model

The Distilled StateCharts (DSCs) formalism [12] is derived from the Harel's Statecharts through a distillation process, purposely carried out for the modeling of lightweight mobile agent behavior, which led to the following structural/semantics differences between Statecharts and DSCs:

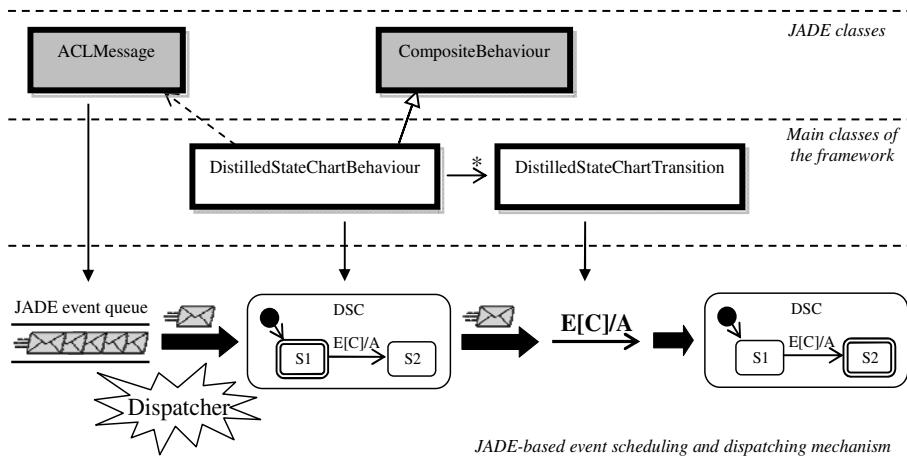
- State entry actions, exit actions and activities are empty so actions can be only hooked to transitions;
- Each composite state has a pseudo initial state from which the default entrance of the composite state originates;

- Transitions (apart from default entrance and default history entrances) are always labeled by an event;
- Default entrance and default history entrances can only be labeled with an action;
- And-decomposition of states and related synchronization modeling elements are not used as DSCs are intended for the behavioral modeling of single-threaded agents;
- Run-to-completion step semantics defined according to the UML state machines semantics [19].

A DSC-based agent behavior relies on an enhanced basic template built according to the FIPA agent lifecycle [20] which JADE agents are compliant with (see Figure 1). In particular, the ACTIVE state, in which an agent carries out its goal-oriented tasks, is always entered through a deep history entrance ( $H^*$ ) whose default history entrance targets the active DSC (ADSC), which actually models the active agent behavior. The default entrance of ACTIVE targeting  $H^*$  allows restoring the agent execution state after agent migration and, in general, after agent suspension.

## 2.2 A Framework for Programming DSC-Based JADE Agents

A new JADE behavior, named DistilledStateChartBehaviour, has been defined to program JADE agents through the DSC formalism. In particular, the DistilledStateChartBehaviour is defined by enhancing the HSMBehaviour [10, 6] with the DSC mechanisms, specifically implementing the history mechanisms which allow a partial (through shallow history H) or full (through deep history  $H^*$ ) recovery of the state history after re-entering into any state previously exited.



**Fig. 2.** Simplified class diagram of the JADE DistilledStateChartBehaviour and event-handling architectural scheme

Figure 2 shows a simplified UML class diagram of the DistilledStateChartBehavior along with a schema of the JADE-based event scheduling and dispatching mechanism integrated into the DistilledStateChartBehavior . The DistilledStateChartBehavior inherits from the JADE CompositeBehaviour and includes a set of nested DistilledStateChartBehaviors and other Behaviours which represent the *states* of the DSC. It maintains the list of *transitions*, represented by the DistilledStateChartTransition class, and handles the event-driven mechanism for transition firing which also determines the “current” state of the DSC state machine at run-time. As it is shown in Figure 2, an event E, instance of the ACLMessage class, is fetched from the JADE event queue by the dispatcher component of the DistilledStateChartBehavior and delivered to the DSC current state (S1) so triggering a state transition to a new state (S2) if the guard C holds.

In the following a detailed description of the main mechanisms (behavior scheduling, event handling, transition firing and history entrances) of the DistilledStateChartBehavior is presented.

**Behaviour scheduling.** The DistilledStateChartBehavior receives the thread of control from the JADE run-time system through the invocation of the *action* method according to the cooperative concurrency mechanism of JADE. The *action* method of the DistilledStateChartBehavior , in turn, invokes the *action* method of the current state; the DistilledStateChartBehavior starts executing the initial state and, successively, activates other states by following the fired transitions and, finally, terminates when completes one of its final states. On the invocation of the *action* method of the current behavior, the Wrapper object, which encapsulates each simple state, allows checking and executing (through the *findAndFireTransition* method) possible transitions outgoing from the current state. This mechanism allows implementing the UML state machine rule: “as soon as a transition is able to fire, it does”. Indeed, the actual implementation is based on the single-threaded model of JADE, which does not support preemption of an action execution.

**Event handling.** An important feature of the DSC state machines is the event driven mechanism for triggering transitions. An event can be represented as a regular JADE ACLMessage or as a DistilledStateChartEvent (extending ACLMessage) so enabling the reuse of the message queuing mechanism of JADE (see Figure 2): when the DistilledStateChartBehaviour is checking for a transition firing, the *receive* method of JADE is invoked to fetch the first message in the queue, which is then passed to the transitions to check if one of them can be fired. The main issue of such mechanism is the integration of behaviors as states. As an event message in queue is fetched through the regular *receive* method, programmers of the DistilledStateChartBehavior can use the *receive* method inside the *action* method of states so possibly interfering with the transition firing mechanism. Moreover, if a message/event is received in a state in which the event is not expected, two handling solutions are possible: (i) the event is put back into the queue so that it could be fetched by another state that is able to handle it; (ii) the event is not put back so it is discharged. This option can be set in the DistilledStateChartBehaviour constructor. The same mechanism can be also used when an agent has multiple behaviors for the purpose of not losing events. In this case, the message template mechanism based on selective filters for events can be

used. In particular, each behavior performs a *receive* operation with a different message template so as to fetch only the events it is able to handle.

**Transition firing.** A transition is represented by the DistilledStateChartTransition class and is added through the *addTransition* method which takes as parameters the transition to be added and the source state. The target state is defined at DistilledStateChartTransition creation and can be at any level of the hierarchy so supporting the specification of inter-level state transitions. The DistilledStateChartTransition unifies trigger event and guard mechanisms into the *trigger(Behaviour source, ACL-Message event)* method, where *source* is the transition source state and *event* is the transition triggering event. The *trigger* method checks for the transition firing and, if the check is positive, the *action* method of DistilledStateChartTransition, which can contain the action hooked to the transition, is invoked. The check based on both the *trigger* and *findAndFireTransition* methods not only involves the current state but also all the states encapsulating it in the order from the inner to the outer states.

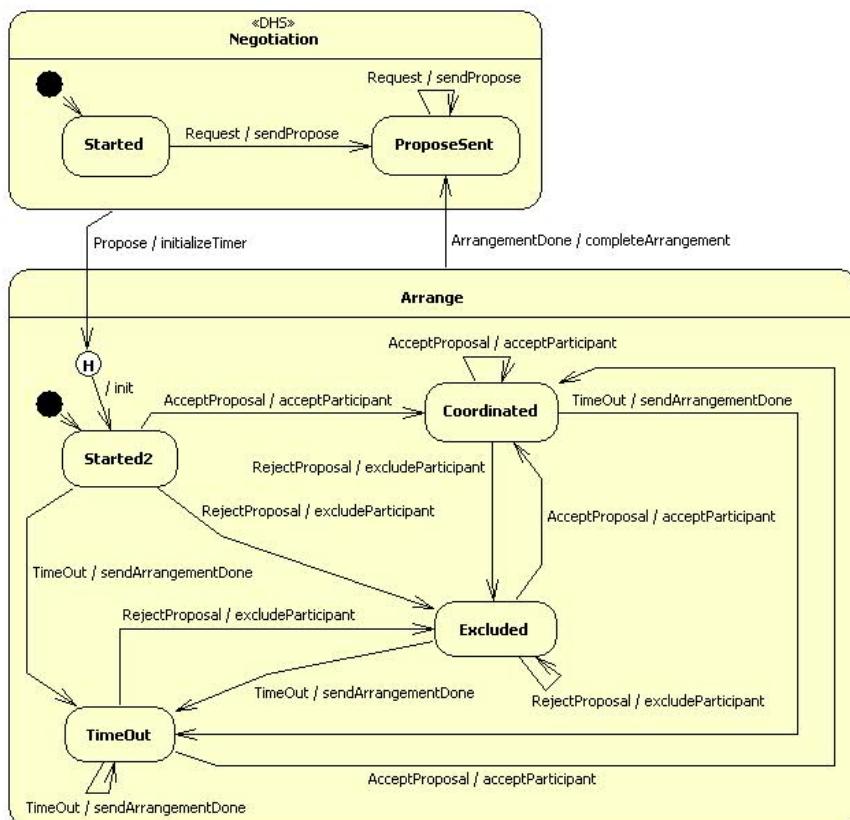
**History entrances.** The DistilledStateChartBehaviour includes the *defaultDeepHistoryEntrance* and the *defaultShallowHistoryEntrance* referring to the states (or behaviors) associated to the deep and shallow history entrances, respectively. To restore the state history, the *lastState* variable of a composite state of the DistilledStateChartBehaviour type, which stores a reference to the last visited state before exiting the composite state, is used. Moreover, the DistilledStateChart Transition includes the two constants *DEEP\_HISTORY* and *SHALLOW\_HISTORY* which indicate that the composite state that is target of the transition is to be entered through the deep or shallow history.

### 2.3 An Agent-Based Meeting Organization System

In this section a DSC-based design of an agent-based meeting organization system, which uses agents to coordinate and arrange meetings in an intelligent way, is proposed. The designed system is derived from a case study based on a meeting participant protocol proposed in [21, 11]. The designed MAS is based on three types of agents: MeetingRequesterAgent (MRA), MeetingBrokerAgent (MBA), MeetingParticipantAgent (MPA). The MBA (whose behavior is reported in Figure 3), manages the meeting arrangement requests sent by the MRA (the meeting organizer), and coordinates the MPAs (the meeting participants). The MRA sends a Request event to the MBA containing all needed information (potential participants, minimum number of participants, meeting topic, and chosen date) related to the appointment to arrange and waits for the meeting confirmation. Upon the reception of the Request event, the MBA sends itself and all the MPAs a Propose event containing the appointment to schedule and then starts a timer. The MPAs send the MBA an AcceptProposal event to accept the appointment or a RejectProposal event to refuse it. On the basis of the received responses, the MBA accepts or excludes the participants and, when it receives all the responses or when the timeout associated to the set timer expires, sends an ArrangementDone event to itself to carry out the final operations (in the *completeArrangement* action) for the current appointment as follows:

- If at least M MPAs have accepted the appointment, the meeting organization is successfully done; then, the MBA sends a Confirm event to the MRA and to the accepting MPAs, which schedule the appointment in their rosters.
- If the appointment has been accepted by less than M MPA and it is not yet reached the maximum limit of N requests of new participants sent to the MRA, the MBA issues a request of new participants to the MRA by sending it an AskForRequest event. Then, the MRA sends a new Request event to the MBA indicating new participants for the same appointment. This way, the MBA can retry to schedule the appointment involving the new provided participants.
- If the appointment has been accepted by less than M MPA and it is reached the maximum limit of N requests of new participants sent to the MRA, the appointment is canceled and a Cancel event is sent to the accepting MPAs.

Table 1 summarizes the event-based interaction relationships among agents, specifying the event source agent, which generates the event, and the event target agent, which receives and handles the event.



**Fig. 3.** DSC-based behavior of the MBA

The shallow history entrance (H) provides a powerful modeling solution when the Arrange state is to be re-entered due to a new Request related to the same appointment. In particular, when a new Request event is received, the MBA, which is in the ProposeSent state, goes into the most recently left simple state of the Arrange state, recovering exactly the same state variable and DSC state so continuing from the previous arrangement state without discontinuity.

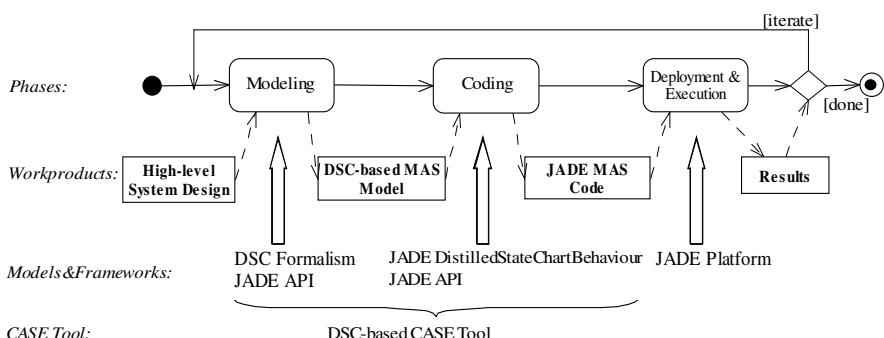
**Table 1.** Event-based interaction relationships among agents

TARGET SOURCE \	MRA	MBA	MPA
MRA		Request	
MBA	Confirm, AskForRequest	Propose, ArrangementDone	Propose, Confirm, Cancel
MPA		AcceptProposal, RejectProposal	

### 3 CASE Tool-Driven Development of DSC-Based JADE Agents

The development of DSC-based JADE agents relies on the process reported in Figure 4 which is organized in the following three phases:

- The *Modeling* phase produces the DSC-based MAS Model on the basis of the High-Level System Design which can be defined either ad-hoc or by means of other methodologies which also support the analysis and high-level design phases [17,18,16]. In particular, the DSC-based MAS Model is specified through the DSC formalism and the JADE API.
- The *Coding* phase works out the DSC-based MAS Model and automatically produces the JADE MAS code according to the DistilledStateChartBehaviour.
- The *Deployment and Execution* phase is fully supported by the JADE Platform to run the developed MAS. A careful evaluation of the obtained Testing Results (e.g. execution traces, performance indices, etc) with respect to the functional and non-functional requirements could lead to a further iteration step which starts from a new (re)modeling activity.

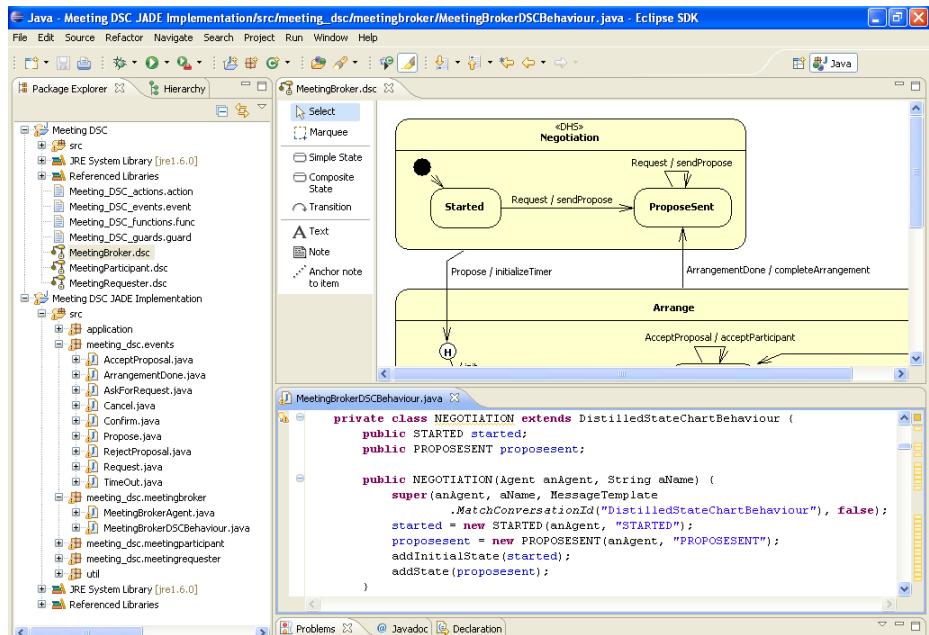


**Fig. 4.** The CASE-driven development process

The first two phases are fully supported by the a DSC-based CASE tool that makes it available (i) the visual modeling of the DSC-based behavior of the agents composing the MAS under-development and (ii) the automatic translation of the modeled agent behaviors into ready-to-be-executed JADE code according to the DistilledStatechartsBehaviour framework.

The CASE tool is obtained by enhancing the ELDATool [7], a graphical tool for visual specification, automatic code translation and simulation of ELDA-based systems, with a new developed component named CodeGeneratorForJADE embedded into the ELDAEditor plug-in. This important facility, which is not offered by the HSMBehaviour graphical tools [11], makes the programming of Statecharts-based JADE agents easier than manual programming of the HSMBehaviour and Distilled-StateChartBehaviour based on complex programming patterns.

As the ELDATool is based on the ELDA agent model [7] which is different from the JADE agent model, event modeling is carried out by carefully considering such difference. In particular, the only event types, provided by the ELDATool, exploitable for the modeling phase are: (i) the ELDAEventMSG, which represents asynchronous messages; (ii) the ELDAEventInternal, which represents self-triggering events. Both kinds of events correspond to the ACLMessage class of JADE or to user-defined classes extending ACLMessage. Moreover it is worth noting that the specification of state variables, actions, guards, events and functions is based on the Java language and the JADE API.



**Fig. 5.** A snapshot of the CASE tool showing the developed system

Figure 5 reports a screenshot of the CASE tool containing the fully developed system described in section 2.3. In particular, in the package explorer there are two folders: (i) *Meeting DSC* containing the set of graphical DSC agent behaviors (*MeetingBroker.dsc*, *MeetingParticipant.dsc*, *MeetingRequester.dsc*) and their related actions, events, functions, and guards; (ii) *Meeting DSC JADE Implementation* containing the generated source code (*src* package). In the central panel, the *MeetingBroker.dsc* is visualized (the complete diagram is reported in Figure 3). Finally in the bottom panel, an excerpt of the generated code of the *MeetingBrokerDSCBehaviour* is reported. The *DistilledStateChartBehaviour* framework class code along with the generated source code of the Meeting DSC example complete of all the involved agents is available at [22].

## 4 Conclusion

This paper has proposed programming techniques and tools based on Statecharts for the rapid development of JADE MASs. In particular, a new JADE behavior, named *DistilledStateChartBehaviour*, has been defined which is based on the Distilled State-Charts formalism providing hierarchical state machines including history mechanisms and features for enabling an automatic restoring of the agent execution state. The *DistilledStateChartBehaviour* has been obtained on the basis of the *HSMBehaviour* JADE add-on purposely debugged and optimized. Moreover, the availability of a CASE tool supporting the specification phase of JADE agent behaviors based on the *DistilledStateChartBehaviour* and their automatic translation into code, facilitates programming and enables rapid prototyping. As the JADE platform is one of the most used agent platform in the AOSE community to program and execute distributed agent systems, the paper proposal contributes to (i) enrich already existing agent-oriented methodologies having JADE as target platform with tools for further automating MAS development and (ii) foster a wider introduction and exploitation of Statecharts-based techniques for agents.

Future work is geared at (i) integrating the code translator for JADE into the ELDA process for automatically generating JADE code after validation through simulation of an ELDA-based MAS; (ii) integrating Statecharts-based modeling and the defined techniques within an MDD-driven agent-oriented methodology such as INGENIAS; (iii) releasing the *DistilledStateChartBehaviour* as JADE add-on; (iv) defining a reverse engineering technique to obtain the DSC-based agent visual model from the agent source code compliant to the *DistilledStateChartBehaviour*.

## References

1. Zambonelli, F., Omicini, A.: Challenges and Research Directions in Agent-Oriented Software Engineering. *Autonomous Agents and Multi-Agent Systems* 9(3), 253–283 (2004)
2. Ambler, S.W.: *The Elements of UML 2*. Cambridge University Press, Cambridge (2005)
3. Harel, D., Gery, E.: Executable Object Modeling with Statecharts. *IEEE Computer* 30(7), 31–42 (1997)
4. Luck, M., McBurney, P., Preist, C.: A manifesto for agent technology: towards next generation computing. *Autonomous Agents and Multi-Agent Systems* 9(3), 203–252 (2004)

5. Bellifemine, F., Caire, G., Greenwood, D.: *Developing Multi-Agent Systems with JADE*. Wiley, Chichester (2007)
6. Griss, M., Fonseca, S., Cowan, D., Kessler, R.: SmartAgent: Extending the JADE agent behavior model. In: Proc. of the Agent Oriented Software Engineering Workshop, Conference in Systems, Cybernetics and Informatics, Orlando, Florida (July 2002)
7. Fortino, G., Garro, A., Mascillaro, S., Russo, W.: Using Event-driven Lightweight DSC-based Agents for MAS Modeling. *International Journal on Agent Oriented Software Engineering* 4(2) (2010)
8. Boloni, L., Marinescu, D.C.: A component agent model – from theory to implementation. In: Müller, J., Petta, P. (eds.) *Proc. of the Second International Symposium from Agent Theory to Agent Implementation*; Trappi, R. (ed.): *Proc. of Cybernetics and Systems*, pp. 663–639. Austrian Society of Cybernetic Studies, Vienna (2000)
9. Bellifemine, F., Poggi, A., Rimassa, G.: Developing multi agent systems with a FIPA-compliant agent framework. *Software Practice and Experience* 31, 103–128 (2001)
10. Griss, M., Fonseca, S., Cowan, D., Kessler, R.: Using UML State Machines Models for More Precise and Flexible JADE Agent Behaviors. In: AAMAS AOSE workshop, Bologna, Italy (July 2002)
11. Kessler, R., Griss, M., Remick, B., Delucchi, R.: A Hierarchical State Machine using JADE Behaviours with Animation Visualization. Technical report, University of Utah (2004)
12. Fortino, G., Russo, W., Zimeo, E.: A statecharts-based software development process for mobile agents. *Information and Software Technology* 46(13), 907–921 (2004)
13. Fortino, G., Garro, A., Mascillaro, S., Russo, W.: ELDATool: A Statecharts-based Tool for Prototyping Multi-Agent Systems. In: Proc. of the Workshop on Objects and Agents (WOA 2007), Genova, Italy, September 24–25 (2007)
14. Nwana, H., Nduma, D., Lee, L., Collis, J.: ZEUS: a toolkit for building distributed multi-agent systems. *Artificial Intelligence Journal* 13(1), 129–186 (1999)
15. Cost, R., Finin, T., Labrou, Y., Luan, X., Peng, Y., Soboroff, I., Mayfield, J., Boughanam, A.: Jackal: A Java-Based Tool for Agent Development. In: Working Notes of the Workshop on Tools for Developing Agents, AAAI 1998 (1998)
16. García-Magariño, I., Gómez-Sanz, J.J., Fuentes-Fernández, R.: Model Transformations for Improving Multi-agent Systems Development in INGENIAS. In: Proc. of the 10th International Workshop on Agent-Oriented Software Engineering, AOSE 2009 (2009)
17. Cossentino, M.: From Requirements to Code with the PASSI Methodology. In: Henderson-Sellers, B., Giorgini, P. (eds.) *Agent-Oriented Methodologies*. Idea Group Inc., Hershey (2005)
18. Caire, G., Coulier, W., Garijo, F., Gómez-Sanz, J., Pavón, J., Kearney, P., Massonet, P.: The Message Methodology. In: Bergenti, F., Gleizes, M.-P., Zambonelli, F. (eds.) *Methodologies and Software Engineering for Agent Systems The Agent-Oriented Software Engineering Handbook*, vol. 11, pp. 177–194. Springer, Heidelberg (2006)
19. Eshuis, R.: Reconciling statechart semantics. *Science of Computer Programming* 74(3), 65–99 (2009)
20. FIPA (Foundation for Intelligent Physical Agents), FIPA Agent Management Support for Mobility Specification, Document FIPA DC00087C (2002/05/10) (2002), <http://www.fipa.org/>
21. Fonseca, S., Griss, M., Letsinger, R.: Agent Behavior Architectures A MAS Framework Comparison, Technical report, N. HPL-2001-332, University of Utah (2001)
22. The JADE DistilledStateChartBehaviour (2010) documentation and software at, <http://plasma.deis.unical.it/software/DSC4JADE/>

# Telco Agent: Enabler of Paradigm Shift towards Customer-Managed Relationship

Vedran Podobnik and Ignac Lovrek

University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia  
`{vedran.podobnik, ignac.lovrek}@fer.hr`

**Abstract.** In this article we propose an agent-based solution for enabling paradigm shift from CRM (*Customer Relationship Management*) to CMR (*Customer-Managed Relationship*). Namely, Telco Agent is a software agent which represents telecom operator in interactions with its customers. Customers provide Telco Agent with their profiles and afterwards Telco Agent uses its mechanisms for semantic matchmaking of customer profiles and creation of customer social network to facilitate autonomous CMR.

**Keywords:** customer-managed relationship, software agent, telecom service provisioning, social networking.

## 1 Introduction

Telecom service provisioning paradigm is shifting from CRM (*Customer Relationship Management*) to CMR (*Customer-Managed Relationship*) [1]. The CMR is a relationship in which a company (i.e., telecom operator – telco) uses a methodology, software and Internet capabilities to encourage its customers (i.e., telecom service users) to manage information pertaining to them (e.g., their profiles). Additionally, CMR paradigm allows customers to define how they want to communicate with the company, what types of services or products they want to purchase, and how they want to pay. It can be noted that described paradigm shift markedly magnifies the dynamics of customer-telco interactions. Consequently, appropriate computing solution should be applied for enabling presented shift towards CMR, to prevent negative effects for both customers and telco (primarily from the perspective of time consumption). We propose software agents as a solution for enabling CMR.

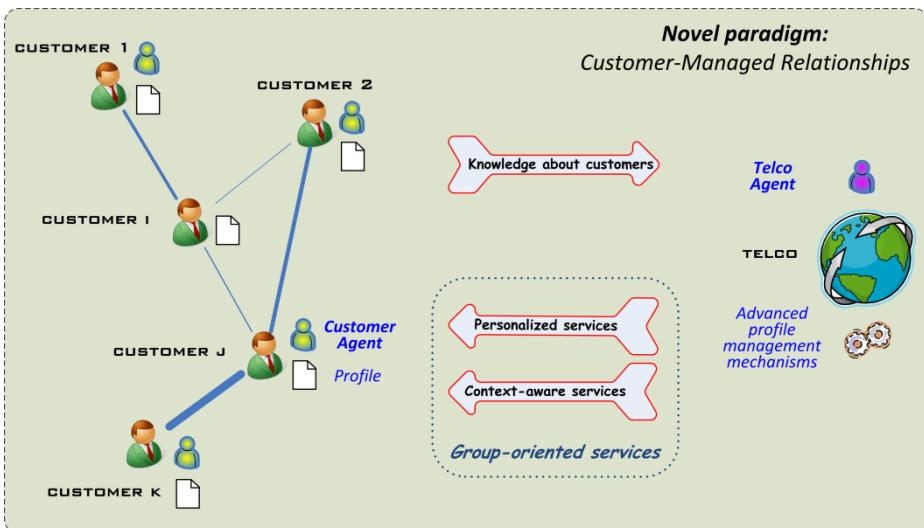
Software agents and multi-agent systems have proven to be very suitable technology for customer profile management and telecom processes enhancements [2][3]. A software agent is a program which autonomously acts on behalf of its principal, while carrying out complex information and communication tasks that have been delegated to it. From the owner's point of view, agents improve her/his efficiency by reducing the time required to execute personal and/or business tasks.

This article is organized as follows. In Section 2, we explain an idea of agent-based CMR in telecom industry. Section 3 describes the Telco Agent, our proposal for enabling the presented paradigm shift. Section 4 concludes the paper and gives ideas for future research work.

## 2 Agent-Enabled Customer-Managed Relationship

Shift towards CMR requires from telco redefining its role in the market and developing novel business models to create new revenue streams and gain competitive advantage. Solution lies in transforming into business entities that are characterized by orientation towards customers and high levels of innovation [4]. In such a scenario is not only important to optimize processes related to transport data through the network, but also great emphasis is put on business processes that describe how to provide telecom service and coordinate the relationship between each customer and telco. To achieve that it is very important for telco to efficiently manage knowledge about its customers – this means that it is not enough just to collect and store knowledge about customers, but it is necessary to use this knowledge in a way that allows the improvement of business processes [5].

The above mentioned requisites brought up by the CMR paradigm considerably influence the design of telco's ICT (*Information and Communication Technologies*) system: ICT solutions must be upgraded with mechanisms which enable advanced customer profile management. Namely, creation, storage and continuous update of customer profiles, as well as autonomous matchmaking of those profiles should be enabled [6] [7] [8].



**Fig. 1.** Customer-managed relationship enabled by software agents

The idea of a solution proposed in this paper is shown in Fig 1. The main motivation for such a solution was replacement of the existing approach where a telecom service is offered to all customers in the same form and under same conditions, regardless the fact that every customer has unique characteristics. In the proposed novel approach the telco observes characteristics of every individual customer and tailors a service to a group of customers that are similar according to certain criteria (i.e., *personalized* [9] and

*context-aware service*). To enable that telco must ensure that each customer has attached a *profile* used for collecting and storing customer-related information. Moreover, customer profiles represent grounding for building knowledge about customer interests, preferences, context and other customer characteristics relevant for service offering and provisioning. Finally, telco can explore this knowledge in order to define structures that reflect relations among its customers and build implicit social networks.

The presented solution is implemented as a multi-agent system (MAS) consisting of two types of agents: *Customer Agents* and *Telco Agent*. Customer Agents represent telco's customers by defining and maintaining a profile of their respective owners. On the other hand, telco is represented by Telco Agent which communicates with a set of Customer Agents and thus has access to customer profiles (i.e., collection of knowledge about customers).

### 3 Telco Agent

In Fig. 2 the general model of a software agent [10] is presented, while emphasizing properties essential for creation of a Telco Agent:

- **Intelligence.** Intelligence of a Telco Agent is grounded on its reasoning mechanisms. Namely, Telco Agent implements mechanisms for semantic matchmaking of customer profiles and creation of customer social network;
- **Autonomy.** Telco Agent executes tasks autonomously without interventions from its owner. Its autonomy is manifested through following two properties – *reactivity* and *proactivity*;
- **Reactivity.** Telco Agent is able to respond to changes in its environment (e.g., appearance of a new customer or a modification of an existing profile);
- **Proactivity.** Telco Agent does not respond only to the excitations from its environment, but takes the initiative in accordance with its tasks (e.g., customer social network analysis prior a new telecom service is introduced or customer clustering as preparation for providing group-oriented service).

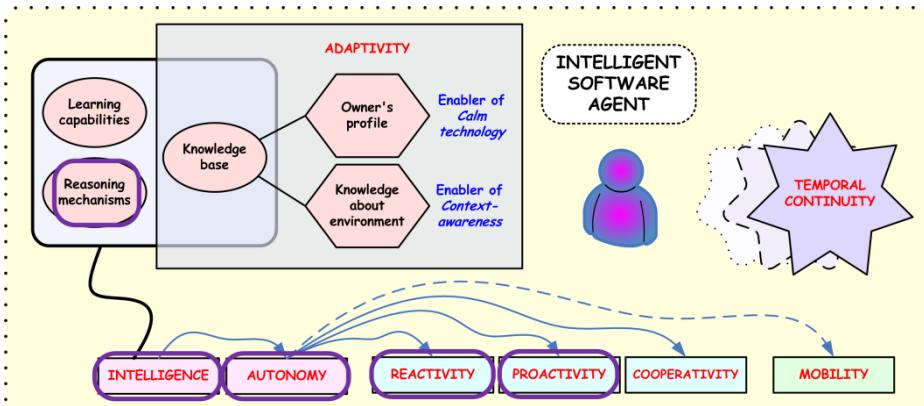


Fig. 2. Telco Agent – the model

### 3.1 Formal Description

Telco  $t$  is represented with its Telco Agent  $\alpha_t$ , which is defined as follows:

$$\alpha_t = (\mathcal{A}_{\mathcal{K}_t}, \text{us}\ddot{\text{s}}\text{p}, \text{id}\ddot{\text{m}}), \quad (1)$$

where  $\mathcal{A}_{\mathcal{K}_t}$  represents a set of Customer Agents with whose owners telco  $t$  has a business relationship, while  $\text{us}\ddot{\text{s}}\text{p}$  and  $\text{id}\ddot{\text{m}}$  represent two mechanisms implemented by a telco: mechanism for semantic matchmaking of customer profiles and mechanism for creation of customer social network, respectively.

### 3.2 Semantic Matchmaking of Customer Profiles

The Semantic Web [11] allows organisation of knowledge into conceptual spaces according to meaning, and replacement of keyword-based searches by semantic query answering [12]. Semantic Web languages, such as *Resource Data Framework (RDF)*, *RDF Schema (RDFS)* and the *Web Ontology Language (OWL)*, can be used to maintain detailed customer profiles. With the help of various query languages, based on *Structured Query Language (SQL)* syntax, it is possible to perform semantic profile matchmaking once the profiles have been created according to a certain standard (e.g., ontology whose part is shown in Fig. 3). Namely, our implementation uses RDQL (*RDF Data Query Language*) and SeRQL (*Sesame RDF Query Language*) queries for extracting knowledge stored in the OWL-upgraded [13] Sesame repository [14]. The mechanism for semantic matchmaking of customer profiles is defined as:

$$\text{us}\ddot{\text{s}}\text{p}(\mathcal{P}_{k_i}, \mathcal{P}_{k_j}) : \mathcal{P} \times \mathcal{P} \rightarrow [0,1], \quad (2)$$

where  $\mathcal{P}$  denotes the set of all customer profiles, while  $\mathcal{P}_{k_i}$  and  $\mathcal{P}_{k_j}$  being two elements of that set (profile  $\mathcal{P}_{k_i}$  describes characteristics of customer  $k_i$ , which is represented by agent  $\alpha_{k_i} \subseteq \mathcal{A}_{\mathcal{K}_t}$ ). The matchmaking result  $\text{us}\ddot{\text{s}}\text{p}(\mathcal{P}_{k_i}, \mathcal{P}_{k_j}) = 0$  means that  $\mathcal{P}_{k_i}$  and  $\mathcal{P}_{k_j}$  are completely different. The higher matchmaking result means that  $\mathcal{P}_{k_i}$  and  $\mathcal{P}_{k_j}$  are more similar, while the maximal result  $\text{us}\ddot{\text{s}}\text{p}(\mathcal{P}_{k_i}, \mathcal{P}_{k_j}) = 1$  means that  $\mathcal{P}_{k_i}$  and  $\mathcal{P}_{k_j}$  are identical.

While creating customer profile we extended the *Composite Capabilities/Preferences Profile (CC/PP)* and *User Agent Profile (UAPerf)* specifications and mapped them to the OWL ontology (whose part describing customer device is shown in Fig. 3) [15]. An example profile (Tab. 1) consists of five different types of customer information:

- Every profile is an instance of a *UserDeviceProfile*. The profile is further described with 20 attributes, as follows;
- Five attributes defining the customer device *hardware*;
- Three attributes defining the customer device *software*;
- Six attributes defining customer *preferences*;
- Six attributes defining the customer *context*.

**Table 1.** An example of customer profile matchmaking

	Attribute	Type	$p_{k_i}$	$p_{k_j}$	Score	Weight
1	<b>ProfileClass</b>	class	MobilePhoneProfile	LaptopProfile	0.250	0.3
2	AvailableMemory	integer	18000	1000000	0.018	<i>Hardware</i>
3	HorizontalResolution	integer	180	1600	0.113	
4	VerticalResolution	integer	230	1050	0.219	
5	BitsPerPixel	integer	16	32	0.500	
6	Imei	string	35461002-303538-0-34		0.000	
7	Os	instance	BasicOs	WindowsVista	0.500	<i>Software</i>
8	Browser	instance	SonyEricssonBrowser	MozillaFirefox	0.500	
9	JavaVersion	integer	15	16	0.940	
10	InformationType	instance	PlainText	Avi	0.333	<i>Preferences</i>
11	InformationService	instance	CroatiaPoliticsInstance	MoviesInstance	0.333	
12	Language	instance	English	Hrvatski	0.500	
13	Genre	instance	RockMusic	ThrillerMovie	0.333	
14	QoS	instance	Silver	Gold	0.500	
15	DeliveryType	instance	NonStreaming	Streaming	0.500	
16	Environment	instance	InnerSpace	InnerSpace	1.000	<i>Context</i>
17	Location	instance	Ina	TrgBanaJelacica	0.250	
18	CoordinatesX	float	50.21459	50.21779	0.990	
19	CoordinatesY	float	48.21344	48.74144	0.990	
20	Time	instance	Night	Night	1.000	
21	SocialActivity	instance	WritingPresentation	CoffeDrinking	0.250	
			<i>Profile similarity</i>		0.427	

In our implementation we do not use standard distance metrics to compare profiles, but rather take advantage of an approach based on semantic matchmaking. Profile class and each attribute in the profile are asserted individually, while the final result is the weighted arithmetic mean of individual scores (in an example of customer profile matchmaking presented by Tab. 1 all attribute weights are set to same value and attribute comparison is worth 70% of final profile similarity, while profile class similarity is worth remaining 30% of final profile similarity). The pseudo-code of semantic matchmaking procedure is given in List. 1. It can be noted from List. 1 that, from the matchmaking perspective, there exist two groups of attributes:

- *Primitive attribute types (float, integer and string).* A comparison result of two ordinal attribute values (float and integer) is a ratio between the smaller and the greater value – e.g., when comparing VerticalResolution (line 4 in Tab. 1) the similarity score is  $230/1050 = 0.219$ . On the other hand, a comparison result of two strings can be *true* or *false* (i.e., 1 or 0; line 6 in Tab. 1 – comparison result is 0 because laptops do not have *International Mobile Equipment Identity (IMEI)* number). The pseudo-code for calculating semantic similarity between ordinal attributes is given in List. 2.;

- *Attributes defined by ontology.* Each profile is an instance of a certain class from the ontology shown in Fig. 3 (line 1 in Tab. 1). Additionally, some attributes can also be defined by ontology rather than by primitive type (e.g., `DeliveryType` from line 15 in Tab. 1). Fig. 3 describes how class hierarchy position is transformed into a real number that represents the similarity between two classes – greater distance between two classes implies smaller similarity. We can see that the `LaptopProfile` and `MobilePhoneProfile` classes are four “steps” away from each other in the hierarchy. The similarity score is calculated by division of 1 by the number steps (in this case 4, so the similarity score equals  $1/4 = 0.25$ ). The pseudo-code for calculating semantic similarity between classes is given in List. 3.

```

Input:  $p_{k_i}, p_{k_j} \in \mathcal{P}$ ,  $profileClassWeight \in [0, 1]$ ,  $attributeWeights \in List[float]$ 
Output:  $u\ddot{s}p(p_{k_i}, p_{k_j}) = finalSimilarity \in [0, 1]$ 
Function calculateProfileSim( $p_{k_i}, p_{k_j}$ ,  $classWeight$ ,  $attributeWeights$ )
DO
  Query findAttributespki = SELECT ?x WHERE (pki, <is>, ?x);
  Query findAttributespkj = SELECT ?x WHERE (pkj, <is>, ?x);
  List attributespki = performQuery(findAttributespki);
  List attributespkj = performQuery(findAttributespkj);
  Float allAttributesSimilarity = 0;
  Float weightSum = 0;
  Class classpki = pki.attributeValue(attributespki.next());
  Class classpkj = pkj.attributeValue(attributespkj.next());
  Float profileClassSimilarity = calculateClassSim(classpki, classpkj);
REPEAT
  Attribute currentAttribute = attributespki.next();
  IF(attributespkj.contains(currentAttribute))
    Value attributeValuepki = pki.attributeValue(currentAttribute);
    Value attributeValuepkj = pkj.attributeValue(currentAttribute);
    Float attributeWeight = attributeWeights.getWeight(currentAttribute);
    Query findAttributeType = SELECT ?x WHERE (attribute, <rdf:type>, ?x);
    Float attributeSimilarity = 0;
    Type attributeType = performQuery(findAttributeType);
    IF(attributeType == string OR integer OR float)
      attributeSimilarity = calculatePrimitiveSim(attributeValuepki, attributeValuepkj, attributeType);
    ELSE IF(attributeType == object)
      Query findAttributeClasspki = SELECT ?x WHERE (attributeValuepki, <serql:directType>, ?x);
      Query findAttributeClasspkj = SELECT ?x WHERE (attributeValuepkj, <serql:directType>, ?x);
      Class attributeClasspki = performQuery(attributeClasspki);
      Class attributeClasspkj = performQuery(attributeClasspkj);
      attributeSimilarity = calculateClassSim(attributeClasspki, attributeClasspkj);
      allAttributesSimilarity += attributeSimilarity * attributeWeight;
      weightSum += attributeWeight;
    UNTIL(attributespki.next() == null);
Float profileClassSimilarityWeighted = profileClassSimilarity * profileClassWeight;
Float allAttributesSimilarityWeighted = (allAttributesSimilarity/weightSum) * (1 - profileClassWeight);
Float finalSimilarity = profileClassSimilarityWeighted + allAttributesSimilarityWeighted;
RETURN finalSimilarity;

```

**Listing 1.** Pseudo-code for semantic matchmaking of customer profiles

```

Input: attributeValuepki, attributeValuepkj ∈ attributeType
Output: primitiveSim ∈ [0, 1]
Function calculatePrimitiveSim(attributeValuepki, attributeValuepkj, attributeType)
DO
  Float primitiveSim = 0;
  IF(attributeType == string)
    primitiveSim = (attributeValuepki == attributeValuepkj);
  ELSE IF(attributeType == (integer OR float))
    IF(attributeValuepki > attributeValuepkj)
      primitiveSim = attributeValuepkj / attributeValuepki;
    ELSE primitiveSim = attributeValuepki / attributeValuepkj;
  RETURN primitiveSim;

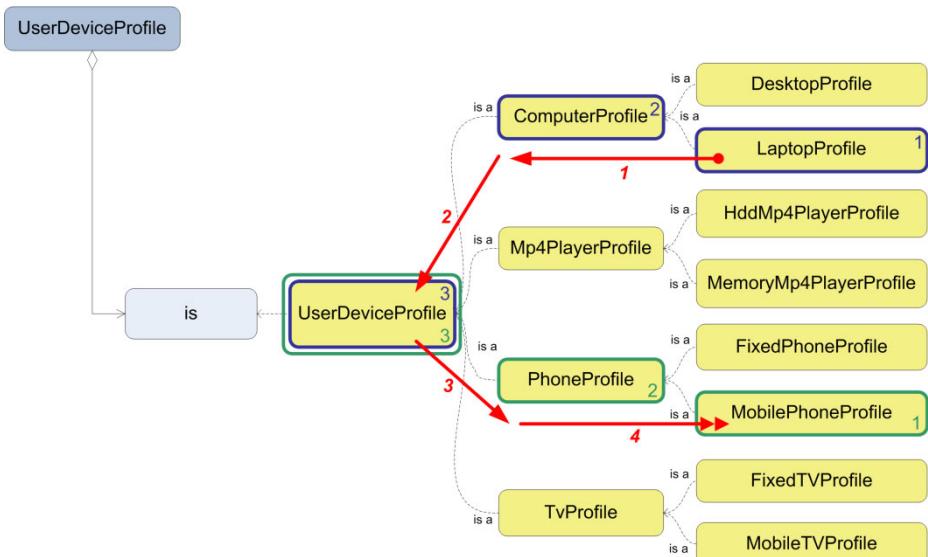
```

**Listing 2.** Pseudo-code for calculating semantic similarity between primitive attributes

```

Input: classi, classj ∈ Class
Output: classSim ∈ [0, 1]
Function calculateClassSim(classi, classj, distanceAntiWeight)
DO
  Query findSuperClassesi = SELECT ?x WHERE (classi, <rdf:type>, ?x);
  Query findSuperClassesj = SELECT ?x WHERE (classj, <rdf:type>, ?x);
  List superClassesi = performQuery(findSuperClassesi);
  List superClassesj = performQuery(findSuperClassesj);
  Integer identicalSuperClasses = findIdenticalClasses(superClassesi, superClassesj);
  Integer classDistance = size(superClassesi) + size(superClassesj) - 2 * identicalSuperClasses;
  Float classSim = 1 / classDistance;
RETURN classSim;

```

**Listing 3.** Pseudo-code for calculating semantic similarity between classes**Fig. 3.** The part of ontology used for creation of customer profiles

### 3.3 Creation of Customer Social Network

The mechanism for building a customer social network is formally defined as a function that takes the matrix of mutual similarities of all customer profiles (there are  $|\mathcal{K}_t|$  such customers, causing similarity matrix dimensions of  $|\mathcal{K}_t| \times |\mathcal{K}_t|$ ) as an argument. The output from the function is graph  $\mathcal{G}_{dm_t}$ :

$$\text{idm}(\text{mat}_{u\ddot{s}p_{p_{k_{|\mathcal{K}_t|}}}}) : \begin{bmatrix} u\ddot{s}p(p_{k_1}, p_{k_1}) & \cdots & u\ddot{s}p(p_{k_1}, p_{k_{|\mathcal{K}_t|}}) \\ \vdots & \ddots & \vdots \\ u\ddot{s}p(p_{k_{|\mathcal{K}_t|}}, p_{k_1}) & \cdots & u\ddot{s}p(p_{k_{|\mathcal{K}_t|}}, p_{k_{|\mathcal{K}_t|}}) \end{bmatrix} \rightarrow \mathcal{G}_{dm_t}. \quad (3)$$

The graph  $\mathcal{G}_{dm_t}$  represents customer social network and is defined as follows:

$$\mathcal{G}_{dm_t} = (\mathcal{K}_t, \mathcal{E}), \quad (4)$$

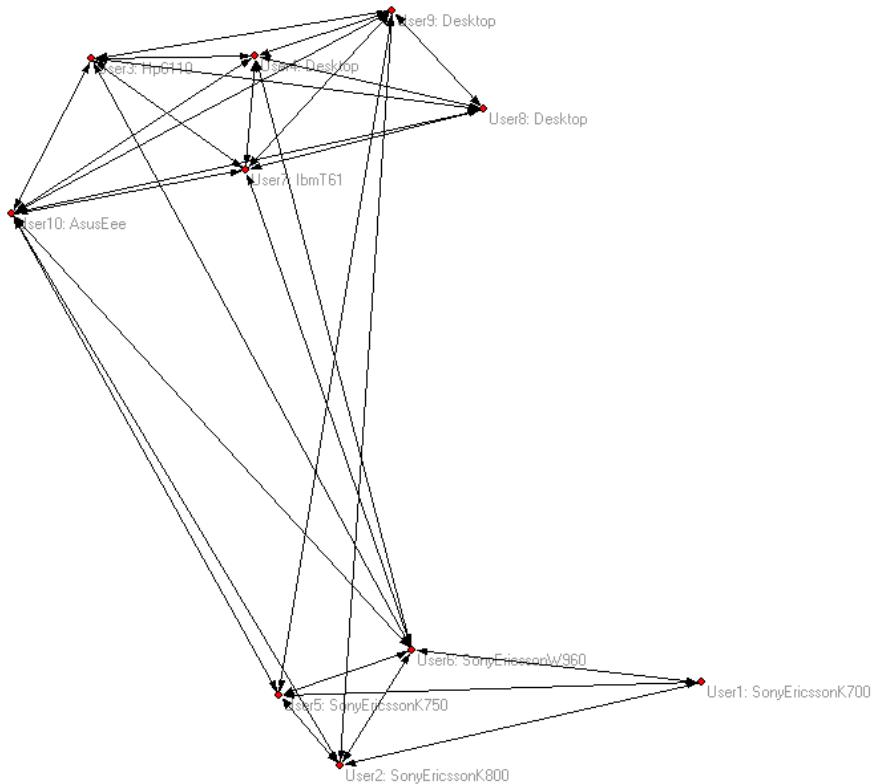
what means that graph has  $|\mathcal{K}_t|$  vertices (one vertex for every customer) and vertices are connected with edges (a set of all edges is denoted with  $\mathcal{E}$ ). The edge weight denotes the connection strength between customers this edge is linking in created social network. From the For. (3) it can be noted that the edge weight between vertices  $k_i$  and  $k_j$  is calculated as  $u\ddot{s}p(p_{k_i}, p_{k_j})$ . This means that the stronger connection strength between customers is consequence of higher similarity between corresponding customer profiles, and vice-versa.

**Table 2.** An example of similarity matrix (10 customer profiles)

	$k_1$	$k_2$	$k_3$	$k_4$	$k_5$	$k_6$	$k_7$	$k_8$	$k_9$	$k_{10}$
$k_1$	1.000	0.786	0.775	0.765	0.465	0.451	0.456	0.447	0.461	0.486
$k_2$	0.786	1.000	0.917	0.850	0.470	0.477	0.526	0.456	0.487	0.520
$k_3$	0.775	0.917	1.000	0.907	0.477	0.481	0.529	0.496	0.493	0.522
$k_4$	0.765	0.850	0.907	1.000	0.512	0.482	0.499	0.502	0.510	0.538
$k_5$	0.465	0.470	0.477	0.512	1.000	0.843	0.837	0.835	0.868	0.800
$k_6$	0.451	0.477	0.481	0.482	0.843	1.000	0.839	0.810	0.846	0.737
$k_7$	0.456	0.526	0.529	0.499	0.837	0.839	1.000	0.819	0.808	0.785
$k_8$	0.447	0.456	0.496	0.502	0.835	0.810	0.819	1.000	0.864	0.816
$k_9$	0.461	0.487	0.493	0.510	0.868	0.846	0.808	0.864	1.000	0.822
$k_{10}$	0.486	0.520	0.522	0.538	0.800	0.737	0.785	0.816	0.822	1.000

For calculating 2D-coordinates of vertices from similarity matrix we use multidimensional scaling mechanism. An example of similarity matrix (10 customer profiles is compared) is given in Tab. 2, while Fig. 4 shows corresponding customer social network (edges with small weights are removed). Additionally, from Tab. 2 it can be noted that matchmaking of identical profiles gives the result equal to 1 (i.e.,  $u\ddot{s}p(p_{k_i}, p_{k_i}) = 1$ ), as well as that matchmaking function is symmetric (i.e.,  $u\ddot{s}p(p_{k_i}, p_{k_j}) = u\ddot{s}p(p_{k_j}, p_{k_i})$ ).

Telco can use a created customer social network to cluster its customers into the groups of similar characteristics and enable provisioning of *group-oriented services*.



**Fig. 4.** An example of customer social network (10 customer profiles)

An example of group-oriented service is *energy-efficient collaborative downloading* [16]. The main idea behind this service is that mobile customers, who are interested in the same content, download some parts of that content directly from a service server and other parts afterwards locally exchange among themselves to reduce overall energy consumption.

## 4 Conclusion and Future Work

This paper presents agent-based solution for enabling paradigm shift from CRM (*Customer Relationship Management*) to CMR (*Customer-Managed Relationship*). We have demonstrated how Telco Agent can make use of telco's knowledge about its customers and build customer social network, a structure that facilitates introduction of CMR paradigm into telco operations. Once telco created customer social network, it can apply various SNA (*Social Network Analysis*) methods for enhancing its service offering and provisioning by means of CMR.

Future work will include study of clustering algorithms and distance metrics, as well as processes suitable for partitioning a set of customers, which are described by a set of attributes included in the profile, into clusters.

**Acknowledgement.** The authors acknowledge the support of research project “Content Delivery and Mobility of Users and Services in New Generation Networks” (036-0362027-1639), funded by the Ministry of Science, Education and Sports of the Republic of Croatia.

## References

1. Moutinho, L.: Futurecast in Consumer (Mis)behaviour. In: Proceedings of 10th International Conference on Telecommunications (ConTEL 2009), Zagreb (Croatia), p. 5 (2009)
2. Podobnik, V., Lovrek, I.: Multi-Agent System for Automation of B2C Processes in the Future Internet. In: 27th IEEE Conference on Computer Communications (INFOCOM) Workshops, pp. 1–4. IEEE Press, Phoenix (2008)
3. Podobnik, V., Petric, A., Trzec, K., Galetic, V., Jezic, G.: Agent-based Provisioning of Group-oriented Non-linear Telecommunication Services. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) ICCCI 2009. LNCS, vol. 5796, pp. 198–204. Springer, Heidelberg (2009)
4. Yoon, J.-L.: Telco 2.0: A New Role and Business Model. Communications 45(1), 10–12 (2007)
5. Svensson, M., Soderberg, J.: Machine-learning technologies in telecommunications. Ericsson Review 2008(3), 29–33 (2008)
6. Bonnin, J., Lassoued, I., Hamouda, Z.B.: Automatic multi-interface management through profile handling. Mobile Networks and Applications 14(1), 4–17 (2009)
7. Panayiotou, C., Samaras, G.: mPERSONA: personalized portals for the wireless user: An agent approach. Mobile Networks and Applications 9(6), 663–677 (2004)
8. Frkovic, F., Podobnik, V., Trzec, K., Jezic, G.: Agent-Based User Personalization Using Context-Aware Semantic Reasoning. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part I. LNCS (LNAI), vol. 5177, pp. 166–173. Springer, Heidelberg (2008)
9. Podobnik, V., Matijasevic, M., Lovrek, I., Skorin-Kapov, L., Desic, S.: Agent-based Framework for Personalized Service Provisioning in Converged IP Networks. In: Kowalczyk, R., Vo, Q.B., Maamar, Z., Huhns, M. (eds.) SOCASE 2009. LNCS, vol. 5907, pp. 83–94. Springer, Heidelberg (2009)
10. Jennings, N., Sycara, K., Wooldridge, M.: A Roadmap of Agent Research and Development. Journal of Autonomous Agents and Multi-Agent Systems 1(1), 7–36 (1998)
11. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web, vol. 284(5), pp. 34–43. Scientific American (2001)
12. Antoniou, G., van Harmelen, F.: Semantic Web Primer. MIT Press, Cambridge (2004)
13. Kiryakov, A., Ognyanov, D., Manov, D.: OWLIM - A Pragmatic Semantic Repository for OWL. In: Dean, M., et al. (eds.) WISE 2005 Workshops. LNCS, vol. 3807, pp. 182–192. Springer, Heidelberg (2005)
14. Broekstra, J., Kampman, A., van Harmelen, F.: Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. In: Horrocks, I., Hendler, J. (eds.) ISWC 2002. LNCS, vol. 2342, pp. 54–68. Springer, Heidelberg (2002)
15. Vrdoljak, L., Bojic, I., Podobnik, V., Kusek, M.: The AMiGO-Mob: Agent-based Middleware for Group-oriented Mobile Service Provisioning. In: Proceedings of the 10th International Conference on Telecommunications (ConTEL 2009), Zagreb (Croatia), pp. 97–104 (2009)
16. Bojic, I., Podobnik, V., Kusek, M.: Agent-enabled Collaborative Downloading: Towards Energy-efficient Provisioning of Group-oriented Services. In: Jędrzejowicz, P., Nguyen, N.T., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2010. LNCS, vol. 6071, pp. 62–71. Springer, Heidelberg (2010)

# Multi-attribute Auction Model for Agent-Based Content Trading in Telecom Markets

Ana Petric and Gordan Jezic

University of Zagreb  
Faculty of Electrical Engineering and Computing  
Unska 3, HR-10000, Zagreb, Croatia  
[{ana.petric,gordan.jezic}@fer.hr](mailto:{ana.petric,gordan.jezic}@fer.hr)

**Abstract.** The advent of the Internet and the development of the Next Generation Network (NGN) have enabled the development of value added services (VAS), while operators' investments in licenses and their desire to stay competitive on the market have triggered the development. When forming VAS, special attention needs to be paid to the purchase of resources (e.g., transport capacity and content) needed for the service creation. As the number of participants on the telecom market increases, the need for automation of transactions carried between them arises. In this paper, we identify stakeholders on the telecom content e-market and propose an appropriate model which captures their transactions. Since content is not a commodity we propose a multi-attribute auction model for content trading which prevents sellers from manipulating the auction outcome by offering unnecessarily high values of some (often less important) attributes in order to compensate for unreasonably low values of other (more important) ones. A multi-agent system which uses the multi-attribute auction model as a negotiation protocol is presented and an illustrative example of content trading in telecom markets is provided.

**Keywords:** content trading, B2B telecom e-market, multi-attribute auctions, multi-agent system.

## 1 Introduction

Provisioning of basic telecommunication services (i.e., fixed and mobile communication, data transfer) is no longer enough to keep existing customers, let alone attract new ones [1] so telecom operators are pursuing innovations and launching new value-added services (VAS) [2] in order to increase revenue. There are two types of resources needed for the creation of VAS. They are the information resources (i.e., content) the service is based on and the transport capacities needed for service provisioning [3]. The term content encompasses movies, songs, news, images and text, in other words data and information within various fields [4]. The Next Generation Network (NGN) brings its own new added value into the market and one of these added values is multimedia content composed of several types of content (e.g., audio, video, data...) [5].

The telecom market is divided into two submarkets, the B2B (Business-to-Business) market and the B2C (Business-to-Consumer) market. Telecom operators buy resources on the B2B market [6], from those resources they create VAS which are then sold to users on the B2C market [7]. The research problem addressed in this paper concerns the automation of business processes related to content trading on the B2B telecom electronic market (e-market) by using multi-attribute auctions.

The paper is organized as follows. Section 2 identifies stakeholders on the telecom content e-market and presents the phases of the proposed Content e-Trading Transaction Model. Section 3 presents a Multi-Attribute Auction Model used during the content trading negotiation process. Section 4 illustrates the use of the Multi-Attribute Auction Model and compares it with a few other multi-attribute decision making approaches, while Section 5 concludes the paper and gives an outline for future work.

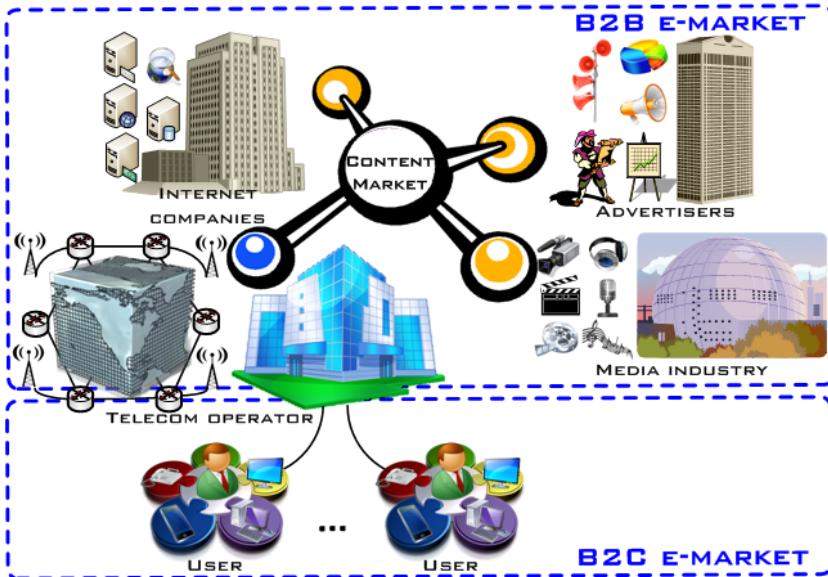
## 2 Telecom e-Market

The telecom content e-market includes participants from the media, Internet, advertising and telecom world [8,9]. *Media companies* provide professionally produced content (e.g., music videos, movies, TV shows) which is used to create VAS. New business models enable *advertisers* to sponsor content and receive valuable feedback from the users. Also, since mobile phones carry diverse context information about their owners, the opportunity for target advertising arises. *Internet companies* (e.g., search engines) help users to find potentially interesting content. They also have the possibility to create a new generation of applications which use context information from users' mobile phones. *Telecom operators* generate new revenue streams by increasing the number of VAS which they offer to their users and at the same time they increase the traffic going through their network. *Users* profit from a greater selection of VAS and from the rise of payment opportunities for the use of VAS. So we can say that on the telecom content e-market, shown on Figure 1, all participants are on the win.

The users buy content packed in VAS from Telecom operators on the *B2C e-market* while all the other previously mentioned companies do business on the *B2B e-market*. The proliferation of e-business and the dynamic nature of business transactions conducted on the Internet, argues for the development of intelligent trading agents which act on behalf of human traders (i.e., buyers and sellers) [10,11,12]. Intelligent trading agents can also be used to impersonate stakeholders in the environment of the NGN in order to enable automated interactions and business transactions on the telecom markets [7].

### 2.1 CeTT Model

The CeTT (Content e-Trading Transaction) Model systematically analyses process of content trading on the telecom content e-market. It was made by adjusting the BBT (Business-to-Business Transaction) Model [13], CBB (Consumer Buying Behaviour) Model [13] and the Sourcing Process [14] to the specifics of the



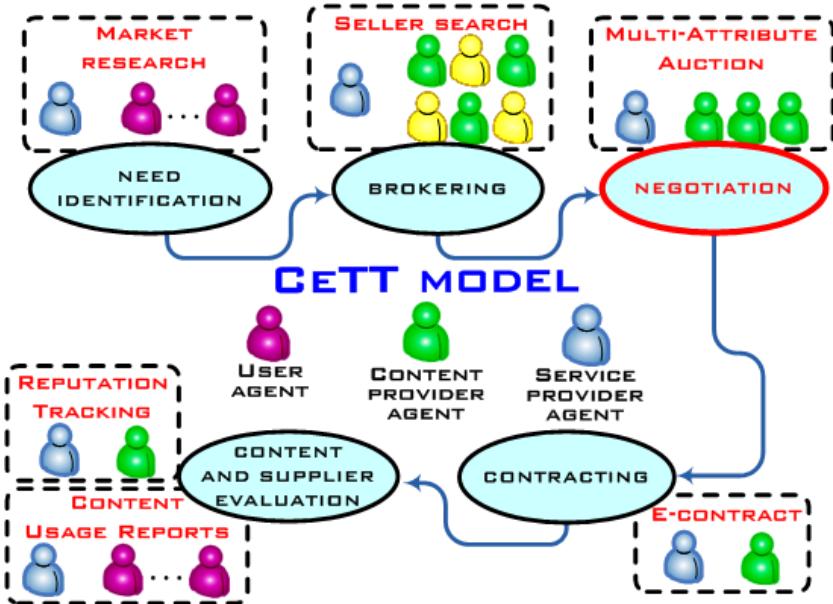
**Fig. 1.** Telecom content market

telecom content e-market. We can formally identify five steps which must be executed in order to successfully complete one content trading transaction on the B2B e-market. These steps are as follows (Figure 2): 1) need identification, 2) brokering, 3) negotiation, 4) contracting, and 5) content and supplier evaluation.

The goal of the *need identification* phase is to specify the type and the appropriate values of the content that the telecom operator would like to purchase. Both the type and the parameters are determined based on the history of users' content consumption as well as on market research concerning latest releases of new and popular content. In the CeTT Model the *Service Provider Agent (SPA)* which represents the telecom operator's department for creating and providing VAS, contacts various *User Agents (UAs)* which represent operator's users in order to find out their preferences. The SPA also searches through operator's database with past transactions and tries to predict which kind of content would users like.

The main role of the *brokering* phase is to match the SPA with content providers that sell the type of content needed for the creation of a new service or upgrades of an old one. The SPA searches the market and identifies a group of potential business partners which are represented by their *Content Provider Agents (CPAs)*.

When studying B2B e-markets, a special attention is paid to the *negotiation* phase since the outcome (i.e. financial efficiency) is still the premier performance measure for most businesses [15]. Negotiation is a process which tries to reach an agreement regarding one or more content attributes (e.g., price, quality, etc.). Each stakeholder in the negotiation process is represented by an intelligent trading agent



**Fig. 2.** Content e-Trading Transaction Model

that negotiates in his behalf (e.g., SPA trades in behalf of telecom operator) [10]. We use the Multi-Attribute Auction Model described in Section 3 for content trading in the CeTT Model since multiple issues need to be settled and the parties involved have different preferences towards these issues.

In the *contracting* phase the negotiated terms are put into legally binding electronic contract [16] and the conditions, clauses and activity sets that satisfy those negotiated terms are specified [17]. The contracting phase of the CeTT Model starts with the contract preparation activities which are then followed by the contract fulfilment process which includes contract execution and execution monitoring. The SPA and the winning CPA agree on the content delivery terms, payment deadlines and penalties in case that one of the parties does not respect the negotiated terms. Later, the SPA checks if the parameters of the delivered content match the negotiated ones and are all deadlines met.

In the *content and supplier evaluation* phase the SPA uses the information from the monitoring part of the contracting phase in order to calculate CPA's reputation based on his fulfilment of the negotiated terms [18]. Content is evaluated after a certain period by tracking its popularity with telecom operator's users. UAs report back to SPA with the latest users content consumption information.

### 3 A Multi-attribute Auction Model for Content Trading

Due to their well defined protocols, auctions are suitable enablers of negotiations in e-markets and as such will be used in the negotiation phase of the CeTT

Model. Participants in the auction (i.e., telecom operator and media companies) will be represented by their agents (i.e., SPA and CPAs). Complex items (i.e., content) often require negotiation of several attributes, and not just the price [19]. They are sold in multi-attribute auctions [20] which are a special case of procurement auctions. Procurement auctions are also called reverse auctions since there are multiple sellers and only one buyer that purchases items.

Existing models of multi-attribute auctions use different approaches to determine the winning offer (e.g., by defining various utility functions [20][21][22], by using fuzzy multi-attribute decision making algorithms [23], by introducing pricing functions and preference relations for determining acceptable offers [24], by calculating the ratio of deviation from the ideal offer and the deviation from the anti-ideal offer [25]).

The prerequisite for conducting the multi-attribute auction is for the SPA to specify the preferences of the content he wishes to purchase. This step is conducted in the need identification phase of the CeTT Model. Preferences are usually defined in the form of a scoring function based on the SPA's utility function [22]. The SPA sends a request to all CPAs identified as potential business partners in the brokering phase of the CeTT Model. The CPAs then reply by sending bids. The winner of the multi-attribute auction is the CPA that provided the highest overall utility for the SPA. Our model is based on reverse auctions and takes into account the price, as well as other non-monetary attributes of the purchased content.

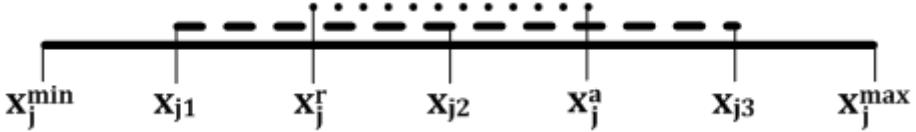
A multi-attribute auction can be defined as a tuple  $\langle b, S, t \rangle$ , where

- $b$  is the buyer agent (i.e., SPA);
- $S$  (of size  $s$ ) denotes the set of all seller agents (i.e., CPAs) that participate in buyer  $b$ 's multi-attribute auction;
- $t : \mathbb{R}^s \rightarrow \mathbb{R}$  is the winner determination function.

The winner determination function ranks CPAs' offers based on the values assigned to them and determines the auction outcome. A description of all negotiable content attributes as well as the functions that assign values to CPAs' offers are defined in the *content evaluation model* which is represented with a tuple  $\langle x, w, U, d_p \rangle$ , where

- $x = \{x_1, \dots, x_j, \dots, x_n\}$  is the set of attributes used to describe the content; each attribute  $j$  has a reserve and aspiration value, denoted as  $x_j^r$  and  $x_j^a$ , respectively, determined by the SPA;
- $w = \{w_1, \dots, w_j, \dots, w_n\}$  is a set of weights that determines the importance of each attribute from  $x$  for the SPA, where  $w_j$  is the weight of attribute  $j$ ;
- $U : \mathbb{R}^{s \times n} \times \mathbb{R}^n \rightarrow \mathbb{R}^s$  is a utility function that calculates the SPA's utility of CPAs' offers;
- $d_p : \mathbb{R}^{s \times n} \times \mathbb{R}^n \rightarrow \mathbb{R}^s$  is a deviation function that calculates the SPA's positive deviation of CPAs' offers.

As shown in Figure 3, there are several relevant values of an attribute that can be used to determine SPA's utility for that attribute. An attribute  $j$  has the lowest

**Fig. 3.** Attribute values

and highest possible value,  $x_j^{min}$  and  $x_j^{max}$ , respectively. The CPAs place offers between those values (e.g.,  $x_{j1}, x_{j2}, x_{j3}$ ). Reserve value  $x_j^r$  marks the lowest value of an attribute  $j$  that is acceptable to the SPA while the aspiration value  $x_j^a$  is the highest value of an attribute  $j$  that the SPA is interested in. An offer with at least one attribute value worse than the reserve value (e.g.,  $x_{j1}$ ) is disqualified. In a single attribute auction, a value better than the aspiration value (e.g.,  $x_{j3}$ ) is usually accepted since it clearly brings additional benefit to the SPA (e.g., lower price than the one that the SPA was ready to pay presents clear savings for the SPA). To the best of our knowledge, the existing multi-attribute auction models do not distinguish between  $x_j^a$  and  $x_j^{max}$  and accordingly do not consider the situation where  $x_j^a < x_j \leq x_j^{max}$ .

The utility function  $U(x)$  was designed in such a manner to prevent CPAs from significantly increasing the total utility of their offers by assigning unnecessarily high values to some (often less important) attributes in order to compensate for unreasonably low values of other (more important) ones. However, the additional benefit that the SPA gets from the value higher than the aspiration value  $x_j^a$  is not completely ignored. After determining the utility of an offer the winner determination function  $t$  also takes into account additional gain that the SPA obtains from each offer before declaring the winner of the auction.

Utility function  $U(x_i)$  takes as input an offer  $x_i$  placed by  $CPA_i$  and, together with the set of weights  $w$  maps it to a real value. Function  $U(x_i)$  can be defined as an additive scoring function that assumes the existence of mutual *preferential independence* between attributes [22]. In order to calculate the utility of offered content, reserve values and weights for each attribute need to be considered [21]. Function  $U(x_i)$  is defined as follows:

$$U(x_i) = \sum_{j=1}^n w_j U(x_{ij}), \text{ where } \sum_{j=1}^n w_j = 1 \quad (1)$$

$$U(x_{ij}) = \begin{cases} \frac{x_{ij} - x_j^r}{x_j^a - x_j^r}, & x_j^r \neq x_j^a \text{ and } x_j^r \leq x_{ij} < x_j^a \\ N.A., & x_{ij} < x_j^r \\ 1, & x_{ij} \leq x_j^a \end{cases} \quad (2)$$

In our model,  $U(x_{ij})$  depends on the reserve and aspiration values,  $x_j^r$  and  $x_j^a$ , respectively, that the SPA defines for each attribute  $j$ . Value N.A. in Equation (2) marks a non-acceptable value for an attribute, i.e., it is worse than the reserve value  $x_j^r$ . An offer is rejected if the utility of at least one attribute is

N.A. Values offered higher than the aspiration value are acceptable, but their utility cannot be higher than 1. Positive deviation function  $d_{p,i}$  compares an offer  $x_i$  placed by  $CPA_i$  with the aspiration offer  $x^a = (x_1^a, \dots, x_j^a, \dots, x_n^a)$  and maps the comparisons to a real value. Function  $d_{p,i}$  is defined as follows:

$$d_{p,i} = \sqrt{\sum_{j=1}^n d_{p,ij}^2}, \quad d_{p,ij} = \begin{cases} w_{ij} \frac{x_{ij} - x_j^a}{x_j^{max} - x_j^a}, & \text{if } x_{ij} > x_j^a \text{ and } x_j^a \neq x_j^{max} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

In our model, the  $d_{p,i}$  depends on the aspiration and highest possible attribute values,  $x_j^a$  and  $x_j^{max}$ , respectively. It calculates the SPA's additional benefit as the positive deviation of an offer  $x_i$  from the aspiration offer  $x^a$ . The function  $d_{p,i}$  takes into account only attributes with values higher than the values of the aspiration offer  $x^a$ . An offer with  $d_{p,i} > 0$  brings additional benefit to the SPA beyond the utility he expects to get.

The primary objective of the winner determination function is to maximize SPA's utility while the secondary objective is to maximize additional benefits that some offers bring. The problem arises when the objectives are in conflict so the function  $t$  tries to find the compromise between them by setting the weight of additional benefits  $w_{bonus}$  low enough to prevent CPAs with average offers from manipulating the auction outcome in their favour, but at the same time  $w_{bonus}$  should be high enough to reward the CPAs with very good offers which also bring additional benefit to the SPA. The  $t$  is defined as follows:

$$t = \max_i T(i), \quad \text{where } T(i) = w_{bonus} d_{p,i} + (1 - w_{bonus}) U(x_i) \quad (4)$$

## 4 An Illustrative Example

The proposed Multi-Attribute Auction Model for determining the auction winner was implemented and the results are presented in this section. The application domain of the model was content trading on the B2B telecom e-market. One buyer (SPA) and 5 sellers (CPAs) participated in a sealed-bid multi-attribute reverse auction. The SPA needs to buy new and popular songs which he sells to his users as a ringtone or a music video. First, he gathers information from UAs and does some market research. The SPA determines that it already offers the first five songs from the music charts so its aspiration value is set on the 6th place on the charts. It also does not want a song ranked under the 60th place so the reservation value is 60. Since songs from more popular artists are sold more often than the ones from new or less popular artists, when purchasing new content, the SPA also takes into account the popularity of the artist that is performing the song.

After the SPA found CPAs that provide the kind of content it wants to purchase, the negotiation phase begins. The agents negotiated on the following attributes:  $x_1$  - the percent of the profit from each sold ringtone or music video that the SPA will get,  $x_2$  - the current position of the song on the music charts,

**Table 1.** Attribute values and SPA's attribute valuations

	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
minimum value ( $x_j^{min}$ )	0	100	100	1	1
maximum value ( $x_j^{max}$ )	100	1	1	10	100
weight ( $w_j$ )	0.30	0.25	0.20	0.15	0.10
reservation value ( $x_j^r$ )	20	60	60	2.5	30
aspiration value ( $x_j^a$ )	50	6	6	6.5	75

$x_3$  - the popularity index of the singer in the previous year,  $x_4$  - the music reviewers' grade of the song and  $x_5$  - the time period that the SPA has the right to sell the song. The Table 1 contains the minimum (i.e., worst) and maximum (i.e., best) possible attribute values as well as SPA's valuations (i.e., weights), reservation and aspiration values for each attribute. We assume that the song will not be on the music charts (i.e., interesting enough to users for them to buy the ringtone or music video) longer than 100 days (roughly three months) so we set the  $x_5^{max}$  on 100 days even though  $x_5^{max}$  can actually be indefinite.

The Table 2 contains the offers placed by CPAs, utilities and positive deviations of those offers as well as ranking of offers according to our model and previously mentioned winning offer determination approaches ([22][24][25][23]). After a set of experiments we determined that  $w_{bonus} = 0.05$  is low enough to prevent compensation of attribute utilities. From the rankings with other approaches we can see that CPAs were able to compensate the lower utility of a certain attribute with the higher utility of another attribute and win in the auction due to SPA's lack of distinguishing between  $x_j^a$  and  $x_j^{max}$ . Since our Multi-Attribute Auction Model prevents CPAs with average offers from manipulating the auction outcome in their favour we plan to use it in the negotiation phase of the CeTT Model.

**Table 2.** Sellers' offers and offer rankings

	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$U(x)$	$d_p$	$T(i)$	Rank	Rank [22]	Rank [23]	Rank [24]	Rank [25]
$CPA_1$	35	12	10	9.5	95	0.807	0.151	0.775	3	1	2	2	2
$CPA_1$	40	20	12	6.5	80	0.812	0.020	0.773	4	4	4	1	4
$CPA_3$	45	25	14	7.5	100	0.832	0.109	0.796	2	2	1	4	1
$CPA_4$	35	7	6	6.5	85	0.845	0.040	0.805	1	3	5	3	5
$CPA_5$	75	46	33	2.5	100	0.564	0.180	0.546	5	5	4	5	4

## 5 Conclusion

In this paper, we identified stakeholders on the telecom content e-market and proposed a model which captured all stages related to transactions carried out on the telecom content e-market. Phases of the introduced CeTT (Content e-Trading Transaction) Model are described and the roles and tasks of intelligent

software agents in the model are defined. The Multi-Attribute Auction Model which defines a protocol for content trading conducted in the negotiation phase of the CeTT Model is presented. A comparison of the proposed model with several other multi-attribute auction models which use different approaches to determine the winning offer was conducted. An example presented in Section 4 illustrated how, unlike the other models, our Multi-Attribute Auction Model prevents sellers (e.g., Content Provider Agents) with average offers from manipulating the auction outcome in their favour. The model maximizes the buyer's (e.g., Service Provider Agent's) utility of placed offers while taking into account additional benefits that some offers bring and it also discourages sellers from offering unnecessary high values of some attributes with the purpose of compensating for unreasonably low values of the other ones.

For future work, we plan to implement the remaining phases of the CeTT Model and integrate them with the Reputation Tracking Reverse Auction Model [18] used in the content and supplier evaluation phase of the CeTT Model.

**Acknowledgments.** The work presented in this paper was carried out within the research project 036-0362027-1639 "Content Delivery and Mobility of Users and Services in New Generation Networks", supported by the Ministry of Science, Education and Sports of the Republic of Croatia.

## References

1. Olla, P., Patel, N.V.: A value chain model for mobile data service providers. *Telecommunications Policy* 26(9-10), 551–571 (2002)
2. Damsgaard, J., Marchegiani, L.: Like Rome, a mobile operator's empire wasn't built in a day!: a journey through the rise and fall of mobile network operators. In: Janssen, M., Sol, H.G., Wagenaar, R.W. (eds.) *Proceedings of the 6th international conference on Electronic commerce*, pp. 639–648. ACM, New York (2004)
3. Podobnik, V., Petric, A., Trzec, K., Jezic, G.: Software Agents in New Generation Networks: Towards the Automation of Telecom Processes. In: Jain, L.C., Nguyen, N.T. (eds.) *Knowledge Processing and Decision Making in Agent-Based Systems*, pp. 71–99. Springer, Heidelberg (2009)
4. Subramanya, S., Yi, B.K.: Utility Model for On-Demand Digital Content. *Computer* 38, 95–98 (2005)
5. Mampaey, M., Ghys, F., Smouts, M., Vaaraniemi, A.: *3G Multimedia Network Services, Accounting, and User Profiles*. Artech House, Inc., Boston (2003)
6. Trzec, K., Lovrek, I., Mikac, B.: Agent Behaviour in Double Auction Electronic Market for Communication Resources. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) *KES 2006. LNCS (LNAI)*, vol. 4251, pp. 318–325. Springer, Heidelberg (2006)
7. Podobnik, V., Lovrek, I.: Multi-agent system for automation of B2C processes in the future Internet. In: Gracanin, D. (ed.) *IEEE INFOCOM Workshops 2008*, pp. 1–4 (2008)
8. Phillipson, J.: Multimedia is a team sport. *Ericsson Business Review* 3(1), 30–34 (2008)
9. LeClerc, M.: Swimming with the sharks. *Ericsson Business Review* 2(3), 18–22 (2007)

10. Maes, P., Guttman, R.H., Moukas, A.: Agents That Buy and Sell. *Communications of the ACM* 42(3), 81–91 (1999)
11. Fasli, M.: Agent Technology For E-Commerce. John Wiley & Sons, Chichester (2007)
12. Podobnik, V., Petric, A., Jezic, G.: An Agent-Based Solution for Dynamic Supply Chain Management. *Journal of Universal Computer Science* 14(7), 1080–1104 (2008)
13. He, M., Jennings, N.R., Leung, H.F.: On Agent-Mediated Electronic Commerce. *IEEE Transactions on Knowledge and Data Engineering* 15(4), 985–1003 (2003)
14. Gattiker, T.F., Huang, X., Schwarz, J.L.: Negotiation, email, and Internet reverse auctions: How sourcing mechanisms deployed by buyers affect suppliers' trust. *Journal of Operations Management* 25(1), 184–202 (2007)
15. He, S., Cattelan, R.G., Kirovski, D.: Modeling viral economies for digital media. In: Sventek, J.S., Hand, S. (eds.) 3<sup>rd</sup> ACM European Conference on Computer Systems – EuroSys 2008, pp. 149–162. ACM, New York (2008)
16. Angelov, S., Grefen, P.: The 4W framework for B2B e-contracting. *International Journal of Networking and Virtual Organisations* 2(1), 78–97 (2003)
17. Krishna, P.R., Karlapalem, K.: Electronic Contracts. *IEEE Internet Computing* 12, 60–68 (2008)
18. Petric, A., Jezic, G.: Reputation Tracking Procurement Auctions. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) ICCCI 2009. LNCS, vol. 5796, pp. 825–837. Springer, Heidelberg (2009)
19. Do, V.T., Halatchev, M., Neumann, D.: A Context-Based Approach to Support Virtual Enterprises. In: Nguyen, N.T., Kowalczyk, R., Chen, S.M. (eds.) Proceedings of the 33rd Hawaii International Conference on System Sciences, vol. 6, p. 6005. IEEE Computer Society, Los Alamitos (2000)
20. Bichler, M., Kalagnanam, J.: Configurable offers and winner determination in multi-attribute auctions. *European Journal of Operational Research* 160(2), 380–394 (2005)
21. Bui, T., Yen, J., Hu, J., Sankaran, S.: A Multi-Attribute Negotiation Support System with Market Signaling for Electronic Markets. *Group Decision and Negotiation* 10(6), 515–537 (2001)
22. Bichler, M.: An experimental analysis of multi-attribute auctions. *Decision Support Systems* 29(3), 249–268 (2000)
23. Tong, H., Zhang, S.: A Fuzzy Multi-attribute Decision Making Algorithm for Web Services Selection Based on QoS. In: Yuan, H., Xu, Y. (eds.) Proceedings of the 2006 IEEE Asia-Pacific Conference on Services Computing, pp. 51–57. IEEE Computer Society, Los Alamitos (2006)
24. Bellosta, M.J., Brigui, I., Kornman, S., Vanderpooten, D.: A multi-criteria model for electronic auctions. In: Haddad, H., Omicini, A., Wainwright, R.L., Liebrock, L.M. (eds.) Proceedings of the 2004 ACM Symposium on Applied Computing, pp. 759–765. ACM, New York (2004)
25. Cheng, C.B.: Solving a sealed-bid reverse auction problem by multiple-criterion decision-making methods. *Computers and Mathematics with Applications* 56(12), 3261–3274 (2008)

# Applying Possibility and Belief Operators to Conditional Statements

Grzegorz Skorupa and Radosław Katarzyniak

Wrocław University of Technology, Wybrzeże Wyspiańskiego 27  
[{grzegorz.skorupa,radoslaw.katarzyniak}@pwr.wroc.pl](mailto:{grzegorz.skorupa,radoslaw.katarzyniak}@pwr.wroc.pl)

**Abstract.** In this paper we outline a model for grounding of natural language statements being conditional statements extended with modal operators of possibility and belief. This work extends a detailed theory of modal language grounding proposed elsewhere. At first, some comments and intuitive semantics of modal conditional statements are presented. At second, a general structure of formal model, that covers this semantics is proposed. This model refers to the idea of possible worlds semantics. The grounding of modal conditional statements is defined by means of original concepts of epistemic satisfaction relation. The work sets up theoretical basis for further analytic and experimental evaluations.

## 1 Introduction

Communication between agents is modelled to imitate human behaviour. Many of the current trends concentrate on speech act theory [2], dealing with typical human speech behaviours. The theory specifies how agents can inform each other, make commitments, ask questions, make proposals and negotiate. Within this work we are considering some of conditional natural language statements that can be expressed using ‘inform’ performative from speech act theory.

We are considering an cognitive agent [6] that observes some environment and describes currently observed situation using a non-extensional modal language. The language allows agent to form some of natural language statements. Modal character of the language allows the agent to enrich her expression with uncertainty. Hence the agent is able to state that she finds something possible or believes something, but is not sure of it.

We wish the agent to choose these statements in a similar way human would do. Hence the agent should be able to find out which statement is adequate to described situation and her knowledge. We are considering what a human listener typically implies about speakers attitude towards situation. We try to model an agent, so that she considers listeners conclusions.

Simple modal statements and also statements with conjunction and alternative have already been analysed in described context [8]. Within this work we are considering some of conditional statements, that can have modal operators of possibility and belief.

We try to answer what has to be speakers’ attitude towards observed situation in order to be allowed to use a conditional statement. In other words we wish to

know when the speaker agrees that given conditional statement is applicable to particular situation.

Main features of a statement's proper usage are described in section 2. In section 3 we specify considered statements and define a formal language. Conditional statements' meaning is analyzed in section 4. In section 5 we describe and model required abilities, that have to be implemented within an agent to be able to use conditional statements properly. Section 6 contains a method of choosing the right statement based on the proposed model.

## 2 Statement Proper Usage

If one wants to be truthful and precise when speaking he has to choose proper statements that are adequate to situation and consistent with his knowledge.

'Being truthful' means, that one does not lie or mislead the listener. In order to say some statement the speaker has to 'internally' agree with it according to the knowledge he has developed. Since we are considering an agent that describes current situation, the agent has to choose statements that are consistent with what she sees and thinks. This means that the statements truthfulness has to be evaluated based on agent's observations and subjective thoughts about the situation.

'Being precise' means, that one chooses best suited statements according to his knowledge. For example, if agent knows something, she will not tell that she finds it merely possible. If agent knows that the sun is shining, she will not tell: "If it is day now, the sun is shining.". Applying "If it is day now" would force the listener to conclude, that agent does not know whether the sun is shining or not.

This leads to conclusion that adequacy of a statement does not entirely rely on classical logical truth but also on other factors. While interpreting a message, listener implies additional information about the speaker's attitude and knowledge. This attitude doesn't have to be embedded directly in the meaning of words but can rely on the statement structure or even a conversational context. For example if the speaker says: "Some athletes smoke", the listener implies, that the speaker thinks, that not all athletes smoke or at least doesn't know if all of them do. Otherwise he would use the sentence: "All athletes smoke". It is conventional to imply, that there are athletes that do not smoke. Grice introduced a technical term 'implicature' to distinguish what is said from what is meant [3]. Ajdukiewicz, when analysing logical truth table of implication, came to conclusion that one has to distinguish statement's logical truth from its proper usage [1]. According to his work it is true that "If the moon is a piece of cheese, I will die on a day with an even date.", but it is not proper to state such a statement.

In order to use a statement, speaker has to consider conventional implicature and proper usage. This means that some additional constraints have to be met. These constraints have to be checked with agents observations and attitude towards situation, that shall be described. While it is quite clear what observations are, the term attitude is still very fuzzy. By attitude we mean some thoughts

and knowledge used by agent to create a subjective mental representation of situation. We shall call this representation a mental state. A mental state must be autonomously generated and somehow embodied within an agent. It must have some crucial properties that are required to check imposed constraints.

The way of checking the constraints is called ‘grounding’, because it connects a statement to external world by means of agent’s internal world representation. Hence the defined method solves the grounding problem [4][7] for statements of considered structure and meaning.

### 3 Considered Statements

We assume the agent is living in some world and uses a predefined formal language to describe currently observed environment state. The formal language is non-extensible and modal. It allows agent to formulate some of natural language statements. Exemplary statements are of the form:

- I know that the apple is red.
- I believe that Alice is blond or brunette.
- If the cage has a roof, then the bird can fly.
- If Fluffy is a home animal, then it is possible that it is a cat.

The term modal means that agent is able to express some uncertainty by adding special operators. Modal operators are of the form: “I believe that” or “It is possible that” and are called belief and possibility operators, respectively.

Many of the language statement types have already been formally defined and grounded in [8]. Within this work we focus on conditional statements of the form: “If  $o_1$  is  $q_1$ , then  $o_2$  is  $q_2$ .”. Where  $o_1, o_2$  are some objects observed by the agent and  $q_1, q_2$  are properties of these objects.

Conditional statements can be divided into a few groups based on their meaning. The meaning of a conditional statement relies on its grammatical structure and context. For example a second conditional statement: “If he were rich, he would buy a Ferrari” means that described person is not rich and can’t afford a luxurious car. A statement: “If he is rich, he will buy a Ferrari” means that the speaker thinks, that a described person will buy Ferrari, under condition that he has money. This implies, that the speaker does not know whether the described person is rich or not.

There are also many statements related to logical implication, that have a similar, but not the same, meaning. These are the statements of the form: “ $o_2$  is  $q_2$  because  $o_1$  is  $q_1$ ”, “ $o_1$  being  $q_1$  implies  $o_2$  being  $q_2$ ”, “ $o_1$  being  $q_1$  causes  $o_2$  being  $q_2$ ”.

We are considering only zero conditional statements. Statements of the form: “If  $o_1$  is  $q_1$ , then  $o_2$  is  $q_2$ ”. We assume statements are used to describe a situation observed by an agent. It is also assumed there is no dialog based context imposed on the statement meaning.

We assume that inconvenience always stresses the consequent not the antecedent. Hence the modal operator is placed in the second half of a statement.

In example, a statement: “If the apple is green, then I believe it is not ripe” is considered within this work. While a statement: “I believe, that if the apple is green, then it is not ripe” is not considered.

### 3.1 Language Syntax

We define only modal conditional statements, because these will be analysed further in this paper.

The alphabet of the language  $L$  consist of the following classes of symbols:

- $O = \{o_1, o_2, \dots, o_M\}$  to represent atomic individuals (objects),
- $Q = \{q_1, q_2, \dots, q_K\}$  for predicates (objects properties),
- symbol ‘ $\neg$ ’ for negation, symbol ‘ $\rightarrow$ ’ for conditional statements, additional bracket symbols ‘(’ and ‘)’,
- symbols  $Pos$ ,  $Bel$  for modal operators of possibility, belief and knowledge.

Now we have to define proper formulas. The formulas are divided into two classes: atomic formulas ( $L^A$ ) and complex formulas ( $L^M$ ).

**Atomic formulas  $L^A$ :**

Let:  $k \in \{1, 2, \dots, K\}$  and  $m \in \{1, 2, \dots, M\}$ . Any statement of the form  $q_k(o_m)$  or  $\neg q_k(o_m)$  is a proper statement of the language  $L$ .

**Complex formulas  $L^B$ :**

Let  $\phi, \psi \in L^A$ . Any statement of one of the following forms:  $\phi \rightarrow \psi$ ,  $\phi \rightarrow Bel(\psi)$ ,  $\phi \rightarrow Pos(\psi)$  is a proper statement of the language  $L$ .

### 3.2 Intuitive Semantics

The intuitive semantics of the statements are presented in table [II](#)

**Table 1.** Language semantics

Formula	Meaning
$q_k(o_m)$	$o_m$ is $q_k$ . (Object $o_m$ has property $q_k$ .)
$\neg q_k(o_m)$	$o_m$ is not $q_k$ . (Object $o_m$ does not have property $q_k$ .)
$\phi \rightarrow \psi$	If $\phi$ , then $\psi$ .
$\phi \rightarrow Bel(\psi)$	If $\phi$ , then I believe, that $\psi$ .
$\phi \rightarrow Pos(\psi)$	If $\phi$ , then it is possible, that $\psi$ .

Interpretation considers common sense understanding of the defined language and is not a formal logical semantics interpretation considering classical truth tables. The interpretation of modal messages is consistent with common understanding of belief and possibility [\[5\]](#). The presented interpretation is not a formal logical definition of modal logic based on Kripke possible worlds [\[9\]](#). Implications are zero conditionals used to describe current situation.

## 4 Conditional Statements Meaning

As proven in paragraph 2 the statement told by the speaker implies not only the pure logical meaning. The listener can also reason about speaker's mental state. Hence agent has to obey conventional implicatures when telling a statement. Within this work we are considering modal conditional statements. It is therefore important to emphasise these implicatures related to them.

### 4.1 Statements without Modal Operator

Most of results presented in this sub-paragraph have been already presented in 10. Here we quote conclusions that are important in the context of the rest of this work.

We are considering conditional statement of the form: "If  $\phi$ , then  $\psi$ ." told by an agent to describe observed situation. Suppose an agent tells such a statement. Let us figure out what the human listener can conclude from such a statement.

Firstly the speaker does not know whether  $\phi$  holds or not. If the speaker knew that  $\phi$  holds, he would immediately know that  $\psi$  holds and would rather tell ' $\psi$ , because of  $\phi$ ' or simply ' $\psi$ '. On the other hand, if speaker knew that  $\phi$  does not hold, it would be pointless of him to say such a statement. Since  $\phi$  does not hold, we can't use the implication to figure out anything about  $\psi$ .

Secondly the speaker does not know whether  $\psi$  holds or not. If he knew that  $\psi$  holds, there would be no point in telling the conditional statement because  $\psi$  already is known to hold. Similarly the speaker does not know that  $\psi$  does not hold. This would imply that the speaker also knows that  $\phi$  does not hold. Hence he would rather say 'Not  $\psi$ , because of not  $\phi$ .' or 'If there were  $\phi$ , there would be  $\psi$ '. The understanding of speaker's knowledge is consistent with results presented in 11.

Lastly, the speaker informs, that he has reasoned about both situations (where  $\phi$  holds or not) and found out that  $\psi$  is guaranteed to hold only when  $\phi$  holds. In fact the speaker has reasoned about four possible situations, and found out that situation where  $\phi$  holds and  $\psi$  does not hold is impossible. Hence the speaker is ready to infer  $\psi$ , if he found out that  $\phi$  holds. But as long as he does not know  $\phi$  he is unable to tell much more about  $\psi$ .

### 4.2 Statements with Belief Operator

Let us consider a situation where the statement "If  $\phi$ , then I believe  $\psi$ ." is told. This is formally written as  $\phi \rightarrow Bel(\psi)$ . The phrase 'I believe that' alarms the listener is uncertain whether  $\psi$  really holds in case of  $\phi$ . The statement means, that in a case where  $\phi$  holds, the speaker believes  $\psi$  holds. On the other hand, if  $\phi$  does not hold, the speaker can't tell that he believes  $\psi$  holds. The speaker has noticed that  $\phi$  changes chance of  $\psi$  happening. But it is not that  $\phi$  guarantees  $\psi$ . As an example, imagine a situation when the speaker says: "If the animal is a bird, then I believe it can fly". The speaker knows that not all birds fly, but most of them do.

The question is how the chance of  $\psi$  happening has to change depending on  $\phi$ . We state that this chance has to rise when  $\phi$  holds. When hearing a statement of the form  $\phi \rightarrow Bel(\psi)$  one implies that, if not  $\phi$ ,  $\psi$  is much less probable. We wouldn't say: If it is a cheap restaurant, then I believe they have metal cutlery". Most of restaurants have metal cutlery, but a cheap restaurant may have cutlery made of plastic. Hence exemplary statement seems at least weird, since fancy restaurants always have metal cutlery.

### 4.3 Statements with Possibility Operator

Now let us consider statement "If  $\phi$ , then I find it possible that  $\psi$ ", formally written as  $\phi \rightarrow Pos(\psi)$ . The phrase 'I find it possible that' means, that the speaker allows situation where both  $\phi$  and  $\psi$  hold and situation where  $\phi$  holds and  $\psi$  does not hold is also very probable.

On the other hand the listener can imply, that the speaker does not want to allow a situation where  $\phi$  does not hold and  $\psi$  holds. Otherwise the speaker would not reason on the value of  $\psi$  basing on  $\phi$ . Similarly to *Bel* modal operator the chance of  $\psi$  happening has to rise when  $\phi$  holds. One wouldn't say: "If the apple is green, then I find it possible to be ripe". This would suggest that red apples shouldn't be ripe. He would rather use a sentence: "If the apple is green, then I believe it is not ripe".

Table 2 presents the meaning of all considered conditional statements.

**Table 2.** What the speaker may think about  $\psi$  when telling a conditional statement

$\phi$	$\psi$	$\phi \rightarrow \psi$	$\phi \rightarrow Bel(\psi)$	$\phi \rightarrow Pos(\psi)$
assuming that $\phi$ does not hold:				
no	no	probable	very probable, must be	should/must be
no	yes	probable, impossible	little probable, impossible	improbable, can't be
assuming that $\phi$ holds:				
yes	no	mustn't be	little probable, rather is not	quite possible, rather is
yes	yes	must be	very probable, rather is	may be, probable

## 5 Mental State

Agent makes a statement based on her mental state, an internal and subjective representation of how the world is and may be. In paragraph 4 we specified how the listener interprets speaker's thoughts based on expressed conditional statement. Speaking agent has to be designed to obey these conventional implicatures. It is important to model agent's internal structures, so that they are able to provide data required to check which statement can be expressed.

If the agent wants to tell one of considered conditional statements, she can't know whether  $\phi$  or  $\psi$  holds. Hence agent has to be able to verify, if she knows

any of the objects' features. If at least one of them is known, none of considered conditional statements can be expressed.

In order to choose the right sentence agent has to be able to measure how both features are interrelated. Agent must consider different cases of how the situation may be. Later she has to estimate what  $\phi$  and  $\psi$  combinations are possible and how probable they are. Further the agent has to choose one of the statements based on these estimations.

### 5.1 Mental State Model

Agent, as an autonomous entity, has to reason about the environment. Some predictions, possible flows of events and evaluations are a result of agent's observations, knowledge and reasoning. We assume this reasoning leads to a mental model that is a set consisting of different possible situations and some evaluations on how probable each situation is. Such mental state model can be obtained by reinforcement learning algorithms like [11]. A mental state model is a set:

$$W = \{(w^{(1)}, p^{(1)}), (w^{(2)}, p^{(2)}), \dots, (w^{(S)}, p^{(S)})\} \quad (1)$$

where  $w^{(s)}$ ,  $s = 1, 2, \dots, S$  is a possible world and  $p^{(s)}$  is a chance of this world being an actual, unknown to the agent, world. Mental model is created by an agent autonomously based on her knowledge. Mental model represents agent's knowledge on how the world may be at a given time moment. Every possible world represents some possible state of the world at one fixed time moment. One of possible worlds should be the actual, real world state. Agent is not omnipotent and does not know which of the worlds is the real one. She can only evaluate how probable the world is to be the actual one.

It is assumed that  $\sum_{s=1}^S p^{(s)} = 1$  and  $p_s = P(w^{(s)})$  ( $s = 1, 2, \dots, S$ ) defines a probability distribution over  $W$ . We assume that every world has positive probability ( $\forall_{s \in \{1, 2, \dots, S\}} p_s > 0$ ). It is not necessary for  $p^{(s)}$  to be probability in its strict definition. It should be some estimation of probability created by the agent. The greater  $p_s$ , the more probable  $w_s$  is.

Each possible world consists of information about objects properties:

$$w^{(s)} = (Q_1^{(s)}, Q_2^{(s)}, \dots, Q_K^{(s)}), \quad s = 1, 2, \dots, S \quad (2)$$

Knowledge about objects having given property is contained within respective set  $Q_k^{(s)}$ . Let  $o \in O$  be an object. We say that:

If  $o \in Q_k^{(s)}$ , then object  $o$  is assumed to exhibit property  $q_k$  in world  $s$

If  $o \notin Q_k^{(s)}$ , then object  $o$  is assumed to not exhibit property  $q_k$  in world  $s$

If, for example,  $o \in Q_k^{(s)}$  in all possible worlds, then agent is sure object  $o$  has property  $q_k$  at given time moment. If  $o \notin Q_k^{(s)}$  in all possible worlds, then agent is sure object  $o$  does not have property  $q_k$ . If there are some worlds where  $o \in Q_k^{(s)}$  and some where  $o \notin Q_k^{(s)}$ , then agent is not sure whether object  $o$  has or does not have property  $q_k$ . If none of the worlds allows the situation, where  $o \in Q_k^{(s)}$  and  $o \in Q_l^{(s)}$  then agent thinks that it is impossible for object  $o$  to have

both properties  $q_k$  and  $q_l$  at once<sup>1</sup>. Overall probability of  $o$  being  $q_k$  at given time moment is a sum of probabilities of all worlds where  $o \in Q_k^{(s)}$  (compare to equation 3).

## 6 Grounding Conditional Statements

We are considering statements of the forms:  $\phi \rightarrow \psi$ ,  $\phi \rightarrow \text{Bel}(\psi)$  and  $\phi \rightarrow \text{Pos}(\psi)$ . Let us assume that:  $\phi = q_k(o_m)$  is an arbitrary chosen property  $q_k \in Q$  for object  $o_m \in O$  and  $\psi = q_l(o_n)$  is an arbitrary chosen property  $q_l \in Q$  for object  $o_n \in O$ .

In order to find out which sentence is adequate to the situation, agent has to verify a statement against the model. In following subsections we define an epistemic relation ' $\models^E$ ', that tells whether a statement can be used in a given situation. The verification process is called grounding, because it connects statements to the world based on agents internal representation of that world. We say that a statement is properly grounded if and only if the epistemic relation holds.

Further we shall be using probabilistic model, to determine if a given statement can be spoken. Please note that required probabilities can be calculated from mental state model in the following manner:

$$P(\phi) = \sum_{s \in I_\phi} p^{(s)}, \quad \text{where } I_\phi = \{s : o_m \in Q_k^{(s)}\} \quad (3)$$

$$P(\psi \wedge \phi) = \sum_{s \in I_{\phi\psi}} p^{(s)}, \quad \text{where } I_{\phi\psi} = \{s : o_m \in Q_k^{(s)} \wedge o_n \in Q_l^{(s)}\} \quad (4)$$

$$P(\psi \wedge \neg\phi) = \sum_{s \in I_{\neg\phi\psi}} p^{(s)}, \quad \text{where } I_{\neg\phi\psi} = \{s : o_m \notin Q_k^{(s)} \wedge o_n \in Q_l^{(s)}\} \quad (5)$$

$$P(\neg\phi) = 1 - P(\phi), \quad P(\psi|\phi) = \frac{P(\psi \wedge \phi)}{P(\phi)}, \quad P(\psi|\neg\phi) = \frac{P(\psi \wedge \neg\phi)}{P(\neg\phi)} \quad (6)$$

### 6.1 Grounding Statements without Modal Operator

Below we define the epistemic relation for a statement of the form: "If  $\phi$ , then  $\psi$ ."

**Definition 1.** *Epistemic relation  $\models^E \phi \rightarrow \psi$  holds iff all following conditions are met:*

- a.  $\underline{\alpha} < P(\phi) < \bar{\alpha}$
- b.  $P(\psi|\phi) = 1$
- c.  $P(\psi|\phi) > \beta P(\psi|\neg\phi)$

where  $0 < \underline{\alpha} < \bar{\alpha} < 1$ ,  $\beta > 1$  are fixed parameters.

---

<sup>1</sup> For example ball can't be red and blue all over, at the same time moment.

Condition *a.* comes from previously justified assumption that value of  $\phi$  can't be known. There must be some realistic chance that  $\phi$  holds and also some chance that  $\phi$  will not hold. We assume that the chance must be realistically probable, hence  $\underline{\alpha}$  can be greater than zero. This way we ignore situations where  $\phi$  is probable, but it is slightly probable.

In case of a simple conditional statement agent has no doubts, that  $\psi$  happens whether  $\phi$  happens, therefore probability in condition *b* must be equal to one.

Condition *c* ensures two required features. Firstly  $\phi$  must positively affect chance of  $\psi$  happening. Secondly  $\psi$  can't happen all the time regardless of  $\phi$ .

The  $\beta$  parameter should be greater than one. This parameter tells how much more often  $\psi$  happens in case of  $\phi$  compared to a situation where  $\phi$  does not hold. For example  $\beta = 2$  means that  $\psi$  should happen at least twice as often when  $\phi$  holds. The higher the  $\beta$  parameter, the greater the impact of  $\phi$  must be and the less willing to use an implication agent is.

## 6.2 Grounding Statements with Belief Operator

We are defining an epistemic relation for a statement of the form: "If  $\psi$ , then I believe, that  $\phi$ ".

**Definition 2.** Epistemic relation  $\models^E \phi \rightarrow Bel(\psi)$  holds iff all following conditions are met:

- a.  $\underline{\alpha} < P(\phi) < \bar{\alpha}$
- b.  $\underline{\alpha}_{Bel} \leq P(\psi|\phi) < \bar{\alpha}_{Bel}$
- c.  $P(\psi|\phi) > \beta P(\psi|\neg\phi)$  where  $0 < \underline{\alpha} < \bar{\alpha} < 1$ ,  $0 < \underline{\alpha}_{Bel} < \bar{\alpha}_{Bel} < 1$ ,  $\beta > 1$  are fixed parameters.

Conditions *a* and *c* are the same as in a case of no modal operators.

Belief operator means, that agent suggests, that  $\psi$  happens often in case of  $\phi$ . Condition *b* ensures that  $\phi$  will happen often, but not always. Parameter,  $\bar{\alpha}_{Bel}$  should be set close to one to ensure high probability of  $\psi$ . The higher the  $\underline{\alpha}_{Bel}$  parameter, the more restrictive agent is when deciding to tell that she believes something.

## 6.3 Grounding Statements with Possibility Operator

Lastly we are defining an epistemic relation for a statement of the form: "If  $\phi$ , then I find it possible, that  $\psi$ ".

**Definition 3.** Epistemic relation  $\models^E \phi \rightarrow Pos(\psi)$  holds iff all following conditions are met:

- a.  $\underline{\alpha} < P(\phi) < \bar{\alpha}$
  - b.  $\underline{\alpha}_{Pos} < P(\psi|\phi) < \bar{\alpha}_{Pos}$
  - c.  $P(\psi|\phi) > \beta P(\psi|\neg\phi)$
- where  $0 < \underline{\alpha} < \bar{\alpha} < 1$ ,  $0 < \underline{\alpha}_{Pos} < \bar{\alpha}_{Pos} < 1$ ,  $\beta > 1$  are fixed parameters.

Conditions *a* and *c* are the same as in a case of no modal operators.

Condition  $b$  is similar to respective condition for belief operator. But it has its own  $\underline{\alpha}_{Pos}$ ,  $\overline{\alpha}_{Pos}$  parameters. To ensure proper understanding of possibility operator,  $\underline{\alpha}_{Pos}$  should be close to zero.

It is best when  $\overline{\alpha}_{Pos} \leq \underline{\alpha}_{Bel}$ , so that agent can't say at the same time moment that she believes and finds the same thing possible. It is also best when  $\overline{\alpha}_{Pos} \geq \underline{\alpha}_{Bel}$ . In such case there is no situation when agent can't use possibility operator because consequent is too probable and on the other hand she can't use belief operator because the same consequent is too little probable. Hence an optimal solution is when  $\overline{\alpha}_{Pos} = \underline{\alpha}_{Bel}$ .

## 7 Summary

We have extended an agent's language with conditional statements. Further we have analysed these statements meaning and interpretation by humans.

We marked most important and required features of agent's model. Our model can be used to verify whether given statement is applicable and meets its conventional meaning.

We assumed a model where strict probabilities, whole possible worlds and binary features are required. For every non-trivial domain it is very difficult and computationally complex to model environment in such a way. In the nearest future we plan to loosen up model requirements to allow partially known possible worlds and imprecise probability estimations.

## References

1. Ajdukiewicz, K.: Conditional sentence and material implication. *Studia Logica* 4(1), 135–153 (1956)
2. Cohen, P., Levesque, H.: Communicative Actions for Artificial Agents. In: Proc. of the 1st International Conference on Multi-agent Systems, San Francisco (1995)
3. Grice, H.P.: Meaning. *Philosophical Review* 88, 377–388 (1957)
4. Harnad, S.: The Symbol Grounding Problem. *Physica D* 42, 335–346 (1990)
5. Hintikka, J.: Knowledge and belief. In: An introduction to the logic of the two notions. Cornell University Press (1962)
6. Huhns, N., Singh, M.: Cognitive Agents. *IEEE Internet Computing* 2(6), 87–89 (1998)
7. Katarzyniak, R.: The language grounding problem and its relation to the internal structure of cognitive agents. *Journal of Universal Computer Science* 11(2), 357–374 (2005)
8. Katarzyniak, R.: Gruntowanie modalnego języka komunikacji w systemach agentowych. Exit, Warsaw (2007) (in Polish)
9. Kripke, S.: Semantical Analysis of Modal Logic. *Zeitschrift für Mathematische Logic und Grundlagen der Mathematik* 9, 67–96 (1963)
10. Skorupa, G., Katarzyniak, R.: Extending modal language of agents' communication with modal implications. *Information systems architecture and technology*, 127–136 (2009)
11. Strehl, A.L., Li, L., Wiewiora, E., Langford, J.: Pac. model-free reinforcement learning. In: Proc. 23nd ICML 2006, pp. 881–888 (2006)

# A Computer Adaptive Testing Method for Intelligent Tutoring Systems<sup>\*</sup>

Adrianna Kozierkiewicz-Hetmańska and Ngoc Thanh Nguyen

Institute of Informatics, Wroclaw University of Technology, Poland  
*{adrianna.kozierkiewicz,ngoc-thanh.nguyen}@pwr.wroc.pl*

**Abstract.** The growth of popularity of computers increases interest of adaptive testing in tutoring systems. Computer adaptive testing is a form of educational measurement that is adaptable to examinee proficiency. In a procedure of adaptive testing it is required to determine a selection of the first item, a method of estimation of student's proficiency, a method of selection of the next item and a termination criterion. In this paper the original algorithm of adaptive testing with all basic steps is proposed. The level of difficulty of the first item is set using user's profile. Such solution allows to start a test, where the first item is suitable for student's preferences. In our method 2-parameter IRT model is applied to choose the next item.

## 1 Introduction

The learning process is closely connected to evaluation processes. The testing of student's knowledge motivates students to learn, gives students and teachers feedback on students' learning progress and allows modifying learning strategies.

The most popular form of evaluation are traditional tests solved by a learner in a paper-and-pencil mode. In this test the number of items, their sequence, content and timing are the same for all learners. The development of information technology in the last years involves computer in an educational measurement. In [2] authors distinguished 4 different generations of testing: Computer Testing (CT), Computer Adaptive Testing (CAT), Continuous Measurement (CM) and Intelligent Measurement (IM). The first generation uses administering conventional tests by a computer. Computer testing uses different number of items, different item sequence and timing for different students but the selection does not depend on examinees' responses. Scientific foundations are usually based on a classical test theory. The second generation is computer adaptive testing. In CAT presentation of the next item or the decision to stop is adaptive and depends on student's performance on the previous items. Scientific foundations are based on Item Response Theory (IRT). In the generation CM student's achievement is estimated and learner profile is taken into account during the evaluation process. Scientific foundations are based on extensions of IRT and learner profiles. The last generation is IM which produces an intelligent scoring, an interpretation of

---

\* This research was financially supported by the Polish Ministry of Science and Higher Education under the grants no. 0419/B/T02/2009/37 and N N519 407437.

individual profiles, advices for learners and teachers based on models of expert knowledge.

So far there exist researches focused on CT and CAT. The main advantages of computer adaptive testing [4], [11] is to reduce opportunities to cheat by the possibility of drawing items from a big item pool, quick feedback: students are given the test results immediately after it is finished. CAT tests are shorter by 50% and give a higher level of precision than static tests. CAT tests reduce students' level of stress and increase the level of motivation because the items are adaptively selected and are not too hard and not too easy for examinee.

The computer testing and computer adaptive testing do not consider learner profile during an evaluation process. In this paper we propose a model of a computer adaptive testing, which uses the learner profile. It is a new generation of a computerized educational measurement.

Before a student starts to learn he has to register in the E-learning system. Some data is directly provided by the learner such as: login, name, gender, educational level etc. Data about a learner can be obtained via psychological questionnaires. It is assumed that E-learning systems store two types of data: user data and usage data [6]. User data contains demographic data (login, name, telephone, e-mail, age, sex, educational level, IQ), learning style (related to perception, receiving, processing and understanding of information by student), abilities (verbal comprehension, word fluency, computational ability, spatial visualization, associative memory, perceptual speed, reasoning), personal character traits (concentration, motivation, ambition, self-esteem, level of anxiety, locus of control, open mind, impetuosity, perfectionism) and interests (humanistic science, formal science, the natural science, economics and law, technical science, business and administration, sport and tourism, artistic science, management and organization, education). Usage data contains information about completed lessons, results of tests, opening and final scenarios, time spent on reading lessons and other information collected during user's interaction with the system [7].

Some elements of user data should be taken into account during evaluation processes. It can be observed that younger and intuitive students are impatient so they are quickly bored by the big number of test items. Learners who are older, reflective, with low level of concentration and high level of anxiety need more time for solving tests. Active users like big number of test items. Sensitive students and people characterized by the high level of anxiety, impetuosity and low level of ambition and self-esteem need to be presented with easier tasks. Users with high level of concentration, low level of motivation, high level of ambition or self-esteem and perfectionists are better in more complicated tasks, questions and exercises. Sequential students often prefer solving easier tasks and continue educational measurements with more and more complicated test items. Global users, on the other hand, like harder test items at once.

In our research we propose using learner profiles to create the model of computer adaptive testing. This solution allows students to be more confident and effective because they can get test items suitable for their preferences and needs. The computer adaptive testing procedure requires establishing several elements which will be done in this paper:

- The choice of the first item presented to an examinee
- The estimation of student's ability based upon all prior answers

- The choice of the optimal item based on the current estimate of the examinee's ability
- The choice of stopping criterion

The proposed procedures take an advantage of student's profiles and IRT. Before starting a test, the system tries to assess the best difficulty level of the first item based on the user's profile. After student's answering the system updates student's proficiency level and selects the optimal item based on a learner's knowledge level. The procedure is finished when the termination criterion is satisfied.

In the next section an overview of different methods applied in computer adaptive testing procedures with demonstrated systems is presented. Section 3 describes the proposed model of computer adaptive testing using learner profiles. Section 4 contains conclusions and further works.

## 2 Related Works

The growth of popularity of using a computer in education contributes to development of methods for computer adaptive testing.

Educational measurement has always been a very important step in the learning process. Therefore, in intelligent tutoring systems more and more often a new technology of students' knowledge assessment is applied. The reliable testing is necessary for determination of an optimal learning scenario because it is necessary to acquire knowledge about the current learner's proficiency level and the potential need of repetitions or additional explanations. Many systems for e-learning proposes still use only static tests such as a very popular and used at the many universities WebCT [14] or Moodle [15] but some tutoring systems propose different strategies of adaptive testing based on IRT or Bayes' Theorem.

In ELM-ART [13], one of the first adaptive web-based systems, the authors proposed very easy strategy for computer adaptive testing. Each test item contains a difficulty parameter which can occur in different test groups. Test groups are collections of test items referred to some unit. The final test starts presenting one test item from the test group of medium difficulty. If students answer incorrectly, system selects another test item with lower difficulty. Otherwise, two test items from the test group of higher difficulty are chosen. Usually about 6-10 items are presented to the examinee.

System INSPIRE [5] incorporates IRT. The test item in INSPIRE is associated with different parameters such as: the level of difficulty or a number of times that the question has been answered correctly or incorrectly by any learner, etc. An adaptive assessment algorithm is conducted in a few steps. Firstly, system has to assume the initial student's proficiency as moderate. Then test items are selected based on the current estimation and presented to learner. The selection of questions is based on IRT. Item Characteristic Curve (ICC) depends on two parameters (a difficulty of the question and a guessing factor) and the Item Information Function (IFF) of each item is calculated. The question with the highest value of IFF for the current estimation of the learner's proficiency is chosen. After student's answering system updates the examinee's proficiency level. The last two steps are repeated until the assumed number of questions are presented or the estimation of the learner's proficiency level reaches a desired value.

GenTAI is a module of ELSA system [10] which allows conducting computerized adaptive testing. In this system 3-parameter logistic IRT model is applied.

In [8] authors used 1-parameter logistic IRT model also called the Rasch Model. Selection of the next item is based on student's proficiency level and the set of construction rules.

Another approach used in adaptive testing is the Bayesian model. In CAT Bayesian model is used to compute the probability of a correct answer to a question, using previous answers to previous questions. The Bayesian item selection criteria are described in [12].

SIETTE [3] is a complete tool for creating adaptive tests where questions are selected as suitable for student's level of knowledge. SIETTE allows creating adaptive tests using different strategies of item selection and different termination criteria. Adaptive tests in SIETTE typically consist of the following steps: after examinee answers, the new ability level is estimated and, basing on current knowledge proficiency level, the next question is selected. This procedure is stopped when a termination criterion is met. In SIETTE, it is possible to choose between three different item selection procedures. The first of them, Bayesian procedure selects the item which minimizes the posterior standard deviation. The second method is dependent on selection of item that gives the minimum distance between the mean of the ICC and the mean of the current student's knowledge distribution. The last procedure selects items randomly. Student's knowledge proficiency level is computed using Bayes' Theorem. SIETTE allows choosing some stopping criterion: all questions are posed, the assumed number of questions were presented to student, the standard deviation of the distribution of learner's knowledge is smaller than a fixed value or proficiency level is greater than assumed.

In our work we want to create a procedure of computer adaptive testing considering student's profile. To our best knowledge, so far such solution has not been proposed.

### **3 Computer Adaptive Testing**

The presentation of each test item in computer adaptive tests is adapted to the student's proficiency level. Therefore, the computer adaptive tests evaluate examinees more precisely and in the shorter time than traditional tests. We assumed that item pool is calibrated and contains multiple-choice questions. Each question is associated with two parameters: the level of difficulty which describes how difficult a question is and discrimination which denotes how well the question is able to discriminate between examinees with slightly different ability.

The difficulty level and the discrimination parameter are initially assigned by the teacher. The procedure of estimating item parameters is described in details in [1].

The algorithm of adaptive testing consists of the following procedures: the first item selection method, the next item selection method and the knowledge level estimation. In the following subsections all procedures will be described in details.

#### **3.1 The First Item Selection Method**

The selection of the first item depends on some student's learning style and personal traits. It was mentioned in Section 1 that some data stored in the user profile should be considered during evaluation processes.

During the registration process system collects information about a student. We assumed that global, old, perfectionist students and students with high level of concentration, motivation, self-esteem, ambition and low level of anxiety are better in complicated tasks. The easier items should be offered to young, sequential, impetuous, sensitive learners and those with low level of concentration, self-esteem, ambition and the high level of anxiety.

**Table 1.** Table of scores

Attribute name	Attribute's value	Score ( $s_j$ )
Concentration	High level	1
	Low level	-1
Self-esteem	High level	1
	Low level	-1
Ambition	High level	1
	Low level	-1
Anxiety	Low level	1
	High level	-1
Motivation	High level	1
Perception	Sensitive	-1
Perfectionism	Yes	1
Impetuosity	Yes	-1
Understanding	Global	1
	Sequential	-1
Age	Old	1
	Young	-1

Before presenting the first item to a student the index  $FI$  is calculated as the sum of scores  $s_j$  (if a value of an attribute from Table 1 occurs in user's profile then the appropriate score  $s_j$  is taken from Table 1 otherwise  $s_j$  is equal to 0):

$$FI = \sum_{j=1}^J s_j \quad (3.1)$$

where:  $J$ -the number of possible values of attributes.

The score of  $FI$  influences the selection of the first item. If  $FI$  belongs to interval [-8, -6], then student should start a test with a very easy, random item. If  $FI$  belongs to interval [-4, -2], then student should be offered a random, easy test item. If  $FI$  is equal to 0, then student is fairly balanced and he is presented a random item of medium level of difficulty. If calculated  $FI$  belongs to [2, 4] or [6, 8] then student should start a test with a random hard or very hard item, respectively.

### 3.2 The Item Selection Method

The item selection method takes advantage of Item Response Theory. For each question the Item Response Function (IRF) is calculated. The IRF represents the probability that a person with given proficiency level will answer to questions correctly. The IRF depends on two item parameters: the discrimination  $a_i$  and the difficulty of the question  $b_i$  [1]:

$$P_i(X_i = \text{correct} | \theta) = \frac{1}{1 + e^{-a_i \cdot (\theta - b_i)}} \quad (3.2)$$

where:

$\theta$  - the student's knowledge level,

$X_i$  - answer to question  $i$ ,  $i \in (1, \dots, M)$ ,

$M$  - the cardinality of set of item pool.

Next, the Item Information Function (IIF) for each question is assessed in the following way [1]:

$$I_i(\theta) = a_i^2 P_i(\theta)[1 - P_i(\theta)] \quad (3.3)$$

The information is an index representing the item's ability to differentiate among individuals. It is obvious that tests should be the most informative. Therefore, the selection of the item is based on the maximum information criterion [9]. The item is selected so that it maximizes the item information for the proficiency level estimated until that moment. Figure 1 and Figure 2 present Item Response Function and Item Information Function for  $a = 2$  i  $b = 0$ , respectively.

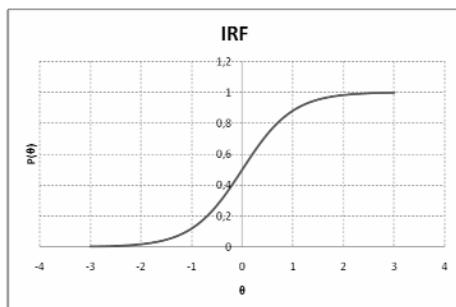


Fig. 1. Item Response Curve

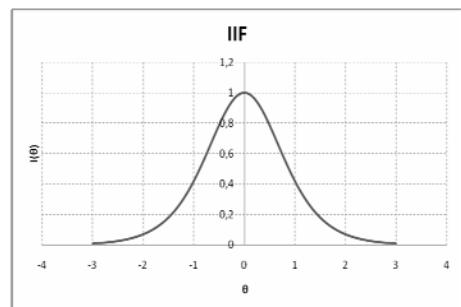
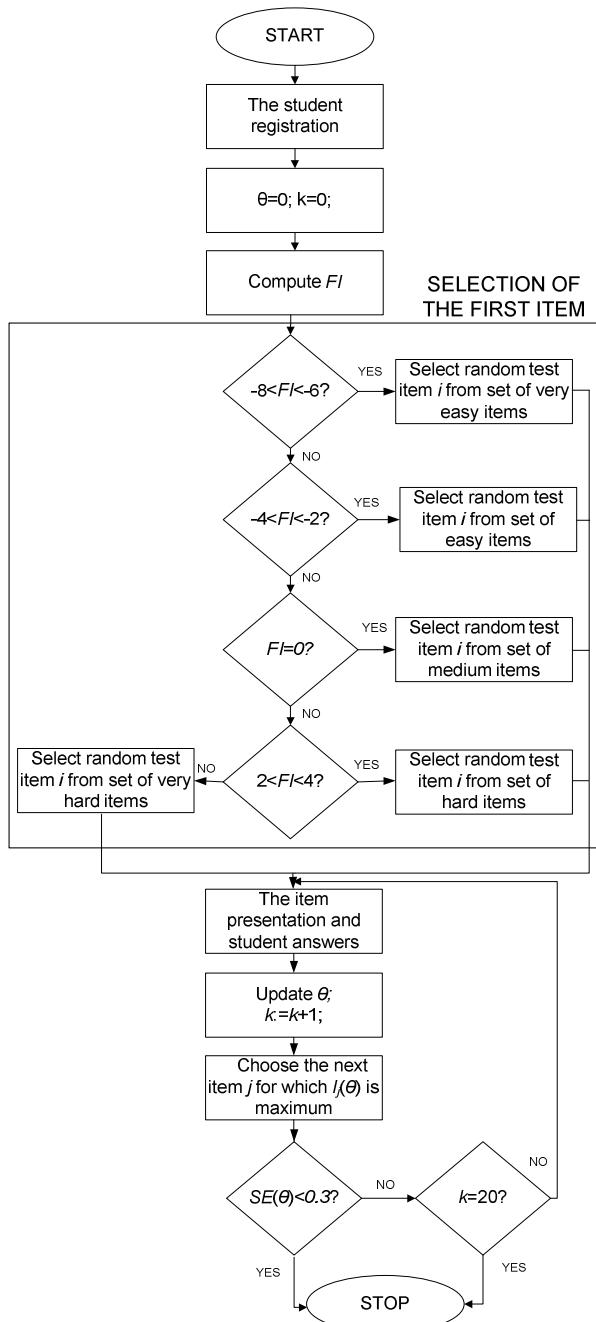


Fig. 2. Item Information Function

### 3.3 The Knowledge Level Estimation

After each student's response it is needed to estimate examinee's proficiency level. Estimation of user's knowledge level is an iterative process. Initially, the student's proficiency is assumed as moderate. After each item presentation and student's answer proficiency level is updated using the following formula [1]:

**Fig. 3.** Procedure of adaptive testing

$$\theta_{s+1} = \theta_s + \frac{\sum_{i=1}^M a_i [u_i - P_i(\theta_s)]}{\sum_{i=1}^M a_i^2 P_i(\theta_s) [1 - P_i(\theta_s)]} \quad (3.4)$$

where:

- $u_i$ - the answer to item,
- $u_i=1$  for correct response and  $u_i=0$  for incorrect response,  $i \in (1, \dots, M)$ ,
- M- the cardinality of set of item pool,
- s- number of iteration

### 3.4 The Termination Criteria

The selection of the next item and estimation of student's proficiency level is repeated until the termination criteria is reached. The test is terminated when the standard deviation of the distribution of student's knowledge is smaller than a fixed value ( $< 0.3$ ) or the number of presented items exceed the assumed number ( $> 20$ ). The standard deviation is calculated using the following equation [1]:

$$SE(\theta) = \frac{1}{\sqrt{\sum_{i=1}^M I_i(\theta)}} \quad (3.5)$$

The procedure of computer adaptive testing is presented in Figure 3.

## 4 Conclusions and Further Work

The learning process consists of several steps [7]. One of the most important tasks is learning result evaluation process. For centuries evaluation process was very stressful for students. It was changed, when adaptive testing appeared. In CAT each question is selected intelligently to fit student's proficiency level, so examinees are not bored with easy tasks and do not feel stressed with hard items.

In this paper the procedure of adaptive testing is proposed. The worked out method is original and innovative because it takes advantage of user profiles in selection of tasks of the first test item. Depending on examinee's profile, user starts a test with the very easy, easy, medium, hard or very hard item. The selection of the next item is based on 2-parameter IRT model. The maximum information criterion is assumed. The procedure is stopped after a fixed number of questions or if the standard deviation of the distribution of student's knowledge is smaller than a threshold.

In the future work the proof of correctness of the procedures described in Section 3 will be dealt with. We plan to run experiments with two groups of users: the experimental group and the control group. The first group will be tested by using the worked out method, the second group will be assessed by using typical computer adaptive tests without considering their users' profiles. It is expected that people from the experimental group will achieve better results and in shorter time than people from the control group.

## References

1. Baker, F.: The Basics of Item Response Theory. In: ERIC Clearinghouse on Assessment and Evaluation, 2nd edn. University of Maryland, College Park (2001)
2. Bunderson, V., Inouye, D., Olson, J.: The four generations of computerized educational measurement. In: Linn, R.L. (ed.) Educational measurement. Macmillan, New York (1988)
3. Conejo, R., et al.: SIETTE: A Web-Based Tool for Adaptive Testing. International Journal of Artificial Intelligence in Education 14, 1–33 (2004)
4. Fetzer, M.: The Next Evolution of Computer Adaptive Testing. In: Talent Managment, p. 5 (2009)
5. Gouli, E., Kornilakis, H., Papanikolaou, K., Grigoriadou, M.: Adaptive Assessment Improving Interaction in an Educational Hypermedia System. In: Proceedings of the PanHellenic Conference with International Participation in Human-Computer Interaction, pp. 217–222 (2001)
6. Kobsa, A., Koenemann, J., Pohl, W.: Personalized Hypermedia Presentation Techniques for Improving Online Customer Relationships. The Knowledge Engineering Review 16(2), 111–155 (2001)
7. Kozierkiewicz, A.: Content and structure of learner profile in an intelligent E-learning system. In: Nguyen, N.T., Kolaczek, G., Gabrys, B. (eds.) Knowledge Processing and Reasoning for Information Society. EXIT Warsaw (2008)
8. Kwan, R., Wong, K., Yu, C., Tsang, P., Kat Leung, K.: Building an Assessment Learning System on the Web. In: Fong, J., Kwan, R., Wang, F.L. (eds.) ICWL 2008. LNCS, vol. 5169, pp. 91–102. Springer, Heidelberg (2008)
9. Linden van der, W.J., Glas, C.A.W. (eds.): Computerized adaptive testing: theory and practice. Kluwer Academic Publishers, Boston (2000)
10. López-Cuadrado, J., Armendariz, A., Pérez, T.A.: Adaptive evaluation in an E-Learning System Architecture. In: Mendez-Vilas, A., et al. (eds.) Current Developments in Technology-Assisted Education, Formatex, Badajoz, Spain, pp. 1507–1511 (2006)
11. Sztejnberg, A., Hurek, J.: Zastosowanie osiągnięć technologii komputerowej w pomiarze edukacyjnym. Komputerowe testowanie w pełni adaptacyjne. In: Bilek, M. (ed.) Aktualni Otazky Vyuky Chemie XII, ICT ve vyuuce chemie a v priprave uctitela chemie, Univerzita Hradec Kralove, Gaudemus, Hradec Kralove, pp. 235–238 (2002)
12. Veldkamp, B.P.: Bayesian Item Selection in Constrained Adaptive Testing Using Shadow Tests. Psicologica 31, 149–169 (2010)
13. Weber, G., Peter Brusilovsky, P.: ELM-ART: An Adaptive Versatile System for Web-based Instruction. W: International Journal of Artificial Intelligence in Education 12, 351–384 (2001)
14. WebCT, <http://www.blackboard.com/>
15. Moodle, <http://moodle.org/>

# Combining Patient Metadata Extraction and Automatic Image Parsing for the Generation of an Anatomic Atlas

Manuel Möller<sup>1,2</sup>, Patrick Ernst<sup>2</sup>, Michael Sintek<sup>1</sup>, Sascha Seifert<sup>3</sup>, Gunnar Grimnes<sup>1</sup>, Alexander Cavallaro<sup>4</sup>, and Andreas Dengel<sup>1,2</sup>

<sup>1</sup> German Research Center for Artificial Intelligence, Kaiserslautern, Germany

<sup>2</sup> University of Kaiserslautern, Germany

<sup>3</sup> Integrated Data Systems, Siemens CT, Erlangen, Germany

<sup>4</sup> University Hospital, Erlangen, Germany

**Abstract.** We present a system that integrates ontology-based metadata extraction from medical images with a state-of-the-art object recognition algorithm for 3D volume data sets generated by Computed Tomography scanners. Extracted metadata and automatically generated medical image annotations are stored as instances of OWL classes. This system is applied to a corpus of over 750 GB of clinical image data. A spatial database is used to store and retrieve 3D representations of the generated medical image annotations. Our integrated data representation allows us to easily analyze our corpus and to estimate the quality of image metadata. A rule-based system is used to check the plausibility of the output of the automatic object recognition technique against the Foundational Model of Anatomy ontology. All combined, these methods are used to determine an appropriate set of metadata and image features for the automatic generation of a spatial atlas of human anatomy 

## 1 Introduction

During the last decades a lot of effort went into the development of automatic object recognition techniques for medical images. Today there is a huge variety of algorithms available solving this task very well. The precision and sophistication of the different image parsing techniques have improved a lot to cope with the increasing complexity of medical imaging data. There are numerous advanced object recognition algorithms for the detection of particular objects on medical images. However, the results of the different algorithms are neither stored in a common format nor comprehensively integrated with patient and image metadata.

At the same time the biomedical informatics community managed to represent huge parts of medical domain knowledge in formal ontologies. Today,

<sup>1</sup> This research has been supported in part by the THESEUS Program in the MEDICO Project which is funded by the German Federal Ministry of Economics and Technology under the grant number 01MQ07016. The responsibility for this publication lies with the authors. Björn Forcher contributed the QPL bridge to XSB Prolog.

comprehensive ontologies cover large parts of the available taxonomical as well as mereological (part-of) knowledge of human anatomy.

With the shift to the application of digital imaging techniques for medical diagnosis, the volume of digital images produced in clinics increased tremendously. An informal study at the Department of Radiology at the Mayo Clinic Jacksonville, FL, revealed an increase from 1,500 radiological images made per day in 1994 to 16,000 images in 2002 [1]. Our clinical partner, the University Hospital Erlangen, Germany, has a total of about 50 TB of medical images. Currently they have about 150,000 medical examinations producing 13 TB data per year. To cope with this data increase our approach is to fuse (A) state-of-the-art medical object recognition techniques and metadata extraction with (B) medical domain knowledge represented in formal ontologies. This common representation is complemented by (C) a digital anatomical atlas generated from a large corpus of 3D volume data sets. This combination allows leveraging medical domain knowledge for data analysis across different abstraction levels.

In the following we present the results of an application of this approach to a huge corpus of clinical image data. We illustrate the benefits of our approach by a number of examples. By leveraging knowledge from the medical domain ontologies we can automatically assess the semantic plausibility of the object detection results, e. g., by checking if the prostate has been detected in a data set of a female patient.

Our main contribution is the design of a generic model for the representation of medical annotations and image metadata using Semantic Web standards. Our approach fuses existing technologies—automatic object recognition and formal ontologies—into a generic system. This system exhibits completely new features which are not available when using these technologies separately.

## 2 Related Work

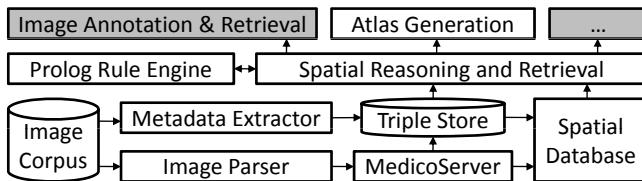
In [7] the authors describe a hybrid approach using metadata extracted from the medical image headers in combination with low-level image features. However, their aim is to speed up content-based image retrieval by restricting the search space by leveraging metadata information.

A study by Güld et al. from 2002 [8] already has dealt with the quality of DICOM metadata information. In contrast to our automatic approach they performed a manual expert assessment of the correctness of different metadata attributes specifying the body region. They concluded that information from DICOM images produced in clinical practice is by no means reliable and thus not suitable for automatic image retrieval. Our results point into the same direction.

Quantitative spatial models are the foundation of digital anatomical atlases. Fuzzy logic has been proven as an appropriate formalism which allows quantitative representations of spatial models [3]. In [9] the authors expressed spatial features and relations of object regions using fuzzy logic. In [4] and [2] Bloch et al. describe generalizations of this approach and compare different options to express relative positions and distances between 3D objects with fuzzy logic.

### 3 System Architecture

Fig. 1 shows an abstraction of the distributed system architecture. It is roughly organized in the order of the data processing horizontally from left to right. Applications benefit from the extracted data and the data integration along the vertical axis. For the latter the data flow is from bottom (backend system) to the top of the illustration. The description in the following sections is oriented on this order. In this paper we focus on data processing and reasoning and do not deal with clinical user interface aspects which would exceed the limitations of this publication.



**Fig. 1.** System architecture overview

**Metadata Extractor:** The entire image acquisition is based on a common semantic model which is represented in the MEDICO Ontology hierarchy [12]. We are using this ontology to define all information sources occurring during the acquisition. The representation on a common abstract level significantly eases the integration of disparate data sources. The most important part in the context of this publication is the DICOM ontology. The Digital Imaging Communications in Medicine (DICOM) standard is the most widely accepted standard in radiological imaging. It specifies the management, storage, and transfer of medical images [6]. In DICOM not only the format and attributes of the raw images are specified, but also a binary header encoding defining a rich set of metadata attributes.

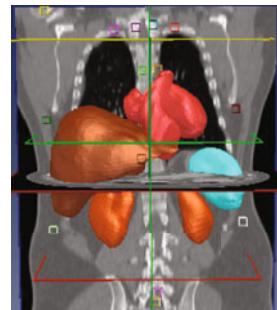
Each metadata element in the standard has a fixed so-called DICOM tag identifying it in the file header. To control the mapping process from DICOM metadata fields to properties in our ontology we annotated those properties with the respective tag using the annotation property `dicomTag`. The extracted metadata contains—amongst others—extensive details about the patient’s demographics. This includes gender, age, smoking status, etc.

**Image Parser:** To represent the results of the automatic object recognition algorithms in the format of our ontology we had to integrate rather disparate techniques into a hybrid system. The automatic object recognition performs an abstraction process from simple low-level features to concepts represented in the Foundational Model of Anatomy (FMA) [14], the ontology we use as our primary source of anatomical domain knowledge.

For automatic object recognition we use a state-of-the-art anatomical landmark detection system described in [15]. Fig. 2 shows a visualization of organs

detected in an abdomen CT scan. The image parsing algorithm generates two fundamentally different output formats: *Point3D* for landmarks and *Mesh* for organs. Apart from their geometric features they always point to a certain anatomical concept which is hard-wired to the model that the detection/segmentation algorithm has used to generate them. A landmark is a point in 3D without spatial extension. Usually they represent extremal points of anatomical entities with a spatial extension. In some cases these extremal points are not part of the official FMA. In these cases we modeled the respective concepts as described in Fig. 1. In total we were able to detect 22 different landmarks from the trunk of the human body. Examples are the bottom tip of the sternum, the tip of the coccyx, or the top point of the liver.

Organs, on the contrary, are approximated by polyhedral surfaces. Such a surface, short *mesh*, is a collection of vertices, edges, and faces defining the shape of the object in 3D. For the case of the urinary bladder, the organ segmentation algorithm uses the prototype of a mesh with 506 vertices which are then fitted to the organ surface of the current patient. Usually, vertices are used for more than one triangle. Here, these 506 vertices form 3,024 triangles. In contrast to the Point3D data, meshes are used to segment organs. For our test, the following organs were available: left/right kidney, left/right lung, bladder, and prostate.



**Fig. 2.** Organs in a thorax & abdomen CT scan

**MedicoServer:** Fig. 1 shows the overall architecture of our approach for integrating manual and automatic image annotation. One of the main challenges was to combine the C++ code for volume parsing with the Java-based libraries and applications for handling data in Semantic Web formats. We came up with a distributed architecture with the *MedicoServer* acting as a middleware between the C++ and Java components using CORBA [13].

**Triple Store:** In recent years there has been great interest in storage, querying, and reasoning on assertion box (abox) instances, for which several Semantic Web frameworks for Java have been proposed. We chose Sesame [5] because of its easy online deployment and fast built-in persistence strategy. Deployed to a central application server Sesame provides the system with a central RDF repository for storage and retrieval of information about the medical domain, clinical practice, patient metadata, and image annotations.

**Spatial DBMS:** As we have seen in the section about the image parsing algorithms, the automatic object recognition algorithms generate several thousand points per volume data set. Storage and efficient retrieval of this data for further processing made a spatial database management system necessary. Our review of available open-source databases with support for spatial data types revealed that most of them now also have support for 3D coordinates. However, the built-in operations ignore the third dimension and thus yield incorrect results, e.g., for distance calculations between two points in 3D. Eventually we decided

to implement a light-weight spatial database supporting the design rationals of simplicity and scalability for large numbers of spatial entities.

**Prolog Rule Engine:** We used the QPL library implemented by Björn Forcher at DFKI in Kaiserslautern. It connects to XSB Prolog<sup>2</sup> by mapping OWL/RDF statements of the form <Subject, Predicate, Object, Context> to Prolog facts of the form `true(Subject, Predicate, Object, Context)`. This allows reasoning with Prolog rules based on knowledge represented in OWL ontologies (see Section 5).

## 4 Generation of the Anatomical Atlas

**Corpus:** The volume data sets of our image corpus were selected primarily by the first use-case of MEDICO which is support for *lymphoma* diagnosis. The selected data sets were picked randomly from all available studies in the medical image repositories of the University Hospital in Erlangen, Germany. The selection process was performed by radiologists at the clinic.

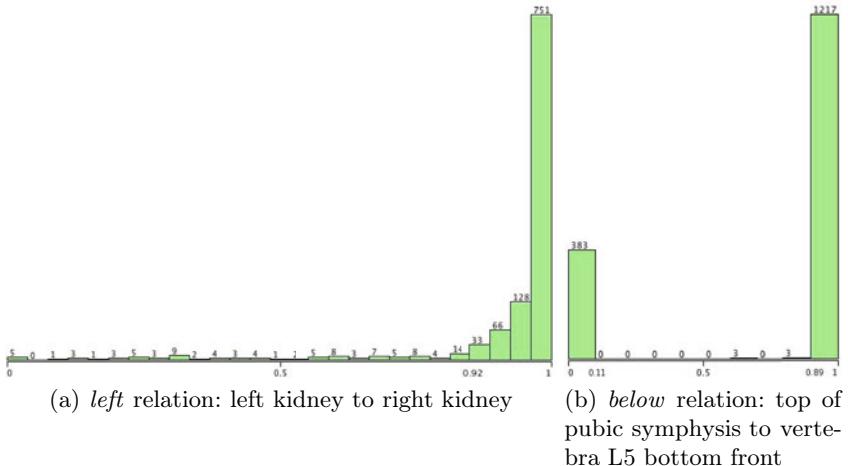
**Table 1.** Summary of corpus features

volume data available	777 GB
number of distinct patients	341
volume data sets	w/o duplicates 2,924 ⊂ 3,604 parseable ⊂ 6,611 total
detection results	60,145 total = 47,283 + 12,862 organs
distinct anatomical concepts	22 landmarks, 6 organs

Table 1 summarizes major quantitative features of the available corpus. Out of 6,611 volume data sets in total only 5,180 belonged to the modality CT which is the only one currently processable by our volume parser. Out of these, the number of volumes in which at least one anatomical entity was detected by the parser was 3,604. This results from the rationale of the parser which was in favor of precision and against recall. In our subsequent analysis we found that our corpus contained several DICOM volume data sets with identical Series ID. The most likely reason for this is that an error occurred during the data export from the clinical image archive to the directory structure we used to store the image corpus. To guarantee for consistent spatial entity locations, we decided to delete all detector results for duplicate identifiers. This further reduced the number of available volume data sets to 2,924.

**Quantitative Spatial Model:** We are using the relative position of anatomical entities in 3D space to generate a quantitative spatial atlas. To express the position of an entity  $E_1$  as a function of the location of entity  $E_2$  the six standard directional relations *above*, *below*, *left*, *right*, *in front*, and *behind* are modeled

<sup>2</sup> <http://xsb.sourceforge.net/>



**Fig. 3.** Histograms for stable and unstable relations

using fuzzy sets. The membership function of a relation is thereby determined by the angles  $\alpha_1, \alpha_2$  between the centroid of  $E_1$  and  $E_2$  as follows:

$$\mu_{rel}(\alpha_1, \alpha_2) = \begin{cases} \cos^2(\alpha_1) \cos^2(\alpha_2) & \text{if } \alpha_1, \alpha_2 \in [-\frac{\pi}{2}, \frac{\pi}{2}] \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

For the representation of a specific relation another angle  $\theta$  is needed stating the main direction, e.g., the main direction for *left* is  $\alpha_1 - \pi$  resulting in the membership function:

$$\mu_{left}(\alpha_1, \alpha_2) = \mu_{rel}(\alpha_1 - \pi, \alpha_2) \quad (2)$$

Due to space limitation we can only refer to [4] for more details of this approach. For every volume data set the pairwise relations between all detected anatomical entities are calculated to determine their mean values with the particular standard deviations. The examination of the results shows that most of the relations (1,005) are stable whereas 129 highly vary. We define a relation as stable if the standard deviation is lower than 0.2. For example, Fig. 3a depicts the stable *left* relation between the kidneys. The huge majority of values lie within the range of 0.92 and 1. By contrast, Fig. 3b shows that the degree of *below* between a landmark of the pubic symphysis and the vertebra L5 has a much higher variability between 0...0.11 and 0.89...1.

With this knowledge it is possible to detect anatomical anomalies by computing the difference *diff* between the output of a relation and the mean stored in the model and comparing the result with the standard deviation *stddev*. The conformity for relations is thereby represented as a fuzzy set using the membership function:

$$\mu_{conf}(diff, stddev) = \begin{cases} 1 & \text{if } |diff| \leq stddev \\ \frac{|stddev - diff|}{stddev} & \text{otherwise} \end{cases} \quad (3)$$

As long as  $diff$  is in the range of the standard deviation the instance is considered to be conforming to the model. If it is greater, the proportion between the two values will be assigned as degree of membership. This design reflects the fact that small changes are tolerable. Furthermore, this allows us to cope with relations varying at a high degree. Relations with a large variability, i.e., the probability that the difference is within the standard deviation is high. Therefore, the membership degree is also high. Based on these results it is possible to perform a simple anatomical conformity check which verifies the spatial relations in a volume parser result. The conformity of a volume is computed by computing the conformity checks' arithmetic mean of all spatial relations mentioned above and comparing this value with a threshold manually determined.

Thus, it was possible to identify 758 volumes which have anomalies, i.e., the conformity is less than 0.9. An examination of the volumes showed that either completely wrong entities are detected or that their locations are incorrect. Another potential source of error is the DICOM metadata. For some volumes, a wrong orientation matrix was specified in the header. This matrix is necessary to determine the patient's position during image acquisition and allows transforming the volume accordingly before parsing. If a volume is not in the normal position, all directional relations are rotated and the volume parsing returns incorrect positions.

**DICOM Metadata and Spatial Models:** So far we described the generation of an anatomical atlas based on sub-symbolic spatial features extracted from the image *content*. However, the DICOM standard describes *metadata* attributes which we assumed to provide additional useful features of a patient's anatomy. We analyzed gender, patient and study ID in order to determine the variability of the anatomy within a particular gender, along different individuals (inter-patient variability) and within the same individual (intra-patient variability) as in [10]. Therefore, the volumes were represented as feature vectors containing the euclidean distances and angles between the entities' centroids. We performed a cluster analysis and found that the metadata attributes had no visible correlation to the other features. To better understand this, we decided to perform a systematic comparison of metadata attributes and image content features which is discussed in the Evaluation section.

## 5 Evaluation

This section presents results from a systematic comparison of image metadata attributes with features extracted from the image content. At the same time this evaluation is an example for the versatility of ontology-based reasoning combined with integrated semantic metadata and image content representation. To our knowledge there exists no system so far allowing a similar comparison of such heterogeneous data as DICOM metadata attributes with high-level anatomical features extracted from the image content. Thus, we can only present our result and cannot compare them to an established gold standard.

**Gender Consistency:** This check compares the patient's gender as it is stored in the image volume's metadata header with evidence for a certain gender that can be deduced from the results of the volume parsing. If the data from these two sources differ either the volume parsing algorithm generated an incorrect result or the metadata is incorrect due to mistakes of the medical personnel who entered this data. In either case, a difference in the gender values can be used to give a hint to the user that it is not reliable and help to avoid potential misunderstandings.

This check makes use of the Prolog rule engine introduced in Section 3. The FMA already contains certain sets of concepts which are clearly gender specific. The concept **Female genital system** and all its transitive parts are a simple example. A similar concept exists for male subjects. The members of the **Set of male pelvic viscera** and the **Set of female pelvic viscera** are a little more complicated since they have some entries in common. Only those concepts occurring exclusively in one of the sets are gender specific. In our test we check for all detected organs and landmarks of a given volume, if any of the detected anatomical entities is gender specific. By means of a few Prolog rules we are able to express the conditions for this test in a very generic form. Instead of hard coding them the source code, the conditions are expressed in declarative rules which are easy to read and extend:

```

hasRegionalPart(X,Y) :- true(Y,uri('fma:regional_part_of'),X,_).
hasRegionalPart(X,Y) :- true(Z,uri('fma:regional_part_of'),X,_), hasRegionalPart(Z,Y).
male_entities(X) :- true(X,uri('fma:member_of'),uri('fma:Set_of_male_pelvic_viscer'),_),_),
\+ true(X,uri('fma:member_of'),uri('fma:Set_of_female_pelvic_viscer'),_).
male_entities(X) :- hasRegionalPart(uri('fma:Male_genital_system'),X).
female_entities(X) :- true(X,uri('fma:member_of'),uri('fma:Set_of_female_pelvic_viscer'),_),_),
\+ true(X,uri('fma:member_of'),uri('fma:Set_of_male_pelvic_viscer'),_).
female_entities(X) :- hasRegionalPart(uri('fma:Female_genital_system'),X).
hasGender(X, male) :- male_entities(X).
hasGender(X, female) :- female_entities(X).
```

At the time of writing the only gender specific organ which our volume parser can detect is the male prostate. But due to the generic approach, our technique can immediately make use of additional detectable anatomical entities without any changes to the checking code or the rules. In 386 ( $\approx 13\%$ ) of 2,924 volumes in total the reasoning based on the detected anatomical entities claimed for a male image content whereas the respective metadata attribute was set to female.

**Body Region:** Another example for the application of reasoning to check metadata attributes is the detection of the body region(s) which is (are) covered by a volume data set. The DICOM standard specifies the attribute *Body Part Examined* for this information. It is important to note that this attribute is defined as optional in to official standard. We found that out of 2,924 for 150 ( $\approx 5\%$ ) this attribute was not set at all. Based on the information of the FMA and the Prolog rules shown below, we additionally compared to compare the body region of a given volume data set using the results of the volume parser. For each anatomical entity detected by the image parser, we check to which body region it belongs, collect the output and compare it the the metadata attribute.

```

thoracic_organ(X) :- true(X,uri('fma:member_of'),uri('fma:Set_of_thoracic_viscerा'),_).
abdominal_organ(X) :- true(X,uri('fma:member_of'),uri('fma:Set_of_viscerा_of_abdomen'),_).
pelvic_organ(X) :- true(X,uri('fma:member_of'),uri('fma:Set_of_pelvic_viscerा'),_).
body_region(X, abdomen) :- abdominal_organ(X).
body_region(X, thorax) :- thoracic_organ(X).
body_region(X, pelvis) :- pelvic_organ(X).

```

In 522 cases ( $\approx 18.5\%$ ) the image content analysis yielded a body region that was contradictory to the metadata. In most of these cases the body region was specified as “NECK” or “HEAD”. Both are body regions for which we are not able to detect any anatomical entity. Due to the use of statistic algorithms the results of the volume parser are per se not absolutely reliable. However, a discrepancy between the body region deduced from the image content and the image metadata can be used to inform the user of a potential problem with either the metadata or the parsing result.

## 6 Conclusion and Future Work

We presented an approach for the integration of medical image metadata and the results of content analysis using Semantic Web technologies. This combination allows leveraging information from existing medical ontologies to detect inconsistencies between metadata attributes and image content. These checks are expressed separately from the application logic using Prolog rules. This makes them easy to maintain and extend. Using this approach we were able to detect both missing metadata attributes and potentially incorrect attribute values.

Our evaluation is based on a large corpus of over 750 GB of clinical image data. From the output of the volume parsing system we generated a qualitative atlas of human anatomy. Imprecision is expressed using fuzzy membership functions. So far this atlas is limited to directional information between the centroids of pairs of anatomical entities. However, it has already been shown to be able to identify anatomical anomalies.

For the future we plan to use surface models instead of centroids for the anatomical atlas. Furthermore, more sophisticated conformity checks will be applied for finding inconsistencies. We also aim to represent distances using fuzzy relationships in addition to the directional information. Once these relations are available in an abstract semantic format they become available for the Prolog reasoning approach discussed above. Eventually, our goal is to combine fuzzy distance and relational information with anatomical knowledge from medical ontologies for spatio-anatomical reasoning. But this is still work in progress.

Among our next steps is also a user evaluation of clinical applications making use of the reasoning, e. g., to support radiologists by suggesting anatomical concepts and relations during manual image annotation. Furthermore, our approach could be used to generate warnings for manually generated image annotations in case they do not conform to the spatial anatomical model. A clinical evaluation of these features is planned for the near future.

## References

1. Andriole, K.P., Morin, R.L., Arenson, R.L., Carrino, J.A., Erickson, B.J., Horii, S.C., Piraino, D.W., Reiner, B.I., Seibert, J.A., Siegel, E.: Addressing the Coming Radiology Crisis – The Society for Computer Applications in Radiology Transforming the Radiological Interpretation Process (TRIP<sup>TM</sup>) Initiative. *Journal of Digital Imaging* 17(4), 235–243 (2004)
2. Bloch, I.: On fuzzy distances and their use in image processing under imprecision. *Pattern Recognition* 32(11), 1873–1895 (1999)
3. Bloch, I.: Fuzzy spatial relationships for image processing and interpretation: a review. *Image and Vision Computing* 23(2), 89–110 (2005)
4. Bloch, I., Ralescu, A.: Directional relative position between objects in image processing: a comparison between fuzzy approaches. *Pattern Recognition* 36(7), 1563–1582 (2003)
5. Broekstra, J., Kampman, A., van Harmelen, F.: Sesame: A generic architecture for storing and querying RDF and RDF Schema. In: Horrocks, I., Hendler, J. (eds.) ISWC 2002. LNCS, vol. 2342, pp. 54–68. Springer, Heidelberg (2002)
6. D. S. Committee: Strategic document. Technical report (May 2009)
7. da Luz, A., Abdala, D.D., Wangenheim, A.V., Comunello, E.: Analyzing dicom and non-dicom features in content-based medical image retrieval: A multi-layer approach. In: 19th IEEE International Symposium on Computer-Based Medical Systems, CBMS 2006, pp. 93–98 (2006)
8. Güld, M., Kohnen, M., Keysers, D., Schubert, H., Wein, B., Bredno, J., Lehmann, T.: Quality of DICOM header information for image categorization. Proc. SPIE 4685(39), 280–287 (2002)
9. Krishnapuram, R., Keller, J.M., Ma, Y.: Quantitative analysis of properties and spatial relations of fuzzy image regions. *IEEE Transactions on Fuzzy Systems* 1(3), 222–233 (1993)
10. Lorenz, C., Krahnstover, N.: Generation of point-based 3d statistical shape models for anatomical objects. *Computer Vision and Image Understanding* 77(9), 175–191 (2000)
11. Möller, M., Folz, C., Sintek, M., Seifert, S., Wennerberg, P.: Extending the foundational model of anatomy with automatically acquired spatial relations. In: Proc. of the International Conference on Biomedical Ontologies (ICBO) (July 2009)
12. Möller, M., Regel, S., Sintek, M.: Radsem: Semantic annotation and retrieval for medical images. In: Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., Simperl, E. (eds.) ESWC 2009. LNCS, vol. 5554, pp. 21–35. Springer, Heidelberg (2009)
13. Object Management Group, Inc: Common object request broker architecture: Core specification. Version 3.0.3 (March 2004) (online)
14. Rosse, C., Mejino, J.L.V.: The Foundational Model of Anatomy Ontology. In: Anatomy Ontologies for Bioinformatics: Principles and Practice, vol. 6, pp. 59–117. Springer, Heidelberg (December 2007)
15. Seifert, S., Kelm, M., Möller, M., Mukherjee, S., Cavallaro, A., Huber, M., Comaniciu, D.: Semantic annotation of medical images. In: Proc. of SPIE Medical Imaging, San Diego, CA, USA (2010)

# Parallel Processing with CUDA in Ceramic Tiles Classification

Tomislav Matić and Željko Hocenski

University J.J. Strossmayer, Faculty of Electrical Engineering, Kneza Trpimira 2b,  
Osijek, Croatia  
`{tomislav.matic, zeljko.hocenski}@etfos.hr`  
<http://zapr.etfos.hr>

**Abstract.** This paper describes the implementation of an algorithm for surface error detection on ceramic tiles in CUDA (Compute Unified Device Architecture). It compares the differences between the CPU and the GPU algorithm implementation, analyzes the features of CUDA GPU and summarizes the general programming model of CUDA. Paper presents the speed up gained in favor of the GPU algorithm implementation. Implemented algorithm used in this paper written in C is relatively simple, and for test results version for the CPU was made and the GPU version. The results show the speed up of the computation compared with the CPU that increases as the image size increases, with the maximum speed up of 4,89 times.

**Keywords:** image processing, parallel programming, CUDA, ceramic tiles, processing speed up, error detection, quality control.

## 1 Introduction

Driven by the demand for real-time, high-definition 3D graphics, the programmable GPU has evolved into a highly parallel, multithreaded; many core processor with tremendous computational horsepower and very high memory bandwidth. GPU is specialized for compute-intensive, highly parallel computation and therefore designed such that more transistors are devoted to data processing than the data caching and flow control [1].

GPU is especially well-suited to address problems that can be expressed as data-parallel computations with high arithmetic intensity (the ratio of arithmetic operations to memory operations) [2]. Because the same program is executed for each data element, there is a lower requirement for sophisticated flow control. For these reasons many problems can be computed in the GPU: push-relabel algorithm for min-cut/maxflow algorithm for graph-cuts [3], fast sorting algorithms of large lists [4], two-dimensional fast wavelet transform [5], and other examples in image and media processing applications, general signal processing, physics simulation, etc.

One example is a typical molecular dynamics simulation of the polymer model. Molecular dynamics simulations are inherently parallel and therefore suitable for CUDA implementation. Speed up of 32 times has been accomplished in favour of CUDA on a GeForce 8800 GTX graphics card relative to 3.0-GHz Xeon processor [2].

The emergence of CUDA technology can meet the demand of general purpose programming on the GPU. CUDA brings the C-like development environment to programmers for the first time, which uses a C compiler to compile programs, and replaces the shader languages with C language and some CUDA extended libraries. Users needn't map programs into graphics APIs, so program development becomes more flexible and efficient. More than one hundred processors resided in CUDA graphics card, schedules hundreds of threads to run concurrently, resolving complex computing problems [6].

For the need of speeding up the process of finding surface defects on ceramic tiles and based on the past examples [7] CUDA program was developed, implemented and tested in NVIDIA GeForce 9800GT graphic card.

Rest of the paper is organized as follows. Section 2 covers CUDA parallel architecture, it's organization and specifications. Section 3 describes reasons for automation of ceramic tile quality control. Section 4 gives detailed explanation of the implemented algorithm in the CPU and the GPU. Testing of the algorithm is explained in section 5. Experimental results are analyzed in section 6, and section 7 summarizes the work and concludes the paper.

## 2 CUDA Architecture

The CUDA architecture is built around a scalable array of multithreaded Streaming Multiprocessors (SMs). A multiprocessor consists of eight Scalar Processor (SP) cores, two special function units for transcendentals, a multithreaded instruction unit, and on-chip shared memory.

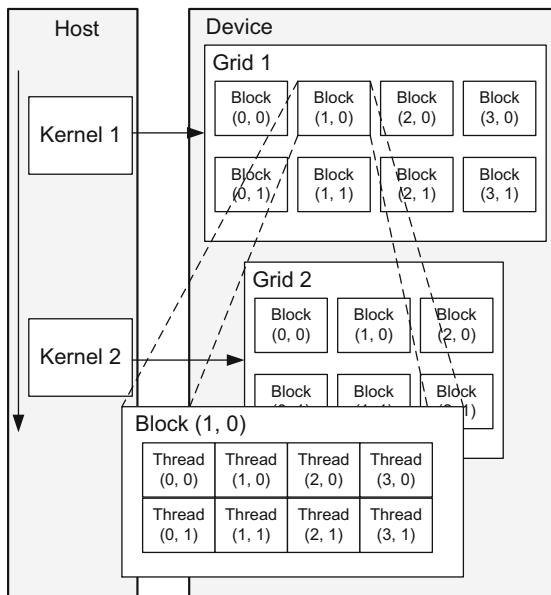
CUDA extends C by allowing the programmer to define C functions, called kernels, that, when called, are executed N times in parallel by N different CUDA threads, as opposed to only once like regular C functions. A kernel is defined using the `_global_` declaration specifier and the number of CUDA threads for each call is specified using a new `<<<...>>>` syntax [1].

Threads are organized in blocks, and blocks in a grid Fig. 1. All threads of a block have consecutive IDs and reside on the same multiprocessor. Threads within a block can cooperate among themselves by sharing data through shared memory and synchronizing their execution to coordinate memory accesses (`_syncthreads()`). The streaming multiprocessor creates, manages, and executes concurrent threads in hardware with zero scheduling overhead. It implements the `_syncthreads()` barrier synchronization intrinsic with a single instruction. The streaming multiprocessor employs a new architecture, SIMT (single-instruction, multiple-thread) akin to SIMD (Single Instruction, Multiple Data) vector organizations. The multiprocessor SIMT unit creates, manages, schedules, and executes threads in groups of 32 parallel threads called *warps*. It splits thread blocks into warps containing threads of consecutive, increasing thread IDs with the first warp containing thread 0.

There are some limitations. Block can have a maximum of 512 threads. Multiprocessor can have 16KB of shared memory and 8192 registers. The maximum number of active blocks per multiprocessor is 8, 24 active warps and 768 active threads. Maximum number of active blocks also depends on the size of shared memory and registers needed [1].

It also needs to be taken into account the speed of the data transfer from the Host to the Device (Device memory). For PCI-E 2.0 16x the max speed is 8 GB/s.

Data transfer (Host to the Device memory) is performed using DMA (Direct Memory Access) and can take place concurrently with kernel processing. Transferred data is persistent on the GPU memory remaining available for subsequent kernels [8].



**Fig. 1.** Programming model and thread organization

### 3 Ceramic Tiles Quality Control

In modern ceramic tiles industry most of production stages are automated. One of the most demanding stages is quality control inspection and classification. This stage includes visual inspection of ceramic tile surface defects and dimensional discrepancy, edge and corner defects, presence of cracks and scratches, dot shaped and blob defects etc. The human vision is used, because quality control process is very complex and demands are changeable on line during production. But human vision is not always reliable and is subjective.

Automation of quality control stage leads to machine vision systems. Such systems are consisted of digital cameras for image acquisition and one or more computers with algorithms for image processing, failure detection and product classification [9, 10].

In this paper algorithm implemented in CUDA is used for speed comparison with the CPU. The algorithm only restricts on the surface defects (spots) of one colour and light randomly textured ceramic tiles. The future work will take into account other problems concerning visual inspection and classification of ceramic tiles [11].

## 4 Algorithm Description

The implemented algorithm is based on the assumption that the surface errors on the ceramic tile are some spots whose brightens values are different from the overall brightens values.

The algorithm is using .bmp file format of the grey scale image. Values of the pixels are read into a memory vector. Then the picture is divided into elements 4x4, 8x8 or 16x16 pixel size. Average value of the pixel elements is calculated and saved, and also the total sum of all elements average value Fig. 2. The Total Sum is then divided by the number of elements to get the Average Sum.

User needs to set the threshold of the defects. Threshold is determined by testing the image with no defects using the algorithm below and selecting the smallest value of the threshold that didn't produce any errors on that image (for multiple defect free images threshold is determined as the maximum of the smallest values that didn't produce any errors). Threshold is used for detecting errors in this way:

```
for k=0 to number_of_elements-1
    if(Avg(Ek) < Average Sum - threshold) defect++;
    if(Avg(Ek) > Average Sum + threshold) defect++;
loop
```

$E_k$  is the  $k$ -th element of the image Fig. 2. This loop gives the number of errors detected on the surface of the ceramic tile (one or more errors can make a defect).

Algorithm for the CPU and the GPU is written in C language and compiled with the NVCC compiler developed by NVIDIA. Code compiled with the NVCC is separated to the CPU and GPU.

### 4.1 CPU Implementation

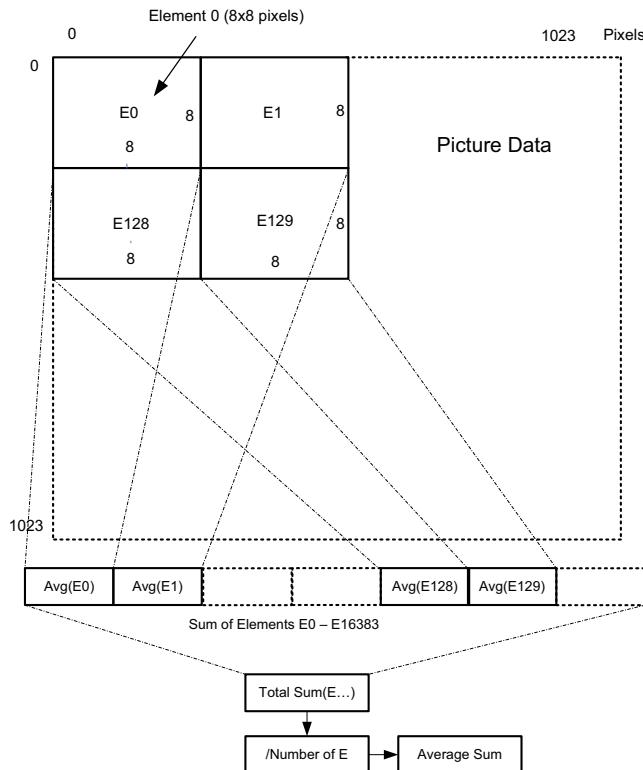
CPU implementation of the algorithm is done using one function *CPUTDefect*. The function receives: picture data, width in pixels, height in pixels, element height, element width and the threshold.

Function returns the number of errors found. It does all of the calculation in 5 loops, calculation of the average values of elements, total sum of average values of elements and the average sum as shown in Fig. 2.

To reduce the number of iterations for the calculation of average value of the element (64 iterations are needed for size of elements 8x8) every iteration in the dedicated loop does 4 additions (number of iteration decreases to  $64/4=16$ ). The way the loop does these additions is shown in Fig. 3.

### 4.2 GPU Implementation

Because algorithm needs to be parallelized GPU implantation of the algorithm is done using three kernels. First kernel (GPUeAvg) calculates the average value of elements, second kernel (GPUTotSum) calculates total sum of the average values (based on the results of GPUeAvg kernel) and the third kernel (GPUCompare) does the comparing



**Fig. 2.** The image processing algorithm description

using the threshold value and returns the number of errors. Image data is copied from the host memory to the device global memory (data is persistent in the device global memory until all three kernels are finished with the computation) and the result is copied from the device memory to the host.

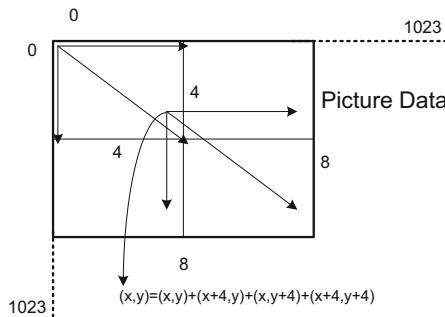
If the element size is 8x8 then the GPUeSum kernel has the block size 4x4=16 (number of threads in a block). Number of blocks is calculated using the formula:  $\text{picture\_width} * \text{picture\_height} / (8x8)$ .

Function first does the four addition method shown in Fig. 3 then uses the reduction method to sum all of the pixel values and finally calculates the average value of the element [12].

GPUMTotSum takes the vector that holds all average values of elements (from the GPUeAvg kernel) and using the reduction method calculates the total sum and based on the number of elements calculates the average sum.

GPUCompare function takes the average values of elements vector, average sum and the threshold and returns the number of errors. This value is copied back to the PC main memory and displayed to the user.

The calculation of the functions is done in parallel, but the three functions are done sequentially. First the GPUeAvg is executed then GPUTotSum and finally GPUCompare.



**Fig. 3.** The four addition method

## 5 Testing of the Algorithm

Testing of the algorithm was done on a desktop computer with installed NVIDIA GeForce 9800GT graphics card and Intel Core 2 Quad Q6600 processor (one core used).

9800GT graphics card has 14 multiprocessors, 57,6 GB/s memory bandwidth, 512 MB of memory and PCI-E 2.0 16x bus [13]. Intel Core 2 Quad Q6600 processor has 4 cores on 2,4 GHz clock speed, with 8 MB of L2 cache and 1066 MHz bus speed [14].

Three types of images were made that simulate surface of the ceramic tile for the analysis: white, light blue (in text blue) and light grey textured (in text textured). All of the images were converted to 256 colour grey scale and all of them have defects (black spots of different sizes). Three sizes of the images were used: 1024x1024, 2048x2048 and 4096x4096 pixels.

For this algorithm image width must be divisible by the elements width, and image height must be divisible by the elements height.

Measurement of computation time, data transfer time (host-device) and the number of errors found were taken for 5 iterations for the size of elements 4x4, 8x8, 16x16. Based on gathered results average speed up value was calculated.

Figure 4 shows the testing stages. In stage 1 the original picture with some defects was loaded to the device memory (1024x1024 pixels textured image). Defects are black spots (circles) with radius pixel size 11, 6 and 3 marked with the red circle (on the bottom of Fig. 4 defects are magnified). If the ceramic tile size is 20x20 cm, sizes of radii of the defects on the Fig. 4 are 2.15 mm, 1.17 mm and 0.56 mm. Stage 2 represents dividing the original image into elements and calculating the average value of the elements. The result of this stage is a smaller image size (1024/element width)x(1024/element height). Because of the averaging the image is blurred but the defects are still visible and marked with the red circles. In stage 3 errors are found based on the set threshold, the total average sum and the data from stage 2. Image in stage 3 represents found errors-black pixels on white background. The errors correspond with the defects in the original image (marked with red circles). The images in stage 2 and 3 were made for purpose of preliminary testing and are not included in the calculation of the speed up.

**Table 1.** Algorithm testing results for 1024x1024 pixels image

1024x1024 Blue	#	4x4 elements		8x8 elements		16x16 elements	
		Time	Errors	Time	Errors	Time	Errors
	1	3,6	8	2,5	5	2,3	3
	2	3,6	8	2,5	5	2,3	3
Nvidia 9800GT	3	3,5	8	2,6	5	2,4	3
	4	3,6	8	2,5	5	2,3	3
	5	3,6	8	2,6	5	2,3	3
Average GPU:		3,58	8	2,54	5	2,32	3
	1	9	8	6	5	5	3
Intel Core 2 Quad Q6600	2	9	8	5	5	5	3
1 Thread	3	8	8	6	5	5	3
	4	9	8	6	5	5	3
	5	8	8	6	5	5	3
Average CPU:		8,6	8	5,8	5	5	3
Speed Up:		2,40		2,28		2,16	
Avg. PC memory - GPU memory transfer time		0,9		0,9		0,9	
Threshold		1		1		1	

**Table 2.** Algorithm testing results for 4096x4096 pixels image

4096x4096 Texture	#	4x4 elements		8x8 elements		16x16 elements	
		Time	Errors	Time	Errors	Time	Errors
	1	39,8	93	22,1	31	17,9	9
	2	39,8	93	22,1	31	17,8	9
Nvidia 9800GT	3	39,7	93	22,2	31	17,9	9
	4	39,8	93	22,1	31	17,9	9
	5	39,7	93	20,7	31	18	9
Average GPU:		39,76	93	21,84	31	17,9	9
	1	144	93	97	31	89	9
Intel Core 2 Quad Q6600	2	143	93	98	31	85	9
1 Thread	3	141	93	100	31	85	9
	4	141	93	95	31	88	9
	5	141	93	96	31	91	9
Average CPU:		142	93	97,2	31	87,6	9
Speed Up:		3,57		4,45		4,89	
Avg. PC memory - GPU memory transfer time		9		9		9	
Threshold		51		39		25	

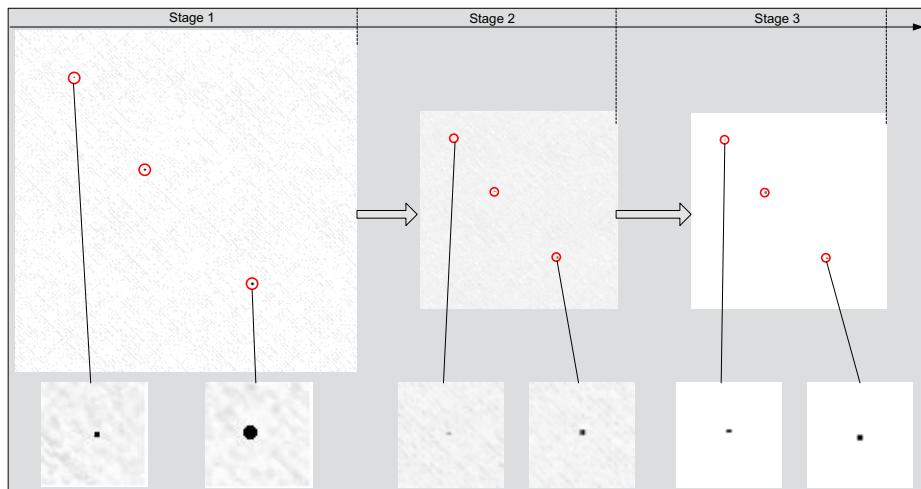
## 6 Results

The results of the testing are shown in Table 1, Table 2 and Fig. 5. The tables show all 5 iterations of the calculations, and the Fig. 5 shows all of the average speed up values of the tested images. Time values in the tables are expressed in ms. The Errors column shows the number of errors detected by the algorithm. All images have had 3 defects, black spots. In Table 2 it can be seen that the number of errors for element

size 4x4 pixels is 93 and in Table 1 only 8. It can be explained by the fact that the results in Table 2 are for image size 4096x4096 pixels, and Table 1 for the image size 1024x1024 pixels and accordingly the size in pixels of the defects radii is greater (for the same defect, number of errors is greater). Number of errors also decreases with the increase of the elements size and in Table 1 for the element size 16x16 pixels number of errors is the same as the number of defects (one error per defect).

The threshold was chosen manually, testing images of the same type (textured, blue, white) that didn't have any defects. The minimum threshold that didn't produce any errors was chosen. Same threshold value was used for the CPU and the GPU implementation and, as expected, number of errors is also the same. In Table 1 threshold value is 1 because the image with no errors had a unified one colour fill. For the textured image threshold value changes with the element size because of the average brightness change in the elements. Intended use of the algorithm is for unified one colour ceramic tiles and for light randomly textured ceramic tiles. For other types of surface morphology and other types of defects different algorithms need to be tested to get the best execution time and defect detection (statistical approach, wavelet analysis, neural networks etc.).

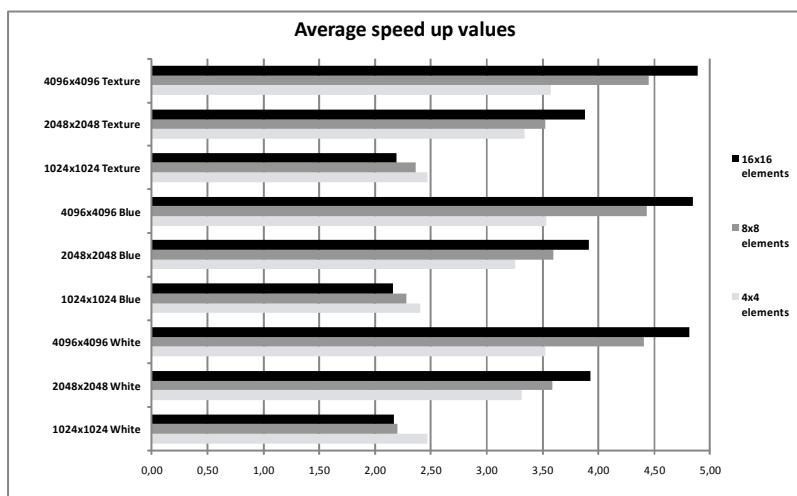
It can be seen from the Fig. 5 that the average speed up is increasing as the image size increases because of the amount of data that needs to be calculated, parallel calculation has greater effect. Also the speed up increases as the element size increases. The reason for this can be found in the block size (block size increases as the element size increases) in the CUDA kernels explained in heading 2. Only discrepancy from the above mentioned are the result of the average speed up for the image size 1024x1024 pixels, it doesn't increase with the element size. For this example CPU calculation time change is smaller as the element size increases than the GPU calculation time change because of the smaller amount of data that needs to be processed.



**Fig. 4.** Algorithm testing stages

During the testing phase implemented algorithm managed to find all the defects on the unified colour images. For the textured images algorithm missed the smallest defect (3 pixels size) for the image size 1024x1024 pixels and the element size 16x16 pixels and 8x8 pixels.

The results of the speed up are only for the computational time, data copying times (host – device) are not included in the average speed up results. Data copying times for image sizes 1024x1024 pixels are 0,9 ms, 2048x2048 pixels are 2,7 ms and for 4096x4096 pixels 9 ms and it does not have a great effect on the final result of the average speed up (data copying time is increasing linearly with the amount of data to be copied, data transfer speed is the same).



**Fig. 5.** Average speed up values for all tested images

## 7 Conclusion

Created algorithm tested in this paper was used for the comparison of the computational speed between the CPU and the GPU version. Algorithm returns the number of errors found on the image. All of the errors are consistent with the defects on the image and no additional defects are found (this is a case when the appropriate value for the threshold was chosen). For the unified colour pictures algorithm managed to find all of the defects no matter what was the size of the images. For the textured image, size of the image depends on the smallest defect (for the image size 1024x1024 pixels and element size 8x8 and 16x16 pixels algorithm didn't detect errors for the 3 pixels radius defect). When the size of the image increased, algorithm detected all defects for all element sizes.

Average speed up values are increasing as the size of the images increases. Reasons for this can be found in the amount of parallelization achieved – algorithm calculations (kernels) are executed by more threads. Also as the size of the element increases the

block size increase Fig. 1 (algorithm calculations for the element is done by more threads) and the average speed up is larger.

Maximum average speed up value is 4,89 calculated for image size 4096x4096 pixels. It can be concluded as the number of calculation for the elements increase the speed up values also increase (elements data are in shared memory and very little time is used for the memory transfer, most of it is spent on arithmetic operations). More complex algorithms should be implemented in CUDA that can detect different surface and structural defects. Also the stages of image preparation and classification that are compute intensive needs to be implemented in CUDA. On the basis of the gathered results it can be concluded that further development of CUDA based algorithms will give promising results in ceramic tiles image processing, especially when the technology further develops.

CUDA has shown to be capable of speed ups for the simple algorithm, future work will include the implementation of the whole system for ceramic tile quality control inspection and classification.

## References

1. NVIDIA Corporation, NVIDIA CUDA Programming Guide Version 2.1, NVIDIA Corporation (2008)
2. Garland, M., Le Grand, S., Nickolls, J., Anderson, J., Hardwick, J., Morton, S., Phillips, E., Zhang, Y., Volkov, V.: Parallel Computing Experiences With Cuda. IEEE Micro 28, 13–27 (2008)
3. Vineet, V., Narayanan, P.J.: CUDA Cuts: Fast Graph Cuts on the GPU. In: Proc. Computer Vision and Pattern Recognition Workshops, Anchorage, pp. 1–8 (June 2008)
4. Sintorn, E., Assarsson, U.: Fast parallel GPU-sorting using a hybrid algorithm. Journal of Parallel and Distributed Computing 68, 1381–1388 (2008)
5. Franco, J., Bernabe, G., Fernandez, J., Acacio, M.E.: A Parallel Implementation of the 2D Wavelet Transform Using CUDA. In: 17th Euromicro International Conference on Parallel, Distributed and Network-based Processing, Weimar, pp. 111–118 (2009)
6. Yang, Z., Zhu, Y., Pu, Y.: Parallel Image Processing Based on CUDA. In: 2008 International Conference on Computer Science and Software Engineering, Wuhan, vol. 3, pp. 198–201 (2008)
7. Hocenski, Ž., Aleksi, I., Mijaković, R.: Ceramic Tiles Failure Detection Based on FPGA Image Processing. In: Proc. Int. Conf. on Industrial Electronics- ISIE 2009, Seoul, pp. 2169–2174 (2009)
8. Che, S., Boyer, M., Meng, J., Tarjan, D., Sheaffer, J.W., Skadron, K.: A performance study of general-purpose applications on graphics processors using CUDA. J. Parallel Distrib. Comput. 68, 1370–1380 (2008)
9. Hocenski, Ž., Keser, T., Baumgartner, A.: A Simple and Efficient Method for Ceramic Tile Surface Defects Detection. In: Proc. 2007 International Symposium on Industrial Electronics, Vigo, pp. 1606–1611 (2007)
10. Hocenski, V., Hocenski, Ž.: Sustainable Development Technology in Ceramic Tiles Industry. In: Proc. 32nd Int. Conf. on Automation Technology, Automation 2009, Tainan, Taiwan, R.O.C: Chinese Institute of Automation Engineering, National Cheng Kung University, pp. 283–288. University of Cincinnati (2009)
11. Keser, T.: Automated Intelligent System For Ceramic Tiles Classification, University J.J.Strossmayer in Osijek, Faculty of Electrical Engineering, thesis, Osijek (2009)

12. Harris, M.: Optimizing Parallel Reduction in CUDA, NVIDIA Developer Technology, [http://developer.download.nvidia.com/compute/cuda/1\\_1/Website/projects/reduction/doc/reduction.pdf](http://developer.download.nvidia.com/compute/cuda/1_1/Website/projects/reduction/doc/reduction.pdf)
13. NVIDIA Corporation, Specifications NVIDIA GeForce 9800GT, NVIDIA Corporation, [http://www.nvidia.com/object/product\\_geforce\\_9800gt\\_us.html](http://www.nvidia.com/object/product_geforce_9800gt_us.html)
14. Intel Corporation, Intel Core2 Quad Processor Specifications, Intel Corporation, <http://www.intel.com/products/processor/core2quad/specifications.htm>

# Signal Receiving and Processing Platform of the Experimental Passive Radar for Intelligent Surveillance System Using Software Defined Radio Approach

Boguslaw Szlachetko and Andrzej Lewandowski

Institute of Telecommunication, Teleinformatics and Acoustics,

Wroclaw University of Technology,

Wroclaw, Poland

{boguslaw.szlachetko, andrzej.lewandowski}@pwr.wroc.pl

<http://www.itta.pwr.wroc.pl/>

**Abstract.** This document presents a signal receiving and processing platform for an experimental FM radio based multistatic passive radar utilizing Software Defined Radio. This radar was designed as a part of the intelligent surveillance system. Our platform consists of a reconfigurable multi-sensor antenna, radio frequency (RF) front-end hardware and personal computer host executing modified GNU Radio code. As the RF hardware (receiver and downconverter) the Universal Software Radio Peripherals were used. We present and discuss different approaches to construct the multichannel receiver and signal processing platform for passive radar utilizing USRP devices and GNU Radio. Received signals after downconverting on the FPGA of the USRP are transmitted to the PC host where second stage of data processing takes place: namely digital beamforming. Digital beamforming algorithm estimates echo signals reflected from flying target. After estimating echo signals Range-Doppler surfaces can be computed in order to estimate the target position.

**Keywords:** passive radar, multistatic radar, software defined radio, GNU Radio, beamforming.

## 1 Introduction

The concept of bistatic passive radars [1] are well known from the beginning of the radar research, however, just now they do receive the deserved attention. The main reasons are advances in hardware and digital signal processing, which makes it possible to use bistatic passive radars in real-world applications.

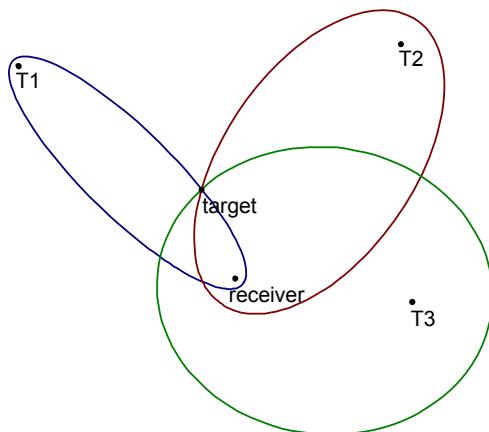
In bistatic passive radars various types (e.g.: VHF, UHV, FM, GSM) of illuminators of opportunity [2][3] can be used which decreases stealth capabilities of targets [4]. The other advantage stems from the fact, that the receivers are passive so they cannot be detected by the Anti Radiation Missiles [5].

The goal of our research was to create a receiver and signal processing system of FM radio based passive radar. This passive radar was designed as a front-end of an experimental intelligent surveillance system (under development). Our platform utilizes Software Defined Radio approach. i.e. the main part of its functions (down-converting, filtering, beamforming etc.) is implemented in software. This approach offers high reconfigurability because in this case functionality of the system is defined by the software, not the hardware [6]. One of the most popular and widely used SDRs is GNU Radio [7][8]. GNU Radio software development kit consists of signal processing blocks written in high level languages C++ and Python. This software is hardware independent the main assumption regarding hardware is having sufficient resources (processor speed and memory).

## 2 Signal Receiving and Processing System of the Experimental Passive Radar

Contours of constant bistatic range are ellipses. So, in order to estimate target location, the target must be illuminated by at least three independent transmitters (FM radio stations located on different azimuth angles) which gives the system known as a multistatic radar. Given position of the receiver the location of the target can be determined by the crossing of the bistatic range ellipses as shown in (Fig. 1). Because of constant changes of target position the adaptive target tracking system have to be used as described in [2].

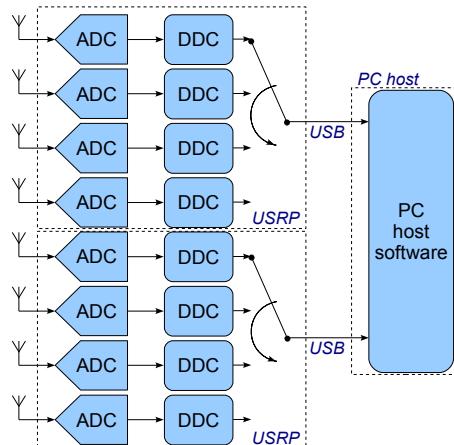
In a classical approach, an omnidirectional antenna receives the signal arriving directly from the transmitter and the signal reflected by the target object simultaneously. Hence, a critical limiting factor of such a system is the unwanted interference in the echo signal due to the direct reception of the FM radio signal. Moreover, the received direct signal is much stronger than the signal incoming



**Fig. 1.** Contours of constant bistatic ranges

from the target. Our solutions of this problem is to employ digital beamforming algorithm. Given the positions of transmitters, the antenna pattern can be formed in a such way that the zero null depths will be located directly on the transmitters' azimuth angles. As a result, the unwanted direct signals will be suppressed.

During our research we create the signal receiving and processing platform (Fig. 2) for a multistatic FM radio based passive radar. This platform consists of a multisensor antenna, a multichannel FM receiver and a personal computer (PC) host. The multisensor array can be equipped with from four to eight dipoles of  $\lambda/2$  length, that can be placed in the uniform linear or circular array. The array is easily reconfigurable: one can change the antenna pattern, the number of sensors and the distance between them. As a receiver (i.e. RF frontend, downconverting and preprocessing hardware) the Universal Software Radio Peripheral (USRP) devices were chosen. Received signals after downconverting are transferred to the PC host where after beamforming adaptive signal processing algorithms can be used for estimating positions of target objects.



**Fig. 2.** Experimental passive radar system utilizing two USRPs and PC-host computer

## 2.1 Multichannel Receiver

As mentioned earlier, we used the Universal Software Radio (USRP) devices [9] as a hardware platform of the receiver. The USRP is a hardware device designed for digital acquisition of radio signals, downconverting them and sending baseband signals via USB to the personal computer (PC) [10]. The PC host can perform additional processing (e.g. beamforming or demodulation) of the baseband signals. The USRP is commonly used as a digital receiver in GNU Radio projects.

The USRP consists of a motherboard and up to four daughterboards. The motherboard is equipped with: four 12-bit 64MS/s analog-to-digital converters

(ADC), four 14-bit, 128MS/s digital-to-analog converters DAC, a Cyclone FPGA from Altera and an USB 2.0 controller for communication with the PC host. The daughterboards are the RF frontend hardware of a USRP. There are several daughterboards available: receivers, transmitters or transceivers. In our project we use Basic RX daughtercards (1-250 MHz receiver).

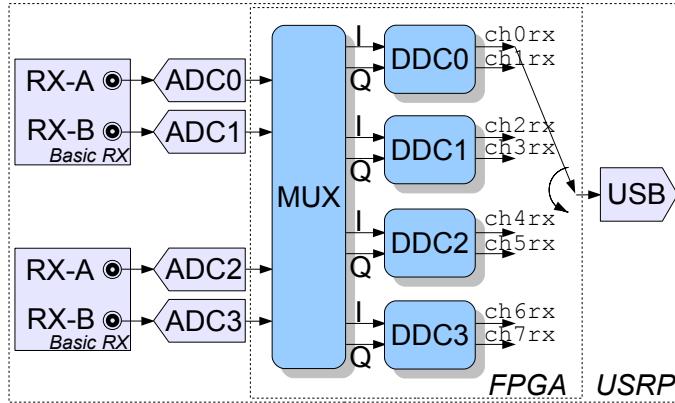
Signals from A/D converters are downconverted to the baseband in Digital Down Converter (DDC) blocks implemented on the USRP's FPGA. DDC blocks are implemented with 4 stages cascaded integrator-comb (CIC) filters [11]. DDC converts the signal of desired FM station down to the baseband, then filters it using halfband filter (HBF) and decimates the signals. The decimation rate of DDC block can be chosen from 8 to 256. As a result the in-phase -  $I$  and quadrature -  $Q$  signals are obtained.

It seems that the simplest method to achieve simultaneous acquisition on more than four channels is utilizing two USRP platforms. Those two USRPs have to use the synchronized sample clock according to [9]. This solution however is not as straightforward as it may seem. Single USRP can provide simultaneous sampling on four RX (receiving) channels. Unfortunately if one tries to synchronize two USRP platforms then using standard unmodified GNU Radio software only 4-channel system can be achieved. There are two reasons of such surprising result. First, one channel in each USRP is used for the special counter synchronizing two independent sample streams from two USB ports (to synchronize counters of the USRPs one have to connect pin io15 of the connector J25 on the daughterboard in the slot A of the USRP with the same pin on the second USRP). Second, the original GNU Radio code allows developers to use only two RX channels of one USRP synchronized with another USRP device.

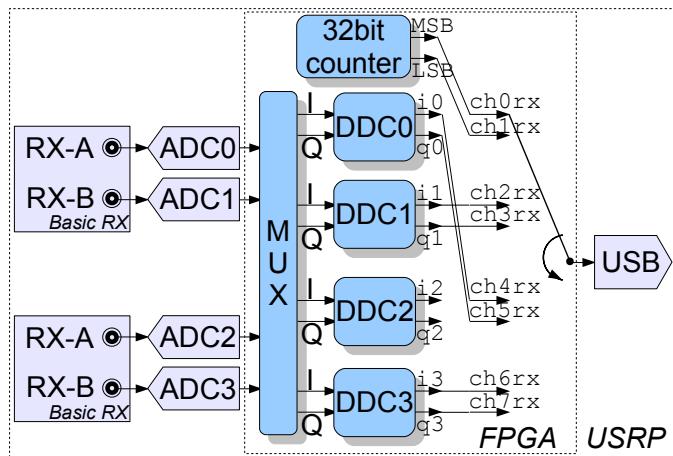
**6-Channel Receiver Using Two USRPs.** In Fig. 3 the standard FPGA's configuration of a single USRP device for simultaneous receiving signals on four channels is presented. Default configuration settings can be found in the file `config.vh` in the GNU radio project folder. This default configuration let us compile project with two receiving channels and two transmitting channels. One have to uncomment the proper line of the file `config.vh` to configure four receiving channels only - in this case there are no transmitting channels (as presented in Fig. 3). Each channel consists of 16-bit in-phase signal  $I$  and 16-bit quadrature signal  $Q$ , so in Verilog code there are eight 16-bit output signals: `ch0rx, ..., ch7rx`.

The GNU Radio project subfolder `usrp/fpga/toplevel/usrp_multi` contains the Verilog code for multiple USRP configuration. This code bases on the `usrp/fpga/toplevel/usrp_std`. The `usrp_multi.v` file defines that `ch0rx` and `ch1rx` are connected to the outputs of 32-bit counter (two most significant bytes - MSB and two least significant bytes - LSB respectively). This counter with remaining three channels are transmitted via USB port to the PC host. Modified multi-USRP functional diagram of the USRPs FPGA configuration is presented in Fig. 4.

The original code of the PC part of the GNU Radio application for multi-USRP configuration can be found in `gr-usrp/src/` subfolder of the GNU Radio

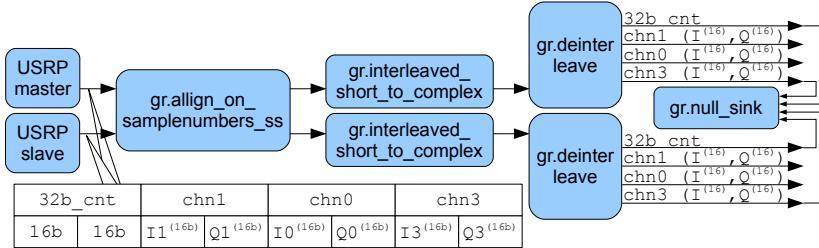


**Fig. 3.** Standard USRP configuration with four receiving channels enabled



**Fig. 4.** Modified USRP configuration with 32-bit counter on channel three

project. In Fig. 5 the data flow according to the `usrp_multi.py` file is presented. Each box represents a software object. The USRP master and USRP slave are objects that collect data from the USB port of the hardware USRP platforms. The data streams from USRP master and slave objects consist of four RX channels (eight 16-bit numbers) where the first channel is the 32-bit counter. Each of received signal consists of  $I$  and  $Q$  part. The `gr.align_on_sampenumber_ss()` object aligns two independent streams from master and slave USRPs using 32-bit counter in order to produce consecutive frames containing input samples with identical counters. Then each frame is converted to the format containing four 32-bit complex numbers ( $I$  and  $Q$  part). The deinterleaver `gr.deinteleave` forms four complex signals. Deinterleavers signal `32-bit_cnt` is connected to `gr.null_sink()` object.



**Fig. 5.** Multi-USRP dataflow of the PC part software

In the original GNU Radio code the signal `chn3` (containing input samples from a RX channel) from the deinterleaver is also connected to `gr.null_sink()`. So only the deinterleaved signals `chn1` and `chn0` of the USRPs' output streams are sent to the output. The signals 32-bit `cnt`(counter) and `chn3` (input samples) are omitted. As a result original GNU Radio multi-USRP configuration allows one to built not six but only four-channel receiver.

One can recover received signal `channel13` by connecting the signal `chn3` from deinterleaver objects to the output of `usrp_multi()` object instead of connecting it to the `gr.null_sink()` object in the `usrp_multi.py` file. Therefore using two USRP devices it is possible to construct the 6-channel receiver.

**9-Channel Receiver Using Three USRPs.** Connection of three USRPs is more complicated. According to USRP clock notes [10] it is possible to use one USRP as a master source and synchronize clock of slave USRPs to the master, but the best solution is to connect an external stable clock source to the `clock_in` input of all three USRPs. Of course pins io15 on daughterboards in the slots A of all USRPs have to be connected, as it was in the case of two USRPs.

The Python code of the `usrp_multi.py` file was written originally for two USRP data streams (as it is presented in Fig. 5) so a new object for the third data path have to be added. The code for `gr.align_on_samplenumber_ss()` object allows to connect many `usrp.source()` objects. The number of aligned USRP sources is limited only by the PC host hardware. After converting, deinterleaving and connecting `chn3` signal to the output (as in the configuration with two USRPs described before) three RX channels for each USRP can be obtained. Hence using presented techniques allows us to build 9-channel receiver platform with three USRPs working synchronously.

**8-channel Receiver Using Two USRPs.** Using two USRPs it is possible to construct 8-channel receiver because each USRP can receive up to four channels simultaneously. However there is a problem with aligning of samples from both USRPs. If one uses 32-bit counters on both USRPs one channel is wasted on each USRP so only six channels are usable. To overcome this limitation we performed deep modifications of Verilog code of USRPs FPGA in order to provide alignment of samples at the hardware not the software level. In this case all four channels

of each USRP contain input samples. This approach will be described in details in our future publications.

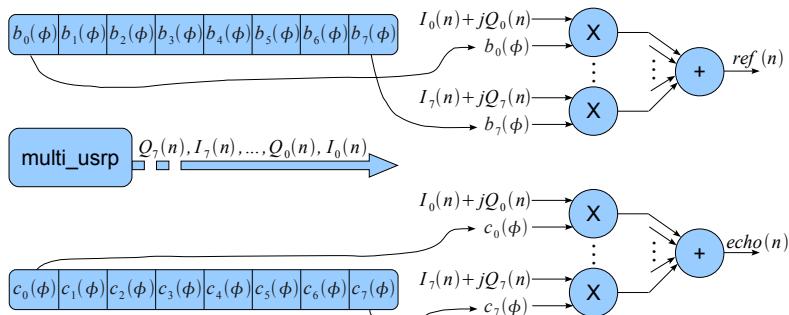
### 3 Digital Signal Processing Subsystem

Baseband signals are transmitted to the PC host via USB port. On the PC our digital signal processing algorithms for estimating signals reflected from objects are implemented. The algorithms are implemented in Python/C++ and they use some DSP sources from GNU Radio. Our application computes the reference and echo signals for three selected FM transmitters in order to obtain three Range-Doppler surfaces in the next stage of data processing.

#### 3.1 Multichannel Digital Beamforming

Digital beamforming algorithm simultaneously estimates two signals for each FM transmitter of the passive radar, namely: the reference signal (i.e., the signal arriving directly from the FM station) and the echo signal reflected from a target object. In order to extract the reference signal, beamforming forms the major lobe at the transmitter's angle  $\phi$ . For the target echo signal, beamforming forms the pattern with null depth at the transmitter's angle. In fact, the two beamformers with two sets of complex coefficients have been implemented for given transmitter azimuth angle: for the direct  $ref(n)$  and reflected  $echo(n)$  signal.

In Fig. 6 the beamformer's structure for one azimuth angle  $\phi$  and 8-channel acquisition is depicted. The  $Q_7(n)I_7(n)\dots Q_0(n)I_0(n)$  denote the baseband signals transferred to the PC from USRPs. The  $b_7(\phi)\dots b_0(\phi)$  and  $c_7(\phi)\dots c_0(\phi)$  are precomputed complex coefficients based on the known positions of the receiver and all three illuminators. There are three sets of these coefficients for three utilized FM radio stations respectively.



**Fig. 6.** An 8-channels digital beamformer

### 3.2 Additional Signal Processing

After beamforming additional adaptive algorithm for cancellation of the direct signal from the echo can be applied. Currently an adaptive lattice algorithm, Schur algorithm and NLMS algorithm are being investigated. At the last stage of data processing Range-Doppler surfaces have to be computed to estimate the target position. The signal processing will be augmented in the future by applying an intelligent target tracking algorithm.

## 4 Results

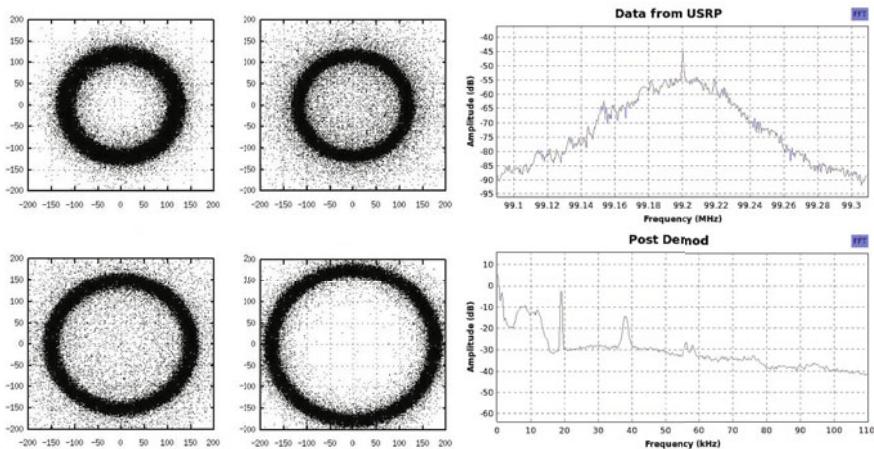
The main result of our research is the multichannel signal receiving and processing SDR-based platform (Fig. 7) we have developed and constructed for the experimental passive radar. The receiver platform utilizes commercial, off-the-shelf (COTS) USRP devices and modified GNU Radio software (as described in section 2.1).



**Fig. 7.** Multichannel signal receiving and processing SDR-based platform

During testing of the system we performed registrations of real-world FM signals. The registered signals captured in different scenarios are being stored in a signal database for a research on DSP algorithms of the passive radar. An example of the I/Q plots for one of the registered FM signals (for 4-channel aquisition) and the spectrum of the down-converted stereo FM radio signal are shown in the Fig. 8. From 0 up to about 16kHz one can see the left plus right ( $L+R$ ) signal, the stereo pilot tone at 19kHz and the left minus right ( $L-R$ ) stereo centered at 38kHz.

Small differences in amplitudes observed for various sensors are probably results of mutual coupling and/or reflections from static objects. The width of the



**Fig. 8.** I/Q plots and spectrum of a registered FM signal

I/Q circle is caused mainly by ground object clutter. These problems together with adaptive direct signal cancellation, computation of Range-Doppler surfaces, estimation of targets' positions and target tracking will be studied in details in our future research.

## 5 Summary

In this paper the new architecture of the signal receiving and processing platform of the multistatic FM-based passive radar for intelligent surveillance purposes was presented. We proposed new approach utilizing Software Defined Radio (GNU Radio) together with COTS and low cost Universal Software Receiver Peripheral devices.

Using unmodified GNU Radio software and two synchronized USRP devices enables one to built only 4-channel receiver which is usually insufficient. We developed different solutions to overcome this limitations. After described modifications of Verilog code one can increase number of channels to six on two USRPs. Using three synchronously working USRPs it is possible to achieve 9-channel configuration it requires appropriate modifications of Python code of the GNU Radio. The last presented solution required some serious modifications of Verilog code configuring the FPGA device of the USRP it enables us to build 8-channel receiver using only two USRP devices.

In our project we use omnidirectional antenna therefore in order to estimate reference and target echo signals additional signal processing have to be performed. We implemented on PC host of the system digital beamforming algorithm that recovers echo and reference signals. In the last stage of processing these signals will be used for computing of Range-Doppler surfaces in order to estimate the position of the target.

Proposed Software Defined Radio approach to constructing the system gives unmatched reconfigurability to the functionality of the whole platform. Digital processing of registered signals implemented on the FPGA of receivers and on the PC host can be easily modified, replaced, scaled and updated according to demand.

**Acknowledgments.** This work was supported by the Polish Ministry of Science and Higher Education from sources for science in the years 2007-2010 under Commissioned Research Project PBZ-MNiSW-DBO-04/I/2007.

## References

1. Willis, N.: *Bistatic Radar*. SciTech Publishing, NC USA (2005)
2. Howland, P.E., Maksimiuk, D., Reitsma, G.: FM radio based bistatic radar. *IEE Proceedings Radar, Sonar and Navigation* 152(3), 107–115 (2005)
3. O'Hagan, D.W., Baker, C.J.: Passive Bistatic Radar (PBR) using FM radio illuminators of opportunity. In: *New Trends for Environmental Monitoring Using Passive Systems*, pp. 1–6. IEEE Press, Los Alamitos (2008)
4. Baker, C.J., Griffiths, H.D.: Bistatic and Multistatic Radar Sensors for Homeland Security. In: *Advances in Sensing with Security Applications*, vol. 2, pp. 1–22. Springer, Netherlands (2006)
5. Johnsen, T., Olsen, K.E.: Bi- and Multistatic Radar. *Advanced Radar Signal and Data Processing*, Educational Notes RTO-EN-SET-086, 4.1–4.34 (2006)
6. Mitola, J.: The Software Radio Architecture. *IEEE Communications Magazine* 33(5), 26–38 (1995)
7. The GNU Radio GNU FSF project website,  
<http://www.gnu.org/software/gnuradio>
8. RadioWare Project, <http://radioware.nd.edu>
9. GNU Radio - USRP docs, <http://gnuradio.org/redmine/wiki/gnuradio/USRP>
10. USRP documentation, <http://gnuradio.org/redmine/wiki/gnuradio/UsrpFAQ>

# Automated Anticounterfeiting Inspection Methods for Rigid Films Based on Infrared and Ultraviolet Pigments and Supervised Image Segmentation and Classification

Michael Kohlert<sup>1,2</sup>, Christian Kohlert<sup>2</sup>, and Andreas König<sup>1</sup>

<sup>1</sup> Institute of Integrated Sensor Systems, Department of Electrical Engineering and Information Technology, University of Kaiserslautern, 67655 Kaiserslautern, Germany

<sup>2</sup> Department of Research & Development, Klöckner Pentaplast GmbH & Co. KG, 56412 Heiligenroth, Germany  
mkohlert@gmx.net,  
c.kohlert@kpfilms.com,  
koenig@eit.uni-kl.de

**Abstract.** Anticounterfeiting of, e.g., pharmaceutical products guarantees customers the delivery of original products. The rigid film industry is developing automated systems to detect product piracy. For this aim, an approach to an automated inspection system for rapid and reliable product verification based on infrared (ir) and ultraviolet (uv)  $\mu\text{m}$ -pigments in rigid films has been developed and statistically meaningful samplesets have been extracted. The industrial manufacturing process has been enhanced for optimized insertion of pigments in rigid films with regard to size, type, and density. The pigments are activated with infrared or ultraviolet light in an encapsulated laboratory system specially developed here. Filter on illumination sources and colour cameras limit the activation and the emission range. Due to optimized film manufacturing and measurement system, the evaluation for uv-pigments can be achieved by a two-stage process of state of the art supervised colour segmentation and blob analysis. The recognition results of the conceived intelligent engineering system fully meets the industrial specification requirements.

**Keywords:** anticounterfeiting, rigid film industry, infrared/ultraviolet pigments, colour image segmentation, blob analysis, automated inspection.

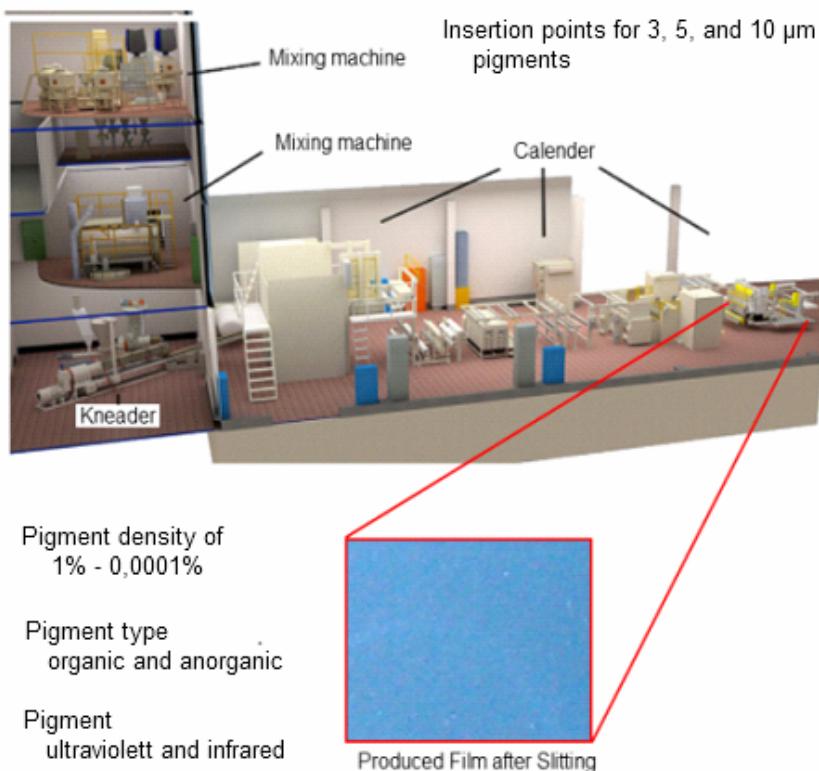
## 1 Introduction

Product piracy causes a loss of 600 bn USD world wide every year. [1] A product copy due to plagiarism is difficult to differentiate from originals. The danger of unknowingly consuming vitally important pharmaceutical products that are either not effective or even poisonous is a stringent societal problem. The financial situation of product markets show the problem of capital investment in new products when no profit is realized due to illegal copies. To disclose illegal copying of original products the rigid film industry develops reliable application and authentification systems to make product packages and their content safer. As one part of the first stage of the supply chain the rigid film industry is able to achieve this goal for the whole supply chain.

The rigid film industry wants to secure the path from the producer to the customer to avoid replication at low additional costs. The package appearance should not change and the whole system has to be integrated easily in the packaging-process.

This work's aim is to develop measuring systems with costs lower than 2000 € for visual detection of pigments and authentification of the product package. The variable cost of the pigment has to be lower than 0,1 Cent/kg of foil. The reproducibility of equal distribution has to be guaranteed during the production process.

The specifications of the research and development department of Klöckner Pentaplast are, to be able to measure the uv- and ir-pigment concentration in films by using light sources, filters and cameras, and a correct separation of pigments and fibres.



**Fig. 1.** Insertion points for pigments in the process of rigid film production: 1. mixing machine, 2. kneader, 3. calender

In this work, the process of calendering is described in section 2, focusing on the insertion of light emitting pigments for anticounterfeiting. A short description on the kinds of pigments used, their activation under infrared and ultraviolet light sources and related data acquisition will be shown in section 3. A hierarchical, two-step classification approach, that serves to segmentate the acquired images of pigmented films is explained in section 4. An overview of the whole anticounterfeiting process is presented in section 5, before concluding.

## 2 Basic Process and Application Specific Adaptations

The standard production of marked transparent and coloured films can be described in a complete process diagram (Fig. 1). Additional pigments can be inserted in a master batch, which is mixed in two mixing machines with high and low temperature, and the kneader, which plastifies the batch. The insertion of pigments in the calender where the film becomes formed is inefficient. The produced foil is slittered at the end of the calendering process for the transport stage. Further changes on the film will be made in a printing firm or other industrial plants after transportation.

The pigments dedicated to product identification were inserted in the recipe of the mixing machines. Anorganic and organic types have been investigated. Anorganic pigments are FDA approved and do not cause any problems for industrial usage. Thus, polymeric films can be doped with fluorescent pigments (FDA approved, anorganic) for protection of polymeric packages. Available diameters of the pigments are 3 $\mu\text{m}$ , 5 $\mu\text{m}$ , and 10 $\mu\text{m}$ .

An inserted pigment density (compound) of 0,0001% up to 1% is used in this work. The intensity and distribution of the pigments is depending on the thickness of the film. In this work the following permutations have been explored. About 1500 pigmented films were produced with 20 different pigments of three different sizes, 3 $\mu\text{m}$ , 5 $\mu\text{m}$ , and 10 $\mu\text{m}$ . Successful laboratory tests with films of 60 $\mu\text{m}$  up to 800 $\mu\text{m}$  thickness allow the use of product identification even for multilayered films.

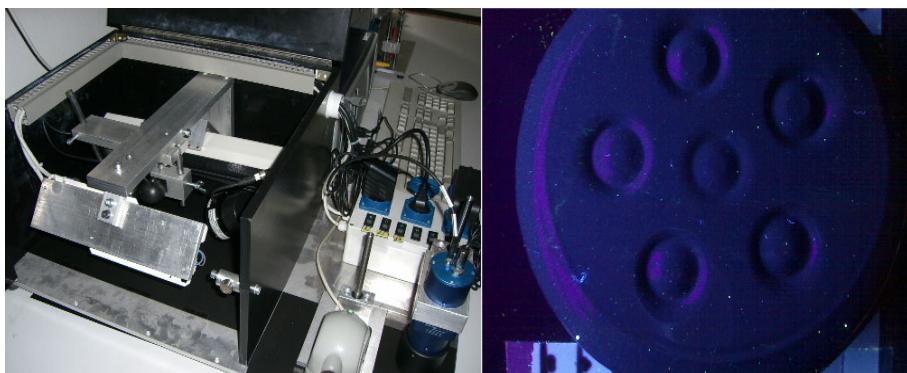
Dedicated measurement laboratory equipment is developed for the detection of the pigments in the films. Ultraviolet fluorescent tubes (UV-A, UV-B), and light emitting diodes (UV-A, UV-B) activate pigments, and other materials. Infrared light emitting diodes (880nm – 1050nm), and laserdiodes (980nm) activate the infrared pigments used in the film.

## 3 Measurement Setup and Data Acquisition

For the required measurements, four different measuring devices (UV lamp system, UV-IR LED system, IR-LED system, and laserdiode system) were developed in this work. The ultraviolet illumination is generated by fluorescent lamps and light emitting diodes at activation wavelengths of 300nm to 370nm. Depending on the pigment emission a bandpass- filter is used to detect the pigment wavelength with a colour camera. It is very efficient to use filters for camera and light source to minimize reflections and other influences.

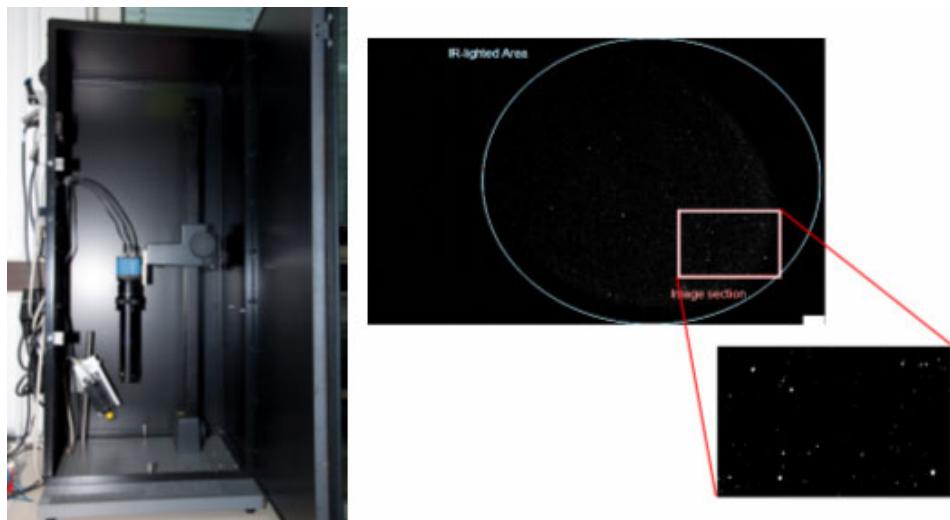
Pigments are activated in ultraviolet or infrared light and emit in lower or higher wavelengths (Fig. 2 shows an example). These pigments are called up- and down-converters. This feature is used for the following measuring devices.

Blisters are prototypes of pharmaceutical packages that show the produced film in the way they are used later for products. The activation light source for the film is placed in 45° above the object. A camera captures the image in 90°. The coloured image acquisition allows a colour-space segmentation as a next step. In this work about 20 blisters (thermoformed films) were produced and captured.



**Fig. 2.** Measuring device (left) for ultraviolet pigments in a blister (right)

The second system is an infrared recognition system, which detects pigments with an activation wavelength from 900nm to 1000nm. (Fig. 3)

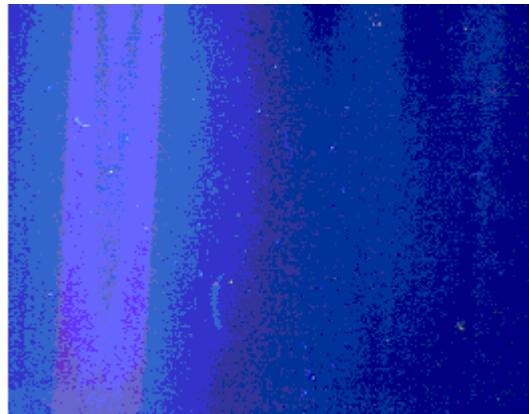


**Fig. 3.** Image acquisition on 10cm<sup>2</sup> with a 11-mega-pixel-camera (b/w), image section of 2cm<sup>2</sup> with a fluorescent infrared pigment concentration of 0,001%

A laserdiode with an optical spreading device is able to illuminate an area of 10 cm<sup>2</sup> in an appropriate normalised distribution for image acquisition. The infrared lighted area and the selected image section with fluorescent infrared pigments are shown in Fig. 3. The illuminated pigments are captured by the black/white camera. In this work, 100 testing films have been captured for ensuing analysis.

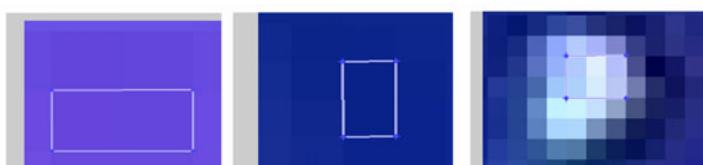
## 4 Two-Stage Recognition System and Results

Based on the optimized rigid films and measurement setups the ensuing evaluation is already feasible by state of the art techniques applied in a two stage approach. The first stage of classification separates the light emitting pigments from the background and the reflections caused by the light source of the measuring device by supervised segmentation. This stage is called here *microclassification*. To separate the pigments from contaminations (fibres) a second classification is carried out on the segmented images, denoted here as *macroclassification*.

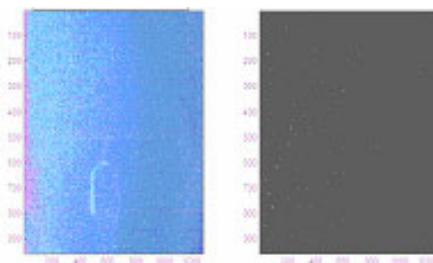


**Fig. 4.** 640x480 pixel RGB image of a rigid film under uv (302nm)

Supervised segmentation of RGB images (see Fig. 4) generates binary images with candidate pigment pixels highlighted. Expert knowledge is integrated during data selection and labelling in the design process. For manually selected relevant sample pixels from 30 specially fabricated films (type: anorganic, size:  $5\mu\text{m}$ , density: 0,001%, not thermoformed), class affiliations were interactively generated (Fig. 5). From the RGB images Lab colour space images [2] of the same size were computed, and from the selected pixel coordinates 1200 labelled three dimensional vectors have been acquired balanced for both classes and randomly split into training and test sets of 600 vectors each.

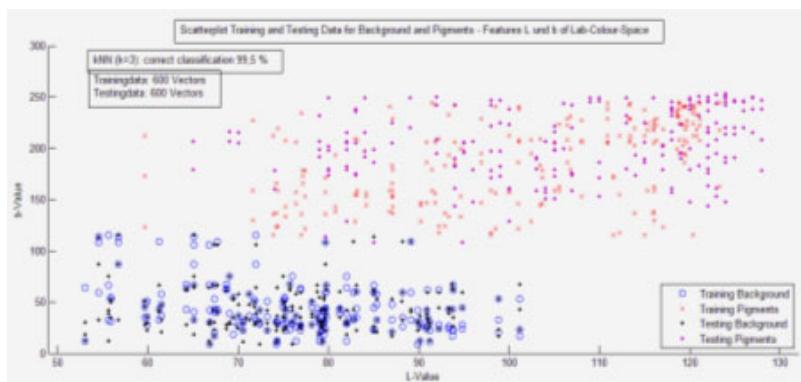


**Fig. 5.** Manual selection of example pixels representing the two classes reflections/background (left, middle) and pigments (right) for segmentation classifier training



**Fig. 6.** Segmentation result: RGB image (left) and binary pigment location image (right)

The classifier achieves supervised segmentation by assigning a class to each pixel of the RGB image (Fig. 6).



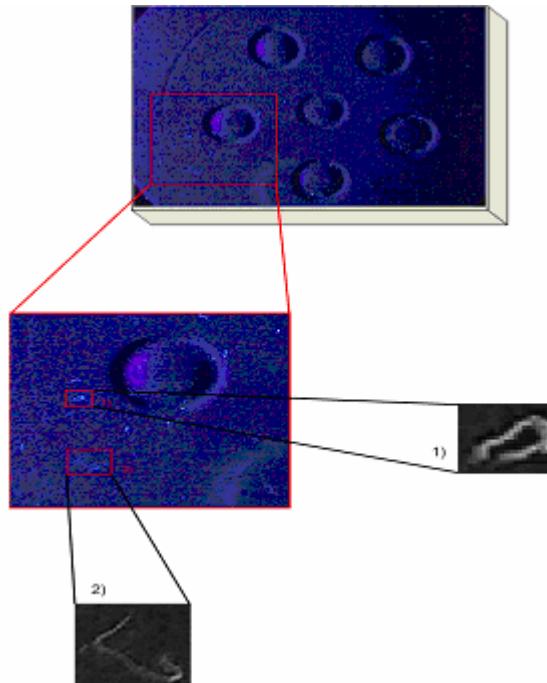
**Fig. 7.** Scatterplot of 600 training- and 600 test vectors (kNN)

For the underlying problem complexity, a non-parametric classifier should be chosen for optimum classification results (Fig. 7 k-nearest-neighbor). The classification results of the training data of 600 vectors (features: L,a,b of colour space LAB) and the test data of 600 vectors are shown in Tab. 1.

**Table 1.** Microclassification results

	KNN	PNN	RNN
Classification rate for test set using hold-out approach	k=1 99.5 % k=2 99,5 % k=3 99,5 % k=5 99,5 % k=6 99,5 % k=8 99% k=10 99 %	$\sigma$ : 0.1 99 % $\sigma$ : 0.03 99 % $\sigma$ : 0.001 60 %	99 %

Thus, established k-nearest-neighbor (kNN), reduced-nearest-neighbor (RNN) and probabilistic neural network (PNN) classifier have been applied, and found already to be satisfactorily working. [3]



**Fig. 8.** Separation of fibres from true pigments after segmentation

The segmented images from the 99.5% correct (kNN) microclassification are used for blob analysis, which has the potential to eliminate residual erroneous pixel classifications by the context of pixel cluster shapes. The macroclassification, based on geometrical features of blob analysis, separates light emitting pigments from fibres. (Fig. 8)

Fibres originate from contamination or clustered particles (Fig. 9). The geometrical structure of the fibres is different from the structure of the pigments. From eight blob



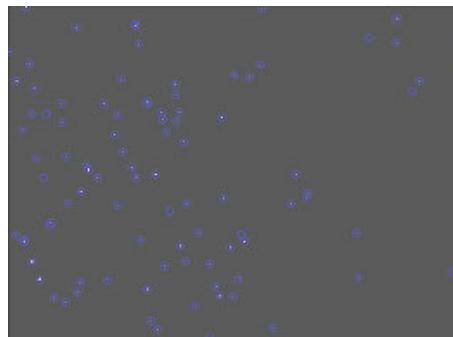
**Fig. 9.** Subregions containing fibres and pigments for training and test data

**Table 2.** Classification results for the blob analysis

	kNN	PNN	RNN
Classification rate for test set using hold-out approach	k=1 100 % k=2 100 % k=3 100 % k=5 100 % k=6 100 % k=8 100 % k=10 100 %	$\sigma$ : 0.1 100 % $\sigma$ : 0.03 100 % $\sigma$ : 0.02 72.5 % $\sigma$ : 0.001 52.5 %	100 %

features, perimeter, area, balance point, compactness, outer circle, inner circle, symmetry, limitative rectangle, the most significant ones (area, rectangle) are automatically chosen by feature selection to distinguish between both classes.

Out of the preselected 300 vectors for pigments, the training set of 100 feature vectors of pigments, 100 feature vectors of fibres, and the test data consisting of 20 feature vectors of pigments and 20 feature vectors of fibres from the same images were employed. (see Tab. 2).

**Fig. 10.** Separated pigments marked blue

The result of the k-nearest neighbor classification is an image only consisting of pigments (Fig. 10 separated pigments). Classes like background, reflections or fibres are not included any more. [4], [5]

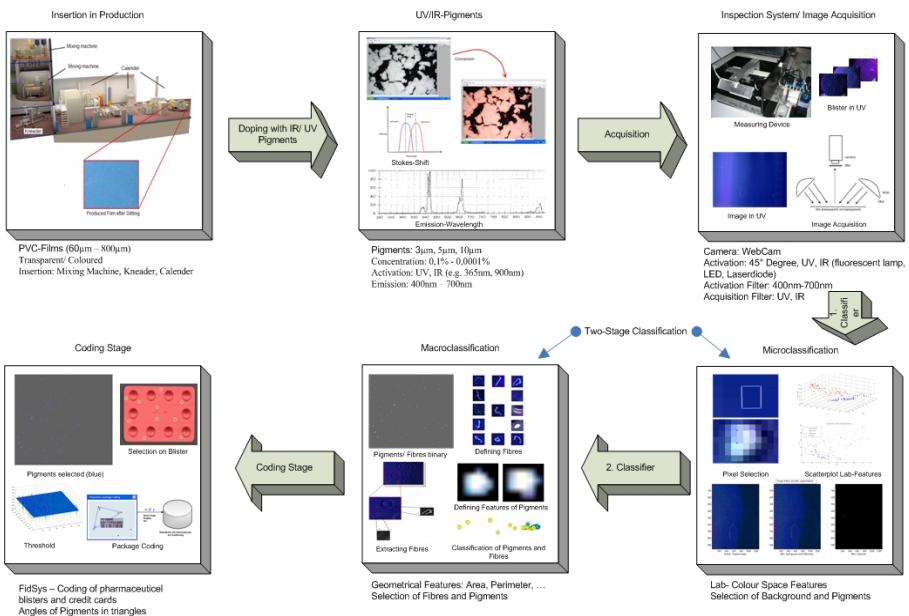
**Fig. 11.** Coding process

After detection and elimination of the fibres, the located pigments can serve for a higher-level authentication process in a following step (Fig. 11). [6] This has been experimentally verified on the detection results for three randomly chosen rigid films.

## 5 Complete Anticounterfeiting System

Using ultraviolet or infrared pigments for anticounterfeiting is a novel approach in the rigid film industry. The insertion in polymeric recipes allows high level security for the whole supply chain.

**Process Diagram for Pigment Insertion, Image Acquisition, Two-Stage Classification and Coding**



**Fig. 12.** Processing diagram for pigment insertion, image acquisition, two-stage classification, and coding stage

The insertion of pigments can be done at the mixing machines, the kneader or the calender. The mixing machines, as the best insertion point, consisting of an mixture of particles (batch) allows a fast distribution of the pigments within the batch.

The choice of the pigment type depends on the measuring device. For optimal results infrared pigments are better to distinguish from other batch particles. Most batch particles emit in ultraviolet ranges, so the ultraviolet emission can be ten times higher than the defined insertion.

A two-stage classification separates pigments from other materials in an image with a correct rate of 100%. (Fig. 12).

For anticounterfeiting of rigid films the specially developed software authenticates the polymeric package via recording and reading of random pigment structures

correctly. It calculates the geometrical positions of the pigments with 99% accuracy and authenticates a film, or a blister within 6 seconds.

## 6 Conclusion

In this work, a viable and economic implementation of a product authentication system was developed, that is able to recognise and locate infrared and ultraviolet pigments in specially fabricated polymeric films. A special feature of the approach is, that film manufacturing and measurement where optimized with regard to recognition accuracy, so that rather basic methods could be employed in the back-end for successful identification., which is also important for aspired low-cost detection devices.

The modification of the polymeric standard process of film production by insertion of pigments was implemented in two of three possible ways. The mixing machine showed best results for optimized distribution. An insertion during the calendering process causes no optimal results. The kneader is another possibility to insert pigments and will be tested in further tasks.

For optimal detection and separation from fibres anorganic infrared pigments with a density of 0,001% and a size of 5 µm showed best results in comparison with ultraviolet pigments.

Four different measurement systems, an UV lamp system, an UV-IR LED system, a non visual IR-LED system, and a laserdiode system were established..

A hierarchical two-stage recognition system with state of the art methods for pigment detection and localization of 100% accuracy has been achieved.

The industrial specification of the company with regard to an homogenous insertion of infrared and ultraviolet pigments in the calendering process, and a separation from fibres is fully achieved. In future work, new and more extensive datasets will be generated, in particular for blob analysis, to ensure practical applicability. Other classification techniques, e.g. Support Vector Machine, will be investigated with existing and new data sets to improve system reliability for viable inline-process measuring and portable inspection systems. The real time behaviour and resource demands of the proposed system has to be regarded carefully and potentially optimized for hand-held inspection devices.

## References

1. Barbieri, A.: Getting real about fake designer goods. In: Bankrate.com (2004), <http://www.bankrate.com/brm/news/advice/scams/20040929a1.asp>
2. Hunter Labs: Hunter Lab Color Scale. Insight on Color 89: Hunter Associates Laboratories, Reston (August 1-15, 1996)
3. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification (2001)
4. Gonzales, R.C., Woods, R.E.: Digital Image Processing (1992)
5. König, A.: Dimensionality Reduction Techniques for Interactive Visualisation, Exploratory Data Analysis and Classification. In: Pal, N.R. (ed.) Pattern Recognition in Soft Computing Paradigm. FLSI Soft Computing Series, vol. 2, pp. 1–37. World Scientific, Singapore (2001) ISBN 981-02-4491-6
6. Kohlert, C., Schmidt, B., Egenolf, W., Zistjakova, T.: Verpackungsfolie für Produktauthentifizierung, Authentifizierungsverfahren und –system. DE 10 2008 032 781 A1 (2010)

# Vowel Recognition by Using the Combination of Haar Wavelet and Neural Network

Mohammad Mehdi Hosseini, Abdorreza Alavi Gharahbagh,  
and Sedigheh Ghofrani

Islamic Azad University, Shahrood branch, Shahrood, Iran  
P.O. Box 36155/163

[hosseini\\_mm@iau-shahrood.ac.ir](mailto:hosseini_mm@iau-shahrood.ac.ir),  
[R\\_alavi@iau-shahrood.ac.ir](mailto:R_alavi@iau-shahrood.ac.ir)

**Abstract.** The lips movements are important in speech recognition and the Lip image segmentation has a significant role in image analysis. In this paper we present a novel technique to recognize Persian Vowels. The method is based on face detection and pupil location. First we perform the lip localization, then the color space CIE L\*U\*V\* and CIE L\*a\*b\* is used in order to improve the contrast between the lip and the other face regions. After that, the lip segmentation by using the Haar wavelet has done and the feature vectors has been extracted from the Haar wavelet result. Finally, the extracted feature vector has been used as neural network inputs and the vowels recognized. The proposed method has been applied on 100 tested images and the accuracy is about 79%.

**Keywords:** Haar wavelet, lip reading, neural network, vowel recognition, segmentation.

## 1 Introduction

Both lips localization and segmentation accurately are important steps in various applications such as automatic speech reading, MPEG-4 compression, special effects, facial analysis and emotion recognition. In this context, the lip analysis is useful for verifying speech and in the same manner lip synchronization as well as in order to minimize the scope for fraudulent access to services controlled by multimodal biometric personal-identity-verification systems. In addition, it has been shown that the lip image analysis can provide a control mechanism for selecting the most appropriate face model (e.g. open mouth or closed mouth) when ever our aim is either face verification or face recognition. However, the lip localization and segmentation in images or videos is a very challenging problem owing to unknown face poses, varying lighting conditions, low image qualities, background clutters. Thereby, such lip a gesture has been extracted from the above referred lip image sequence. Various techniques have been developed to achieve a good and robust segmentation so far. Zhang [1] used hue and edge features to achieve mouth localization and segmentation; Eveno et al [2] detected characteristic points in the mouth region and then used a parametric model to fit the lip. Recently, the statistical methods have been developed to extract face features and particularly the mouth. Coots et al. [3] introduced active shape model (ASM) and active

appearance model (AAM). Wang [4] and Liew [5] proposed FCMS<sup>1</sup> and SFCM<sup>2</sup> algorithms for the lip extraction. Both methods integrate the color information along with different kind of spatial information into a fuzzy clustering structure. However, the above methods should be used in indoor situations with controllable lighting condition.

Different methods have been developed to recognize vowels according to the audio and visual features. Neural network has widely been used in many of these methods. A simple neural network is used for lip reading in [6]. Time delay neural network (TDNN) and recursive neural network for lip reading used in [7, 8].

Although the wavelet transform has been successfully applied in edge detection, there are few reports on using this method for lip segmentation. The problem arises because the outer labial contour of the mouth has very poor color distinction in compare to its skin background. In this paper we propose the automatic lip localization and segmentation algorithm based on the captured full-frontal faces with perfect quality. The Lip features extraction has done by Haar Wavelet and they used as input parameters to a neural network system for vowel recognition. The suggested algorithm is based on the Wavelet transform, while we use the color space, CIEL\*U\*V\* and CIEL\*a\*b\* as well.

## 2 Proposed Method

### 2.1 Lip Localization

The lip localization is performed in two steps. The first step is face detection and the second step is extraction the lip Region-Of-Interest (ROI) from the image.

#### 2.1.1 Face Detection

We have used Bayesteh [9] method for face detection. The proposed algorithm includes two stages at first according to color segmentation, the algorithm search potential face region. Then, it involves face verification by using a cascade classifier. Thereby the face is determined as a rectangle with width W and height H, Fig 1.



**Fig. 1.** (I) input image (II) face detector outputs

---

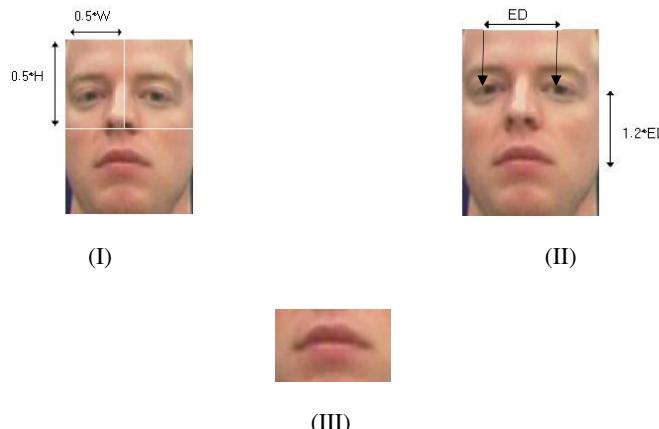
<sup>1</sup> Spatial Fuzzy C-Means Clustering.

<sup>2</sup> Fuzzy C-means With Shape Function.

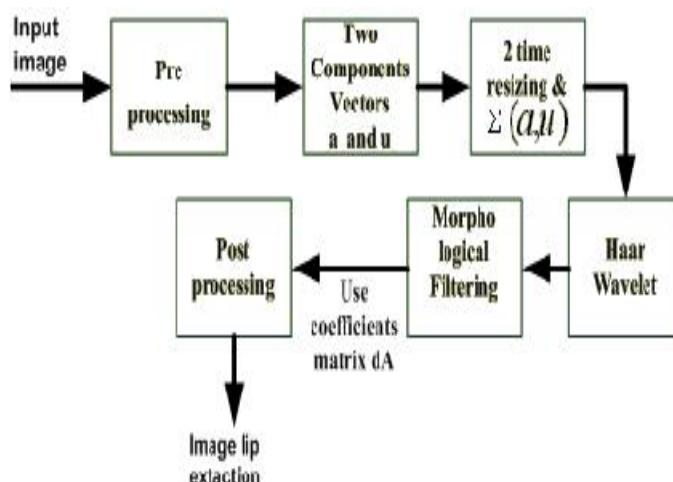
### 2.1.2 Lip Region of Interest

After face detection, the following steps should be followed for lip extraction:

- The face is divided to two equal halves, up half and down half. We can ensure that pupils located in up half and lip located in down.
- The up half divided to two equal, left and right regions. After that by using horizontal and vertical histograms, pupils' position can be obtained. The distance between pupils is named ED.
- From the middle of pupils, move down a length of  $1.2 \times ED$ , Then with this center we plot a rectangle of length ED and width of  $0.7 \times ED$ . Fig 2.



**Fig. 2.** I) step 1 and 2 II) step 3 III) lip region



**Fig. 3.** Block diagram of the lip segmentation algorithm

## 2.2 Lip Segmentation

After finding the lip location, the lip segmentation algorithm in five steps is used Fig 3.

### 2.2.1 Pre Processing

The lightning condition is changed to normalize the Luminance level in original image by the log-function [10]:

$$g(x, y) = k + \frac{\log(f(x, y) + 1)}{d \times \log(t)} \quad (1)$$

Where  $f(x, y)$  is the original and  $g(x, y)$  is the pre-processed image. The set parameters ( $K$ ,  $d$ ,  $t$ ) are adjusted to control the location and the shape of the curve. In this study ( $K$ ,  $d$ ,  $t$ ) are evaluated (12, 0.5, 2) respectively.

### 2.2.2 Color Transform

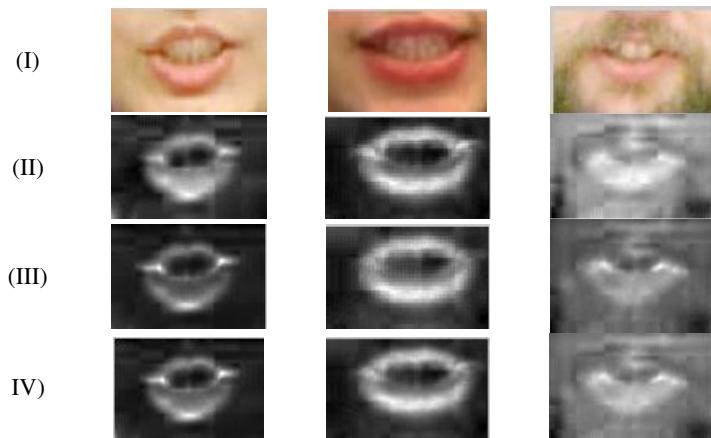
The colors of the lip and the skin region usually overlap, so an especial color space should be chosen to show the small variations. As we know, the distance between any two points in color space is proportional to the perceived color difference, a uniform color space is required. Therefore, we transform the RGB image to the CIEL\*U\*V\* and CIEL\*a\*b color space. Although the vector color  $\{L^*, a^*, b^*, v^*, u^*\}$  for any image can be determined by using the equations that have been addressed by [11], we have used only Parameters  $\{a^*, u^*\}$  in our study.

Although in that case there will be more statistical dependence between two vectors  $\{a^*, u^*\}$ , but note that the main difference between lip and face area is the reddish lip color of all races and in the two selected vectors this color is the most effective. Other vectors usually vary in different races because color of face is different. We tested this assumption in different images and in compare to all cases these selected vectors showed better results. So the other remained parameters can be waived. See fig 4.

### 2.2.3 Lip Segmentation

After pre-processing and transforming the image to the CIEL\*U\*V\* and CIEL\*a\*b color space, the lip segmentation procedure is done as follows:

- The two vector components  $\{a^*, u^*\}$  are added to each other and obtained the image resize to be matched with the original image.
- The 2-D Haar wavelet transform and product is performed and four different matrixes, (dA, dH, dV, dD), are determined. The dA matrix is sufficient for the lip region extraction. So we discard the three matrixes, i.e. dH, dV, dD. In following, the morphological filtering and the post processing are employed to increase the accuracy.



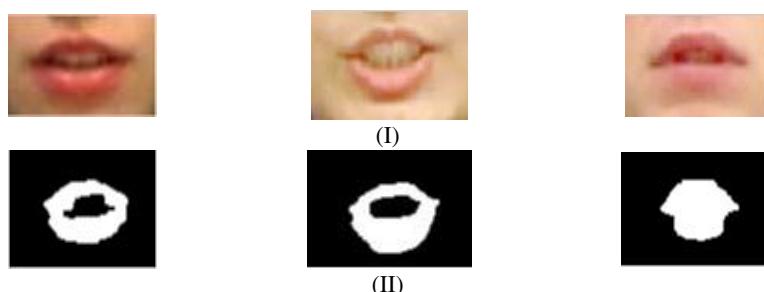
**Fig. 4.** I) picture of the lips II) vector color  $a^*$  III) vector color  $u^*$  IV) output of sum  $a^*$  and  $u^*$

#### 2.2.4 Morphological Filtering

The Grayscale morphological closing and opening with an 8-neighborhood structuring element is used to smooth membership map and eliminate small erroneous blobs and holes.

#### 2.2.5 Post Processing

The output of previous step should be converted to a binary image. first of all we scan all image pixels, assigning preliminary labels to nonzero pixels and recording label equivalence in a union-find table. Note that resolving the equivalence classes has done by using the union-find algorithm; reliable pixels based on the resolved equivalence classes, then a Gaussian filter is used to smooth the image and eliminate some under size points. The final result is shown in fig 5 for different cases.



**Fig. 5.** (I) original lip image (II) extracted lip region

### 3 Vowel Recognition

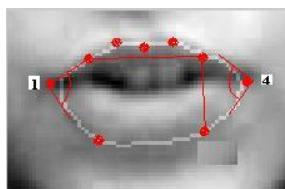
Vowels and consonants are the basic elements of each language. In Persian there are 6 distinct vowels demonstrated as "ا، اے، او، آ، ا، آئ" which are fairly similar to English vowels "a, e, o, â, i and u". In this stage, a new method is proposed for the vowel recognition. In this method, some lip features that have been extracted in former sections apply to an appropriate neural network.

#### 3.1 Feature Vector

Based on a segmented lip image, we are able to extract the key feature points of the lips. We extract the following features:

- Normalized width (mouth opening in horizontal direction).
- Height (mouth opening in vertical direction).
- Average vertical distance of points 5-3 and 5-9.
- Angle between points corner left and right lips.

Fig. 6 shows the selected feature points in a lip contour.



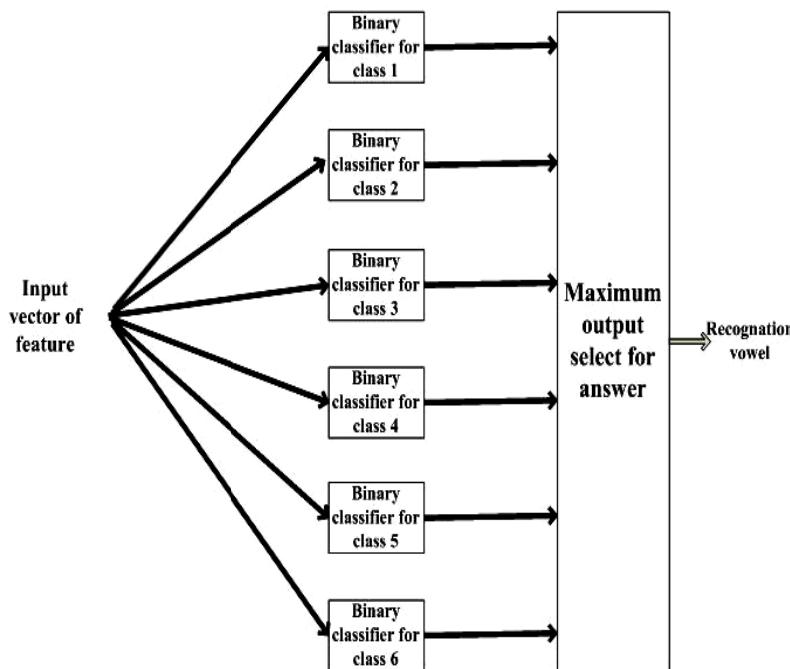
**Fig. 6.** Mouth feature point

#### 3.2 Neural Network

The neural networks have been widely used in various problems such as speech classification and recognition [12]. The problem is a multiclass classification, so we have used one-vs-all method. This kind of strategy has been used in machine learning widely with named the entity recognition. A classifier for each character has been supposed in a way that the network has six classifiers. The input of these classifiers is feature vector and the maximum output of these classifiers has been selected as result. Several experiments showed that the two layer neural network with 20 inputs, 25 hidden layer neurons, 6 outputs (corresponding to 6 persian vowels) and tan-sigmoid activation function had better results than other conditions. The chosen network is a feed forward back propagation network that used in binary mode. Fig.7 shows the neural network structure.

## 4 Material and Methods

In this work, a suitable data base has been prepared. From 6 persons were asked to utter Persian words in which, each word contained a single vowel. We use a subset of



**Fig. 7.** Structure Neural Network used for Classification

**Table 1.** Experiment for classes

Input Vowel Recognition \	ا (a)	ي (e)	و (o)	ا (â)	ى (i)	أ (u)
(a) ا	<b>89</b>	0	0	11	0	0
(e) ي	0	<b>70</b>	21	0	0	9
(o) و	0	0	<b>79</b>	0	0	21
(â) ا	0	0	20	<b>80</b>	0	0
(i) ي	0	0	28	0	<b>72</b>	0
(u) أ	0	0	15	0	0	<b>85</b>

666 image sequences from the database which include the 40 monosyllabic Persian words for each speaker. The sizes of image sequences for the frontal views are 512×384 pixel. The three layer neural network with 10 inputs, 25 hidden neurons and 1 output has been chosen. These numbers of inputs and hidden layers are selected according to the extensive experiments and only tested for the best result in recognition of the 6 classes of vowels. For testing the validity of the proposed method, we choose 80 percent of data as training and 20 percent as testing. For choosing this data, cross validation algorithm is used. It should be notified that the speed of the proposed algorithm and accuracy of the results are better than other methods because classifiers do not conflict with each other and each classifier train immediately in training process.

## 5 Results

The reported results are shown in table 1. The total accuracy in the proposed method is 79%. [14] has done similar work for Persian vowels and obtained accuracy about 64.4%. Their proposed algorithm was evaluated by employing it in recognition of 6 main Persian vowels. [13, 6] have done works with accuracy of 70% but in Japanese vowels. They have classified 5 Japanese vowels uttered by 2 persons.

While the process burden in our method is not changed significantly in comparison with referred methods, the performance in our method is improved.

## 6 Conclusion

In this paper, we propose a new simple and efficient method for lip segmentation, visual lip features extraction, and Persian language vowels recognition. Despite the simplicity of the proposed method, the results show better accuracy in comparison with the other methods with approximately the same process burden.

## References

1. Zhang, X., Mersereau, R.M.: Lip Feature Extraction Towards an Automatic Speechreading System. In: ICIP, Vancouver, BC, Canada (2000)
2. Eveno, N., Caplier, A., Coulon, P.Y.: A Parametric Model for Realistic Lip Segmentation. In: International Conference on Control, Automation, Robotics and Vision, ICARCV (2002)
3. Cootes, T.F.: Statistical Models of Appearance for Computer Vision. Technical report (2004) free to download on, <http://www.isbe.man.ac.uk/bim/refs.html>
4. Wang, S.L., Lau, W.H., Leung, S.H.: Automatic Lip contour extraction from color image. Pattern Recognition, 2375–2387 (2004)
5. Liew, A.W.C., Leung, S.H., Lau, W.H.: Lip contour extraction from color images using a deformable model. Pattern Recognition, Oxford, 2949–2962 (2002)
6. Shinchi, T., et al.: Vowel recognition according to lip shapes using neural networks. Proc. of IEEE, Tottori, 1772–1777 (1998)
7. Matthews, L.: Extraction of Visual Feature For Lipreading. IEEE Transaction on Pattern Analysis and Machine Intelligence 24(2), 198–213 (2002)

8. David Stork., G., et al.: Neural Network Lipreading System for Improved Speech Recognition. In: Pro. IJCNN, pp. 285–295 (1992)
9. Bayesteh, A., Faez, K.: Boosted Bayesian kernel classifier method for face detection. In: Proceeding of Third international conference on natural computation (ICNC), pp. 533–537 (2007)
10. Werblin, F.s.: Control of retinal sensitivity: III. Lateral interactions at the inner plexiform layer. *J.Gen. Physiol.*, Copenhagen 88–110 (1974)
11. Hosseini, M.M., Ghofrani, S.: Automatic Lip Extraction Based On Wavelet Transform. In: IEEE GCIS, China, pp. 393–396 (2009)
12. Yuhas, B.P., et al.: Neural network models of sensory integration for improved vowel recognition. Proc. IEEE, 1658–1668 (1988)
13. Hosseini, M.M.: Vowel Recognition on lip image. Msc.project, University of Islamic azad Iran, pp. 101–105 (2009)
14. Sadeghi, V.S., Yaghmaie, K.: Vowel Recognition using Neural Networks. *IJCSNS International Journal of Computer Science and Network Security* 6(12), 154–158 (2006)

# Bayesian Classification Using DCT Features for Brain Tumor Detection

Qurat-ul Ain<sup>1</sup>, Irfan Mehmood<sup>1</sup>, Syed M. Naqi<sup>2</sup>, and M. Arfan Jaffar<sup>1</sup>

<sup>1</sup> National University of Computer & Emerging Sciences, FAST, Islamabad

<sup>2</sup> Quaid-i-Azam University, Islamabad

qurat-ul-ain@nu.edu.pk, irfanmehmood@gmail.com,  
smnaqi@qau.edu.pk, arfan.jaffar@nu.edu.pk

**Abstract.** Mortality rate by the brain tumor was very high some years before. But now this rate is decreased in the recent years due to the earlier diagnosis and proper treatment. Chances of the long survival of the patient can be increased by the accurate brain tumor diagnosis. For this regard we are proposing more accurate and efficient system for brain tumor diagnosis and brain tumor region extraction. Proposed system first diagnosis the tumor from the brain MR images using naïve bayes classification. After diagnosis brain tumor region is extracted using K-means clustering and boundary detection techniques. We are achieving diagnosis accuracy more than 99%. Qualitative results show that accurate tumor region is extracted by the proposed system. The proposed technique is tested against the datasets of different patients received from Holy Family hospital and Abrar MRI&CT Scan center Rawalpindi.

**Keywords:** Magnetic resonance imaging (MRI), Classification, Image Segmentation, Bayesian Classification, K-Means clustering.

## 1 Introduction

There are many medical imaging modalities are used for visualizing the structure of the brain for tumor diagnosis by the radiologist. These imaging modalities are PET, CT scan and Magnetic Resonance Imaging (MRI). MRI gives high resolution images among all other imaging modalities. Brain MRI is a noninvasive imaging technique which is used for visualizing the brain soft tissues anatomy.

Image intensity in MRI depends upon four parameters. One is proton density (PD) which is determined by the relative concentration of water molecules. Other three parameters are T1, T2, and T2\* relaxation, which reflect different features of the local environment of individual protons [1].

Medical imaging techniques help the radiologist for diagnosing the brain tumor. Computer aided surgeries also required prior analysis of tumor portion inside the brain. So for this purpose brain tumor region extraction is very important after diagnosis of the brain tumor.

In this paper, we proposed a multi step approach for tumor region extraction from the malignant brain MR images. In the proposed technique DCT features[2] of the images are extracted as an initial step. On the basis of these features brain MR images

are classified as normal or tumorous using naïve Bayesian classification [3]. Once the image is determined as tumorous then it is further processed. Segmentation is done using K-means clustering [4] on this tumorous brain MR image which identifies the tumor region from it.

Major contributions of the proposed system are:

- Chances of wrong diagnosis are very less because error rate of this system is very low
- Quite fast as compared to the manual segmentation done by the radiologist for brain region extraction
- Only one time training is required and this system can diagnose tumor from every new dataset provided to this system
- Tumor region is extracted quite accurately even it is located at different location of the brain

The paper is organized as follows. First, Section II contains related work carried out in this field and identifies the problems associated with this field. Detail description of the proposed system is described in Section III. Section IV contains the details of implementation and relevant results. Finally, conclusions and discussions are presented in Section V.

## 2 Related Work

El-Syed et al proposed a hybrid technique for brain MR images classification as tumorous and normal. In this technique discrete wavelet transform coefficients are used as attributes for classification. These attributes are reduced using Principal Component Analysis (PCA) before providing to the classifier. KNN and Feed forward back propagation neural network are used as classifier. The main weakness of the proposed technique is that it requires fresh training each time a new data is arrived [5]. Another adaptive method for brain MRI tissue classification is proposed by Chris et.al. Training dataset is customized by pruning strategy in [6]. By this pruning strategy classification can accommodate the anatomical variability and pathology of brain MRI. Incorrect label samples generated by prior tissue probability maps are reduced by minimum spanning tree. These corrected samples are then used by the KNN classifier for classification of the tissues from brain MR image. It cannot classify accurately tumorous tissues of the brain which is its main drawback.

For brain MR image segmentation Ping et al proposed modified FCM algorithm by modifying membership weighting of every cluster and integrating spatial information. Ping et al applied this method on different MR images and this technique gives appropriate results for MR images of different noise type [7].

Rajeev et al proposed a method for brain tumor detection based on segmentation of the image. Segmentation of the brain MRI is done using watershed algorithm in the proposed technique. The method in [8] does not require any initialization for the segmentation of the brain MRI. Dubey et.al proposed a semi-automatic segmentation technique for brain tumors from MRI. Level set evolution with contrast propagation is used for segmentation in [9] . Pre and post contrast difference images and mixture modeling fit of the histogram are used for calculating probabilities for background

and tumor region. For segmentation of the tumor boundaries level set evolution is initialized using whole image.

M. Masroor et al proposed segmentation of brain MR images for tumor extraction by combining K-means Clustering for grouping tissues and perona malik anisotropic diffusion model for image enhancement [10]. T.bala et al proposed a fuzzy clustering method for the segmentation of Brain MR images. T.bala et al visually provide the result of their proposed technique but quantitatively they does not prove that their results are better and neither they compare their results with some other technique to validate their results [11].

Ping et.al proposed a region based brain MR image segmentation method in [12]. In this technique modified Mumford Shah's segmentation algorithm is incorporated for MR images segmentation into white matter, gray matter and CSF regions. This method add tuning weight concept in conventional Mumford Shah's energy function. This weights needs to be tune for segmentation into three tissues classes. Segmentation of brain MRI in three tissues is not done at a time instead these are done one by one. This method also needs training from manual segmented images for Mumford Shah's energy function parameters initialization. This method is applied only on T1 weighted images. As T1 weighted images are low in contrast thus manual seed is required for initialization purpose. Due to manual seeding and training from manual segmented images this technique cannot be applied for every type of brain MR images like T2 weighted and PD images.

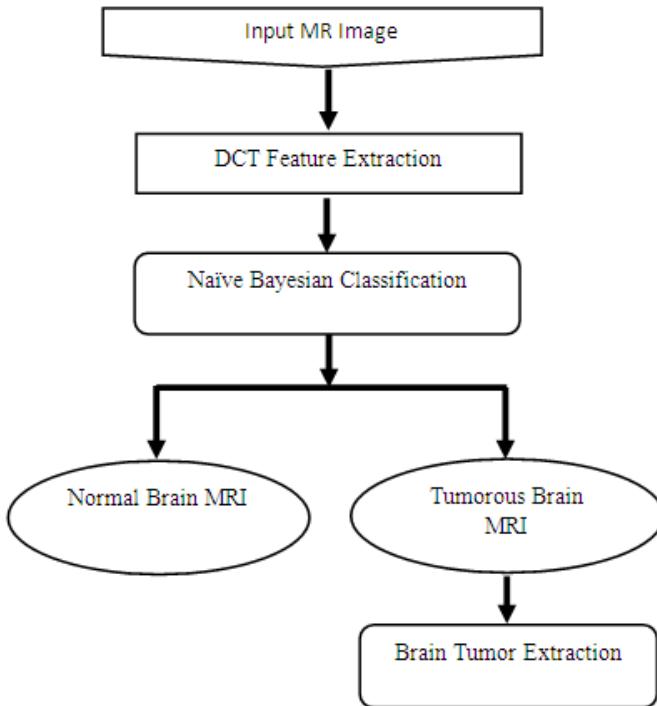
### 3 Proposed Method

The proposed system consists of three major modules. Brain MR Image is provided to the system as an input. First module consists of feature extraction phase using Discrete Cosine Transform [2] from input MR image. Second module of the proposed system classify the input brain MR image as normal or tumorous on the basis of the DCT features extracted in the first phase. Last module of the proposed system is segmentation. This module is required only for tumorous brain MR images. Segmentation extracts the tumor region from the brain MR image which is classifies as tumorous in the second phase.

Complete system architecture of the proposed method is shown in Fig. 1. Details about the major components of the proposed algorithm are discussed in the following subsections one by one.

#### 3.1 DCT Feature Extraction

Discrete cosine transform (DCT) is used for converting the signal into its frequency components. In image processing DCT attempts to decorrelate the image data. DCT has the ability to pack the image data into as few DCT coefficients as possible without any distortion. DCT has the property of separability and symmetry. 2-Dimensional DCT of the input is defined by the following equation [2, 13].



**Fig. 1.** Architecture of the Proposed System

$$C(u, v) = \alpha(u)\alpha(v) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x, y) \cos\left[\frac{\pi(2x+1)u}{2N}\right] \cos\left[\frac{\pi(2y+1)v}{2N}\right] \quad (1)$$

Where  $0 \leq u \leq N$ ,  $0 \leq v \leq N$ , and

$$\alpha(u) = \begin{cases} \sqrt{\frac{1}{N}} & \text{for } u = 0 \\ \sqrt{\frac{2}{N}} & \text{for } u \neq 0 \end{cases} \quad (2)$$

Proposed system calculates 2-D discrete cosine transform of the image using equation (1,2). For the sake of preventing from computational complexity and to make the system more time efficient we just take top seven highest frequency components as features vector for classification.

### 3.2 Classification

For estimating the class of the new data item a probabilistic model is defined that is known as Bayesian classification. Bayesian classifier is a statistical classifier. Bayesian

classification is based on the Bayes theorem. Bayesian classification is used for classifying objects into associated classes based on the attributes of these objects. Attributes of the data/object are considered as independent of each other in Naive Bayes classification [3, 14].

### 3.2.1 Bayes Theorem

Bayesian theory gives a mathematical calculus of degree of belief. Bayes theorem is a way to calculate the conditional probabilities of features. Bayes theorem can be defined as [5, 14]

- Let  $X$  is a data sample/Object.
- Let  $H$  be some hypothesis, such as  $X$  belongs to a specified class  $C$ .
- $P(H)$  is known as the prior probability.  $P(H)$  tells the probability of given data sample belonging to the specified class.
- $P(X|H)$  is the probability that the hypothesis  $H$  holds given the observed data sample/object  $X$ .
- $P(H|X)$  is called the posterior probability. It is based on more information than the prior probability  $P(H)$ , which is independent of  $X$ .

Baye's Theorem provides a way of calculating the posterior probability

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (3)$$

$P(X|H)$  is the posterior probability of  $X$  conditioned on  $H$ .  $P(X)$  is the prior probability of  $X$ . Posterior probabilities are class density estimates. Accuracy of the classifier strongly depends upon this parameter. As much accurate the class density estimate as much higher accuracy is achieved [13].

### 3.2.2 Naïve Bayes Classification

Bayesian classifier estimates the class of unknown data item using probabilistic statistics model. Challenge in the Bayesian classification is to determine the class of data sample which have some number of attributes. Bayesian classification estimates the class of this unknown data item on the basis of known data item which are provided with the class labels for the training purpose of the Bayesian classification [5, 18].

Let  $S$  be the dataset with  $t$  objects such that  $X_1, X_2, \dots, X_t$ . Each object have  $n$  attributes such that  $A_1, A_2, \dots, A_n$ . let there are  $m$  classes  $C_1, C_2, \dots, C_m$ . Naïve bayes classifier predicted the unknown data sample  $X$  which is without the label to the class  $C_i$  if and only if

$$P(C_i|X) > P(C_j|X) \text{ for } 1 \leq j \leq m, j \neq i \quad (4)$$

then  $X$  is assigned to  $C_i$ . This is called Bayes decision rule.

Using Bayes theorem maximum posterior hypothesis can be calculated using

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (5)$$

The class prior probabilities  $P(C_i)$  can be estimated by

$$P(C_i) = s_i/s \quad (6)$$

Where  $s_i$  is the total number of training samples which have class  $C_i$  and  $s$  is the total number of training samples. In order to reduce a computation and to avoid the effect of scarcity of data the naïve assumption of class conditional independence is made. Conditional probability  $P(C_i|X)$  is estimated using the prior probability or estimates given by:

$$P(X|C_i) = \prod_{k=1}^n P(x_k|C_i) \quad (7)$$

The probabilities  $P(x_1|C_i)$ ,  $P(x_2|C_i)$ , ...,  $P(x_n|C_i)$  are estimated from the training samples. An attribute  $A_k$  can take on the values  $x_{1k}$ ,  $x_{2k}$ , ..

$$P(x_k|C_i) = s_{ik}/s_i \quad (8)$$

Where  $s_{ik}$  is the number of training samples of class  $C_i$  having the value  $x_k$  for  $A_k$  and  $s_i$  is the number of training samples belonging to  $C_i$

In Naïve bayes classification dataset is divided into two sets, training and testing respectively. Training dataset is considered as prior information and model is constructed on the basis of this training dataset. Class of the unknown data is determined using this model.

### 3.3 Segmentation

Segmentation is performed for extracting the tumorous brain portion from the brain MRI. This process segments the brain MR image into two portions. One segment contains the normal brain tissues and the second segment contains the tumorous cells. The segment which contains the tumorous cells is the desired region which is known as tumorous region.

#### 3.3.1 K-Means Clustering

K-means clustering is an unsupervised clustering technique which is used to partition the data set  $n$  into  $K$  groups [6]. K-means clustering algorithm initially set centers of each cluster which is known as centroids of clusters. K-means algorithm minimizes the intra cluster distance and maximizes the inter cluster distance. Each instance is assigned to the cluster which has closest center value with it. Each cluster center  $C_j$  is updated to be the mean of its constituent instances. This algorithm aims at minimizing an objective function. The objective function for K-means is

$$\text{Objective Function} = \sum_{j=1}^k \sum_{i=1}^n |x_i - \mu_j|^2 \quad (9)$$

Where  $|x_i - \mu_j|^2$  is a distance measure between a data point  $x_i$  and the cluster centre  $C_j$ , is an indicator of the distance of the  $n$  data points from their respective cluster centers.

### 3.3.2 Tumor Boundary Detection

K-means clustering segments the MR image into two regions. One region contains the normal brain cells and the second region contains the tumorous brain cells. K-means clustering segments the brain MR image on the basis of intensity pixels of the image. Canny edge detection method is used for detection of the tumorous cells edges from the segmented brain MR image. With the help of these edges brain tumor region is extracted from the original brain MR image.

## 4 Experimental Results and Discussion

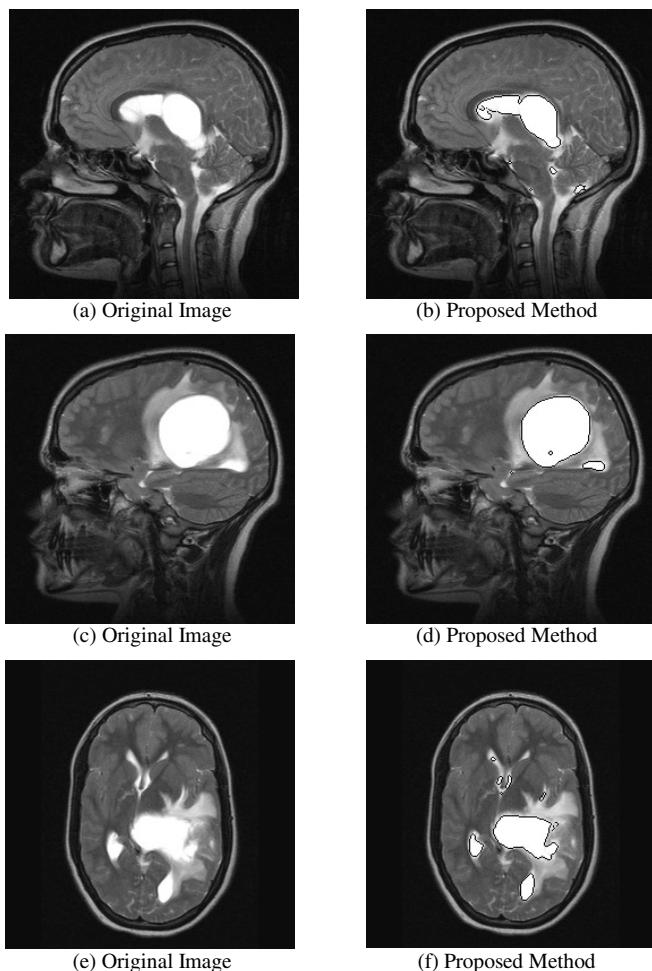
The proposed system is implemented by using the MATLAB environment. Proposed technique is tested on the dataset available at [15] and it is also tested on real brain MRI data set available at [16].

Available brain MRI data is initially classified as normal and tumorous. We have used ten-fold cross validation for training and testing purpose on both data sets for classification. Naïve bayes classifier is used for classifying brain MR images. Performance of the classifier is measured in terms of confusion matrix, sensitivity, specificity and accuracy. We are achieving 99.89% accuracy for dataset of real MR images and 99.64% of accuracy is achieved for AANLIB dataset. This is very high accuracy for classification. Specificity and sensitivity rates for real MRI dataset and for AANLIB dataset are also more than 99%. Error rate of the proposed system for classification phase is also very low that is less than 5%. Classifier accuracies for the testing data is compared with the work proposed in [5] is shown in Table 1. We have also compared our result using the same preprocessing procedure with SOM and SVM and mentioned in the table 1.

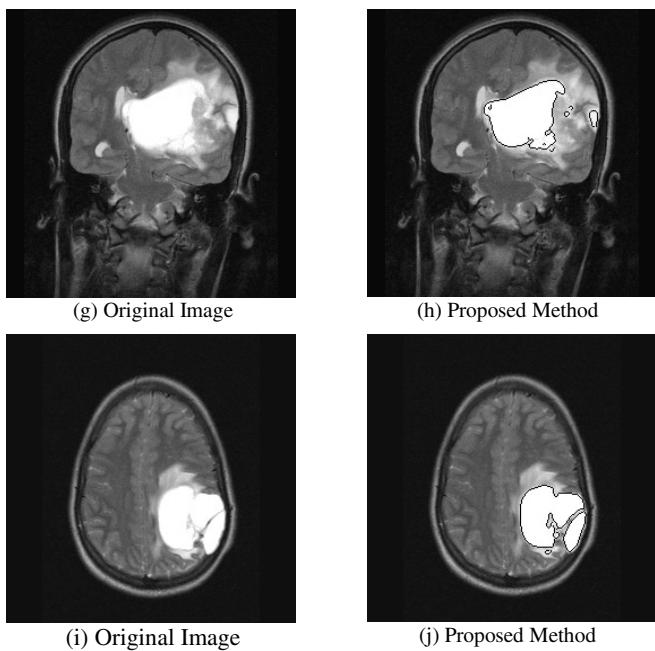
**Table 1.** Performance in term of Accuracy of proposed technique and work in [5]

Technique	AANLIB Data Accuracy (%)	Real MRI Data Accuracy (%)
DCT+ Bayes	<b>99.64</b>	<b>99.89</b>
DWT+PCA+ANN [5]	95.4	94.3
DWT+PCA+KNN [5]	98.2	97.5
DWT+SOM [5]	95.13	94.72
DWT+SVM [5]	97.25	96.64
DWT+ Bayes	97.87	97.24
DCT+SVM	98.23	98.09
DCT+SOM	97.12	96.86

After classification tumorous brain MR images are segmented for tumor region extraction. We show the results of tumor extracted region from brain MR images in Fig.2 All results show that tumor region is extracted quite accurately. In Fig.2 (b), (d), (f), (h), and (j) tumor extracted region is shown .All results show that tumor boundary is quite accurately detected. All tumor cells are identified by the proposed system very accurately. In Fig.2 (j) all tumor region is not connected together. Some tumor cells are disjoint from each other. Tumor is spread over some brain portion. This is quite accurately determined by the proposed system. We validate our segmented results from radiologist. According to the radiologist our results of segmented images are correct and our proposed system accurately extracts the tumor region from affected brain MR images.



**Fig. 2.** Results of our Proposed System

**Fig. 2. (Continued)**

## 5 Conclusion and Future Work

We proposed a multi-phase system for the diagnosis of brain tumor and tumor region extraction. Proposed system performs diagnosis using Naïve Bayes classification on the basis of DCT features of the brain MR images. Segmentation is performed on the tumorous images which are determined in the first step.

K-means clustering is used for segmenting the brain MR image. Boundary of the tumor cells is determined using the ‘Canny’ edge detection method. Proposed system accurately classifies the normal and tumor brain images and this is proved by the results presented. Tumor cells are also quite accurately identified by the proposed system.

Future work of this research is to measure the size and thickness of the tumor extracted region.

## Acknowledgment

The authors would like to thank Higher Education Commission (HEC), Govt. of Pakistan and NU-FAST for providing funds and required resources to complete this work.

## References

- [1] Sonka, M., Tadikonda, S.K., Collins, S.M.: Knowledge-Based Interpretation of MR Brain Images. *IEEE Transaction on Medical Imaging* 15(4), 443–452 (1996)
- [2] Khayam, S.A.: The Discrete Cosine Transform: Theory and Application. WAVES lab technical report (May 2004)
- [3] Mitchell, T.M.: Machine Learning, Bayesian Learning, pp. 154–178. McGraw Hill, New York (1997)
- [4] MacQueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of 5<sup>th</sup> Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press, Berkeley (1967)
- [5] El-Dahshan, E.S.A., Salem, A.-B.M., Younis, T.H.: A hybrid technique for automatic MRI brain images classification. *Studia Univ, Babes Bolyai, Informatica* LIV, 55–67 (2009)
- [6] Cocosco, C.A., Zijdenbos, A.P., Evans, A.C.: A fully automatic and robust brain MRI tissue classification method. *Medical Image Analysis* 7(4), 513–527 (2003)
- [7] Wang, P., Wang, H.: A modified FCM algorithm for MRI brain image segmentation. In: International Seminar on Future BioMedical Information Engineering, pp. 26–29. IEEE computer society, Los Alamitos (2008)
- [8] Ratan, R., Sharma, S., Sharamac, S.k.: Brain tumor detection based on multi-parameter MRI image analysis. *ICGST-GVIP Journal* 9(III), 9–17 (2009) ISSN 1687-398X
- [9] Dubey, R.B., Hanmandlu, M., Gupta, S.K.: Semi-automatic Segmentation of MRI Brain Tumor. *ICGST-GVIP Journal* 9(4), 33–40 (2009) ISSN: 1687-398X
- [10] Ahmed, M.M., Mohamad, D.B.: Segmentation of Brain MR Images for Tumor Extraction by Combining Kmeans Clustering and Perona-Malik Anisotropic Diffusion Model. *International Journal of Image Processing* 2(1), 27–34 (2008)
- [11] Bala Ganesan, T., Sukanesh, R.: Segmentation of Brain Images Using Fuzzy Clustering Method with Silhouutte Method. *Journal of Engineering and Applied Sciences, Medwell Journals*, 792–795 (2008)
- [12] Chen, P.-F., Grant Steen, R., Yezzi, A., Krim, H.: Brain MRI T1-map and T1 weighted image segmentation in a variational framework. In: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 417–420 (2009)
- [13] Strang, G.: The Discrete Cosine Transform. *SIAM Review* 41(1), 135–147 (1999)
- [14] Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification, 2nd edn. John Wiley and Sons, New York (2001)
- [15] Harvard Medical School, Web, <http://med.harvard.edu/AANLIB/>
- [16] <http://www.abrarmriict.com/>

# A New Strategy of Adaptive Nonlinear Echo Cancelling Volterra-Wiener Filter Structure Selection

Pawel Biernacki

Telecom, Acoustic and Computer Science Institute,  
Wroclaw University of Technology,  
Wyb. Wyspianskiego 27, 50-350 Wroclaw, Poland  
 [{pawel.biernacki}@pwr.wroc.pl](mailto:{pawel.biernacki}@pwr.wroc.pl)

**Abstract.** In the article a nonlinear orthogonal echo cancelling filter parameters selection strategy is proposed to maximize echo cancelling quality to computation complexity ratio. Presented approach leads to the orthogonal realization of the nonlinear echo cancelling filter of Volterra-Wiener class which structure changes due to higher-order statistics of the filtered signals.

**Keywords:** Volterra-Wiener class filter, orthogonal filter, filter structure selection.

## 1 Introduction

The problem of computational complexity of the Voltterra-Wiener class filter has been discussed in [3][5][6]. Depending on the application there are given various methods to choose the order of the filter. In the case of echo cancellation (not just the acoustic echo) there is no known publications on the methods of selecting the appropriate filter structure. The suggestions contained in [4][7][8] do not allow one to determine unambiguously the strategy of the nonlinear orthogonal echo cancellation filter structure.

The aim of this paper is to propose strategy for optimizing the filter order and filter nonlinearity degree in terms of its computational complexity, so to obtain the best (maximal) value of the echo reduction to computational complexity ratio.

## 2 Nonlinear Orthogonal Echo Cancellation Filter Complexity

Using the results presented in [9],[10], during the echo cancellation filtering process the following multi-dimensional Fourier series expansion of the original signal ( $x_0$ ) is obtained

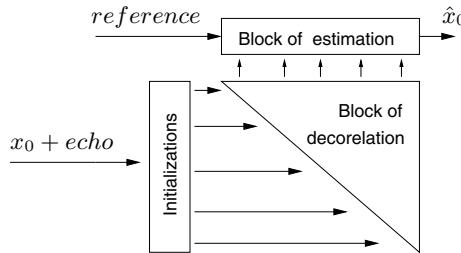
$$\hat{x}_0 = \sum_{i_1=0}^N {}^x \rho^{i_1} r_0^{i_1} + \sum_{i_1=0}^N \sum_{i_2=i_1}^N {}^x \rho^{i_1, i_2} r_0^{i_1, i_2} + \dots$$

$$\dots + \sum_{i_1=0}^N \sum_{i_2=i_1}^N \dots \sum_{i_M=i_{M-1}}^N {}^x\rho^{i_1, \dots, i_M} r_0^{i_1, \dots, i_M} \quad (1)$$

where

$$\begin{aligned} {}^x\rho^{i_1} &= (x_0, r_0^{i_1}) \\ {}^x\rho^{i_1, i_2} &= (x_0, r_0^{i_1, i_2}) \\ &\dots \\ {}^x\rho^{i_1, \dots, i_M} &= (x_0, r_0^{i_1, \dots, i_M}) \end{aligned} \quad (2)$$

are the generalized (multi-dimensional) Fourier [1], [2] (i.e. Schur-type) coefficients. The coefficients (2) can be interpreted as the orthogonal representation of the random variable  $x_0$  in the subspace  $S$  spanned by the orthonormal elements  $\{r_0^{i_1}, \dots, r_0^{i_1, i_2}, \dots, r_0^{i_1, i_2, i_3}, \dots, r_0^{i_1, \dots, i_M}\}$  (where  $i_1 = 0, \dots, N$ ,  $i_2 = i_1, \dots, N$ ,  $i_3 = i_2, \dots, N$ ,  $i_M = i_{M-1}, \dots, N$ ) on one hand, on the other - as the coefficients in the orthogonal realization of the multi-dimensional nonlinear echo-cancelling filter [10] (figure 1, figure (2)).



**Fig. 1.** General structure of the nonlinear orthogonal echo-cancelling filter

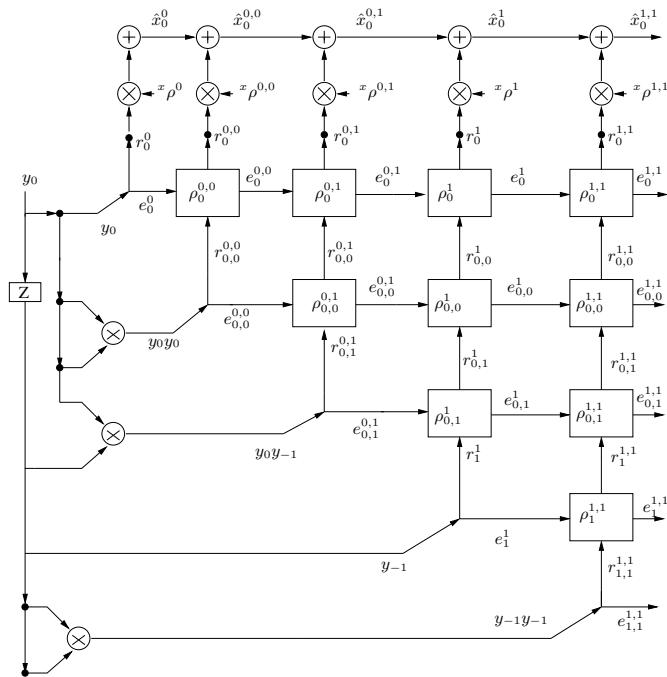
The least-square error of the original signal estimation is equal

$$\begin{aligned} {}^xR_{N,0}^{\{M\}} &= ({}^x\varepsilon_{N,0}^{\{M\}}, {}^x\varepsilon_{N,0}^{\{M\}}) = \dots = \\ &= ||x_0||_{\Omega}^2 - \sum_{j_1=0}^N |{}^x\rho^{j_1}|^2 - \dots - \\ &\quad + \sum_{j_1=0}^N \dots \sum_{j_M=j_{M-1}}^N |{}^x\rho^{j_1, \dots, j_M}|^2 \end{aligned} \quad (3)$$

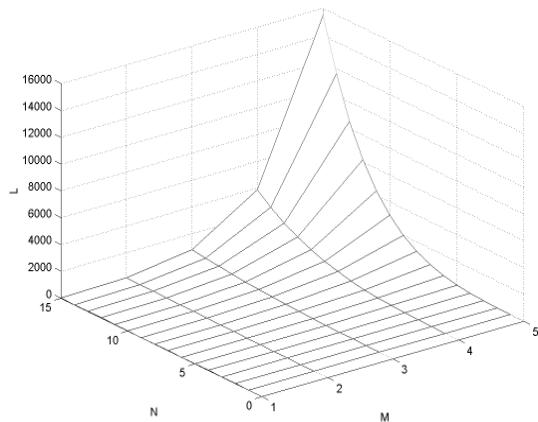
This error depends only on the coefficients  ${}^x\rho^{j_1, \dots, j_q}$

Considering (1) it is easy to notice, there are two parameters which present filter complexity: the filter order (filter memory) and degree of the filter non-linearity. Increasing one of them entails quick growth a computing complexity needed to estimate  $(\hat{x}_0)$ . The number of the elementary section in the estimation block of the echo cancelling filter is

$$K_{N,M} = \sum_{m=1}^M \frac{(N+m-1)!}{m!(N-1)!} \quad (4)$$



**Fig. 2.** The nonlinear orthogonal echo-cancelling filter ( $N = 1, M = 2$ )



**Fig. 3.** The number of elementary sections  $K_{N,M}$  in the estimation block

Figure 3 presents how quick rises the value of (4) when  $N$  or  $M$  is changing. The number of the elementary sections in the decorrelation block of the echo cancelling filter is

$$L_{N,M}^{ns} = \frac{(1 + K_{N,M})K_{N,M}}{2} \quad (5)$$

The realization of the decorrelation block is more laborious. The key is to select such the values of  $N$  and  $M$ , that the minimal number of computed ON basis elements allow to determine the 'good' estimate (achieving the desire value of the error (3)).

It is possible to change  $N$  or  $M$ , or both of them at one time. Real system should be efficient. This aim can be achieve by adaptive filter parameters selection using some intelligent strategy.

### 3 The Criterion of the Echo-Cancelling Filter Parameters Selection Strategy

The aim of increasing degree of filter nonlinearity or its order is a improvement of the echo cancelling quality (minimizing (3)).

To judge the proper values of  $N$  and  $M$  the objective measure is needed, which shows the relative improvement of the echo cancelling quality. It should describe a change of the estimation error causes by extension of filter structure (increasing the number of elementary sections in the decorrelation block). The following cost function is defined

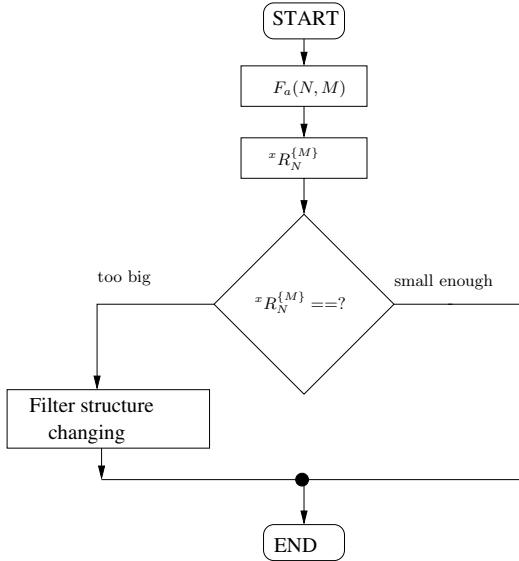
$$FK(N_1, M_1; N_2, M_2) \triangleq -\frac{\frac{L_{N_2, M_2}^{ns} - L_{N_1, M_1}^{ns}}{L_{N_1, M_1}^{ns}}}{\frac{{}^x R_{N_2}^{\{M_2\}} - {}^x R_{N_1}^{\{M_1\}}}{{}^x R_{N_1}^{\{M_1\}}}} \quad (6)$$

where  $N_1, N_2$  are the filter orders,  $M_1, M_2$  are the filter nonlinearity degrees. The value  $L_{N,M}^{ns}$  (4) describes the number of elementary sections in the decorrelation block for filter of order  $N$  and nonlinearity degree of  $M$ .  ${}^x R$  is defined in (3). The equation (6) is interpreted as a relative change of the number of elementary sections in the decorrelation block to relative improvement of a echo cancelling quality (change value of the estimation error) when the parameters  $N, M$  are changing. Because the denominator is always negative the sign minus is used. This cost function allows to judge efficiency of the selected filter structure and the filter complexity needed to improved the estimate  $\hat{x}_0$ .

### 4 The Intelligent Strategy of Estimation Improvement by Selection of Filter Parameters

The pattern of one loop of filtering process is presented in figure 4.

After filtering by a filter  $F_a(N, M)$ , the estimation error is computed  ${}^x R_N^{\{M\}}$ . If its value is small enough the filter structure is unchanged. If the error is too big



**Fig. 4.** One loop of the filtering process

the filter parameters are changing according to some strategy. The block 'Filter structure changing' represents proposed in this article strategy of selecting filter parameters.

Assuming value of the cost function (6) determines the possible values of the parameters  $N$  and  $M$ .

The dominator of (6) is the control element, which depends on the Schur coefficients  ${}^x \rho$ . Adding a new non-zero valued coefficient decreases the value of (6). It is equivalent to changing the filter order or/and its nonlinearity degree. It causes the increasing of filter complexity (especially the decorrelation block).

When one or few coefficients  ${}^x \rho$  are added it is not necessary to compute the whole ON basis anew, but only its new element. The filter  $F_a(N_1, M_1)$  has  $L_{N_1, M_1}^{ns} = K_{N_1, M_1}(K_{N_1, M_1} + 1)/2$  elements in the decorrelation block, where  $K_{N, M}$  is described in (5). The value of cost function (6) after adding a new coefficient  ${}^x \rho_{new}$  (it means adding  $K_{N_1, M_1} + 1$  elementary sections to the decorrelation block) is

$$FK(N_1, M_1; N_2, M_2) = \frac{2}{K_{N_1, M_1}} \cdot \frac{{}^x R_{N_1}^{\{M_1\}}}{|{}^x \rho_{new}|^2} \quad (7)$$

Using (7) and considering presented discussion the following strategy for filter structure selection is proposed

**Strategy 1.** If  $K_{N_1, M_1}$  is the number the coefficients  ${}^x \rho$  for the filter  $F_a(N_1, M_1)$ , then increasing  $K_{N_1, M_1}$  by one (adding the new coefficient  ${}^x \rho_{new}$ ), which is effective in a new filter  $F_a(N_2, M_2)$ , follows when and only when

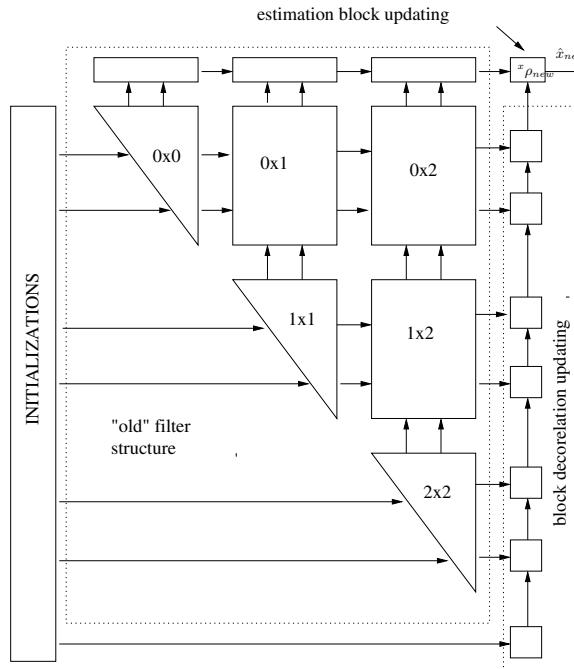
$$\begin{aligned}
 F_a(N_1, M_1) \rightarrow F_a(N_2, M_2) \Leftrightarrow FK(N_1, M_1; N_2, M_2) = \\
 = \frac{2}{K_{N_1, M_1}} \cdot \frac{x R_{N_1}^{\{M_1\}}}{|x \rho_{new}|^2} < \delta \quad (7)
 \end{aligned}$$

'The partial actualization' by a coefficient  $x \rho_{new}$  can be done by changing  $N$  or/and  $M$ . This allows to check the different combinations of  $N$  and  $M$ . The choice for testing the new Schur coefficient depends on the maximal value of  $K_{N,M}$ . Using proposed strategy ② the following rules are proposed:

- for determining  $M$  increase the parameter  $N$  up to reach maximal value  $K_{N,M}$ ; it means the linear, bi-linear, tri-linear,... elements of ON basis are checked, and choosing is one for which  $x \rho_{new}$  meets ③

Selection the values of  $N$  i  $M$  is stopped when the desire value of ③ is reached. Figure 5 presents the process of adding a new coefficient  $x \rho_{new}$  to the filter structure.

Using proposed strategy it is possible to determine the  $x \rho$  to eliminate these which only insignificant minimize the value of ③. It is impossible to determine one way of the parameters selection. Every configuration corrupted signal - reference signal is specific and requires the individual decision process.



**Fig. 5.** Process of updating the filter structure

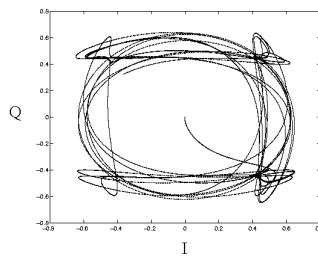
It is easy to notice it is not important if the next Schur coefficient  ${}^x\rho$  is higher order or higher degree. In both cases the new column of elementary sections is added to the decorrelation block.

It should be notice, the new coefficient  ${}^x\rho_{new}$  can not be the neighbor of already existing one [1]. If the last determined  ${}^x\rho$  came from the filter eg:  $N = 2, M = 3$  the new  ${}^x\rho_{new}$  can come from the filter eg:  $N = 5, M = 10$ .

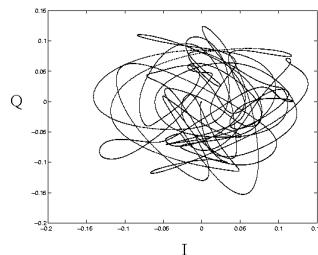
## 5 Simulations

To present proposed strategy in the first simulation a telecommunication signal (MSK modulation) was used (figure 6). Corrupting signal had non symmetric probability density function, and was lowpass. Bandwidths of both signals overlap. Figure 7 presents the IQ graph of corrupted signal. It is impossible to demodulate the original signal because of echo. Figure 8 presents demodulated signal after filtering. The filter of parameters  $M = 3$  and  $N = 10$  was used. Improvement of S/N ratio is about 14dB.

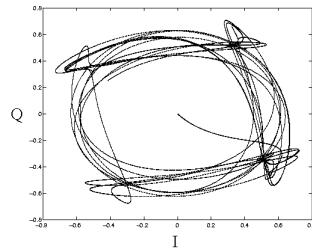
The decorrelation block had 40755 elementary sections and the estimation block had 284 elementary sections. Using criterion (II) and choosing  $\delta = 100$ , for every instant of time the number of the estimation block elementary sections was established, presented in figure 9. It can be seen how the structure of the filter is changing during the time.



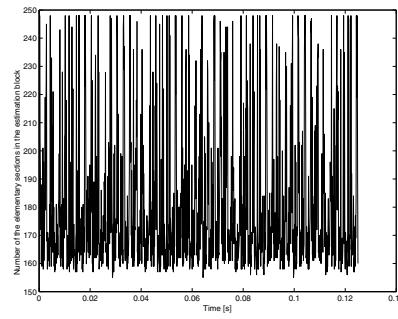
**Fig. 6.** IQ graph of original MSK signal



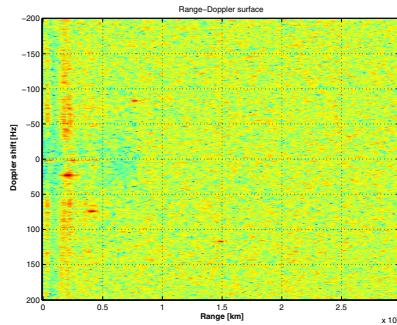
**Fig. 7.** IQ graph of corrupted MSK signal



**Fig. 8.** IQ graph of MSK signal after filtering



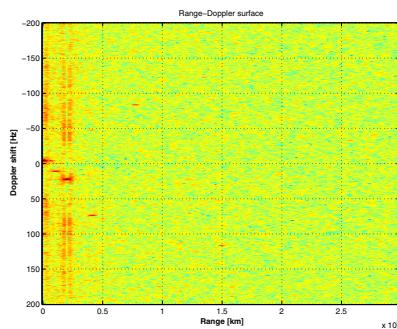
**Fig. 9.** The number of the estimation block elementary sections for strategy I



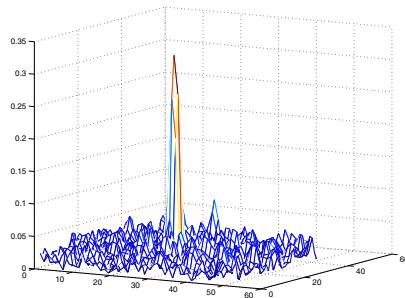
**Fig. 10.** 'Range-doppler' surface - 'full' filter structure

In the second simulation proposed intelligent structure selection algorithm was used for echo cancelling in a passive radar system. First figures show the 'Range-doppler' surface for the radar observed area. Red points mean flying objects.

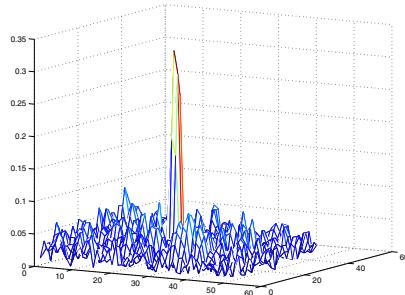
Next figures (12,13) present the fragment of the 'Range-doppler' surface near  $Range = 150\text{km}$  and  $Dopplershift = 120\text{Hz}$  in 3D style.



**Fig. 11.** 'Range-doppler' surface - optimized filter structure



**Fig. 12.** 'Range-doppler' surface - 'full' filter structure (a fragment of the 'Range-doppler' surface in 3D style)



**Fig. 13.** 'Range-doppler' surface - optimized filter structure (a fragment of the 'Range-doppler' surface in 3D style)

It can be seen how high is the peak of the detected object. The measure of the reference signal cancellation performance were established (11.10dB for adaptive selected structure, 11.68dB for 'full' structure). The optimal structure of the filter was 28 times less elements (elementary filter sections) than the full filter structure).

## 6 Conclusion

Presented results allows to draw the following conclusions:

- The structure of presented filters makes possible their rebuilding very easy. Using proposed strategy of filter parameter selection the filter complexity can be adopted to filtered signals.
- The value of the estimation error can be controlled by selecting filter structure during filtering process.

## References

1. Lee, D.T.L., Morf, M., Friedlander, B.: Recursive Least-Squares Ladder Estimation Algorithms. *IEEE Trans. on CAS* 28, 467–481 (1981)
2. Schetzen, S.: The Volterra & Wiener Theories of nonlinear systems. John Wiley & Sons, New York (1980)
3. Dijksma, A., Wanjala, G.: Generalized Schur Functions and Augmented Schur Parameters. In: Operator Theory in Krein Spaces and Nonlinear Eigenvalue Problems, pp. 135–144. Birkhäuser, Basel (2006)
4. Marple, S.L.: Two-dimensional lattice linear prediction parameter estimation method and fast algorithm. *IEEE Signal Process. Lett.* 7(6), 164–168 (2000)
5. Serban, I., Turcu, F., Najim, M.: Schur coefficients in several variables. *J. Math. Anal. Applicat.* 320, 293–302 (2006)
6. Borys, A.: Nonlinear Aspects of Telecommunications: Discrete Volterra Series and Nonlinear Echo Cancellation. CRC Press LLC, Boca Raton (2001) ISBN: 9780849325717
7. Stenger, A., Kellermann, W., Rabenstein, R.: Nonlinear acoustic echo cancellation with 2nd order adaptive volterra filters. In: ICASSP (1999)
8. Stenger, A., Kellermann, W.: RLS-Adapted Polynomial for Nonlinear Acoustic Echo Cancelling. *Signal Processing* 80, 1747–1760 (2000)
9. Biernacki, P.: Orthogonal Schur-type solution of the nonlinear echo-cancelling problem. In: Proc. 14th IEEE International Conference on Electronics, Circuits and Systems. ICECS 2007, Marrakech, Morocco, December 11-14, pp. 318–321. IEEE, Piscataway (2007)
10. Biernacki, P.: Geometric solution of the nonlinear echo cancelation problem. In: Proc. International Conference on Signals and Electronic Systems, ICSES 2006, Lodz, Poland, September 17-20, vol. 1, pp. 171–174 (2006)

# Intelligent System for Commercial Block Recognition Using Audio Signal Only

Pawel Biernacki

Telecom, Acoustic and Computer Science Institute,  
Wroclaw University of Technology,  
Wyb. Wyspianskiego 27, 50-350 Wroclaw, Poland  
 [{pawel.biernacki}@pwr.wroc.pl](mailto:{pawel.biernacki}@pwr.wroc.pl)

**Abstract.** In this article the effective method of a single commercial extracting from a advertising block and its recognition using only the audio signal is presented. Proposed algorithm uses a multidimensional orthogonal audio signal representation for a track parametrization. Simulation results for poor commercial audio signal recording conditions and comparison with the known methods are presented. The proposed solution gives a recognition at the level of 98%. This is the result better than the popular methods based on spectral analysis.

**Keywords:** audio signal recognition, orthogonal filter, multidimensional signal representation.

## 1 Introduction

One of the main function of the media research companies is the radio and television commercial blocks monitoring. To investigate what and when is broadcast the efficient and intelligence methods of a single commercial extraction and recognition are needed. Existing solutions base on Hidden Marcov Models [2], video signals [3] [4] or use different frequency domain 'audio fingerprinting' methods [9][11][10]. The aim of the study is to propose method for identifying advertising spots, which may differ from each other only small piece, such as a single word. In addition, the new method should work well for signals of poor quality recording. Every day about 10000 commercials are broadcasting, which must be recognized. Therefore, improving the quality of diagnosis, even by 1% compared with existing methods is a big achievement.

This article consists of the following parts. In section 2 a single commercial extraction from advertising block is discused. Section 3 describes the proposed method of parameterization of a single audio track. In section 4 the simulation results are presented.

## 2 Single Commercial Extraction

For a commercial start and end points the following elements can be used:

- short silence period after each commercial (no every silence denotes commercial end)
- spectrum differences before and after silence period
- commercial duration is often a multiple of 5 seconds periods

Listed above conditions for start/end commercial point identification have practical applications in TV commercial block division. In the radio broadcasting case the commercials often don't have clear start/end points (e.g. a track or station identifier appears in background). In such situations a single commercial separation is very difficult or even impossible.

Proposed intelligent solution consist of the following three steps:

- **step 1** - silence detection; empirical research shows that silence duration between commercials is from 60 ms to 200 ms long. Different silence periods are very often the parts of the commercials. To detect silence period the two criteria are used:

- audio signal  $x(t)$  magnitudes are between some maximum value

$$|x(t)| < Max, \quad Max - const \quad (1)$$

- signal energy is below some constant value

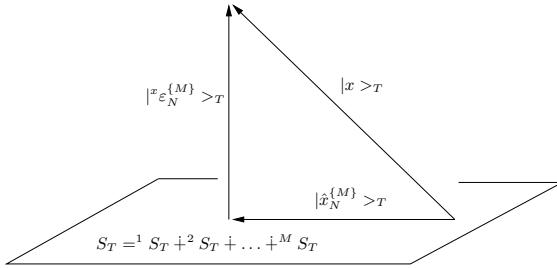
$$Energy = \sum_{t=t1}^{t2} |x(t)|^2 < Energy_{max} \quad (2)$$

where  $t1$  and  $t2$  are start and end points when condition (1) is true.

- **step 2** - power spectrums density (Welch method) of the audio signal fragments of the duration  $T1[s]$  ( $T1$  - algorithm parameter) before and after silence period are calculated. If the distance between the spectrums is too big (there is a small correlation between examined signal fragments) then the detected silence period (step 1) is a potential start or end point of the commercial
- **step 3** - having the set of potential start/ end points of the commercials and assuming the commercial duration as a multiple of 5 seconds periods the proper start points of the following commercials are established (Experimental results allowed to estimate the probability density function of appearing the commercial of the particular duration. This function is used to take a decision of the start/end point appearing).

### 3 Single Commercial Recognition

In proposed solution a single commercial is parametrized, making up 'the commercial fingerprint'. Commercial identification consist in searching examined fingerprint in the commercial fingerprint database.



**Fig. 1.** The estimate  $|\hat{x} >_T|$  of the desired signal

### 3.1 Multidimensional Audio Signal Parametrization

Given a vector  $|y >_T$  of samples  $\{y_0, \dots, y_T\}$  of a commercial track (a voice+music signal), observed on a finite time interval  $\{0, \dots, T\}$ , the signal parametrization problem can be stated as follows (see Figure 1). The estimate of the desired signal

$$|\hat{x}_N^{(M)} >_T| \triangleq \mathbf{P}(S_T)|x >_T \quad (3)$$

is the orthogonal projection of the element  $|x >_T|$  on the space  $S_T$  spanned by the following set of the linear and nonlinear observations

$$|Y >_T| = [|^1 Y >_T |^2 Y >_T \dots |^M Y >_T] \quad (4)$$

where

$$|^m Y >_T = [|y_{i_1} \dots y_{i_m} >_T; i_1 = 0, \dots, N, \\ i_2 = i_1, \dots, N, \dots, i_m = i_{m-1}, \dots, N] \quad (5)$$

for  $m = 1, \dots, M$ . The orthogonal projection operator on  $|Y >_T|$  is defined as

$$\mathbf{P}(S_T) \triangleq |Y >_T \langle Y | Y >_T^{-1} \langle Y |_T \quad (6)$$

If an ON (generalized; i.e., multidimensional) basis of the space  $S_T$  is known, the projection operator on  $S_T$  can be decomposed as

$$\begin{aligned} \mathbf{P}(S_T) &= \sum_{j_1=0}^N \mathbf{P}(|r_0^{j_1} >_T) + \dots + \\ &+ \sum_{j_1=0}^N \dots \sum_{j_M=j_{M-1}}^N \mathbf{P}(|r_0^{j_1, \dots, j_M} >_T) \end{aligned} \quad (7)$$

where  $\mathbf{P}(|r_0^{j_1, \dots, j_M} >_T)$  stands for the orthogonal projection operator on the one-dimensional subspace spanned by the element  $r_0^{j_1, \dots, j_M}, m = 1, \dots, M$  of an ON basis of the space  $S_T$ . Since

$$\mathbf{P}(|r_0^{j_1, \dots, j_w} >_T) = |r_0^{j_1, \dots, j_w} >_T \langle r_0^{j_1, \dots, j_w}|_T \quad (8)$$

the orthogonal expansion of the estimate of the desired signal can be written as

$$\begin{aligned} |\hat{x}_N^{\{M\}} >_T = \mathbf{P}(S_T) |x >_T &= \sum_{j_1=0}^N |r_0^{j_1} >_T < r_0^{j_1} |x >_T + \\ &+ \dots + \sum_{j_1=0}^N \dots \sum_{j_M=j_{M-1}}^N |r_0^{j_1, \dots, j_M} >_T < r_0^{j_1, \dots, j_M} |x >_T \end{aligned} \quad (9)$$

The estimation error associated with the element  $|\hat{x}_N^{\{M\}} >_T$  is then

$$|{}^x\varepsilon_N^{\{M\}} >_T \triangleq \mathbf{P}(S_T^\perp) |x >_T = |x >_T - |\hat{x}_N^{\{M\}} >_T \perp S_T \quad (10)$$

The estimate (3) will be called optimal (in the least-squares sense) if the norm

$$\| |{}^x\varepsilon_N^{\{M\}} >_T \| = {}^x\varepsilon_N^{\{M\}} |{}^x\varepsilon_N^{\{M\}} >_T^{\frac{1}{2}} \quad (11)$$

of the estimation error vector (10) is minimized for each  $T = 0, 1, 2, \dots$

The multidimensional signal parametrization problem can be solved by the derivation of a (generalized) ON basis of the estimation space  $S_T$  (i.e. calculation of the orthogonal representation (the generalized Fourier coefficients [8]) of the vector  $|x >_T$  in the orthogonal expansion (9)).

To derive the desired ON basis of the estimation space  $S_T$ , we employ (consult [5]) the following

**Theorem 1.** *The partial orthogonalization step results from the recurrence relations*

$$\begin{aligned} |e_{i_1, \dots, i_q}^{j_1, \dots, j_w} >_T &= [|e_{i_1, \dots, i_q}^{j_1, \dots, j_w-1} >_T + \\ &+ |r_{i_1, \dots, i_q+1}^{j_1, \dots, j_w} >_T \rho_{i_1, \dots, i_q; T}^{j_1, \dots, j_w}] (1 - (\rho_{i_1, \dots, i_q; T}^{j_1, \dots, j_w})^2)^{-\frac{1}{2}} \end{aligned} \quad (12)$$

$$\begin{aligned} |r_{i_1, \dots, i_q}^{j_1, \dots, j_w} >_T &= [|e_{i_1, \dots, i_q}^{j_1, \dots, j_w-1} >_T \rho_{i_1, \dots, i_q; T}^{j_1, \dots, j_w} + \\ &+ |r_{i_1+1, \dots, i_q+1}^{j_1, \dots, j_w} >_T] (1 - (\rho_{i_1, \dots, i_q; T}^{j_1, \dots, j_w})^2)^{-\frac{1}{2}} \end{aligned} \quad (13)$$

where

$$\rho_{i_1, \dots, i_q; T}^{j_1, \dots, j_w} = - < e_{i_1, \dots, i_q}^{j_1, \dots, j_w-1} | r_{i_1, \dots, i_q+1}^{j_1, \dots, j_w} >_T \quad (14)$$

Proof can be found in [6].

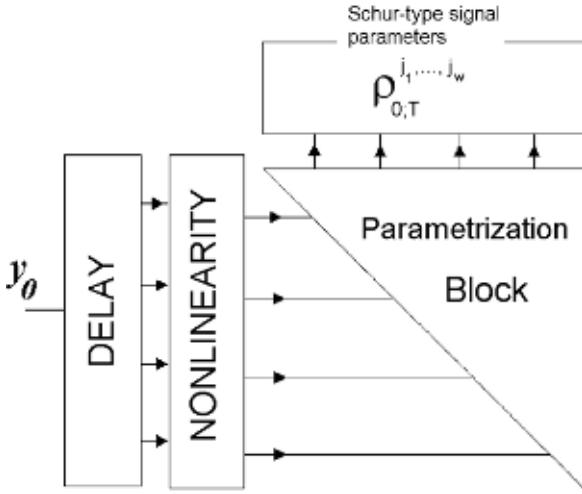
The above relations make it possible to construct an orthogonal parametrization (decorrelation) filter, operating directly on the signal samples (compare with [7]). The coefficients

$$\rho_{0; T}^{j_1, \dots, j_w} \quad (15)$$

can be interpreted as generalized Fourier coefficients which represent parametrized signal (commercial track) in the multidimensional space.

The above recurrence relations (13) actually solve the problem of the real-time derivation of the (generalized) ON basis of the estimation space. The diagram of the signal parametrization filter is presented in Figure 2.

The Schur coefficients (15) can be used for audio signal parametrization. Entire audio commercial parametrization procedure can be done in the tree steps:



**Fig. 2.** Nonlinear orthogonal signal parametrization filter

1. dividing audio commercial signal into the one second long pieces

$$y_{T_k} = \{y_T(n); n = k * f_p, k * F_p + 1, \dots, k * 2 * f_p - 1\} \quad (16)$$

- where  $k$  means  $k$ -th signal section and  $k = 0, 2, \dots, K - 1$ .
2. computing Schur coefficients (15) vector for each piece
  3. construct ID matrix for the audio commercial (using Schur coefficients vectors as its rows)

$$YM = \begin{pmatrix} \rho_{0;T_0}^0 & \dots & \rho_{0;T_1}^{j_1,...,j_w} & \dots \\ \vdots & & \vdots & \vdots \\ \rho_{0;T_{K-1}}^0 & \dots & \rho_{0;T_K}^{j_1,...,j_w} & \dots \end{pmatrix} \quad (17)$$

To identify the one audio commercial the distances between pattern commercial matrix vectors and identifying commercial matrix vectors are computing by

$$d_i(YM_1, YM_2) = \sum_l |YM_1(i, l) - YM_2(i, l)| \quad (18)$$

for  $i = 0, 1, \dots, K - 1$  and  $l$  changes from 1 to matrix row length. If this distance for one or more  $i$  is higher then establish value the recognition is negative.

## 4 Results

The presented algorithm of the audio commercial recognition was implemented in C language and tested in PC environment. The following signals were tested:

- radio FM station commercials

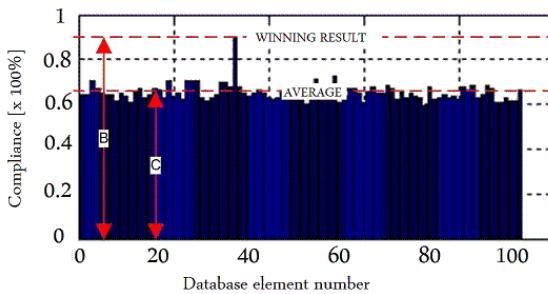
- audio track from TV commercials

All the commercials were 8-bit signals sampled with frequency 11.2kHz. Audio fingerprints database was created, which included 100 audio signals.

In order to compare the proposed solution with existing audio fingerprint methods one selected to test the following audio signal parametrization algorithms:

- Robust Audio Hashing (RAH) [9]
- Normalized Spectral Subband Moments (NSSM) [11]
- Mel-Frequency Cepstrum Coefficients (MFCC) [10][1]

To assess the quality of algorithms in addition to the standard measure of compliance the reference with recognized track (percentage compliance), a new measure was defined - confidence, which determines how much the winning result (for database searching) differs from the average search the database by this winning track. The higher this measure is that we have greater confidence that the recognition was effective. Figure 3 explains the ideas of this measure.



**Fig. 3.** Nonlinear orthogonal signal parametrization filter

$$\text{Confidence} = \frac{B}{C} \quad (19)$$

where  $B$  is the best result and  $C$  is the average result in the whole database.

Table 1 shows influence of the filter parameters (filter order  $N$  and nonlinearity degree  $M$ ) on recognition effectiveness. Increasing values of the filter parameters above some values does not correct the recognition effectiveness, but can make it worse. It seems that filter order about  $N = 20$  and degree nonlinearity  $M = 2$  is the best solution.

Selecting the best result ( $N = 20$ ,  $M = 2$ ) compared it with the results of other methods for parameterisation audio tracks. The results are shown in table 2.

There is no faultless method. Commercials were recorded in conditions different from their patterns, and this could hamper the identification. In addition, ads differ sometimes only one word which increased recognition errors.

**Table 1.** Influence of the filter parameters on the recognition effectiveness

Filter parameters	Recognition [%]
N=10, M=1	84
N=10, M=2	86
N=20, M=1	97
N=20, M=2	98
N=20, M=3	92
N=30, M=2	95

N - filter order, M - degree of nonlinearity

**Table 2.** Comparison of selected methods of parameterization

Method	Recognition [%]	Confidence
RAH	99	1.71
NSSM	92	1.27
MFCC	95	1.36
Proposed method	99	2.09

**Table 3.** Comparison of selected methods of parameterization - echo corruption

Method	Recognition [%]	Confidence
RAH	89	1.21
NSSM	79	1.17
MFCC	78	1.05
Proposed method	88	1.24

**Table 4.** Comparison of selected methods of parameterization - equalization corruption

Method	Recognition [%]	Confidence
RAH	85	1.45
NSSM	81	1.13
MFCC	80	1.16
Proposed method	85	1.49

The best recognition was obtained for the algorithms RAH and the proposed solution. For poor quality signals ( $f_s = 11.025$  kHz 8-bits/sample) algorithms NSSM and MFCC have fared much worse. The proposed in this article solution makes the greatest confidence value. The algorithm does not leave a doubt who is the winner (recognized track). It seems that confidence parameter allows selection of the proposed solution as the best, but we must remember the high computational complexity, which is growing rapidly with increasing parameterization order  $N$  and its nonlinearity degree  $M$ .

There were also conducted quality tests for recognition algorithms for tracks distorted by adding echoes, equalization, recording them through a microphone. Results of studies are shown in the tables 3,4 and 5.

**Table 5.** Comparison of selected methods of parameterization - commercial microphone recording

Method	Recognition [%]	Confidence
RAH	90	1.30
NSSM	89	1.22
MFCC	89	1.19
Proposed method	91	1.44

Above tables show that the proposed method gives a good recognition results. Only for echo distortion (table 3), RAH algorithm is better. If we look at the value of confidence parameter presented in the article solution is the best solution.

The proposed solution can operate in real time. The process of signal parameterization of the average television advertising, which lasts 30 seconds, takes approximately 10 seconds for N=20 and M=2 (result for PC computer, CPU AMD Athlon 2,4GHz). Searching a database consisting of 100 tracks takes approximately one second. So it is possible to keep track of advertisements and identify a single commercial in real time.

## 5 Conclusions

The presented results allow for the following conclusions:

- Presented algorithm is efficient in more than 98% for not corrupted audio tracks
- Presented solution is resistant on the poor recording conditions
- The value of the confidence parameter of the proposed solution very confirms the effectiveness of its recognition properties
- Proper parametrization filter parameters selection is necessary for the high recognition effectiveness

## References

1. Kay, S.M.: Modern Spectral Estimation. Prentice Hall, Englewood Cliffs (1988)
2. Rabiner, L.: Fundamentals of Speech Recognition. Prentice Hall PTR, Englewood Cliffs (1993)
3. Lienhart, R., Kuhmunch, C., Euelsberg, W.: On the detection and recognition of television commercials. In: Proc. IEEE Conf. on Multimedia Computing and Systems, Ottawa, Canada, pp. 509–516 (June 1997)
4. Zabih, R., Miller, J., Mai, K.: A feature-based algorithm for detecting and classifying scene breaks. In: ACM Conference on Multimedia, San Francisco, California (November 1995)
5. Biernacki, P., Zarzycki, J.: Multidimensional Nonlinear Noise-Cancelling Filters of the Volterra-Wiener Class. In: Proc. 2-Nd Int. Workshop on Multidimensional (nD) Systems (NDS-2000), pp. 255–261. Inst. of Control and Comp. Eng. TU of Zielona Gora Press, Czocha Castle (2000)

6. Biernacki, P., Zarzycki, J.: Orthogonal Schur-Type Solution of the Nonlinear Noise-Cancelling Problem. In: Proc. Int. Conf. On Signals and Electronic Systems (ICSES 2000), Ustron, pp. 337–342 (2000)
7. Lee, D.T.L., Morf, M., Friedlander, B.: Recursive Least-Squares Ladder Estimation Algorithms. IEEE Trans. on CAS 28, 467–481 (1981)
8. Schetzen, S.: The Volterra & Wiener Theories of nonlinear systems. John Wiley & Sons, New York (1980)
9. Haitsma, J., Kalker, T., Oostveen, J.: Robust audio hashing for content identification. In: Proc. of the Content-Based Multimedia Indexing, Firenze, Italy (September 2001)
10. Morgan, N., Bourlard, H., Hermansky, H.: Automatic Speech Recognition: An Auditory Perspective. In: Greenberg, S., Ainsworth, W.A. (eds.) Speech Processing in the Auditory System, p. 315. Springer, Heidelberg (2004) ISBN 9780387005904
11. Paliwal, K.K.: Spectral subband centroid features for speech recognition. In: Proc. IEEE ICASSP, pp. 617–620 (1998)

# Viewpoint Insensitive Actions Recognition Using Hidden Conditional Random Fields

Xiaofei Ji<sup>1,2</sup>, Honghai Liu<sup>1</sup>, and Yibo Li<sup>2</sup>

<sup>1</sup> Intelligent Systems and Biomedical Robotics Group

School of Creative Technologies, University of Portsmouth

<sup>2</sup> School of Automation

Shenyang Institute of Aeronautical Engineering

**Abstract.** The viewpoint issue has been one of the bottlenecks for research development and practical implementation of human motion analysis. In this paper, we introduce a new method, *e.g.*, hidden conditional random fields(HCRFs) to achieve viewpoint insensitive human action recognition. The HCRF model can relax the independence assumption of the generative models. So it is very suitable to model the human actions from different actors and different viewpoints. Experiment results on a public dataset demonstrate the effectiveness and robustness of our method.

**Keywords:** Human action recognition, Viewpoint insensitive, Conditional random field, Hidden conditional random field.

## 1 Introduction

Human action recognition from video is an important task due to its potential applications such as visual surveillance, human-robot interaction and content based video retrieval *etc.*. A challenging issue in this research field is the diversity of action information which originates from different camera viewpoints. An effective action recognition system should be correctly recognize human actions while the viewpoint of the camera is changing.

The view-invariant action recognition has received considering attention over the past decade [1]. Graphical models, *e.g.*, Hidden Markov Model(HMM) and its variants have been the dominant tools in view-invariant human motion modelling [2,3,4,5]. However those generative models require that observations are conditionally independent which makes the methods unsuitable for accommodating multiple overlapping features or long-range dependences among observations.

Conditional random field(CRF) and its variants were recently introduced to avoid the independent assumption between observations by using an exponential distribution to model the entire sequence given the observations. They support efficient inference using dynamic programing and their parameters can be learned using convex optimization. Sminchisescu *et al.* [6] first introduced CRFs to classify the diverse human motion activities. On the basis of their work, Quattoni *et al.* [7] proposed the HCRF by incorporating hidden state variables in random field model to extend CRF into the temporal domain. The model combines the

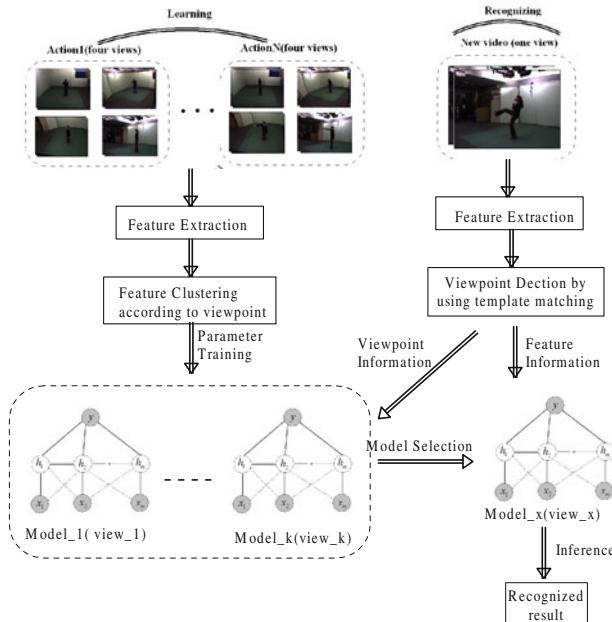
ability of CRFs to use dependent input features and the ability of HMMs to learn latent structure. The results have shown that HCRFs outperformed both CRFs and HMMs in arm gesture recognition. Zhang and Gong [8] formulated a modified HCRF to have a guaranteed global optimum in modeling the temporal action dependencies after the HMM pathing stage. Wang and Suter [9] exploited factorial CRF for activity modeling and recognition. They applied a two-chain CRF to simultaneously perform key-pose classification and activity classification.

In this paper, we introduce discriminate HCRFs for view insensitive human action recognition. No previous work has investigated HCRF in this context. Multiple HCRF models are trained by using different actions from different camera viewpoints to obtain view insensitive action recognition. This proposed method can release the dependence assumption of HMMs. Experiment results on one public dataset have demonstrated the effectiveness and robustness of our method.

The remainder of this paper is organized as follows. Section 2 gives a overview of our method. A brief introduction to HCRF is introduced in Section 3. Action modelling and recognition are proposed in Section 4. The results are presented and discussed in Section 5. The paper is concluded in Section 6 with analysis on future research challenges and directions.

## 2 The Overview of Proposed Method

Human actions involve both spacial(the body pose in each time step) and temporal(the transition of the body poses over time) characters in their representations.



**Fig. 1.** The framework of proposed method

The actual appearance of the spacial-temporal representation varies significantly with camera viewpoint changes. We propose the view-insensitive method based on the assumption that actor's orientation should be constant or the change is slight during they are performing one action. The framework of our method is shown as Fig. 1.

During the action modelling, we extract the motion features from each video. Then the training sequences from the same action are clustered into several different viewpoints. Multiple HCRFs are applied to model different human actions from different camera viewpoints. Each HCRF model is a single multi-class HCRF which is trained to recognize all action class under a particular camera viewpoint.

During the action recognition, we compute the similarity between the input image and the synthetic templates which are rendered from multiple viewpoints using POSER to estimate the possible range of camera viewpoint. The HCRF model with the similar viewpoint is used to calculate the probability. The highest probability is chosen as the recognized result.

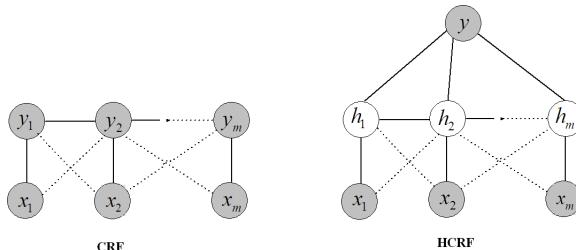
### 3 A Brief Introduction to HCRF

We give a brief introduction of HCRF as described in [10,11,12,8]. The graphical representations of linear-CRF and HCRF is shown in Fig. 2. Given a sequence of  $m$  local observation  $\{x_1, x_2, \dots, x_m\}$  denoted by  $\mathbf{x}$ , and its class label  $y \in Y$ , we want to find a mapping  $p(y|\mathbf{x})$  between them, where  $y$  is conditioned on  $\mathbf{x}$ .

An HCRF is defined as:

$$P(y|\mathbf{x}, \theta) = \sum_{\mathbf{h}} P(y, \mathbf{h}|\mathbf{x}, \theta) = \frac{\sum_{\mathbf{h}} e^{\Psi(y, \mathbf{h}, \mathbf{x}; \theta))}}{\sum_{y' \in Y, h \in H^m} e^{\Psi(y', h, \mathbf{x}; \theta)}} \quad (1)$$

where  $\theta$  is the set of parameter of the model, and  $\mathbf{h} = \{h_1, h_2, \dots, h_m\}$ , each  $h_i \in H^m$  captures certain underlying structure of each class and  $H$  is the set of hidden states in the model. The potential  $\Psi(y, \mathbf{h}, \mathbf{x}; \theta)$  measures the compatibility between a label, a set of observations and a configuration of the hidden state.



**Fig. 2.** The graphical representation of CRF and HCRF(Gray circles are observed variables

The following objective function is used in training parameter  $\theta$ :

$$L(\theta) = \sum_{i=1}^n \log P(y_i | \mathbf{x}_i, \theta) - \frac{\|\theta\|^2}{2\sigma^2} \quad (2)$$

where  $n$  is the total number of training sequences.

## 4 HCRFs for Action Recognition

We adopt sequential silhouettes as action cues to describe human actions. Sometimes silhouettes from different actions look quite similar from single camera. So it is difficult to achieve the independence assumption of observations. Owing to the advantage of the HCRF, it is very suitable for modelling and recognizing human actions. In this paper, we propose the view insensitive human action HCRF models which are trained by using the multiple view action sequences. The details are followed.

### 4.1 Silhouette Representation

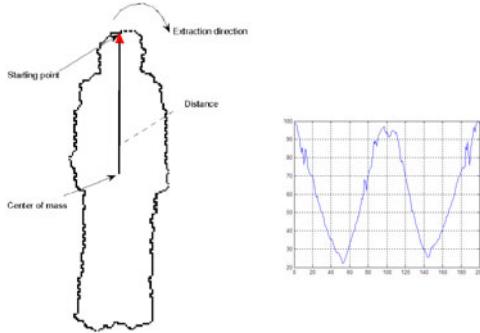
There are some representative shape features of the silhouette image in the previous papers, such as shape context descriptor [13], width feature [14]. We describe the silhouette images using the silhouette distance metric, in that it not only can capture both structural and dynamic information for an action, but it also can be efficiently obtained [15]. An example of the distance signal is shown in Fig.3, which is generated by calculating the Euclidean distance between the center of the mass points and each edge point of the silhouette images in clockwise direction.

In order to obtain image scale and rotation invariance, firstly the principle axis of the silhouette is computed before computing shape feature, the rotation angle of the principle axis is compensated so that the principal axis is vertical. Then the contour of the silhouette image is uniformly re-sampled to 200 edge points, and the distance is normalized into [0, 100].

In order to reduce computational cost, the action feature space is projected into the embedding space including the first 10 principle components by using kernel principle analysis(KPCA) methods. Then those low-dimension features are used to train action models.

### 4.2 Model Training

Our goal is to recognize different actions performed by different actors under different camera viewpoints, with different styles regardless of large variation in manner and speed. It is a challenging task to model so much various characters in one model. So we adopt multiple HCRFs to model human actions from different camera viewpoints. We first classify the training sequences from the same action into several different viewpoints. Then one multi-class HCRF model is trained to recognize all action class under a particular camera viewpoint.



**Fig. 3.** Contour shaper feature of the silhouette image

The long range dependencies is incorporated in our action models. So a window parameter  $\omega$  is introduced in the potential function. It defines the amount to past and future history to be used when predicting the state at time  $t$ . The potential function is defined [10]:

$$\begin{aligned} \Psi(y, \mathbf{h}, \mathbf{x}; \theta, \omega) = & \\ \sum_{(j,k) \in E} \theta_e[y, h_j, h_k] + \sum_{j=1}^m \theta_y[y, h_j] + \sum_{j=1}^m \varphi(x, j, \omega) \cdot \theta_h[h_j] & \end{aligned} \quad (3)$$

Where  $\theta_e[y, h_j, h_k]$  refers to parameters that correspond to class  $y$  and the pair of states  $h_j$  and  $h_k$ .  $\theta_y[y, h_j]$  stands for parameters that correspond to class  $y$  and state  $h_j$ .  $\theta_h[h_j]$  refers to the parameters that correspond to state  $y_j \in Y$ .

Given the training sequences, the parameter  $\theta$  is learn by maximizing the objective function:

$$\theta^* = \operatorname{argmax}_{\theta} L(\theta) \quad (4)$$

Gradient ascent e.g. Quasi-Newton optimization technique is used to search for the optimal parameter value.

### 4.3 Action Recognition

During action recognition, we first detect the possible range of camera viewpoint. Normally, the most of human actions is performed from the standing pose. We compute the similarity between the first frame of the input sequence with the synthetic standing templates. Those templates are rendered from multiple viewpoints using POSER. In our method, we assume that the tilt angle is known and only different pan angles should be considered( $20^\circ$  intervals), as shown in Fig. 4.

After the possible range of camera viewpoint is obtained, the HCRF model with the similar viewpoint is selected to calculate the probability. Given a new testing sequence  $\mathbf{x}$ , we estimate the most probable label sequence  $\mathbf{y}^*$  that maximizes the conditional model:

$$\mathbf{y}^* = \operatorname{argmax}_{\mathbf{y}} P(\mathbf{y} | \mathbf{x}, \theta^*) \quad (5)$$



**Fig. 4.** Multiple-view templates for standing pose

## 5 Experiments

### 5.1 Dataset

We demonstrate the proposed framework on a public dataset, the multiple view IXMAS dataset. It contains 12 actions (i.e., check watch, cross arms, scratch head, sit down, get up, turn around, walk in a circle, wave a hand, punch, kick, point, pick up), each of which was performed 3 times by 12 actors (5 females and 7 males). Each action is recorded by 5 cameras with the frame rate of 23 fps. In this dataset, actor orientations are arbitrary since no specific instructions were given during the acquisition, as shown in Fig 5. So this dataset is very suitable for testing the effectiveness of the view-invariant action recognition algorithm.

It is a challenging task to recognize the actions from this dataset in that the same action looks quite different when observed from different camera viewpoint, in addition to the variation cause by the personal action difference of actors.

Human silhouette extraction from image sequence is relatively easier for current vision techniques. How to achieve the silhouette images is not considered in our framework. Human silhouette images of the observation sequences are provided with the dataset. The quality of the silhouette image is general good



**Fig. 5.** The IXMAS database [16]

but many defects are also presented. So morphological close and open operations are applied to the image to deal with noise due to background subtraction.

## 5.2 Testing

Our proposed method is based on the assumption that actor's orientation should be constant or the change is slight during they are performing one action. we chose 8 actions (i.e., check watch, cross arms, scratch head, sit down, get up, wave a hand, punch, kick), performed by 10 actors, each 3 times, and viewed by 4 cameras (except top camera) as training and testing objects in our experiment. The action sequences were all manually segmented in advance, so no action segmentation was considered. 9 actors were used for model learning each time, another one was used to test the models. During every process, the training sequences from the same action were clustered into 4 different viewpoints. Then a single multi-class HCRF was trained to recognize all action class for one particular camera viewpoint. The highest probability was chosen as the recognize action. When the window parameter was set to 1 and the number of hidden states was 10, the recognition rates for each action class are listed in Table II

**Table 1.** The Recognition rate

action	recognition rate(%)
check watch	78.3
cross arms	76.7
scratch head	78.3
sit down	88.3
get up	86.7
wave a hand	78.3
punch	81.7
kick	81.7
overall	81.3

According to the training data, the pan angle of camera changes from  $0^\circ - 180^\circ$ , and the tilt angle changes from  $0^\circ - 45^\circ$ . In our experiment, we only applied four HCRF models to cover all possible camera viewpoints. Even we did not consider the effect of tile angle, one HCRF model covered almost  $45^\circ$  viewpoint changes. Our system achieves a satisfying average recognition rate of 81.3% using a single camera.

## 5.3 Comparison

**Comparison with CRF.** We evaluated HCRF with varying levels of long range dependencies and compared performance to CRF. We trained a linear

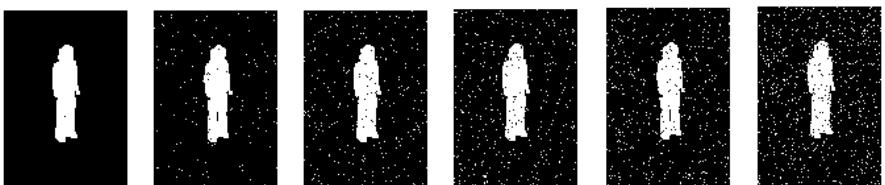
**Table 2.** The Recognition Comparison across Different Models

Models	recognition rate(%)
CRF( $\omega=0$ )	73.3
CRF( $\omega=1$ )	77.6
HCRF( $\omega=0$ )	79.8
HCRF( $\omega=1$ )	81.3

CRF which predicts labels for each frame in a sequence, not the entire sequence. During testing, the sequence label was assigned based on the most frequently occurring action label per frame. The comparison results is shown in the table②

From this table, we can find that: 1) The proposed framework using HCRF or CRF is all effective to recognize the different human actions from different actor and arbitrary viewpoint. 2) Owing to incorporate the hidden state, the performance of HCRF models is much better than CRF models. 3) Incorporating long-rang dependences is useful to improve the performance of CRFs and HCRFs.

**Robustness to Noise.** The proposed method abstracts the human action features from human silhouette images. However in the real world, the human silhouette images are always accompanied by noise from background subtraction which could be caused by change of illumination of scene, shadows *etc.*. In order to test the robustness of the proposed method to noise, the salt & pepper noise with different variances is added to the silhouette. An example of silhouette images with different degrees of noise is shown in the Figure⑥. Then we use the original silhouette sequence for training and the noise-silhouette for testing. The accuracies of action recognition with respect to different noise densities are shown in table③.



**Fig. 6.** Silhouette images with different degrees of noise: from left to right: noise densities are respectively 0, 0.02, 0.04, 0.06, 0.08, 0.1.

As illustrated in table③ that there is no significant effect in the recognition performance owing to the noise. The proposed method is robust to a considerable amount of noise.

**Table 3.** The Recognition Comparison with Different Noise

Noise Density	recognition rate(%)
0	81.3
0.02	80.8
0.04	80.3
0.06	79.6
0.08	79.1
0.10	78.8

## 6 Conclusion

A view insensitive action recognition method using multiple HCRF models has been proposed and tested on a challenge dataset in this paper. Experimental results have demonstrated that the proposed method can effectively recognize human actions performed by different people and different actions types. Furthermore the proposed method exhibited significantly robustness to camera viewpoint changing. Our future work are targeted as follows:

1. Different action features have various discriminative abilities, such as silhouettes, shapes, appearances, optical flow *etc*. It is necessary to study the characterize of those cues and fuse multiple features of action to improve the algorithm's effective and robustness [14][17][9].
2. It is difficult to model the variations of different human actions from different actors and different viewpoints simultaneity by using linear HCRFs, introducing the multi-layer HCRFs [18] into view-invariant human action recognition is our next work.

## References

1. Ji, X., Liu, H.: Advances in view-invariant human motion: A review. *IEEE Trans. System, Man, and Cybernetics-part C: Applications and Reviews* 40, 13–24 (2010)
2. Ahmad, M., Lee, S.: Hmm-based human action recognition using multiview image sequences. In: Proc. Int Conf. Pattern Recognition, vol. 1, pp. 263–266 (2006)
3. Weinland, D., Grenoble, F., Boyer, E., Ronfard, R., Inc, A.: Action recognition from arbitrary views using 3D exemplars. In: Proc. IEEE Conf. Computer Vision, pp. 1–7 (2007)
4. Lv, F., Nevatia, R.: Single view human action recognition using key pose matching and viterbi path searching. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1–8 (2007)
5. Yang, Y., Hao, A., Zhao, Q.: View-invariant action recognition using interest points. In: Proc. Int. Conf. Multimedia information retrieval, pp. 305–312 (2008)

6. Sminchisescu, C., Kanaujia, A., Metaxas, D.: Conditional models for contextual human motion recognition. *Computer Vision and Image Understanding* 104(2-3), 210–220 (2006)
7. Quattoni, A., Wang, S., Morency, L., Collins, M., Darrell, T., Csail, M.: Hidden-state conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2007)
8. Zhang, J., Gong, S.: Action categorization with modified hidden conditional random field. *Pattern Recognition* 43, 197–203 (2010)
9. Wang, L., Suter, D.: Recognizing human activities from silhouettes: motion subspace and factorial discriminative graphical model. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1–8 (2007)
10. Wang, S., Quattoni, A., Morency, L., Demirdjian, D., Darrell, T.: Hidden conditional random fields for gesture recognition. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 1521–1527 (2006)
11. Wang, Y., Mori, G.: Learning a discriminative hidden part model for human action recognition. *Advances in Neural Information Processing Systems (NIPS)* 21, 1721–1728 (2008)
12. Liu, F., Jia, Y.: Human action recognition using manifold learning and hidden conditional Random Fields. In: The 9th International Conference for Young Computer Scientists, pp. 693–698 (2008)
13. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Analysis and Machine Intelligence* 24(4), 509–522 (2002)
14. Cherla, S., Kulkarni, K., Kale, A., Ramasubramanian, V.: Towards fast, view-invariant human action recognition. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1–8 (2008)
15. Dedeoğlu, Y., Töreyin, B., Güdükbay, U., Çetin, A.: Silhouette-based method for object classification and human action recognition in video. In: Proc. European Conf. Computer Vision, pp. 62–77 (2006)
16. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding* 104, 249–257 (2006)
17. Hu, Y., Cao, L., Lv, F., Yan, S., Gong, Y., Huang, T.S.: Action detection in complex scenes with spatial and temporal ambiguities. In: Proc. IEEE Conf. Computer Vision, pp. 1–8 (2009)
18. Natarajan, P., Nevatia, R.: View and scale invariant action recognition using multiview shape-flow models. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1–8 (2008)

# Fuzzy Hyper-Prototype Clustering

Jin Liu and Tuan D. Pham

School of Engineering and Information Technology

University of New South Wales

Canberra, ACT 2600, Australia

j.liu@student.adfa.edu.au, t.pham@adfa.edu.au

**Abstract.** We propose a fuzzy hyper-prototype algorithm in this paper. This approach uses hyperplanes to represent the cluster centers in the fuzzy c-means algorithm. We present the formulation of a hyperplane-based fuzzy objective function and then derive an iterative numerical procedure for minimizing the clustering criterion. We tested the method with data degraded with random noise. The experimental results show that the proposed method is robust to clustering noisy linear structure.

## 1 Introduction

Cluster analysis is one of the most useful techniques in computational data analysis in signal and image analysis. Generally, clustering algorithms can be categorized into two types, namely hierarchical clustering and partitional clustering. Hierarchical clustering aims to build a tree-like partition structure of the data and it can further be divided into agglomerative [1] or divisive [2] according to whether the tree building is a bottom-up or top-down process. The former one starts with all the data and calculates the pairwise similarity between data points and then form a proximity matrix, after that the algorithm merges those closest samples into same group and recalculate the proximity matrix. This procedure repeats until all the data are grouped in one same cluster. Being different from the agglomerative clustering, divisive clustering starts from one cluster with all the data, and divides the cluster into a group of new clusters until that all the clusters contain only one sample. Since for a cluster with  $N$  samples, there will be  $2^{N-1} - 1$  possible divisions, the process is computationally expensive, which prevents divisive clustering from practical use.

Being different, partitional clustering algorithms directly separate the data into a group of clusters according to some criteria. Partitional clustering can be further divided into two categories, namely hard partitional clustering and fuzzy partitional clustering, according to whether the membership of each sample to cluster is a binary or a continuous value. The former one allows each sample to belong to only one cluster, while the later one allows each sample to belong to more than one cluster, and hence suitable for overlapping data sets. Many clustering methods fall into partitional clustering, i.e.,  $c$ -means clustering, partition around mediods [3], self-organizing maps [4], and so on.

Cluster analysis has been applied in a wide range of areas, ranging from biology, medicine, computer science, geology to economy. Despite huge amount of applications of cluster analysis, many of the current clustering algorithms, such as  $c$ -means and fuzzy  $c$ -means, may better suit data with spherical shapes. For some other certain types of data structures, for instance, linear or hyperplane shaped data clusters, most of the current clustering method may not perform well [5]. Meanwhile, although some methods like graph-theoretic methods are capable to detect linear or non-linear cluster structure, there is no explicit representations for the clusters, and hence they are difficult to be implemented to perform classification. Also, in some areas like computer vision, clustering algorithms for structure segmentation may not only involve the cluster centers but also need to consider the linear geometry of the clusters. Moreover, real data like microarray gene expression data sets often overlap with each other. For clustering algorithm, how to take both overlapping and linear subspace structure into account is worth to investigate.

In this paper, we present an unsupervised fuzzy hyper-clustering technique which is a kind of fuzzy clustering with hyperplane-represented cluster centers. This work is an on-going development of our conceptual framework in fuzzy hyper-clustering analysis for handling noisy signals [6]. The aim of the hyper-clustering is to find a partition matrix and a group of hyperplanes that minimize the sum of the weighted distance of all data to cluster centers. The clustering problem can then be considered as a constrained optimization problem. We derive a solution that iteratively updates cluster centers and partition matrix until convergence achieved.

Organization of the paper is as the follows. Section 2 presents some related work. Section 3 describes the proposed fuzzy hyperplane clustering in detail, including the objective function, derivation for minimizing the objective function and description of the result algorithm. In Section 4, we reported the experimental result of FHC and conclusions are given in Section 5.

## 2 Related Work

Some methods which are closely related to the proposed fuzzy hyper-clustering are briefly discussed in the following subsections.

### 2.1 Hyperplane-Based Clustering and Classification

The  $k$ -plane clustering algorithms use hyperplane as cluster centers. The first variant is  $k$ -bottleneck hyperplane clustering ( $k$ -bHPC) [7]. The aim of  $k$ -bHPC is to partition data into several groups, and to find a hyperplane for each group that minimizes the maximum distance between data points to their projections on the hyperplane. The objective function can be written as

$$\min \max \frac{|\mathbf{w}_j^t \mathbf{x}_i - v_j|}{\|\mathbf{w}_j\|_2^2} \quad (1)$$

where the  $\{\mathbf{w}_j, v_j\}$  is the hyperplane of the  $j$ -th cluster and  $\mathbf{x}_i$  is the  $i$ -th sample.

Another variant [8] of the  $k$ -planes clustering uses hyperplanes to represent cluster centers and the objective of the clustering is to minimize the sum of the squared Euclidean distances between data and their projections on the hyperplane. This  $k$ -planes clustering algorithm iteratively updates the partition matrix and clustering hyperplanes until convergence reached. The update of partition matrix is achieved by assigning data to the closest hyperplane and the update for the hyperplane is achieved through eigenvalue decomposition. This  $k$ -planes clustering was extended by considering the fuzzy membership, or by combining with sparse component analysis.

The support vector machines (SVMs) [9] is a well known classification method, which aims to find an optimal separating hyperplane that maximizes the margin between different types of data. Some extensions of SVMs made efforts to reduce the computational burden while maintain the predictive accuracy. In these work, hyperplane is not used to separate but to approximate each type of data. The optimal hyperplane minimizes the sum of squared Euclidean distance of one type data and meanwhile maximizes the sum of squared Euclidean distance of the other type data. The objective function can be written in the form of Rayleigh quotient and the solution can be obtained through generalized eigenvalue decomposition. Some of the approximating hyperplanes are parallel to each other [10,11], some are extended to be non-parallel [12,13,14], and others were extended to perform multi-category classification by using the one-from-the-rest approach[11].

## 2.2 Fuzzy $c$ -Varieties

Fuzzy  $c$ -varieties (FCV) is an extension of fuzzy  $c$ -means clustering proposed by Bezdek et al. [5,15]. Cluster prototypes in FCV are represented as a group of linear varieties, and the clustering criteria is the sum of squared Euclidean distances between data and prototypes.

FCV is reliable in finding linear structure of data when all the clusters are  $r$ -dimensional linear varieties. But as pointed by Bezdek, FCV has the size problem. This is because that the linear varieties are infinite and data clusters far from each other may be grouped into same cluster if they are collinear. To overcome this problem, Bezdek proposed to consider the center of mass  $\mathbf{v}_j$  of each cluster and introduced another combination of objective function.

$$J = (1 - \alpha)J_{V0m}(U, V_{0c}) + \alpha J_{Vrm}(U, V_{rc}) \quad (2)$$

where  $\alpha \in [0, 1]$ , the larger  $\alpha$  the more punishment for data far from the mass center even they are closer to the linear variety.

There have been developments based on the FCV. In [16], the author extended the FCV into kernel Hilbert space and presented a kernelized FCV. Some other work viewed FCV as a combination of local principal component analysis and fuzzy clustering, and extended FCV with independent component analysis [17], or using some other distance measures instead of Euclidean distance to provide a robust version of the FCV [18].

### 2.3 Some Other Related Methods

Some other related methods include the generalized principle component analysis (GPCA) [19], non-negative matrix factorization (NMF) [20], and optimal separating hyperplane [21]. The GPCA is an extension of principle component analysis proposed by Vidal recently, which aims to find a group of basis for the local linear subspaces, the dimension of the subspace can be different. The NMF is a method for low rank matrix approximation which attracts considerable attention in research community recently. It is similar with other matrix factorization methods like PCA, ICA etc., the difference lies in that the NMF requires all the data entries of the matrix to be non-negative, which makes the factorized result easier to interpret. It has also been shown that NMF has close relationship with clustering methods like  $k$ -means and spectral clustering. The optimal separating hyperplanes solves the problem of finding an optimal hyperplane that minimizes the distances from all the misclassified data points to the hyperplane, the distance can be of arbitrary norm.

### 2.4 Trade-Offs

The novelty of the proposed method against the methods listed above is to find a group of hyperplanes that best-fit the hidden data structure. Although GPCA and  $k$ -plane clustering aim to find linear structure in the subspace, they only use crisp membership and data are not allowed to belong to more than one subspace which makes them not suitable for analyzing overlapping data sets, for instance, the microarray gene expression data. Another difference between the proposed method from FCV,  $k$ -plane clustering and GPCA is that in the objective function of the proposed method, the criterion is the sum of distance between data and cluster centers, rather than squared Euclidean distance. The difference in the objective function leads to a different update of membership function, and may give different convergence results. Moreover, the proposed method tries to find a linear subspace spanned according to the direction of minor variance while the FCV tried to find a subspace spanned according to the direction of the largest  $r$  principle variance of the fuzzy scatter matrix. That is also the reason that in some work the FCV was viewed as a combination with local PCA and fuzzy clustering [17].

## 3 Fuzzy Hyper-Clustering

Being motivated by the useful concepts of coupling fuzzy  $c$ -means clustering with hyperplane mapping, we propose herein a fuzzy hyper-clustering (FHC) technique, and discuss the method in the subsequent sections. In the following parts of the paper, all the vectors are column vectors by default and written in bold. Transpose form of the vectors or matrix are written with superscript  $t$ . For instance,  $\mathbf{w}_j$  denotes a column vector, which is the normal vector of the  $j$ -th hyperplane, and  $\mathbf{w}_j^t$  denotes its transpose.

### 3.1 Hyperplane-Based Fuzzy Clustering

Being different from most current clustering techniques such as  $c$ -means and fuzzy  $c$ -means clustering, which represent the cluster centers using  $p$ -dimensional mean vectors of the data, the proposed fuzzy hyper-clustering adopts the geometrical hyperplanes, which have been employed to develop support vector machines and other kernel-based methods [22], to represent its cluster centers  $\mathbf{h}_j = (\mathbf{w}_j, v_j)$ ,  $j = 1, \dots, c$ , where  $c$  is the number of clusters. In the succeeding sections, we refer  $\mathbf{h}_j$  as a hypercluster.

In the proposed clustering technique, sample points are assigned fuzzy memberships to each hypercluster according to its distances to the hyperclusters. The aim of the fuzzy hyper-clustering is to find a fuzzy partition matrix  $\mathbf{U} = [u_{ij}]$ ,  $i = 1, \dots, n$ ,  $n$  is the number of samples; and hyperclusters  $\mathbf{h}_j$ ,  $j = 1, \dots, c$ , that minimizes the sum of the distances from all points to all hyperclusters. Where  $u_{ij}$  is the fuzzy membership of  $i$ -th object data vector to  $j$ -th hypercluster.

The resulting partition matrix and hyperclusters would minimize the following objective function

$$J = \sum_{i=1}^n \sum_{j=1}^c u_{ij}^m d(\mathbf{x}_i, \mathbf{h}_j) \quad (3)$$

where  $\mathbf{h}_j$  is the  $j$ -th hypercluster  $(\mathbf{w}_j, v_j)$ ,

$$\mathbf{h}_j = \{w_{1j}, w_{2j}, w_{3j}, \dots, w_{pj}, v_j\}$$

and  $\mathbf{w}_j = \{w_{1j}, w_{2j}, w_{3j}, \dots, w_{pj}\}$  is a  $p$ -dimensional normal vector to the  $j$ -th hypercluster. The distance from a data point to the hypercluster is

$$d(\mathbf{x}_i, \mathbf{h}_j) = \frac{|\mathbf{w}_j^t \cdot \mathbf{x}_i - v_j|}{\|\mathbf{w}_j\|^2} \quad (4)$$

$$\|\mathbf{w}_j\| = 1; j = 1, \dots, c; \exists w_{ij} \neq 0 \quad (5)$$

$$\sum_{j=1}^c u_{ij} = 1, i = 1, \dots, n, u_{ij} \in [0, 1] \quad (6)$$

where  $\mathbf{w}_j^t \cdot \mathbf{x}_i$  denotes the dot product between vector  $\mathbf{w}_j^t$  and vector  $\mathbf{x}_i$ .

To obtain a solution that minimizes the above objective function  $J$ , we adopt an iterative numerical model which updates the fuzzy partition process until a convergence of the solution is reached. This process is analogous to the fuzzy  $c$ -means clustering algorithm.

### 3.2 Derivation of Iterative Numerical Method

Given the objective function to minimize

$$J = \sum_{i=1}^n \sum_{j=1}^c u_{ij}^m \frac{|\mathbf{w}_j^t \cdot \mathbf{x}_i - v_j|}{\|\mathbf{w}_j\|^2} \quad (7)$$

subject to constraints expressed in Eq. 5 and Eq. 6.

The Lagrangian of the objective function can be written as

$$L = \sum_{i=1}^n \sum_{j=1}^c u_{ij}^m \frac{|\mathbf{w}_j^t \cdot \mathbf{x}_i - v_j|}{\|\mathbf{w}_j\|^2} - \sum_{j=1}^c \lambda_j (\mathbf{w}_j^t \cdot \mathbf{w}_j - 1) - \sum_{i=1}^n \alpha_i (\sum_{j=1}^c u_{ij} - 1) \quad (8)$$

Taking the first derivatives of  $L$  with respective to  $u_{ij}$

$$\partial L / \partial u_{ij} = m u_{ij}^{m-1} \frac{|\mathbf{w}_j^t \cdot \mathbf{x}_i - v_j|}{\|\mathbf{w}_j\|^2} - \alpha_i \quad (9)$$

and setting Eq. 9 equal to 0, we get

$$\begin{aligned} u_{ij}^* &= \left[ \frac{\alpha_i}{m} \frac{\|\mathbf{w}_j\|^2}{|\mathbf{w}_j^t \cdot \mathbf{x}_i - v_j|} \right]^{\frac{1}{m-1}} \\ &= \left[ \frac{\alpha_i}{m} \right]^{\frac{1}{m-1}} \left[ \frac{1}{|\mathbf{w}_j^t \cdot \mathbf{x}_i - v_j|} \right]^{\frac{1}{m-1}} \end{aligned} \quad (10)$$

Substituting Eq. 10 into Eq. 6, we get

$$\sum_{j=1}^c \left[ \frac{\alpha_i}{m} \frac{\|\mathbf{w}_j\|^2}{|\mathbf{w}_j^t \cdot \mathbf{x}_i - v_j|} \right]^{\frac{1}{m-1}} = 1 \quad (11)$$

then we have

$$\left[ \frac{\alpha_i}{m} \right]^{\frac{1}{m-1}} = \sum_{j=1}^k \left[ \frac{1}{|\mathbf{w}_j^t \cdot \mathbf{x}_i - v_j|} \right]^{\frac{-1}{m-1}} \quad (12)$$

Substituting Eq. 12 into Eq. 10, we get

$$u_{ij}^* = \frac{\frac{1}{|\mathbf{w}_j^t \cdot \mathbf{x}_i - v_j|}^{\frac{1}{m-1}}}{\sum_{j=1}^c \frac{1}{|\mathbf{w}_j^t \cdot \mathbf{x}_i - v_j|}^{\frac{1}{m-1}}} \quad (13)$$

It is obvious the Lagrangian in Eq. 8 will reach its minimum at the same  $(\mathbf{w}_j^*, v_j^*)$  with

$$L' = \sum_{i=1}^n \sum_{j=1}^c u_{ij}^m \frac{(\mathbf{w}_j^t \cdot \mathbf{x}_i - v_j)^2}{\|\mathbf{w}_j\|^2} - \sum_{j=1}^c \lambda_j (\mathbf{w}_j^t \cdot \mathbf{w}_j - 1) - \sum_{i=1}^n \alpha_i (\sum_{j=1}^c u_{ij} - 1) \quad (14)$$

Hence, similarly, by taking the first derivatives of  $L'$  with respective to  $v_j$  and considering constraint in Eq. 5, we get

$$\begin{aligned} \partial L' / \partial v_j &= -2 \sum_{i=1}^n u_{ij}^m (\mathbf{w}_j^t \mathbf{x}_i - v_j) \\ &= -2 \left[ \sum_{i=1}^n u_{ij}^m \mathbf{w}_j^t \mathbf{x}_i - \sum_{i=1}^n u_{ij}^m v_j \right] \end{aligned} \quad (15)$$

setting the derivation to 0, we get

$$\sum_{i=1}^n u_{ij}^m v_j = \sum_{i=1}^n u_{ij}^m \mathbf{w}_j^t \mathbf{x}_i \quad (16)$$

hence

$$\begin{aligned} v_j^* &= \frac{\sum_{i=1}^n u_{ij}^m \mathbf{w}_j^t \mathbf{x}_i}{\sum_{i=1}^n u_{ij}^m} \\ &= \frac{\mathbf{w}_j^t \mathbf{X} \mathbf{u}_j^m}{\mathbf{e}^t \mathbf{u}_j} \end{aligned} \quad (17)$$

where  $\mathbf{e}$  is a  $n$ -dimensional column vector with all of its elements equal to one.

By taking the first derivative of  $L'$  with respective to  $\mathbf{w}_j$ , we get

$$\begin{aligned} \partial L' / \partial \mathbf{w}_j &= 2 \sum_{i=1}^n u_{ij}^m (\mathbf{w}_j^t \mathbf{x}_i - v_j) \mathbf{x}_i - 2 \lambda_j \mathbf{w}_j \\ &= 2 \left[ \sum_{i=1}^n u_{ij}^m (\mathbf{x}_i^t \mathbf{w}_j - v_j) \mathbf{x}_i - \lambda_j \mathbf{w}_j \right] \end{aligned} \quad (18)$$

Let Eq. 18 equal to 0 and substitute Eq. 17 into Eq. 18, we can get

$$\sum_{i=1}^n u_{ij}^m \mathbf{x}_i (\mathbf{x}_i^t - \frac{\sum_{i=1}^n u_{ij}^m \mathbf{x}_i^t}{\sum_{i=1}^n u_{ij}^m}) \mathbf{w}_j - \lambda_j \mathbf{w}_j = 0 \quad (19)$$

$\mathbf{w}_j$  is the eigenvector corresponding to eigenvalue  $\lambda_j$  of matrix  $\mathbf{M}_j$

$$\mathbf{M}_j = \sum_{i=1}^n u_{ij}^m \mathbf{x}_i (\mathbf{x}_i^t - \frac{\sum_{i=1}^n u_{ij}^m \mathbf{x}_i^t}{\sum_{i=1}^n u_{ij}^m}) \quad (20)$$

Especially, for the objective function

$$\begin{aligned} J &= \sum_{i=1}^n \sum_{j=1}^c u_{ij}^m \frac{(\mathbf{w}_j^t \cdot \mathbf{x}_i - v_j)^2}{\|\mathbf{w}_j\|^2} \\ &= \sum_{j=1}^c \left( \sum_{i=1}^n u_{ij}^m \mathbf{w}_j^t (\mathbf{w}_j^t \cdot \mathbf{x}_i) \cdot \mathbf{x}_i - 2 \frac{\mathbf{w}_j^t \cdot \mathbf{x}_i \mathbf{w}_j^t \mathbf{X} \mathbf{u}_j}{\mathbf{e}^t \mathbf{u}_j} + \frac{\mathbf{w}_j^t \mathbf{X} \mathbf{u}_j \mathbf{w}_j^t \mathbf{X} \mathbf{u}_j}{\mathbf{e}^t \mathbf{u}_j \mathbf{e}^t \mathbf{u}_j} \right) \\ &= \sum_{j=1}^c \left( \mathbf{w}_j^t \sum_{i=1}^n u_{ij}^m (\mathbf{w}_j^t \cdot \mathbf{x}_i - v_j) \mathbf{x}_i \right) \\ &= \sum_{j=1}^c \lambda_j \end{aligned} \quad (21)$$

Hence, the  $c$  eigenvectors corresponding to the smallest  $c$  eigenvalues of matrix  $\sum_{i=1}^n u_{ij}^m \mathbf{x}_i (\mathbf{x}_i^t - \frac{\sum_{i=1}^n u_{ij}^m \mathbf{x}_i^t}{\sum_{i=1}^n u_{ij}^m})$  are the optimal  $\mathbf{w}_j^*$  that minimize the objective function Eq. 21. As objective function in Eq. 7 will reach its minimum at the same optimal  $\mathbf{w}_j^*$ , hence  $\mathbf{w}_j^*$  is also minimizes objective function Eq. 7.

### 3.3 FHC Algorithm

Computational procedure of the FHC algorithm is summarized as follows.

1. Initialize partition matrix  $\mathbf{U}$  and hyperclusters  $\mathbf{h}_j, j = 1, \dots, c$ .
2. Update the  $\mathbf{w}_j$  through eigenvalue decomposition of matrix  $\mathbf{M}_j$  in Eq. 20 and selecting the eigenvector corresponding to the smallest eigenvalue.
3. Update  $v_j$  according to Eq. 17 under the current partition matrix  $\mathbf{U}(t)$  and  $\mathbf{w}(t)_j$ , where  $t$  is the iteration count, then we get the updated  $\mathbf{h}_j(t+1)$ .
4. Based on the newly updated fuzzy hyperclusters  $\mathbf{h}_j(t+1)$ , update the fuzzy partition matrix  $\mathbf{U}$  according to Eq. 13.
5. If the algorithm converges, then the computation stops. Otherwise, go to Step 2.

We consider the algorithm converges if the maximum change in the partition matrix between iterations is less than a preset positive small number  $\varepsilon$ . The resulting partition matrix  $\mathbf{U}^*$  and hyperclusters  $\mathbf{h}_j^*, j = 1, \dots, c$ , satisfy the solution which minimizes the objective function  $J$ .

## 4 Experiments

To validate the proposed fuzzy hyper-clustering, we carried out an experiment using artificial data and compared the results between our FHC and the conventional FCM.

**Table 1.** Confusion matrix of FHC

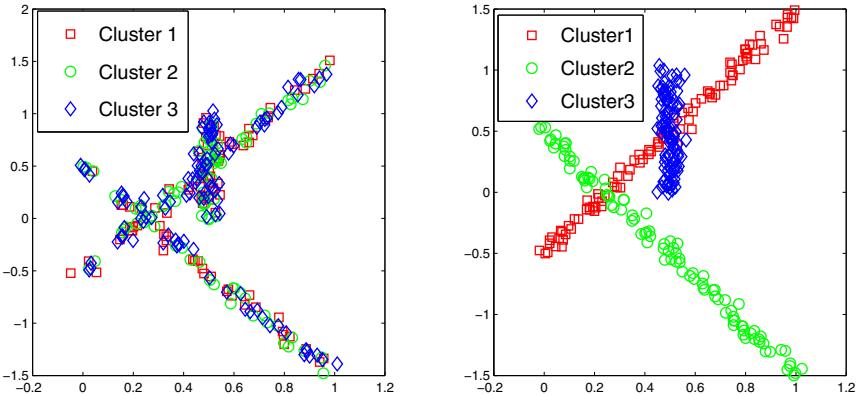
	c1	c2	c3
c1	98	2	9
c2	2	93	0
c3	0	5	91

**Table 2.** Confusion matrix of FCM

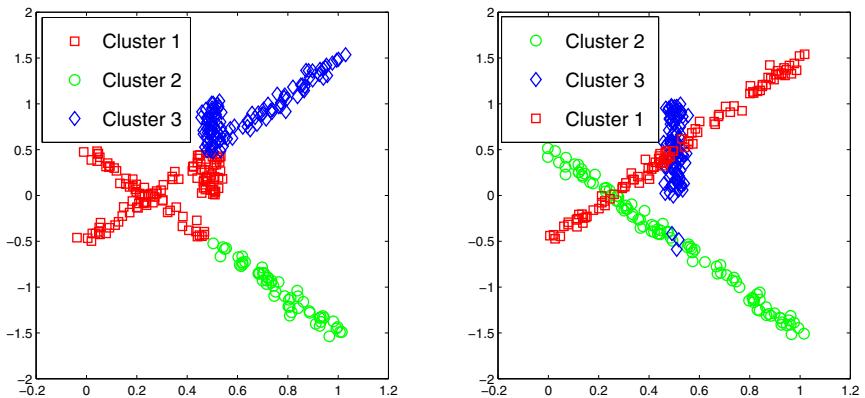
	c1	c2	c3
c1	53	45	33
c2	0	55	0
c3	47	0	67

A two-dimensional dataset of one hundred points was generated and composed of three classes,  $c1 : [x, 2x - 0.5]$ ,  $c2 : [x, -2x + 0.5]$  and  $c3 : [0.5, x]$ , where  $x \in [0, 1]$  and all the observations were added with noise of a two-dimensional item with each of its element randomly drawn from  $[-0.02, 0.02]$ . The fuzzy exponents in both methods were set to 2 and cluster number was assumed to be known as 3.

The left plot of Figure 1 shows the data which were randomly generated. The right plot of Figure 1 shows the original partitions of the original dataset. The left and right plots of Figure 2 show the clustering results of FCM and FHC, respectively. Tables 1 and 2 show the confusion matrices of the two clustering methods. From these figures and confusion matrices, we can see that although some points are misclassified in the FHC, the result is much better than that of the FCM.



**Fig. 1.** Artificial data starting from random initialization (left), original classes of artificial data (right)



**Fig. 2.** FCM clustering result (left), and FHC clustering (right)

Based on the experimental results, we can see that when the clustering subspace has linear structure, the proposed FHC outperforms the FCM, and can better approximate the original data structure.

## 5 Conclusion

We have presented a proposed fuzzy hyper-clustering algorithm which can be useful for pattern classification in signal and image processing. We formulated the objective function for the proposed hyper-clustering and derived an iterative numerical algorithm for minimizing the objective function. Then we tested

the convergence, carried the performance of the proposed method on artificially generated data and compared against the FCM. The experimental results have shown the promising application of the proposed method for handling noisy data of linear structure.

## References

1. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proc. of the National Academy of Sciences of the United States of America 95, 14863–14868 (1998)
2. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. of the National Academy of Sciences of the United States of America 96, 6745–6750 (1999)
3. Dudoit, S., Fridlyand, J.: A prediction-based resampling method for estimating the number of clusters in a dataset. Genome biology 3, 1465–6914 (2002)
4. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., Golub, T.R.: Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proc. of the National Academy of Sciences of the United States of America 96, 2907–2912 (1999)
5. Bezdek, J.C., Coray, C., Gunderson, R., Watson, J.: Detection and characterization of cluster substructure I. Linear structure: Fuzzy  $c$ -lines. SIAM Journal on Applied Mathematics 40(2), 339–357 (1981)
6. Liu, J., Pham, T.D.: Fuzzy hyper-clustering for pattern classification in microarray gene expression data analysis. In: BIOSTEC 2010, pp. 415–418 (2010)
7. Dhyani, K., Liberti, L.: Mathematical programming formulations for the bottleneck hyperplane clustering problem. In: Modelling, Computation and Optimization in Information Systems and Management Sciences, pp. 87–96 (2008)
8. Bradley, P.S., Mangasarian, O.L.:  $k$ -plane clustering. J. Global Optimization 16, 23–32 (2000)
9. Burges, C.J.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2, 121–167 (1998)
10. Fung, G., Mangasarian, O.L.: Proximal support vector machine classifiers. In: ACM-SIGKDD-KDD 2001, pp. 77–86 (2001)
11. Glenn, O.L.M., Fung, M.: Multicategory proximal support vector machine classifiers. Machine Learning 59, 77–97 (2005)
12. Mangasarian, O.L., Wild, E.W.: Multisurface proximal support vector machine classification via generalized eigenvalues. IEEE Trans Pattern Anal. Mach. Intell. 28, 69–74 (2006)
13. Yang, X., Chen, S., Chen, B., Pan, Z.: Proximal support vector machine using local information. Neurocomputing (in-print)
14. Ghorai, S., Mukherjee, A., Dutta, P.K.: Nonparallel plane proximal classifier. Signal Processing 89, 510–522 (2009)
15. Bezdek, J.C., Coray, C., Gunderson, R., Watson, J.: Detection and characterization of cluster substructure ii. fuzzy  $c$ -varieties and convex combinations thereof. SIAM J. Applied Mathematics 40(2), 358–372 (1981)
16. Leski, J.M.: Fuzzy  $c$ -varieties/elliptotypes clustering in reproducing kernel hilbert space. Fuzzy Sets and Systems 141(2), 259–280 (2004)

17. Honda, K., Ichihashi, H.: Fuzzy local independent component analysis with external criteria and its application to knowledge discovery in databases. *International Journal of Approximate Reasoning* 42(3), 159–173 (2006)
18. Honda, K., Ichihashi, H.: Component-wise robust linear fuzzy clustering for collaborative filtering. *International Journal of Approximate Reasoning* 37(2), 127–144 (2004)
19. Vidal, R., Ma, Y., Sastry, S.: Generalized principal component analysis (GPCA). *IEEE Trans. Pattern Anal. Mach. Intell.* 27(12), 1945–1959 (2005)
20. Ding, C., He, X., Simon, H.D.: On the equivalence of nonnegative matrix factorization and spectral clustering. In: Proc. SIAM Data Mining Conf., pp. 606–610 (2005)
21. Mangasarian, O.L.: Arbitrary-norm separating plane. *Operations Research Letters* 24(1-2), 15–23 (1999)
22. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge (2000)

# Clustering Using Difference Criterion of Distortion Ratios

Fujiki Morii

Department of Information and Computer Sciences,  
Nara Women's University, Nara 630-8506, Japan  
[fmorii@ics.nara-wu.ac.jp](mailto:fmorii@ics.nara-wu.ac.jp)

**Abstract.** Clustering using a difference criterion of distortion-ratios on clusters is investigated for data sets with large statistical differences of class data, where K-Means algorithm (KMA) and Learning Vector Quantization (LVQ) cannot necessarily reveal the good performance. After obtaining cluster centers by KMA or LVQ, a split and merge procedure with the difference criterion is executed. Focusing on an interesting data set which is not resolved by KMA or LVQ, some experimental clustering results based on the difference criterion and the split and merge procedure are provided.

## 1 Introduction

Among many partitional clustering methods, K-Means algorithm (KMA) and Learning Vector Quantization (LVQ) provide effective methods for clustering, which is an important and fundamental research issue for pattern recognition, image processing and data mining [1-9].

As defects of KMA and LVQ, we have two main defects. The first defect is that KMA and LVQ have the possibility of obtaining poor convergent cluster centers by the selection of initial cluster centers, whose defect may be resolved by random selections or a split and merge procedure [6] to minimize the squared-error distortion. The second defect is that bad classification may be caused by different statistical distributions of the class data, even when the minimum distortion is attained. To resolve this defect, another criterion being different from the squared-error distortion criterion must be introduced.

The purpose of this paper is to improve the second defect by incorporating a difference criterion of distortion-ratios on clusters into a split and merge procedure for KMA and LVQ, whose procedure realizes clusters with unimodality. This difference criterion composed of the squared-error distortion is an improved version of the references [10,11], where the appropriate value of a threshold to execute or not the split and merge operation can be determined more easily. The criterion offers a wide range of appropriate threshold values by building distortion ratios based on both split operation and merge operation into the criterion.

In this paper, after obtaining convergent cluster centers by KMA or LVQ, the split and merge procedure with the difference criterion is executed to improve the

second defect. Splitting each cluster into two subclusters, a cluster minimizing the distortion-ratio is selected tentatively, whose ratio is defined by the ratio of the sum of the distortion of the two subclusters for the distortion of the cluster. In the merge operation, one of the subclusters and its neighboring cluster maximizing the distortion ratio is merged tentatively. When the difference between the distortion ratio of the merge operation and the ratio of the split operation is greater than a preassigned threshold, the split and merge operation is executed. This split and merge operation is iterated until the difference criterion on the operation is satisfied for all clusters.

Focusing on an interesting data set which is not resolved by KMA or LVQ, the performance on KMA and LVQ without the difference criterion is shown in Section 2. The clustering method with the difference criterion is described in detail in Section 3, and some clustering experiments are provided in Section 4. Introducing the difference criterion to partitional clustering methods, it is demonstrated that the classification performance is improved.

## 2 Performance on KMA and LVQ

Let us treat a data set  $X$  composed of  $n$  samples  $\mathbf{x}_i = (x_{i1}, \dots, x_{iU})^t, i = 1, \dots, n$ , where  $X$  has  $K$  classes and samples are  $U$ -dimensional real vectors. By using a clustering method,  $X$  is partitioned into  $K$  disjoint subsets  $X_k, k = 1, \dots, K$ . These subsets are called clusters.

As important partitional clustering methods, we deal with KMA and LVQ depicted in the following.

### (Method using KMA)

(KMA1) Set values of initial cluster centers  $\{\mathbf{c}_k^{(1)}, k = 1, \dots, K\}$ . Repeat (KMA2) and (KMA3) for  $t = 1, 2, \dots$  until convergent.

(KMA2) For a given  $\{\mathbf{c}_k^{(t)}\}$  and each  $\mathbf{x}_i$ , compute

$$\alpha = \arg \min_k \|\mathbf{x}_i - \mathbf{c}_k^{(t)}\|^2, \quad (1)$$

and determine  $\mathbf{x}_i \in X_\alpha^{(t)}$ , where the notation  $\|\cdot\|$  shows the Euclidean norm.

(KMA3) Compute

$$\mathbf{c}_k^{(t+1)} = \frac{1}{n_k^{(t)}} \sum_{\mathbf{x}_i \in X_k^{(t)}} \mathbf{x}_i, \quad k = 1, \dots, K, \quad (2)$$

where  $n_k^{(t)}$  is the number of samples in  $X_k^{(t)}$ , and  $\mathbf{c}_k^{(t+1)}$  is called the  $(k+1)$ th cluster center of  $X_k^{(t)}$ , whose center corresponds to a representative of the cluster.

(End of KMA)

### (Method using LVQ)

(LVQ1) Set values of initial cluster centers  $\{\mathbf{c}_k^{(1)}, k = 1, \dots, K\}$ . Repeat (LVQ2) and (LVQ3) for  $v = 1, 2, \dots$  until convergent.

(LVQ2) Set

$$\mathbf{c}_l^{(v)} = \arg \min_{\mathbf{c}_k^{(v)}} \| \mathbf{x}^{(v)} - \mathbf{c}_k^{(v)} \| . \quad (3)$$

(LVQ3) Compute

$$\mathbf{c}_l^{(v+1)} \leftarrow \mathbf{c}_l^{(v)} + \alpha^{(v)} [\mathbf{x}^{(v)} - \mathbf{c}_l^{(v)}], \quad (4)$$

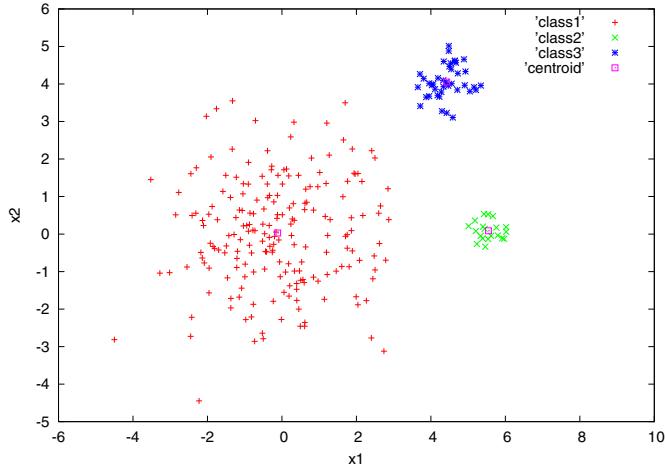
and determine  $\mathbf{x}^{(v)} \in \text{cluster } X_l$ .

(End of LVQ)

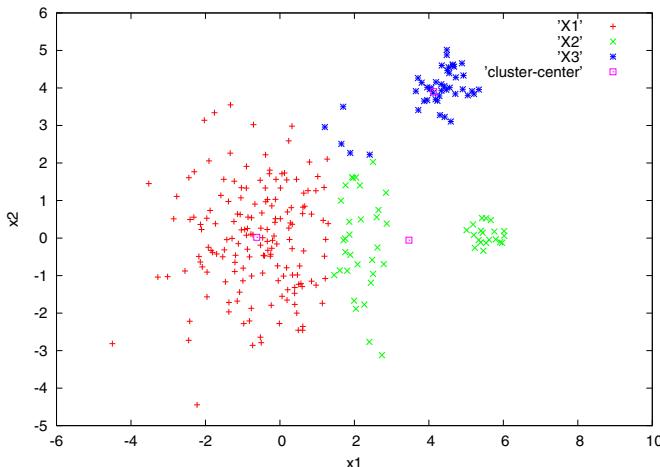
In the LVQ algorithm, we use  $\mathbf{x}^{(1)} \leftarrow \mathbf{x}_1, \dots, \mathbf{x}^{(n)} \leftarrow \mathbf{x}_n, \mathbf{x}^{(n+1)} \leftarrow \mathbf{x}_1, \dots, \mathbf{x}^{(2n)} \leftarrow \mathbf{x}_n, \mathbf{x}^{(2n+1)} \leftarrow \mathbf{x}_1, \dots$ , and a learning rate  $\alpha(v) = g/v$ , where  $g$  is a constant number.

Let us focus on an interesting data set with 3 classes shown by Fig. 1, whose set has large differences between statistical distributions of the data of classes. The data information is provided by Table 1.

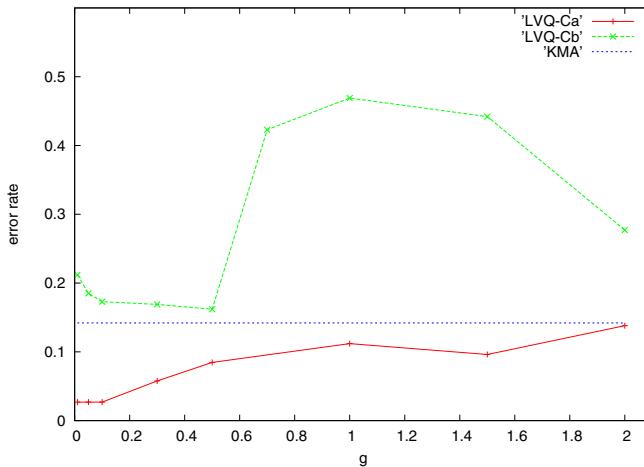
Using a set of good initial cluster centers  $C_a = \{\mathbf{c}_1^{(1)} = (0, 0), \mathbf{c}_2^{(1)} = (5.5, 0), \mathbf{c}_3^{(1)} = (4.5, 4)\}$  and the convergence condition  $\|\mathbf{c}_k^{(t+1)} - \mathbf{c}_k^{(t)}\| \leq 1.0 \times 10^{-5}$  for all  $k$ , Fig. 2 and Table 2 show the typical bad classification result by KMA, where we obtain the clusters  $\{X_k, k = 1, 2, 3\}$ . For another set of initial cluster centers  $C_b = \{\mathbf{c}_1^{(1)} = (-1, -2), \mathbf{c}_2^{(1)} = (4, -1), \mathbf{c}_3^{(1)} = (3, 4)\}$ , we obtain the same result of Table 2.

**Fig. 1.** Data set composed from 3 classes**Table 1.** Data information on classes in Fig. 1

Class	Number	Centroid	Variance	Covariance
Class 1	200	(-0.117, -0.0282)	(2.15, 2.06)	0.120
Class 2	20	(5.55, 0.0858)	(0.0819, 0.0599)	-0.00688
Class 3	40	(4.42, 4.05)	(0.171, 0.186)	0.0359

**Fig. 2.** Clustering by KMA with  $K = 3$ **Table 2.** Data information on clusters  $\{X_k, k = 1, 2, 3\}$  in Fig. 2

Cluster	Number	Cluster center	Distortion	Number of error	Error rate
$X_1$	163	(-0.623, -0.0195)	519	37	0.142
$X_2$	52	(3.46, -0.0561)	202		
$X_3$	45	(4.13, 3.90)	55.7		

**Fig. 3.** Learning constant  $g$  vs. error rate by LVQ for the sets  $C_a, C_b$

When using the sets  $C_a, C_b$  and the above convergence condition, the performance on LVQ for the learning constant  $g$  is provided by Fig.3. Although LVQ reveals the good performance for appropriate values of initial cluster centers and  $g$ , its performance is affected greatly by those values. When using the set  $C_a$  and the maximum iteration number  $10^6$ , we cannot obtain convergent results for  $g \geq 2.5$ . For the set  $C_b$  and  $g \geq 0.7$ , inappropriate convergent results far from class centroids are obtained.

### 3 Clustering Using Difference Criterion of Distortion-Ratios

Let us begin with the situation of already obtaining convergent cluster centers by KMA or LVQ. Then the data set  $X$  is partitioned into  $K$  clusters  $\{X_k, k = 1, \dots, K\}$  through the Voronoi diagram. In order to improve the second defect like Fig.2, we incorporate a difference criterion of distortion-ratios to a split and merge procedure.

Each cluster  $X_k$  is tentatively split into 2 subclusters  $X_{k1}$  and  $X_{k2}$  by using KMA to realize the minimum squared-error distortion. The minimum distortion for  $X_k$  is given by

$$D_k(1) = \min_{\mathbf{c}_k} \sum_{\mathbf{x}_i \in X_k} \|\mathbf{x}_i - \mathbf{c}_k\|^2. \quad (5)$$

The minimum distortion for  $X_{k1}$  and  $X_{k2}$  is also defined as

$$D_k(2) = \min_{\{\mathbf{c}_{kp}\}} \min_{\{X_{kp}\}} \sum_{p=1}^2 \sum_{\mathbf{x}_i \in X_{kp}} \|\mathbf{x}_i - \mathbf{c}_{kp}\|^2. \quad (6)$$

Let us introduce a splitting measure for  $X_k$  given by

$$\rho_k = D_k(2)/D_k(1), \quad (7)$$

which is called the distortion-ratio. The small value of  $\rho_k$  states that  $X_k$  should be split into 2 subclusters. When  $X_k$  has 2 subclusters, the value of  $D_k(1)$  is large because of the strong bimodality of  $X_k$ , but the value of  $D_k(2)$  becomes small since subcluster centers are near to the centroids of the 2 subclusters.

Computing

$$\hat{\rho}_k = \min_k \rho_k, \quad (8)$$

as the most possible candidate to be split,  $X_{\hat{k}}$  is split into 2 subclusters  $X_{\hat{k}1}, X_{\hat{k}2}$  tentatively.

For the merge operation of neighboring  $X_v$  and  $X_w$  in  $\{X_k\}$  including the clusters and subclusters, as a merging measure we introduce

$$\psi_{vw} = (D_v(1) + D_w(1))/D_{vw}(1), \quad (9)$$

where  $D_v(1)$  and  $D_w(1)$  are used according to (5), and the distortion of the merged cluster  $X_v \cup X_w$  is given by

$$D_{vw}(1) = \min_{\mathbf{c}_{vw}} \sum_{\mathbf{x}_i \in X_v \cup X_w} \|\mathbf{x}_i - \mathbf{c}_{vw}\|^2. \quad (10)$$

The large value of  $\psi_{vw}$  represents that  $X_v$  and  $X_w$  should be merged.

Using  $\psi_{vw}$ , let us consider the merging operation of the subclusters  $X_{\hat{k}1}$ ,  $X_{\hat{k}2}$ . Assuming that  $S(X_{\hat{k}p})$  is a set of neighboring clusters to the subcluster  $X_{\hat{k}p}$ , compute

$$X_{\hat{z}} = \arg \max \left\{ \max_{X_z \in S(X_{\hat{k}1})} \psi_{\hat{k}1,z}, \max_{X_z \in S(X_{\hat{k}2})} \psi_{\hat{k}2,z} \right\}. \quad (11)$$

When  $X_{\hat{k}\hat{p}}$  satisfies (11),  $X_{\hat{k}\hat{p}}$  and  $X_{\hat{z}}$  are the most possible pair candidate for merging.

Let us introduce the difference criterion of distortion-ratios defined as

$$\Delta = \psi_{vw} \mid_{w=X_{\hat{z}}}^{v=X_{\hat{k}\hat{p}}} - \rho_{\hat{k}} < \xi \quad (12)$$

for a predetermined threshold  $\xi$ . If (12) is satisfied, the split and merge operation is not executed, otherwise its operation is executed and  $\{X_k\}$  is renewed.  $\rho_{\hat{k}}$  for  $X_{\hat{k}}$  with bimodality reveals small values. When the merged  $X_{\hat{k}\hat{p}} \cup X_{\hat{z}}$  has unimodality,  $\psi_{vw} \mid_{w=X_{\hat{z}}}^{v=X_{\hat{k}\hat{p}}}$  shows large values. When changing from bimodality to unimodality like this, we select an appropriate value of  $\xi$  to satisfy  $\Delta \geq \xi$ . Then, the bad cases in Fig.2 are resolved.

This split and merge operation is iterated until (12) for all clusters is satisfied. It is also assumed that the maximum iteration number  $H$  is predetermined. Our clustering method using the difference criterion of distortion-ratios is summarized as follows.

### (Clustering Using Difference Criterion)

**(CDC0)** Assume that the value of  $\xi$  in (12) and the maximum iteration number  $H$  of the split and merge operation are given.

**(CDC1)** Classify samples in  $X$  into  $K$  clusters  $\{X_k^{(0)}\}$  by KMA or LVQ.

**(CDC2)**  $\{X^{(S_1)}\} \leftarrow \{X_k^{(0)}\}$ , and  $\{X^{(S_2)}\} \leftarrow \{X_k^{(0)}\}$ .

**(CDC3)** Compute  $\rho_{\hat{k}}$  and  $X_{\hat{k}}$  for  $\{X^{(S_2)}\}$ .

**(CDC4)**

If( $\Delta \geq \xi$ ) {

If the iteration number  $h$  is greater than  $H$ , go to (END).

Execute the split and merge operation for  $\{X^{(S_1)}\}$ , and renew  $\{X^{(S_1)}\}$ .

$\{X^{(S_2)}\} \leftarrow \{X^{(S_1)}\}$ .

Go to (CDC3).

}

else {

$\{X^{(S_2)}\} \leftarrow \{X^{(S_2)}\} - X_{\hat{k}}$ .

If( $\{X^{(S_2)}\} = \phi$ ), go to (END).

Go to (CDC3).

}

**(END)**

The value of  $\xi$  is determined empirically by human being. Large values of  $\xi$  lower the possibility of the split and merge operation. Since the range of appropriate values of  $\xi$  is wide, we can determine good values easily in comparison

with the distortion-ratio criterion proposed in the references [10][11], where the execution of the split and merge operation is determined by the condition  $\rho_{\hat{k}} \leq \xi$ .

## 4 Clustering Experiments

Let us begin with the classification result shown by Fig. 2 and Table 2, which is provided by applying KMA to the data set of Fig. 1 and reveals the typical bad performance.

The distortion-ratio of (7) for the clusters  $\{X_k, k = 1, 2, 3\}$  is given by Table 3. Since  $X_3$  satisfies  $\rho_{\hat{k}} = \min_k \rho_k = 0.291$  of (8), the cluster  $X_3$  is split into 2 subclusters  $X_{31}$  and  $X_{32}$  tentatively. The tentative classification result is shown by Fig. 4.

Since  $\psi_{vw} |_{w=X_{\hat{k}\hat{p}}}^{v=X_{\hat{k}\hat{p}}} = 0.892$  with  $X_{\hat{k}\hat{p}} = X_{31}$  and  $X_{\hat{z}} = X_1$  in (12),  $X_{31}$  and  $X_1$  is merged tentatively. When  $\xi = 0.3$  is selected in (12), we obtain  $\Delta = 0.892 - 0.291 = 0.601 > 0.3$ . Then, we determine that  $X_3$  is split into  $X_{31}$  and  $X_{32}$ , and  $X_{31}$  and  $X_1$  is merged. In this case, the range of appropriate values of  $\xi$  is wide, and we can adopt the range  $[0.1, 0.5]$ .

Next, we execute the same operations for the clusters  $X_1 \cup X_{31}$ ,  $X_2$  and  $X_{32}$ , where the merged cluster of  $X_1$  and  $X_{31}$  is denoted by  $X_1 \cup X_{31}$  or  $X_1 m X_{31}$ . Table 4 provides the distortion-ratio of (7) for  $X_1 \cup X_{31}$ ,  $X_2$  and  $X_{32}$ . Since  $X_2$  satisfies  $\rho_{\hat{k}} = \min_k \rho_k = 0.299$  of (8), the cluster  $X_2$  is split into 2 subclusters  $X_{21}$  and  $X_{22}$  tentatively. The tentative classification result is shown by Fig. 5.

Since  $\psi_{vw} |_{w=X_{\hat{k}\hat{p}}}^{v=X_{\hat{k}\hat{p}}} = 0.764$  with  $X_{\hat{k}\hat{p}} = X_{21}$  and  $X_{\hat{z}} = X_1 \cup X_{31}$  in (12),  $X_{21}$  and  $X_1 \cup X_{31}$  is merged tentatively. We obtain  $\Delta = 0.764 - 0.299 = 0.465 > 0.3$ . Then, we determine that  $X_2$  is split into  $X_{21}$  and  $X_{22}$ , and  $X_{21}$  and  $X_1 \cup X_{31}$  is merged.

After all, as the final classification result, we acquire the following correct solution with no error.

$$((X_1 \cup X_{31}) \cup X_{21}) \equiv Class1 \quad (13)$$

$$X_{22} \equiv Class2 \quad (14)$$

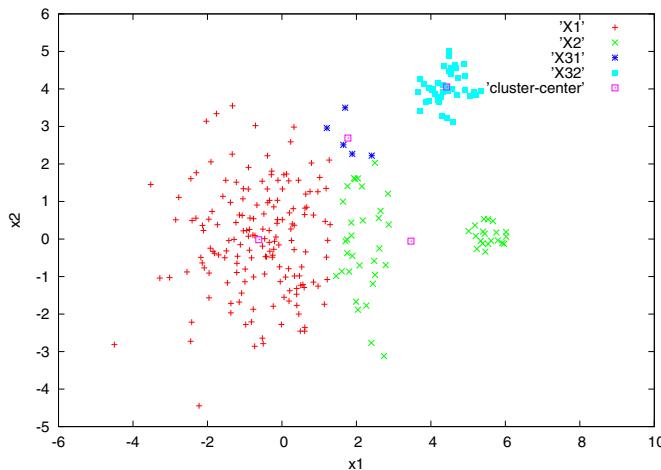
$$X_{32} \equiv Class3 \quad (15)$$

**Table 3.** Distortion-ratio for  $\{X_k, k = 1, 2, 3\}$

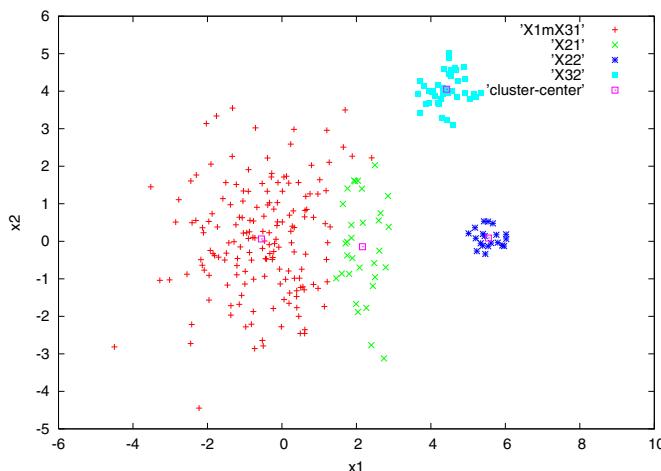
Distortion-ratio	$X_1$	$X_2$	$X_3$
$\rho_k$	0.603	0.299	0.291

**Table 4.** Distortion-ratio for  $\{X_1 \cup X_{31}, X_2$  and  $X_{32}\}$

Distortion-ratio	$X_1 \cup X_{31}$	$X_2$	$X_{32}$
$\rho_m$	0.604	0.299	0.527



**Fig. 4.** Representaion of  $X_1$ ,  $X_2$ ,  $X_{31}$  and  $X_{32}$ , where  $X_3$  is split into  $X_{31}$  and  $X_{32}$  by KMA



**Fig. 5.** Representation of  $X_1 \cup X_{31}$ ,  $X_{21}$ ,  $X_{22}$  and  $X_{32}$ , where  $X_1 \cup X_{31}$  is merged from  $X_1$  and  $X_{31}$ , and  $X_2$  is split into  $X_{21}$  and  $X_{22}$  by KMA

**Table 5.** Distortion-ratio for Classes

Distortion-ratio	Class 1	Class 2	Class 3
$\rho_k$	0.667	0.581	0.527

The values of the distortion-ratio for the 3 classes are provided by Table 5. Through the tentative split and merge operations for Class 1-3, the difference criterion of (12) is satisfied. Hence, the split and merge operations are not executed.

When using LVQ and the difference criterion for the data given by Fig.1 and Table 1, we can also obtain the good performance except inappropriate values of initial cluster centers and  $g$  stated about Fig.3.

## 5 Conclusion

We proposed a new clustering method using the difference criterion of the distortion-ratios with the split and merge procedure to improve the classification performance of the partitional clustering method based on KMA or LVQ etc. Prosecuting some clustering experiments for the data with large statistical differences among classes, it was demonstrated that the proposed method offers the good performance. Concerning to the determination of threshold values in the difference criterion, we can set appropriate values easily because of the wide range of those values. As a future issue, we would like to develop this method to a general and robust method under the consideration of various kinds of data.

It is also an important issue to estimate the number of clusters correctly without assuming probability distributions of data. To obtain the number of clusters, EM algorithm with Akaike information criterion, X-means algorithm [12] and hierarchical clustering etc. have been proposed. In hierarchical clustering, agglomerative clustering with merge operation and divisive clustering with split operation are main approaches [11,13]. This spilt and merge procedure with appropriate cluster validity will be applied to the problem of the number of clusters.

## References

1. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. John Wiley & Sons, INC., Chichester (2001)
2. MacQueen, J.: Some Methods for Classification and Analysis of Multivariate Observations. In: Proc. 5th Berkeley Symp. on Math. Stat. and Prob., vol. 1, pp. 281–297. Univ. of California Press, Berkeley and Los Angeles (1967)
3. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs (1988)
4. Gordon, A.D.: Classification, 2nd edn. Chapman and Hall, Boca Raton (1999)
5. Linde, Y., Buzo, A., Gray, R.M.: An Algorithm for Vector Quantizer Design. IEEE Trans. Commun. 28, 84–95 (1980)
6. Kaukoranta, T., Franti, P., Nevalainen, O.: Iterative split-and-merge algorithm for vector quantization codebook generation. Optical Engineering 37(10), 2726–2732 (1998)
7. Jain, A.K.: Data Clustering: 50 Years Beyond K-Means. In: The King-Sun Fu Prize lecture delivered at the 19th ICPR (December 8, 2008)
8. Kohonen, T.: Self-Organizing Maps, 2nd edn. Springer, Berlin (1997)

9. Pal, N.R., Bezdek, J.C., Tsao, C.-K.: Generalized Clustering Networks and Kohonen's Self-Organizing Scheme. *IEEE Trans. Neural Network* 4(4), 549–557 (1993)
10. Morii, F., Kurahashi, K.: Clustering Based on Multiple Criteria for LVQ and K-Means Algorithm. *JACIII* 13(4), 360–365 (2009)
11. Morii, F.: Clustering Based on Distortion-Ratio Criterion. In: Proc. of IEEE International Symposium on Industrial Electronics (ISIE 2009), pp. 1129–1133 (2009)
12. Pelleg, D., Moore, A.: X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In: Proc. of the 17th International Conf. on Machine Learning, pp. 727–734 (2000)
13. Ding, C., He, X.: Cluster Merging and Splitting in Hierarchical Clustering Algorithms. In: Proc. IEEE International Conference on Data Mining, pp. 139–146 (2002)

# Computer-Generated Conversation Based on Newspaper Headline Interpretation

Eriko Yoshimura, Seiji Tsuchiya, and Hirokazu Watabe

Dept. of Intelligent Information Engineering & Sciences,  
Faculty of Science and Engineering, Doshisha University,  
Kyo-Tanabe, Kyoto, 610-0394, Japan

{eyoshimura,sttsuchiya,watabe}@indy.doshisha.ac.jp

**Abstract.** While some robot-related research has been done on conversational text, based on greetings and other general types of conversation and utilizing extensive conversation knowledge bases; to date, very little work on conversation involving topical newspaper articles has been done. If robots could engage in conversation on subjects pertaining to current affairs, it is likely they would seem much more appealing and approachable. However, if a robot were to merely read out news mechanically it would lack the conversational flexibility to be an effective communication tool, and the approachability of the robot would be significantly limited. In light of this, the present study investigates the automatic generation of highly expandable conversation sentences using cues taken from news stories.

**Keywords:** Computer conversation, Natural languages, Newspaper headline.

## 1 Introduction

In recent years machines have become an integral part of our personal and social lives, and their presence is becoming increasingly indispensable. In view of this fact, they should perhaps be envisioned as “machines (robots) that coexist with humans.” Given that numerous robots capable of walking on two legs, running, and dancing already have been developed, this dream has and continues to be realized to some extent. However, for robots to truly “coexist with humans” it is important that in addition to excellent physical capabilities, they be provided with “intelligence” and that they be capable of using this intelligence to engage in conversation. While current robots execute predetermined responses according to extensive rules [1], they are not able to respond on the basis of their own mechanical intelligence. For this reason, it is likely that anyone who repeatedly uses such a robot will eventually become bored with it, considering it to be nothing more than a mechanical device.

Human beings are social animals. That is, they live, interacting with others, amongst family and friends, and as part of a community, country or other group. Regardless of the size of a community, we actively seek out information from the community to which we belong, and try to express this desire for information outwardly. There are various kinds of information about current affairs that we consider essential, and we receive and pass this on to others—everything from urgent warnings of danger

that need to be communicated directly in the event of a disaster or accident, to cultural information that enriches our lives, for the purpose of pleasure, or entertainment. Thus, in a “conversation,” an intimate form of human communication, current affairs clearly represent an important topic. Information about current affairs can be regarded as an essential element of everyday life that can serve as a lead-in to a conversation, or as a connection to a matter that concerns the conversation partner.

Various linguistic semantic understanding techniques have been proposed. Representative example is a linguistic meaning understanding method based on a case frame dictionary [2]. The dictionary is a large-scale database (corpus) that is constructed on example sentences including words which have meaning label by information of syntax analysis. One of the applications using these techniques is conversational system [3]. A conversational system generates contents of an utterance with pattern matching processing between the templates and the case frame dictionary. However, the techniques have a problem of scarce variation in contents of an utterance because the templates and corpus don’t have enough data.

While some robot-related research has been done on conversational text, based on greetings and other general types of conversation and utilizing extensive conversation knowledge bases; to date, very little work on conversation involving topical newspaper articles has been done. If robots could engage in conversation on subjects pertaining to current affairs, it is likely they would seem much more appealing and approachable. However, if a robot were to merely read out news mechanically it would lack the conversational flexibility to be an effective communication tool, and the approachability of the robot would be significantly limited. In light of this, the present study investigates the automatic generation of highly expandable conversation sentences using cues taken from news stories. The present study was conducted in Japanese, so the following explanations apply to the grammatical characteristics of the Japanese language. The information for conversation topics used in the present study was sourced from newspaper articles [4], because newspaper articles are easy to acquire and likely deal with subjects of common interest to many people. Our objective in this study was to generate conversations that are highly expandable, so we utilized article titles that best expressed the content of the newspaper articles.

Since newspaper headlines are written according to grammatical rules that are particular to news articles, we could not use headlines as conversation sentences without first modifying them. The grammatical rules peculiar to news articles relate to things such as the frequent use of code characters, and the use of verbs, word-endings, and particles for abbreviations. For this reason, it is first necessary to sort out and interpret this information based on the grammatical rules specific to newspaper headlines. Subsequently, after interpreting its meaning, this information can be converted into a conversation sentence. An example of generated conversation text from the present study is shown in Fig. 1.

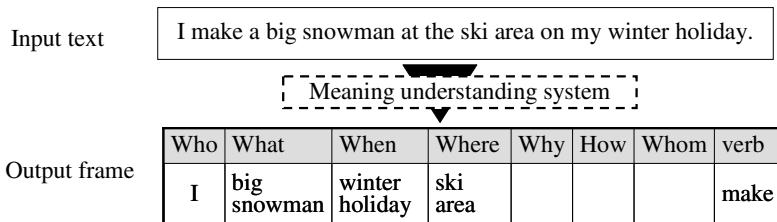
Newspaper headline	Men’s golf: Ishikawa starts 2 strokes off lead in 10th place. Yamashita and Nonaka head field.
Generated sentence	Hey! Ishikawa is in 10th place, 2 strokes off the lead in men’s golf.

**Fig. 1.** Example of a generated conversation text

## 2 Interpreting Newspaper Headline Text

In the present study “interpreting text” refers to the process of distinguishing the semantic function of each word of the text. For this purpose, we created frames for sorting out the semantic functions.

We created frames for newspaper headlines as a model for semantic interpreting frames. A meaning understanding system is a system that analyzes a single sentence input by classifying the elements of the sentence as one of the 6W1H (Who, What, When, Where, Whom, Why, How), or as a verb.



**Fig. 2.** Example of using a meaning understanding system

The newspaper headline frames are based on a meaning understanding system and consist of 9 frames—6W1H, verb, and topic.

The syntax of newspaper headlines follows its own particular rules, so the headline text is generally not grammatically correct. Because the meaning understanding system assumes that the single sentence input is grammatical correct, it is difficult to correctly classify a newspaper headline, unmodified, according to 6W1H. For this reason, it is necessary to clarify the rules of newspaper headline syntax, and then to interpret headlines according to these rules.

An example of newspaper headline frames is shown in Fig. 3.

Men's golf: Ishikawa starts 2 strokes off lead in 10th place. Yamashita and Nonaka head field.

Who	What	When	Where	Why	How	Whom	Verb	Theme
Ishikawa	2-strokes off lead in 10th places						start	Men's golf

**Fig. 3.** Example of newspaper headline frames

### 2.1 Newspaper Headline Syntax Rules

The headlines of daily newspapers make use of particular code characters, like “:” and “.”. After studying the usage of these special codes, we classified newspaper headlines into 12 patterns.

A, B, C, and D are text strings between the code characters, i.e., they define blocks.

The various patterns occur at different frequencies in a newspaper headline.

**Table 1.** Example of newspaper headline pattern classification

	Code construction	Example sentence
1	A: B C	Figure: Oda 2nd in SP France Cup
2	A: B	Pakistan: 41 dead in suicide bombing
3	A: B, C	Child support allowance: Govt., policy to suspend payment
4	A: B... C	J1: Masashi Nakayama dropped from team... Iwata
5	A: B, C D	World Gymnastics Championship: Uchimura, 1st in qualifying Overall leader
6	A:B C D	JR west: leak accident apologize for systematic campaign final report in next month
7	A:B, C...D	Canoe distress: boarding 3 girls, find in rest...Iriomote
8	A:B C,D	Canvass: increasing military detachment 49% disagreed, 47% agreed
9	A:B...C,D	M. Jackson: put out a new song...first release since he died, theme of love
10	A:B...C D	NTT: proceed is less than 5 trillion yen for the first time in 11 years...cell phone at a stand September ad report
11	A:B C D, E	World Gymnastics Championship: parallel bars Tanaka bronze medal Uchimura, 6th in horizontal bar
12	A:B,C,D E	Super-flu: cause death in Tokyo, Hyogo, Aichi Death toll rise to 26

We therefore assessed the frequency of occurrence of each pattern by examining 1,608 newspaper headlines that appeared between Oct. 13 and Nov. 12, 2009. The results for frequency of occurrence are shown in Table 2. As shown in Table 2, patterns 1 to 5 accounted for a total of 88.1% of all headline occurrences. As a result, we decided to focus on these 5 patterns in the present study.

**Table 2.** Frequency of occurrence (%)

Pattern	1	2	3	4	5	6	7	8	9	10	11	12
Frequency of occurrence*	45.1	16.7	10.0	8.7	7.5	4.2	2.3	2.3	1.2	1	0.5	0.4

\*(n = 1,608)

## 2.2 Topic Text Extraction

Newspaper headlines are composed of a single phrase or multiple phrases. In the case of headlines composed of multiple phrases, one of the phrases expresses the main topic, while the other phrases add supplementary information. Making use of the pattern characteristics, we identified and extracted the text expressing the topic from the newspaper headline for generating conversation starter lines.

Using the characteristics of code character syntax, we extracted a coherent single sentence from the newspaper headline to serve as the main topic. We then decomposed this single sentence into frames using a semantic interpretation system. The rules for adding particles and connecting blocks of text to extract a single sentence from newspaper headlines are shown in Table 3.

**Table 3.** List of rules for utilizing the pattern characteristics

Condition	Rule
Patterns 1, 2, 4	Block B is considered to be a single sentence that serves as the topic.
Patterns 3, 5	A particle <i>ga</i> was added between block B and block C, thereby joining the two blocks into a single sentence expressing the topic.

In Japanese, the particle *ga* indicates that the noun immediately preceding this particle is the subject of the sentence. Patterns 3 and 5 are grammatically incorrect because of the omission of *ga* between block B and block C. blocks not explicitly mentioned in Table 3 are understood to represent supplementary information.

## 2.3 Addition of Omitted Words

Verbs and particles are sometimes omitted in newspaper headlines. Thus, to create grammatically correct sentences the text is processed to add the necessary omitted verbs and particles.

### 2.3.1 Adding Verbs

In Japanese-language newspaper headlines, there are two kinds of instances in which verbs are omitted. In one case, the word acting as a verb is nominalized. Thus, according to Japanese grammar, the nominalized verb is turned into a verb. (e.g., development -> (to) develop) In the second case, the verb is omitted. As a rule, Japanese sentences end with a verb, preceded immediately by a particle. If the verb is omitted then the sentence ends with a particle. Thus, if the word at the end of sentence is a particle, an appropriate verb is added. We use extensive ranking frames [5] that are generated automatically from the Web. In Japanese, a verb and the nouns connected with that verb have to be analyzed according to the particular usage for the particle. This task can be accomplished by examining the appearance degree of sets of connections between nouns, verbs, and particles from the Web.

**Table 4.** Extensive ranking frames for “tournament”

Rank	Verb	Degree of appearance
<i>ni</i> (“to”)	Advance	471
<i>ni</i> (“in”)	Participate	288
<i>wo</i> (precedes tr. verb object)	Perform	276
<i>he</i> (“in”)	Advance	71

Using these extensive ranking frames, we obtained highest degree verb in sets of sentence-ending particles and the nouns immediately preceding them. For example, in the case of the phrase “To the Serbia tournament,” the noun “tournament” immediately precedes the particle *he* that ends the phrase (in Japanese syntax), from which we conclude that “advance” is the most appropriate verb to add.

### 2.3.2 Adding Particles

As explained above, in Japanese sentences the sentence-ending verb is preceded by a particle. If this particle that precedes the sentence-ending verb is omitted, a noun occurs immediately before the verb. Therefore, if a sentence ends in this manner, with a noun-verb set, an appropriate particle must be inserted between the noun and verb. To find a suitable particle in this case, we use extensive ranking frames. This can also be done by searching for verbs. An example the verb “start” is shown in Table 5.

**Table 5.** Results for “start” using an extensive ranking frame

Rank	Noun	Degree of appearance
<i>kara</i> (“from”)	Time	26,614
<i>kara</i> (“from”)	Sales	10,774
<i>wo</i> (precedes tr. verb object)	Management	3,963

In this case, the nouns found by the search are qualifying words (modifiers). For example, words such as morning and evening express the concept of time. Therefore, these nouns may not be compatible with the sentence-ending noun. Accordingly, to examine the association between the sentence-ending noun and searched nouns, we use the calculation method [6] described below. When the sentence-ending words are the noun “meeting” and the verb “start,” we examined the degree of association of “meeting” with the searched nouns shown in Table 5. The results are presented in Table 6. The candidate nouns having a degree of association higher than a certain threshold value were regarded as particles to be added.

**Table 6.** Degree of association between “meeting” and candidate “nouns”

Noun	Degree of association
Time	0.25
Sale	0.22
Management	0.33

Degree of association is a method to assess quantitatively between words. It is calculated using Concept Base [7]. Concept base is a knowledge-base consisting of words (concepts) and word clusters (attributes) that express the meaning of these words. This is automatically constructed from multiple sources, such as Japanese dictionaries and contains approximately 120,000 registered words organized in sets of concepts and attributes. An arbitrary concept  $A$ , is defined as a cluster of paired values, consisting of attribute,  $a_i$ , which expresses the meaning and features of the concept, and weight  $w_i$ , which expresses the importance of attribute  $a_i$ , in expressing concept  $A$ :

$$A = \{(a_1, w_1), (a_2, w_2), \dots, (a_N, w_N)\}$$

Attribute  $a_i$  is called the first-order attribute of concept  $A$ . In turn, an attribute of  $a_i$  (taking  $a_i$  as a concept) is called a second-order attribute of concept  $A$ .

Figure 4 shows the elements of the Concept “train” expanded as far as the Secondary Attributes. The method for calculating the Degree of Association involves developing each concept up to second-order attributes, determining the optimum combination of

train	train, 0.36	railroad, 0.10	$a_i, w_i$	Primary Attributes Secondary Attributes
	train, 0.36	railroad, 0.10	$\dots a_{i1}, w_{i1}$	
	railroad, 0.10	subway, 0.25	$\dots a_{i2}, w_{i2}$	
	:	:	:	
	$a_{ij}, w_{ij}$	$a_{2j}, w_{2j}$	$\dots a_{ij}, w_{ij}$	

**Fig. 4.** Example demonstrating the Concept “train” expanded as far as Secondary Attributes

first-order attributes by a process of calculation using weights, and evaluating the number of these matching attributes.

For Concepts  $A$  and  $B$  with Primary Attributes  $a_i$  and  $b_i$  and Weights  $u_i$  and  $v_j$ , if the numbers of attributes are  $L$  and  $M$ , respectively ( $L \leq M$ ), the concepts can be expressed as follows:

$$A = \{(a_1, u_1), (a_2, u_2), \dots, (a_L, u_L)\}$$

$$B = \{(b_1, v_1), (b_2, v_2), \dots, (b_M, v_M)\}$$

The Degree of Identity ( $I(A, B)$ ) between Concepts  $A$  and  $B$  is defined as follows (the sum of the weights of the various concepts is normalized to 1):

$$I(A, B) = \sum_{a_i=b_j} \min(u_i, v_j) \quad (1)$$

The degree of association is weighted by calculating the degree of identify for the entire targeted primary attribute combinations and then determining the correspondence between primary attributes. Specifically, priority is given to determine the correspondence between primary attributes. For primary attributes that do not match, the correspondence between primary attributes is determined so as to maximize the total degree of matching. Using the degree of matching, it is possible to give consideration to the degree of association even for primary attributes that do not match perfectly. When the correspondences are thus determined, the degree of association  $R(A, B)$  between concepts  $A$  and  $B$  is as follows:

$$R(A, B) = \sum_{i=1}^L I(a_i, b_{xi})(u_i + v_{xi}) \times \{\min(u_i, v_{xi}) / \max(u_i, v_{xi})\} / 2 \quad (2)$$

## 2.4 Filling in the Newspaper Headline Frames

The supplemented important sentence is fitted to the newspaper headline frames. The words in the sentence are decomposed to the corresponding frames by a semantic interpretation system. Block A of the newspaper headline is then added to the topic frame.

The other blocks—those aside from the main topic, obtained as described in 2.2 above—represent supplementary information. Thus, if there is any information that can be included in the newspaper headline frame, that information should be included.

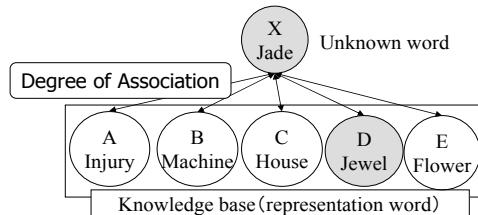
If the Who frame is empty and the supplementary information features a proper noun such as “Tiger Woods” or “White House,” then this block should be included in the Who frame. In addition, if the Where frame is empty, and a block serving supplementary information includes reference to a place, then this block should be included in the Where frame.

### 3 Generating Conversation Text

Based on this understanding, we attempted to generate natural conversation text. The function of individual words can be understood according to the newspaper headline frames, so each word can be suitably arranged according to grammatical rules. In newspaper headlines there is a tendency to not specify the tense of the action. It is therefore quite possible that the generated conversation text comes out quite unnaturally with regard to tense. Thus, if the past tense is considered to be appropriate, the sentence is modified to past tense. Similarly if the future tense is judged to be most appropriate then the text is converted to future tense.

We prepared a knowledge base of words that typically relate to the past, consisting of a total of 37 verbs, such as “die” and “crash.” If a word matches one of these, the text is converted to the past tense to create a more natural sentence.

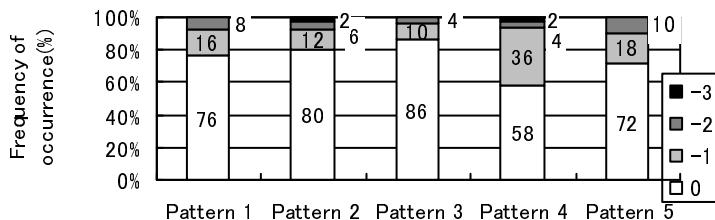
Note, however, that these verbs are typical ones that were created by a modest degree of human effort; and by no means comprise a comprehensive list. Words that are not included in the knowledge base are referred to as “unknown words.” By processing these unknown words [7], our knowledge can be flexibly expanded, building on a small-scale human-created knowledge base. By using the Concept Base and the Degree of Association, an unknown word that doesn’t exist in knowledge base can be processed. The word “unknown word” means a word doesn’t exist in human create knowledgebase and exist in large-scale word database, Concept base (include 90000 concepts). Now, it uses as an example of unknown word “Jade” (figure 5). First, it calculate the Degree of Association between this unknown word “Jade” and nodes of knowledge base and it link a node with the highest degree of association for “Jade”. For “Jade”, the node with the highest degree of association is “Jewel”. By this, it considered that “Jade” has property of “Jewel”. Moreover, attributes of “Jade” is acquired from the Concept base.



**Fig. 5.** Example of unknown word processing

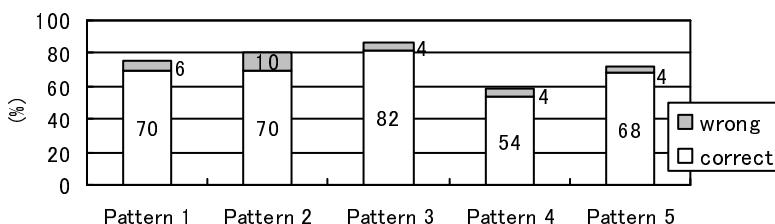
### 4 Assessment

We assessed the precision of frame decomposition. As test data, we took 50 headlines of each pattern from a selection of newspaper articles [4], to prepare a total of 250 sentences. A frame decomposition was then performed on this data, and the results were assumed to represent a correct interpretation of the data. A system frame decomposition that was completely correct was assigned a score of 0; in the case of one error a score of -1 was given, -2 for two errors, -3 for three errors, and so forth.

**Fig. 6.** Frame decomposition precision

The assessment of the results is shown in Fig.6. Since the frequency of occurrence of each pattern is different, the overall precision, taking into account frequency of occurrence, was 75.8%.

We then used the correctly frame-decomposed test data to generate conversation text. After a visual examination by three individuals, if all three agreed on the match, the result was judged “correct,” otherwise the result was judged as “incorrect.” The results of this examination are shown in Fig. 7. The overall precision was 61.3%. The precision for generating a reconstructed conversation from correct frames was 92.2%. Thus, we can conclude that by decomposing from correct frames, it is possible to generate reconstructed conversation text with an accuracy of approximately 90%.

**Fig. 7.** Conversation generation precision

We showed that if frames can be correctly decomposed, it is possible to generate conversation text with a very high degree of accuracy. An example of conversation text generation is shown in Table 7.

**Table 7.** Example of conversation text generation

Correct	Newspaper headline	Mitsukoshi Isetan: Paris Mitsukoshi, to shut at end of September
	Output	It seems that Paris Mitsukoshi will shut at the end of September.
Incorrect	Newspaper headline	Dalai Lama: Visit to White House in Feb
	Output	It seems that the Dalai Lama is going to guide a visit to the White House in February.

Conclusion Most of the incorrect patterns resulted from a failure to supplement the correct words. For example, in the case of the incorrect pattern in Table 7, there is no verb. Based on the noun “visit” and the final particle *he* (“to”), the verb “guide” was added. In this particular text, however, the particle *he* has almost no meaning, serving only to indicate the existence of a plan. In this case, it would have been better to eliminate *he* and simply convert the preceding noun “visit” into the form of a verb, “to visit.”

## 5 Conclusion

In the present study, we proposed a method of automatically generating conversation text based on newspaper headlines. We demonstrated a technique for understanding the meaning of newspaper headlines, and then showed that it is possible to generate conversation text based on the extracted meaning. From the results of assessing this proposed technique, we can conclude that the proposed technique is effective; with conversation text being generated from the interpreted meaning with a high degree of accuracy. In light of this finding, we expect that the generation of topical conversation text from news to create natural conversations offers the promise of exciting new applications that create new topics.

## Acknowledgment

This research has been partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (Young Scientists (B), 21700241).

## References

1. Sekine, S., Inui, K., Torisawa, K.: Corpus Based Knowledge Engineering, <http://nlp.cs.nyu.edu/sekine/CBKE/>
2. Kurohashi, S., Nagao, M.: A method of case structure analysis for Japanese sentences based on examples in case frame dictionary. IEICE Transactions on Information and Systems E77-D(2), 227–239 (1994)
3. Dohsaka, K., Shimaze, A.: A model for incremental utterance production in task-oriented dialogues. Transactions of Information Processing Society of Japan 37(12), 2190–2200 (1996)
4. Mainichi.jp – Mainichi Newspaper, <http://www.mainichi.jp/>
5. Kawahara, D., Kurohashi, S.: Case Frame Compilation from the Web using High-Performance Computing. In: Proceedings of the 5th International Conference on Language Resources and Evaluation, pp. 1344–1347 (2006)
6. Watabe, H., Kawaoka, T.: The Degree of Association between Concepts using the Chain of Concepts. In: Proc. of SMC 2001, pp. 877–881 (2001)
7. Okumura, N., Yoshimura, E., Watabe, H., Kawaoka, T.: An Association Method Using Concept-Base. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part I. LNCS (LNAI), vol. 4692, pp. 604–611. Springer, Heidelberg (2007)

# Using Regression Analysis to Identify Patterns of Non-Technical Losses on Power Utilities

Iñigo Monedero<sup>1</sup>, Félix Biscarri<sup>1</sup>, Carlos León<sup>1</sup>, Juan I. Guerrero<sup>1</sup>, Jesús Biscarri<sup>2</sup>, and Rocío Millán<sup>2</sup>

<sup>1</sup> Department of Electronic Technology, University of Seville,  
C/ Virgen de Africa, 7, 41011 Sevilla, Spain

<sup>2</sup> Endesa, Avda. Borbolla S/N, 41092 Seville, Spain

**Abstract.** A non-technical loss (NTL) is defined as any consumed energy or service which is not billed because of measurement equipment failure or ill-intentioned and fraudulent manipulation of said equipment. This paper describes new advances that we have developed for Midas project. This project is being developed in the Electronic Technology Department of the University of Seville and its aim is to detect non-technical losses in the database of the Endesa Company. The main symptom of a NTL in a customer is an important drop in his billed energy. Thus, a main task for us is to detect customers with anomalous drops in their consumed energy. Concretely, in the paper we present two new algorithms based on a regression analysis in order to detect two types of patterns of decreasing consumption typical in customers with NTLs.

**Keywords:** Non-technical loss, power utility, data mining, regression analysis, Pearson correlation coefficient.

## 1 Introduction

A non-technical loss (NTL) is defined as any consumed energy or service which is not billed because of measurement equipment failure or ill-intentioned and fraudulent manipulation of said equipment. Although in the literature there are many works and researches [1-10], however there is not too much research about NTL detection in power utilities [11-16] although as we have said and it is verified the NTLs are very extended in this field. Thus, current methodology work by the electrical companies in the detection of NTLs is basically of two kinds. The first one is based on making in-situ inspections of some users (chosen after a consumption study) from a previously chosen zone. The second one is based on the study of the users which have null consumption during a certain period. The main problem of the first alternative is the need for a large number of inspectors and, therefore, a high cost. The problem with the second option is the impossibility of detecting users with non-null consumption (these are only the clearest cases of non-technical losses). Nowadays, data mining techniques [17-18] are being applied to multiple fields and detection of NTLs is one field in which it has met with success recently [19-22].

This paper describes new advances in the data mining process included on a prototype for NTL detections from the databases of the Endesa Company. The work is

within the framework of MIDAS project which we are developing at the Electronic Technology Department of the University of Seville with the funding of the electrical company.

We have presented results in MIDAS project using a detection process based on extraction rules and clustering techniques [23-24]. We are currently working on an additional line in order to detect other type of NTLs. The aim of this new line is the identification of patterns of drastic drop of consumption. It is because we know that the main symptom of a NTL is a drop in the billed energy of the customers.

Our algorithms are based on a regression analysis on the evolution of the consumption of the customer. The aim is to search strong correlation between the time (in monthly periods) and the consumption of the customer. The regression analysis makes it possible to adjust the consumption pattern of the customer by means a line with a slope. This slope must be indicative of the speed of the drop of the consumption and, therefore, the degree of correlation. Although the concept is quite simple we have developed this idea and we have reached two more-complex algorithms which make it possible to identify with a high grade of accuracy two type of suspicious (and typically corresponding to NTL) drops. The algorithms were programmed with SPSS Clementine (in version 11) [23-24].

## 2 Selection of Customers for the Analysis

For the development and tests of the algorithms we selected a sample set made by customers with rate 3.0.2 and 4.0. These types of rates are basically assigned by the Endesa Company to identify the enterprises whose contracted power is greater than 15 KW. Besides, concretely inside this set, we used those customers with a very high contracted power ( $>40$  KW). We chose 40 KW as lower limit in order to reach a total number of customers manageable for an analysis in detail and, at the same time, with the highest expected consumptions (and therefore to get in this way that each detected NTL supposed large among of recovered energy). This sample set was reached for the most important region of the Endesa Company: Catalonia (our objective in the future, once completed the validation of the algorithms with this region, will be to apply to all the regions of the Endesa Company as well as the remaining rates).

We configured an analysis period of 2 years which were a time enough to see a sufficiently detailed evolution of the consumption of the customer and, on the other hand, not too long to register along the contract the possible changes of type of business or the changes in the consumption habits of the client. With these customers we generated a table from which included condensed all the information of consumption and type of contract for each customer: reading values of the measurements equipment, bills from the last 2 years, amount of power contracted and the type of customer (private client or the kind of business of the contract), address, type of rate, etc. Thus, with this information in our study we could access to the type of customer as well as the evolution of its consumption in the last two years.

An interesting point of the pre-processing was the one concerning the reading values of the measurement equipment. Normally, the consumption billed is the result of consumption read, but this is not always true. If the company has no access to the data, and there is no doubt consumption has been made, the company experts estimate

the actual consumption, based on the recent historic. Severe and continuous differences between read data and billed data show abnormal behavior. In this sense, a filling up of missing values is performed.

Additionally to the previous selection we carried out a filtering of those customers with:

- Very low consumption (1000 KWs in the two years). This filter was carried out because the study of the consumption pattern of these customers is very limited and, besides, these customers are detected in the inspections of the company.
- Less number of reading values from the measurements equipment (under 10 from the 24 months of the analysis). We filtered those customers because our algorithms would be based on the consumption pattern of each customer and it was very difficult to study with less reading values. Besides, these customers with few reading values were not our objective because the Endesa Company has got its own methods in order to have identified them and to carry out the alerts to its inspectors if it is necessary.
- Without some reading value in the four last months. It was important that the customer had some reading value in these four months because it was necessary in order to carry out a precise process of the previously-described filling up of the reading values (since without final reading values it was not possible to adjust intermediate values).

Once carried out the selection and filtering of the sample set we had with a set of 24771 customers for our analysis.

### **3 Algorithm Based on Regression Analysis**

As we mentioned previously, an evident symptom of an anomalous consumption of the customer and for the detection of NTLs in the customers is a drastic drop of their consumption. These drops can be due to a real slope of the consumptions of the customers (e.g. due to a change of type of contract or by a different use of the consumed energy). But, in turn, these slopes can be due to failures in the measurement equipment or voluntary alterations of this equipment (both cases generates NTLs to the company and therefore loss of money for it).

On the other hand and as it is known, correlation and regression analysis are statistical tools for the investigation of relationships among the evolutions of different variables. Usually, the investigator seeks to ascertain the causal effect of one variable upon another (in our case consumption upon time). The goal of regression analysis is to determine the values of parameters for a function that cause the function to best fit a set of data observations that you provide (and therefore, at the same, to reach a degree of correlation). In linear regression, the function is a linear (straight-line) equation. For example, if we assume the value of an automobile decreases by a constant amount each year after its purchase, and for each mile it is driven.

We developed two complementary algorithms based on a linear regression analysis of the consumption pattern. Our objective was, on the one hand, the detection of those customers with dependence between consumption and time, and on the other hand, if this dependence was with decreasing consumption.

The first algorithm was based on the Pearson correlation coefficient, and the second one was based on a windowed regression analysis of the two years of consumption of each customer.

### 3.1 Algorithm Based on the Pearson Correlation Coefficient

In statistics, the Pearson correlation coefficient ( $r$ ) [25-26] is a measure of how well a linear equation describes the relation between two variables  $X$  and  $Y$  measured on the same object or organism.

The result of the calculus of this coefficient is a numeric value from -1 to 1. A value of 1 shows that a linear equation describes the relationship perfectly and positively, with all data points lying on the same line and with  $Y$  increasing with  $X$ . A score of -1 shows that all data points lie on a single line but that  $Y$  increases as  $X$  decreases. At last, a value of 0 shows that a linear model is inappropriate – that there is no linear relationship between the variables.

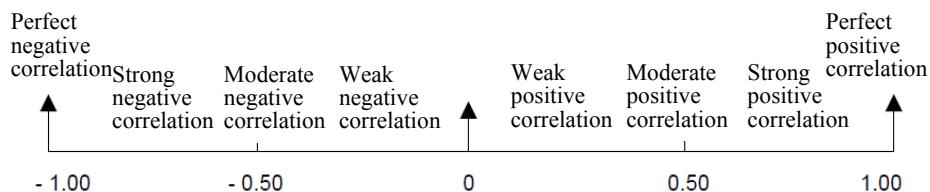
Our objective with this first algorithm was to identify those customers with important continuous drop in their consumption and, therefore, whose pattern of drop was very close-fitting to a linear equation. Thus, with this objective we identified and studied those customers with a Pearson coefficient near to -1.

The Pearson coefficient ( $r$ ) is calculated by means the following equation:

$$-1 \leq r = \frac{Cov(X, Y)}{s_X s_Y} = \frac{\sum_{t=1}^n (X_t - \bar{X}) * (Y_t - \bar{Y})}{\sqrt{\sum_{t=1}^n (X_t - \bar{X})^2} * \sqrt{\sum_{t=1}^n (Y_t - \bar{Y})^2}} \leq +1 \quad (1)$$

Where  $Cov(X, Y)$  is the covariance between  $X$  and  $Y$ .  $s_X s_Y$  is the product of the standard deviations for  $X$  and  $Y$ .

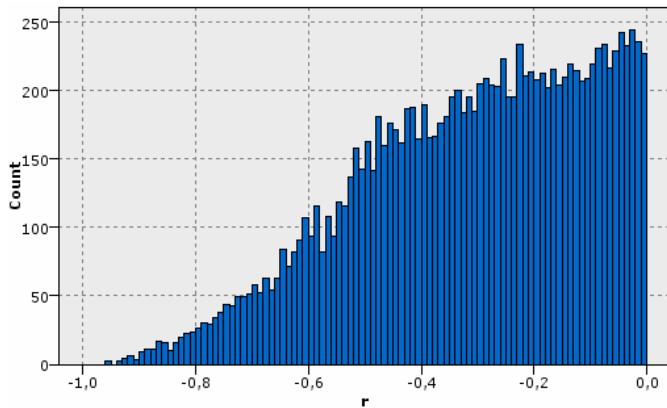
Thus, the result of this coefficient is interpreted as follows:



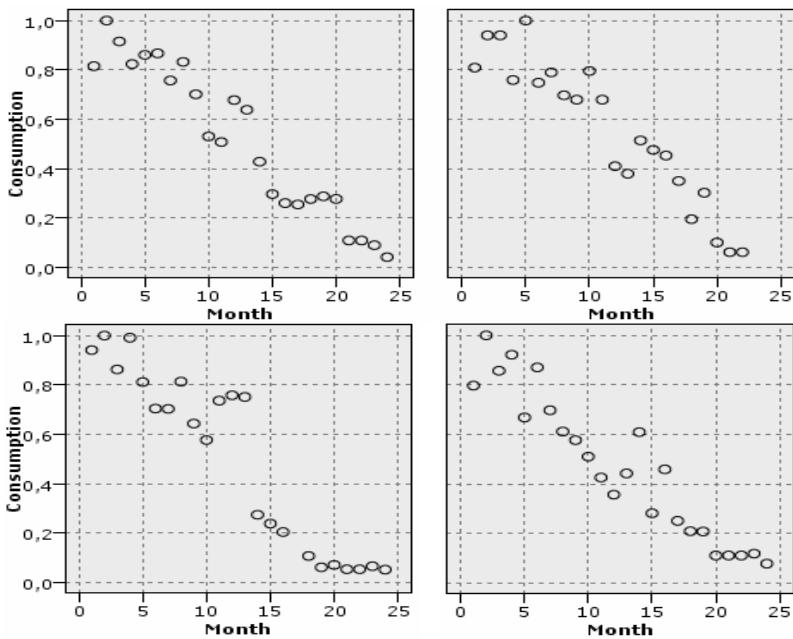
**Fig. 1.** Interpretation of the Pearson correlation coefficient

We applied this coefficient to the sample set, sorting the table by increasing Pearson coefficient and we obtained surprising results. We could observe many customers with strong (and some ones almost perfect) negative correlation. Thus, we could count 331 customers with  $r$  below -0.75 (the resultant histogram for negative values of  $r$  is showed in figure 2).

In figure 3 it is showed the scaled consumption of four customers with strong negative correlation. All the customers with strong correlation were suspicious of having some type of NTL (because is very strange a drop in consumption so pronounced and



**Fig. 2.** Histogram for negative values of  $r$  in the sample set



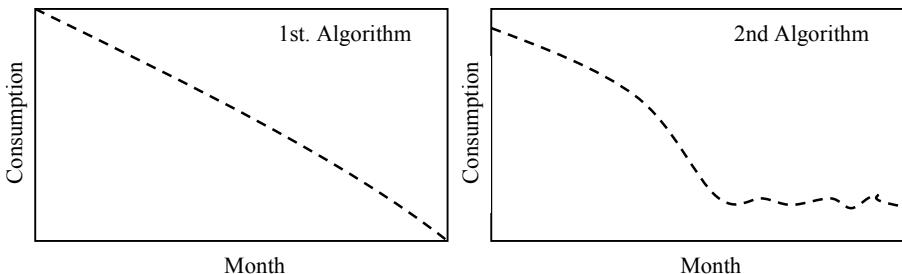
**Fig. 3.** Examples of customers with strong negative correlation

continuous in time) and they were proposed to be inspected in-situ by the inspectors of the Endesa Company (the results are remarked in paragraph 4 of the paper).

Thus, with this first algorithm we could detect with a high grade of accuracy those customers with a continuous drop in their consumption. The issue was that also were interesting those customers that their consumption was steadied with low values after falling. This last type of customers could not be detected with this algorithm. Thus, we developed a complementary method to this first algorithm based on a windowed algorithm.

### 3.2 Algorithm Based on a Windowed Linear Regression Analysis

The objective of this algorithm was to detect customers that their consumption was steadied with low values after falling and therefore with the consumption pattern of figure 4. It is important to emphasize that we were looking for customers with low consumption in the last months (but it was not interesting the customers which have got null-consumption in these months due to that they are already detected by the Endesa Company in its internal inspections).



**Fig. 4.** Consumption patterns searched with first and second algorithm

In order to get our objective we designed an algorithm by means the analysis of the consumption of the customer in two windows (each window with the half of the consumption values of the customer). For the first window we used the Pearson correlation coefficient (searching for values near to -1). On the other hand, for the second window we used a linear regression analysis [26] in which we searched for slopes near to 0 and non-zero offsets for this line. Thus, the Pearson coefficient for the first window (the 12 first consumption values) was calculated with the equation (1). While the values of the linear regression analysis for the second window (corresponding to the 12 last consumption values) were calculated in this way:

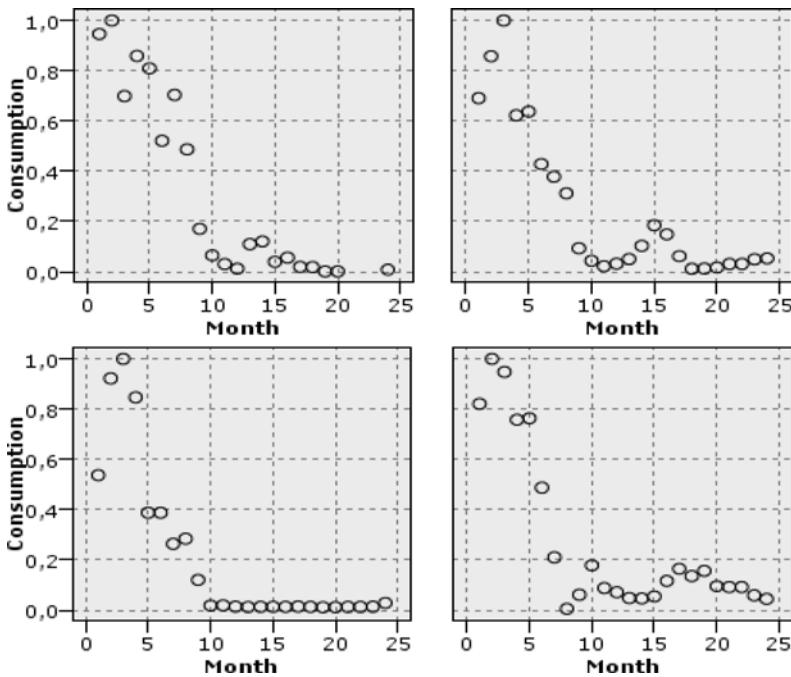
$$y = \alpha + \beta x \quad \beta = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \quad \alpha = \bar{y} - \bar{\beta}\bar{x} \quad (2)$$

Once calculated these values, we applied the following rule in order to extract those customers with a pattern similar to figure 4:

$$\begin{aligned} \text{Abs}(\beta_{w2}) < 30 \text{ and } \text{Average}_{w2} < (\text{Maximum}/5) \text{ and} \\ \text{Average}_{w2} > (\text{Maximum}/100) \text{ and } R_{w1} < -0.5 \end{aligned} \quad (3)$$

Where  $\text{Abs}(\beta_{w2})$  is the absolute value of  $\beta$  for the second window (we took the absolute values in order to identify in this second window those customers without slope in their consumption or very low –positive or negative–),  $\text{Average}_{w2}$  is the average of the consumption for the second window,  $\text{Maximum}$  is the value maximum of the reading values of the customer, and  $R_{w1}$  is the Pearson coefficient for the first window. The meaning of this rule implies an important drop in the first year and a stabilization of the consumption with low values (but not null) with respect to the total consumption of the customer.

Applying this rule on our sample set we obtained 81 customers. Through a display of their consumption, we could verify that these clients had the pattern of Figure 4. In Figure 5 we can observe the patterns for four customers of these 81.



**Fig. 5.** Examples of customers detected with second algorithm

## 4 Results and Conclusions

NTL is an important issue in power utilities because it has a high impact on company profits. Despite this, nowadays the methodology of detection of NTL of the companies is not very advanced as these companies used detection methods that do not exploit the use of data mining techniques. We have developed methods to detect NTLs and we have tested them on a real database supplied by the Endesa Company.

Concretely, in this paper we have presented a line of work based on the detection of consumption drops by means of two algorithms which use regression analysis. Thus, we obtained a list of customers with evident and suspicious drops of consumption. We selected a sample of the 80 more representative customers (40 customers with each one of the two algorithms) who had clearer drop in the consumption. These cases could be due to a drop of electrical demand for their business but never due to a low contract because in that case they would have reading information in their equipment. Therefore, it was interesting as additional information to study the type of business of these suspicious customers in order to know if it was a business in which the demand is currently falling (e.g., currently, the construction business in Spain). Thus,

we studied the business information for each customer in order to be able to control this fact and to avoid unnecessary inspections. It is known by the inspectors of the Company that the following types of business are more likely to have consumption drops innate to their use of the energy (and not due to possible NTL): wells, lightings, irrigation pumps, water purification and construction (previously mentioned). So, from the 80 detected customers, we filtered those with these types of contracts and we obtained a definitive list of 62.

Currently, the Endesa Company is carrying out inspections with a set of customers from the ones who were detected by our methods. Up to now, with the results obtained in the inspections, we have reached an around 38% of success. The total These results are considered very satisfactory taking into account, first, the rate of success of the Company in its routine inspections (less than 10%) and, second, the less input information used in our algorithms (basically the evolution of the consumption of the customer).

To date, the total energy recovered with our prototype stands at about 2 millions of kWh, which implies a large amount of money saved for the Endesa Company. This is allowing us to continue working with guarantees in our project. Thus, in order to improve the filtering process previous to the in-situ inspections, we are currently working on an expert system that takes as input all this information from the database and carries out the task of hand analysis. It is to complete our detections with more information that can be determining the decision to inspect in situ that customer (as for example the type of business, the stationary consumption in some types of business or even the location of the customer).

## Acknowledgements

The authors would like to thank the Endesa Company and Sadiel Company for providing the funds for this project (since 2005). The authors are also indebted to the following colleagues for their valuable assistance in the project: Gema Tejedor, Francisco Godoy and Joaquín Mejías. Special thanks to Jesús Macías, Eduardo Ruizberriz, Juan Ignacio Cuesta, Tomás Blazquez and Jesús Ochoa for their help and cooperation.

## References

1. Wheeler, R., Aitken, S.: Multiple algorithms for fraud detection. *Knowledge based systems* 13, 93–99 (2000)
2. Kou, Y., Lu, C.-T., Sinvongwattana, S., Huang, Y.-P.: Survey of fraud detection techniques. In: Proceeding of the 2004 IEEE International Conference on Networking, Sensing and Control, Taiwan, March 21, pp. 89–95. IEEE press, Los Alamitos (2004)
3. Fawcett, T., Provost, F.: Adaptable fraud detection. *Data mining and Knowledge Discovery* 1, 291–316 (1997)
4. Artís, M., Ayuso, M., Guillén, M.: Modeling different types of automobile insurance frauds behavior in the spanish market. In: *Insurance Mathematics and Economics*, vol. 24, pp. 67–81. Elsevier Press, Amsterdam (1999)

5. Daskalaki, S., Kopanas, I., Goudara, M., Avouris, N.: Data mining for decision support on customer insolvency in the telecommunication business. *European Journal of Operational Research* 145, 239–255 (2003)
6. Brause, R., Langsdorf, T., Hepp, M.: Neural data mining for credit card fraud detection. In: Proceeding 11th IEEE International Conference on Tools with Artificial Intelligence. IEEE press, Los Alamitos (1999)
7. Kirkos, E., Spathis, C., Manolopoulos, Y.: Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications* 32, 995–1003 (2007)
8. Burge, P., Shawe-Taylor, J.: Detecting cellular fraud using adaptative prototypes. In: Proceeding on AI Approaches to Fraud Detection and Risk Management, pp. 9–13. AAAI Press, Menlo Park (1997)
9. Cabral, J., Pinto, J., Linares, K., Pinto, A.: Methodology for fraud detection using rough sets. In: 2006 IEEE International Conference on Granular Computing. IEEE press, Los Alamitos (2006)
10. Denning, D.: An intrusion-detection model. *IEEE transactions on Software Engineering* 13, 222–232 (1987)
11. Yap, K.S., Hussien, Z., Mohamad, A.: Abnormalities and fraud electric meter detection using hybrid support vector machine and genetic algorithm. In: Proceeding of the Third IASTED International Conference Advances in Computer Science and Technology, Phuket, Thailand, April 2-4, Iasted Press (2007)
12. Filho, J., als: Fraud identification in electricity company customers using decision tree. In: IEEE International Conference on Systems, Man and Cibernetics, IEEE/PES, The Hague (2004)
13. Cabral, J., Pinto, J., Gontijo, E.M., Reis, J.: Fraud detection in electrical energy consumers using rough sets. In: 2004 IEEE International Conference on Systems, Man and Cybernetics. IEEE press, Los Alamitos (2004)
14. Cabral, J., Pinto, J., Martins, E., Pinto, A.: Fraud detection in high voltage electricity consumers using data mining. In: IEEE Transmission and Distribution Conference and Exposition T&D, April 21-24, IEEE/PES (2008)
15. Sforza, M.: Data mining in power company customer database. In: Electric Power Systems Research, vol. 55, pp. 201–209. Elsevier Press, England (2000)
16. Jiang, R., Tagiris, H., Lachsz, A., Jeffrey, M.: Wavelet based features extraction and multiple classifiers for electricity fraud detection. In: Transmission and Distribution Conference and Exhibition 2002: Asia pacific, October 6-10. IEEE/PES (2002)
17. Kantardzic, M.: Data mining: concepts, models methods and algorithms, 1st edn. AAAI/MIT Press (1991)
18. Witthen, I., Frank, E.: Data Mining—Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann/Academic Press, New York/San Mateo (2000)
19. Editorial, Recent advances in data mining. *Engineering applications of Artificial Intelligence* 19, 361–362 (2006)
20. McCarthy, J.: Phenomenal data mining. *Communications of the ACM* 43(8), 75–79 (2000)
21. Ramos, S., Vale, Z.: Data mining techniques application in power distribution utilities. In: IEEE Transmision and Distribution Conference and Exposition T&D, April 21-24, IEEE/PES (2008)
22. Nizar, A., Dong, Z., Zhao, J.: Load profiling and data mining techniques in electricity deregulated market. In: Power Engineering Society General Meeting, June 18-22. IEEE/PES (2006)

23. Biscarri, F., Monedero, I., León, C., Guerrero, J.I., Biscarri, J., Millán, R.: A data mining method based on the variability of the customers consumption. In: 10th Int. Conf. on Enterp. Inf. Systs., ICEIS 2008, Barcelona, Spain, June 12-16 (2008)
24. Biscarri, F., Monedero, I., León, C., Guerrero, J.I., Biscarri, J., Millán, R.: A mining Framework to detect non-technical losses in power utilities. In: 11th Int. Conf. on Enterp. Inf. Systs., ICEIS 2009, Milano, Italy, May 6-10 (2009)
25. Pearson, K.: Mathematical contributions to the theory of evolution.—III. Regression, heredity and panmixia. *Philos. Trans. R. Soc. London, ser. A* 187, 253–318 (1896)
26. Moore, D.: Basic Practice of Statistics. W.H. Freeman, San Francisco (2006)

# Enhancing the Symbolic Aggregate Approximation Method Using Updated Lookup Tables

Muhammad Marwan Muhammad Fuad and Pierre-François Marteau

VALORIA, Université de Bretagne Sud, Université Européenne de Bretagne  
BP. 573, 56017 Vannes, France  
[{marwan.fuad,pierre-francois.marteau}@univ-ubs.fr](mailto:{marwan.fuad,pierre-francois.marteau}@univ-ubs.fr)

**Abstract.** Similarity search in time series data mining is a problem that has attracted increasing attention recently. The high dimensionality and large volume of time series databases make sequential scanning inefficient to tackle this problem. There are many representation techniques that aim at reducing the dimensionality of time series so that the search can be handled faster at a lower dimensional space level. Symbolic representation is one of the promising techniques, since symbolic representation methods try to benefit from the wealth of search algorithms used in bioinformatics and text mining communities. The symbolic aggregate approximation (SAX) is one of the most competitive methods in the literature. SAX utilizes a similarity measure that is easy to compute because it is based on pre-computed distances obtained from lookup tables. In this paper we present a new similarity measure that is almost as easy to compute as the original similarity measure, but it is tighter because it uses updated lookup tables. In addition, the new similarity measure is more intuitive than the original one. We conduct several experiments which show that the new similarity measure gives better results than the original one.

**Keywords:** Time Series Data Mining, Symbolic Representation, Symbolic Aggregate Approximation, Updated Minimum Distance.

## 1 Introduction

Similarity search is a fundamental problem in computer science. This problem has many applications in multimedia databases, bioinformatics, pattern recognition, text mining, computer vision, data mining, machine learning and others.

Time series are data types that appear in many medical, scientific and financial applications. Time series data mining includes many tasks such as classification, clustering, similarity search, motif discovery, anomaly detection, and others. One key to performing these tasks successfully is representation methods that can represent the time series efficiently and effectively. Another key is indexing time series in appropriate structures, which direct the query processing towards regions in the search space, where similar time series to the query are likely to be found, which makes the retrieving process faster.

Time series are high dimensional data, so even indexing structures can suffer from the so-called “dimensionality curse”. One of the best solutions to deal with this

phenomenon is to utilize a dimensionality reduction technique to reduce the dimensionality of time series, then to utilize a suitable indexing structure on the reduced space. There have been different suggestions to represent time series, to mention a few; DFT [1,2], DWT [3], SVD [8], APCA [6], PAA [5,13], PLA [11], etc.

Among dimensionality reduction techniques, symbolic representation has several advantages, because it allows researchers to benefit from text-retrieval algorithms and techniques [6].

Similarity between two data objects can be depicted by means of a similarity measure, or a distance metric, which is a stronger mathematical concept. There are quite a large number of similarity measures; some are applied to a particular data type, while others can be applied to different data types. Among the different similarity measures, there are those that can be used on symbolic data types. At first they were available for data types whose representation is naturally symbolic (DNA and proteins sequences, textual data, etc.). But later these symbolic similarity measures were also applied to other data types that can be transformed into strings by using different symbolic representation techniques.

Of all the symbolic representation methods in the times series data mining literature, the symbolic aggregate approximation method (SAX) [4] stands out as one of the most powerful methods. The main advantage of this method is that the similarity measure that it uses is easy to compute, because it uses statistical lookup tables. In this paper we present an improved similarity measure to be used with SAX. It has the same advantages as the original similarity measure used in SAX. But our new similarity measure gives better results in different time series data mining tasks, in return of a small additional computational cost that does not affect the overall complexity which remains  $O(N)$  for both measures.

The rest of this paper is organized as follows: in section 2 we present related work on dimensionality reduction, and on symbolic representation of time series in general, and SAX in particular. The proposed similarity measure is presented in section 3. In section 4 we present the results of some of experiments we conducted. The conclusion is presented in section 5.

## 2 Related Work

### 2.1 Dimensionality Reduction

Time series are generally highly correlated data, so representation methods that aim at reducing the dimensionality of these data by projecting the original space onto a lower dimensional space and processing the query in the reduced space is a scheme that is widely used in time series data mining community.

When embedding the original space into a lower dimensional space and performing the similarity query in the transformed space, two main side-effects may be encountered; *false alarms*, also called *false positivity*, and *false dismissals*. False alarms are data objects that belong to the response set in the transformed space, but do not belong to the response set in the original space. False dismissals are data objects that the search algorithm excluded in the transformed space, although they are answers to the query in the original space. Generally, false alarms are more tolerated than false

dismissals, because a post-processing scanning is usually performed on the candidate answer set using the original data objects and the original distance to filter out the data objects that are not valid answers to the query. However, false alarms can slow down the search time if they are too many. False dismissals are a more serious problem and avoiding them requires more sophisticated procedures.

False alarms and false dismissals are dependent on the transformation used in the embedding. If  $f$  is a transformation from the original space  $(S_{original}, d_{original})$  into another space  $(S_{transformed}, d_{transformed})$  then in order to guarantee no false dismissals this transformation should satisfy:

$$d_{transformed}(f(u_1), f(u_2)) \leq d_{original}(u_1, u_2), \quad \forall u_1, u_2 \in S_{original} \quad (1)$$

The above condition is known as the *lower-bounding lemma*. [13]

If a transformation can make the two above distances equal for all the data objects in the original space, then similarity search produces no false alarms or false dismissals. Unfortunately, such an ideal transformation is very hard to find. Yet, we try to make the above distances as close as possible. The above condition can be written as:

$$0 \leq \frac{d_{transformed}(f(u_1), f(u_2))}{d_{original}(u_1, u_2)} \leq 1 \quad (2)$$

A *tight* transformation is, by definition, one that makes the above ratio as close as possible to 1.

## 2.2 Symbolic Representation

One of the dimensionality reduction techniques in time series data mining is symbolic representation. Symbolic representation of time series uses an alphabet  $\Sigma$  (usually finite) to reduce the dimensionality of the time series. This can be defined formally as follows: given a time series  $TS = t_1, t_2, \dots, t_n$ , the symbolic representation scheme is a map:

$$[t_i, t_j] \xrightarrow{f} \alpha_k ; \alpha_k \in \Sigma \quad (3)$$

Symbolic representation of time series has attracted much attention, because by using this scheme we can not only reduce the dimensionality of time series, but can also benefit from the numerous algorithms used in bioinformatics and text data mining.

The symbolic aggregate approximation method (SAX) is one of the most powerful methods of symbolic representation of time series. SAX is based on the fact that normalized time series have “highly Gaussian distribution” (according to the authors of [4]), so by determining the breakpoints that correspond to the alphabet size, one can obtain equal sized areas under the Gaussian curve.

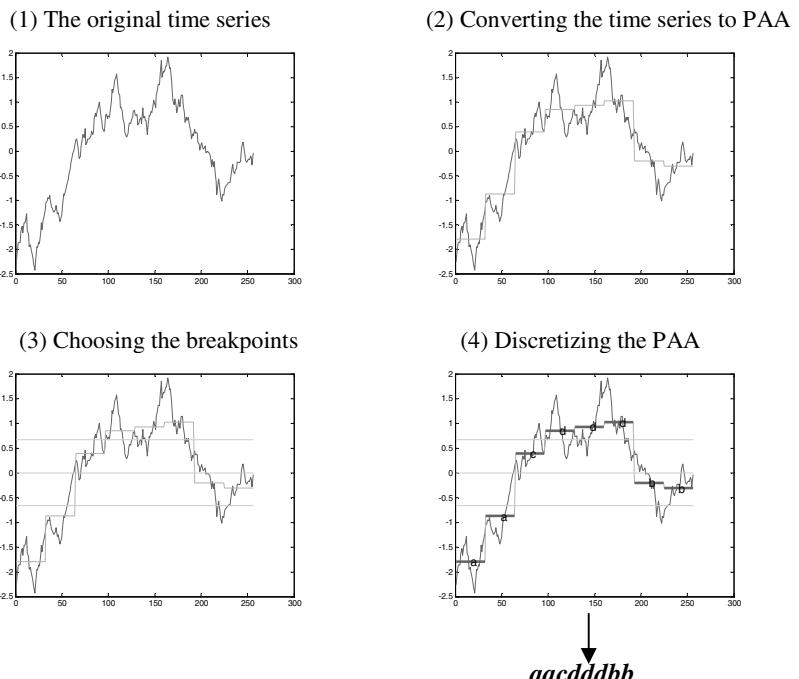
SAX is applied as follows: in the first step the time series are normalized. In the second step, the dimensionality of the time series is reduced by using PAA (Piecewise Aggregate Approximation) [5]. In PAA the times series is divided into equal sized frames and the mean value of the points within the frame is computed. The lower dimensional vector of the time series is the vector whose components are the means

of the successive frames. In the third step, the PAA representation of the time series is discretized. This is achieved by determining the number and location of the breakpoints. The number of breakpoints is related to the desired alphabet size (which is chosen by the user), i.e.  $\text{alphabet\_size} = \text{number}(breakpoints) + 1$ . Their locations are determined by statistical lookup tables so that these breakpoints produce equal-sized areas under the Gaussian curve. The interval between two successive breakpoints is assigned to a symbol of the alphabet, and each segment of the PAA that lies within that interval is discretized by that symbol. The last step of SAX is using the following similarity measure:

$$\text{MINDIST}(\hat{S}, \hat{T}) \equiv \sqrt{\frac{n}{N}} \sqrt{\sum_{i=1}^N (\text{dist}(\hat{s}_i, \hat{t}_i))^2} \quad (4)$$

Where  $n$  is the length of the original time series,  $N$  is the length of the strings (the number of the frames),  $\hat{S}$  and  $\hat{T}$  are the symbolic representations of the two time series  $S$  and  $T$ , respectively, and where the function  $\text{dist}()$  is implemented by using the appropriate lookup table.

Notice that MINDIST is a similarity measure and not a distance metric since it violates two conditions of distance metrics which are the identity condition: ( $x = y \Leftrightarrow d(x, y) = 0$ ), and the triangular inequality condition:



**Fig. 1.** The different steps of SAX

( $d(x, z) \leq d(x, y) + d(y, z)$ ), and it respects only one condition which is the symmetry condition: ( $d(x, y) = d(y, x)$ )

We also need to mention that the similarity measure used in PAA is:

$$d(X, Y) = \sqrt{\frac{n}{N} \sum_{i=1}^N (\bar{x}_i - \bar{y}_i)^2} \quad (5)$$

Where  $n$  is the length of the time series,  $N$  is the number of frames. It is proven in [5] and [13] that the above similarity distance is a lower bounding of the Euclidean distance applied in the original space of time series. This results in the fact that MINDIST is also a lower bounding of the Euclidean distance, because it is a lower bounding of the similarity measure used in PAA. This guarantees no false dismissals. Figure.1 illustrates the different steps of SAX.

### 3 The Updated Minimum Distance (UMD)

The main advantage of SAX, which makes it fast to apply, is that the similarity measure it uses is easy to compute, because it is based on pre-computed distances obtained from corresponding lookup tables. However, MINDIST has a main drawback; in order to be lower bounding this similarity measure ignores all the distances between any successive symbols of the alphabet and considers them to be zero. For instance, the lookup table of the MINDIST for an alphabet size of 3 is the one shown in Table 1.

This drawback has two consequences; the first is that MINDIST is not tight enough, which produces many false alarms. The second consequence can be shown by the following example. Let the symbolic representing of the five time series  $TS_1, TS_2, TS_3, TS_4, TS_5$  using SAX with alphabet size= 4 be:  $TS_1 = aabdd$ ,  $TS_2 = bacdc$ ,  $TS_3 = abbcd$ ,  $TS_4 = bacdd$ ,  $TS_5 = bbbdc$ . The MINDIST between any two of these five time series is zero, which is not only unintuitive, since no two time series of these five are identical, but this also produces many false alarms.

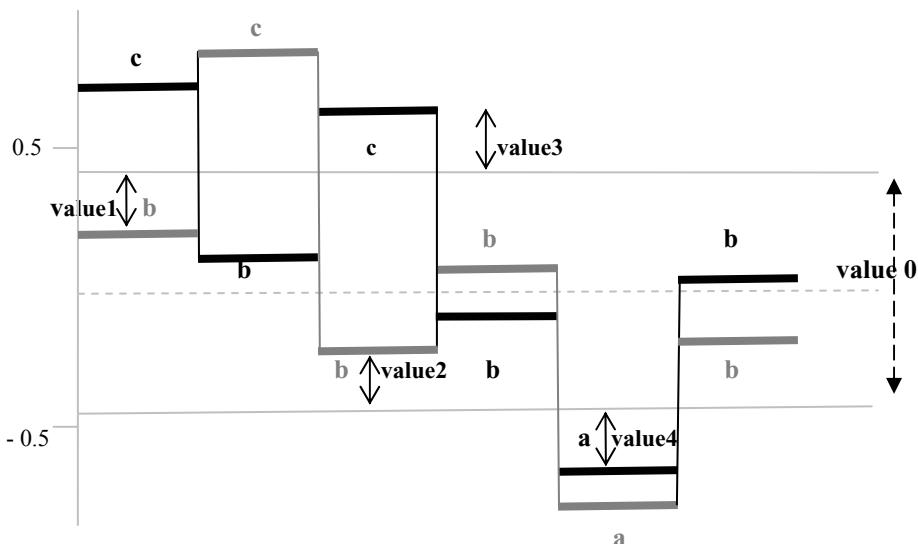
In this section we present a modified minimum distance, which remedies the above problems. The new minimum distance has all the advantages of the original distance, in that it is also a lower bounding of the Euclidean distance, and it is also fast to compute, as it uses pre-computed distances. But the new minimum distance is tighter. It is also intuitive, in that it does not ignore the distances between successive symbols.

**Table 1.** The lookup table of MINDIST for alphabet size equals to 3. All values between any successive symbols are equal to zero. The breakpoints in this case (obtained from statistical tables) are: -0.43 and 0.43. The distance between them is 0.86.

	a	b	c
a	0	0	0.86
b	0	0	0
c	0.86	0	0

The principle of our new minimum distance, which we call the updated minimum distance (UMD) is to recover the distances between any successive symbols, which were ignored in MINDIST. Figure 2 shows an example of the ignored distances in the case of alphabet size equals to 3, and which are recovered in UMD. The breakpoints are -0.43 and 0.43. In this case the only non-zero distance according to MINDIST is  $dist(a,c)$  which is equal to 0.86 (the distance indicated by the dashed arrow). The distances represented by the solid arrows are the distances between the minima or the maxima of all the symbols of the alphabet and the corresponding breakpoint. These distances are ignored in MINDIST, but as we can see they are not equal to zero. So  $dist(a,b)$ , which was zero in MINDIST can be updated to become  $value2+value4$ , and  $dist(b,c)$  which was also zero in MINDIST can be updated to become  $value1+value3$ , and even  $dist(a,c)$  is updated to become  $value4+value3+value0$  (the original value). Lookup tables of different alphabet sizes are updated in a like manner. Obviously, this update of lookup tables results in a tighter similarity distance. For instance, the lookup table shown at the beginning of this section can be updated to become the one shown in Table2.

We can easily notice that this new similarity measure is lower bounding of the PAA similarity measure presented in (5) in section 2.2, since we take the closest distance between two successive symbols among all the distances of all the PAA segments of these two symbols. As a result, our new similarity measure is also a lower bounding of the Euclidean distance (c.f section 2.2). This is the same property that MINDIST has.



**Fig. 2.** The PAA representation of two time series: TS1 =cbcba (boldface black) and TS2=bcbbab (boldface grey). The solid arrows show the ignored distances and the dashed arrow shows the only distance considered by MINDIST :  $dist(a,c)=value0$  (0.86).

**Table 2.** The updated lookup table for alphabet size equals to 3. We can see that the distances between successive symbols are no longer equal to zero, and the distance  $dist(a,c)$  is tighter than  $dist(a,c)$  in Table 1.

	a	b	c
a	0	value2+value4	0.86+ value4+value3
b	value2+value4	0	value1+value3
c	0.86+ value4+value3	value1+value3	0

The other consequence of this update is that UMD, which is based on the updated lookup tables, is intuitive, because it gives non-zero values to successive symbols, so the UMD of any two of the five time series presented at the beginning of this section is not zero, which is what we expect from any similarity measure applied to these time series because they are not identical.

In order to obtain the minimum and the maximum values of each symbol, the SAX algorithm is modified so that at the step where the different segments of the PAA are compared against the breakpoints to decide what symbol is used to discretize that segment, at that step we modify SAX so that it keeps a record of the minimum and maximum values of each segment of that time series. This is performed at indexing

```

procedure dist=UMD(TS1, TS2, TS1MinMax, TS2MinMax,
Alphabet, LookupTable)
// INPUT : TS1,TS2; two input times series presented
// symbolically
// INPUT : TS1MinMax,TS2MinMAX; The Min and Max
// distances between TS1 (TS2) and the corresponding
// breakpoints
// INPUT : Alphabet
// INPUT : LookupTable is the look up table used with
// MINDIST
// OUTPUT : RETURN dist, the UMD distance between TS1,
// TS2
for  $\alpha_k \in$  Alphabet
    mn= min (TS1MinMax( $\alpha_k$ ), TS2MinMax( $\alpha_k$ ))
    mx= min (TS1MinMax( $\alpha_k$ ), TS2MinMax( $\alpha_k$ ))
//update the lookup table for symbol  $\alpha_k$ 
    UpdateTable  $\leftarrow$  Update(Lookuptable,  $\alpha_k$ , mn, mx)

end
return dist=MINDIST(TS1, TS2, UpdateTable)
end procedure

```

**Fig. 3.** Pseudo Code for the UMD distance

time, so it does not include any extra cost at query time. Then when comparing two time series, we take the minimum (maximum) that corresponds to the same symbol of the two times series to find the mutual minimum (maximum) that corresponds to each symbol. These minima and maxima, which are computed at indexing time, are used to update the lookup tables. The update process includes very few addition operations (three for alphabet size= 3, for instance), whose computational cost is very low compared with the cost of computing the similarity measure. So UMD has the same complexity as that of MINDIST. The pseudo code for the UMD is shown in Figure 3.

So, as we can see, the computational cost of UMD is a little bit higher at indexing time, but it has the same complexity as MINDIST at query time.

We also have to mention that UMD is also a similarity measure and not a distance metric. However, it is one-step closer to being distance metric since it violates only one condition of the distance metric which is the triangular inequality condition (c.f section 2.2).

## 4 Empirical Evaluation

We conducted extensive experiments on the proposed similarity measure. In our experiments we tested UMD on all the 20 datasets available at UCR [14] and for all alphabet sizes, which vary between 3 (the least possible size that was used to test MINDIST) to 20 (the largest possible alphabet size). The size of these datasets varies between 28 (Coffee) and 6164 (wafer). The length of the time series varies between 60 (Synthetic Control) and 637 (Lightning-2). So these data sets are very diverse. We also tested these datasets using the Euclidean distance, because this distance is widely used in the time series data mining community [7, 12].

### 4.1 Tightness

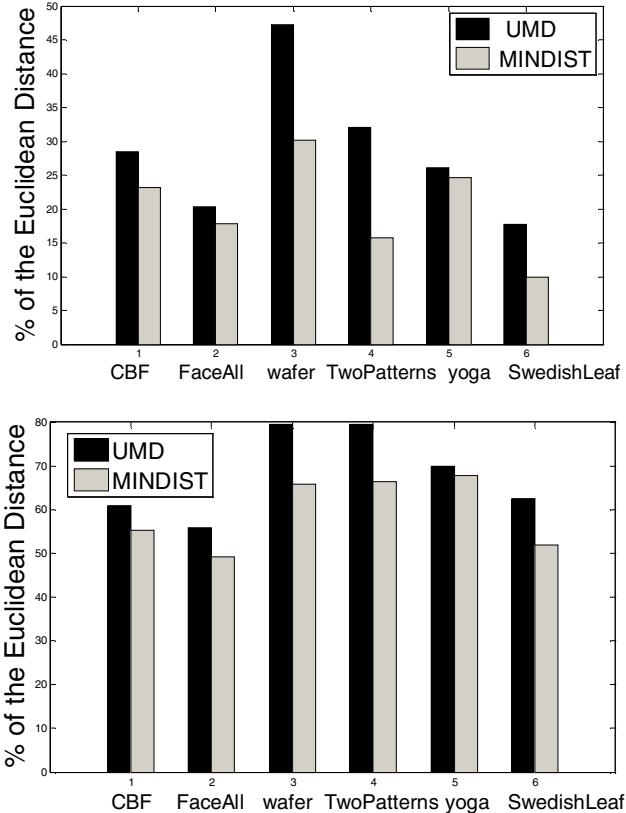
As mentioned in section 2.1, tightness of the similarity distances enhances the search process, because it minimizes the number of false alarms. As a result, it decreases the post processing time.

We compared the tightness of UMD with the tightness of MINDIST, for all the datasets and for all values of the alphabet size. In all the experiments, UMD was tighter than MINDIST. In Figure 4, we present some of the results we obtained for alphabet size equals to 3 and 10, respectively. We chose to report these datasets because they are the largest data sets in the UCR archive, thus the tightness, which is the average of the rate of the corresponding similarity measure to the Euclidean distance between all pairs of time series in the dataset, is more accurate statistically.

The experiments conducted on these datasets using other values of the alphabet size, in addition to the experiments on the other datasets in the UCR archive for all values of the alphabet size, all gave similar results.

### 4.2 Classification

Classification is one of the main tasks in time series data mining. We tested the proposed similarity measure in a classification task based on the first nearest-neighbor rule on all the data sets available at UCR. We used leaving-one-out cross validation.



**Fig. 4.** Comparison of the tightness of UMD with the tightness of MINDIST in 6 data sets and for alphabet size=3 (above) and alphabet size=10 (below). The figure shows that UMD is tighter than MINDIST.

In order to make a fair comparison, we used the same compression ratio (the number of points used to represent one segment in PAA) that was used to test SAX with MINDIST (i.e. 1 to 4). The first version of SAX used alphabet size that varies between 3 and 10. Then in a later version the alphabet size varied between 3 and 20. We conducted two main classification experiments; the first one is for alphabet size (3:10), and the second is for alphabet size (3:20). Each experiment tested all the datasets. In each experiment we start by varying the alphabet size (3 through 10 in the first experiment and 3 through 20 in the second one) on the training set of each dataset to find the optimal value of the alphabet size of that dataset; i.e. the value that minimizes the classification error rate. Then we use that optimal alphabet size on the testing set of that dataset. Table 3 shows the results of our experiments (There is no training phase for the Euclidian distance). Columns 3 and 4 contain the results of the first experiment of UMD and MINDIST, respectively, where the alphabet size varies between 3 and 10. The best result between UMD and MINDIST for that range of alphabet size is highlighted. Columns 5 and 6 contain the results of the second experiment

**Table 3.** The rate error of UMD and MINDIST for  $\alpha$  between 3 and 10 (columns 3 and 4), and for  $\alpha$  between 3 and 20 (columns 5 and 6). Column 2 shows the error rate of the Euclidean distance.

	1-NN Euclidean Distance	UMD ( $\alpha^*$ between 3 and 10)	MINDIST ( $\alpha$ between 3 and 10)	UMD ( $\alpha$ between 3 and 20)	MINDIST ( $\alpha$ between 3 and 20)
<b>Synthetic Control</b>	0.12	0.007	0.033	0.003	0.023
<b>Gun-Point</b>	<b>0.087</b>	0.213	0.233	0.14	0.127
<b>CBF</b>	0.148	0.131	0.104	0.054	0.073
<b>Face (all)</b>	<b>0.286</b>	0.306	0.319	0.305	0.305
<b>OSU Leaf</b>	0.483	0.471	0.475	0.471	0.475
<b>Swedish Leaf</b>	<b>0.213</b>	0.291	0.490	0.242	0.253
<b>50words</b>	0.369	0.338	0.327	0.345	0.334
<b>Trace</b>	<b>0.24</b>	0.34	0.42	0.27	0.35
<b>Two Patterns</b>	0.09	0.076	0.081	0.065	0.066
<b>Wafer</b>	0.005	0.004	0.004	0.004	0.004
<b>Face (four)</b>	<b>0.216</b>	0.273	0.239	0.273	0.239
<b>Lighting-2</b>	0.246	0.230	0.213	0.229	0.148
<b>Lighting-7</b>	0.425	0.411	0.493	0.411	0.425
<b>ECG200</b>	0.12	0.11	0.09	0.11	0.13
<b>Adiac</b>	<b>0.389</b>	0.634	0.903	0.494	0.867
<b>Yoga</b>	<b>0.170</b>	0.193	0.199	0.172	0.181
<b>Fish</b>	<b>0.217</b>	0.366	0.514	0.257	0.263
<b>Beef</b>	0.467	0.367	0.533	0.333	0.433
<b>Coffee</b>	0.25	0.179	0.464	0.071	0.143
<b>Olive Oil</b>	<b>0.133</b>	0.367	0.833	0.3	0.833
<b>MEAN</b>	0.234	0.265	0.348	0.227	0.284
<b>STD</b>	<b>0.134</b>	0.158	0.247	0.147	0.237

(\*:  $\alpha$  is the alphabet size)

of UMD and MINDIST, respectively, where the alphabet size varies between 3 and 20. The best result between UMD and MINDIST for the same range of alphabet size is also highlighted.

The results show that for alphabet size in the interval (3:10), UMD outperforms MINDIST in 14 datasets, and MINDIST outperforms UMD in 5 datasets, and in one case they both give the same result. For alphabet size that varies in the interval (3:20) UMD outperforms MINDIST in 14 datasets and MINDIST outperforms UMD in 4 datasets, and in 2 datasets they both give the same results. The average error of UMD over all the datasets and for both ranges of alphabet size is smaller than that of MINDIST. The standard deviation for UMD is also smaller than that of MINDIST and for both ranges of the alphabet size. The significance of this statistical parameter is that

when the standard deviation is small, the similarity measure is more robust, and can be applied to different kinds of datasets.

The results obtained show that the general performance of UMD is better than that of MINDIST.

It is worth to mention that for the Euclidian distance, there is no compression of data, thus no loss of information, so in some cases it may give better results than symbolic, compressed distances. However, using the Euclidean distance as a reference should not be taken for granted because this distance has a few inconveniences; it is sensitive to noise and to shifts on the time axis [10].

## 5 Conclusion and Future Work

In this paper we presented a new similarity measure to be used with the symbolic aggregate approximation (SAX). The new similarity measure UMD improves the performance of SAX compared with the original similarity measure MINDIST used with SAX. We conducted several experiments of times series data mining tasks. The results obtained show that SAX with UMD gives better results than SAX with MINDIST. Another interesting feature of the new similarity measure is that it has the same complexity as that of MINDIST. Other experiments on the CPU time of both MIDIST and UMD may also be conducted to support the results presented in this paper.

The future work can focus on improving this similarity measure by tracing other values in the original time series. This can make the proposed similarity measure even tighter, which is what we are working on now.

Another direction of future work focuses on modifying SAX to benefit more from UMD. We think this can be achieved by using a representation method other than PAA. This may include more calculations or more storage space at indexing time, but it could give better results.

UMD can also be used as a fast-and-dirty filter to speed up the nearest neighbor search by quickly filtering out the time series which are not answers to the query and terminating the search using the Euclidean distance.

## References

1. Agrawal, R., Faloutsos, C., Swami, A.: Efficient similarity search in sequence databases. In: Lomet, D.B. (ed.) FODO 1993. LNCS, vol. 730. Springer, Heidelberg (1993)
2. Agrawal, R., Lin, K.I., Sawhney, H.S., Shim, K.: Fast similarity search in the presence of noise, scaling, and translation in time-series databases. In: Proceedings of the 21st Int'l Conference on Very Large Databases, Zurich, Switzerland, pp. 490–501 (1995)
3. Chan, K., Fu, A.W.: Efficient Time Series Matching by Wavelets. In: Proc. of the 15th IEEE Int'l Conf. on Data Engineering, Sydney, Australia, March 23-26, pp. 126–133 (1999)
4. Lin, J., Keogh, E.J., Lonardi, S., Chiu, B.Y.-c.: A symbolic representation of time series, with implications for streaming algorithms. In: DMKD 2003, pp. 2–11 (2003)
5. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra: Dimensionality reduction for fast similarity search in large time series databases. J. of Know. and Inform. Sys. (2000)

6. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra: Locally adaptive dimensionality reduction for similarity search in large time series databases. In: SIGMOD, pp. 151–162 (2001)
7. Keogh, E., Kasetty, S.: On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, July 23 - 26, pp. 102–111 (2002)
8. Korn, F., Jagadish, H., Faloutsos, C.: Efficiently supporting ad hoc queries in large datasets of time sequences. In: Proceedings of SIGMOD 1997, Tucson, AZ, pp. 289–300 (1997)
9. Larsen, R.J., Marx, M.L.: An Introduction to Mathematical Statistics and Its Applications, 2nd edn. Prentice Hall, Englewood Cliffs (1986)
10. Megalooikonomou, V., Wang, Q., Li, G., Faloutsos, C.: Multiresolution Symbolic Representation of Time Series. In: proceedings of the 21st IEEE International Conference on Data Engineering (ICDE), Tokyo, Japan, April 5-9 (2005)
11. Morinaka, Y., Yoshikawa, M., Amagasa, T., Uemura, S.: The L-index: An indexing structure for efficient subsequence matching in time sequence databases. In: Cheung, D., Williams, G.J., Li, Q. (eds.) PAKDD 2001. LNCS (LNAI), vol. 2035, pp. 51–60. Springer, Heidelberg (2001)
12. Reinert, G., Schbath, S., Waterman, M.S.: Probabilistic and Statistical Properties of Words: An Overview. Journal of Computational Biology 7, 1–46 (2000)
13. Yi, B.K., Faloutsos, C.: Fast time sequence indexing for arbitrary  $L_p$  norms. In: Proceedings of the 26st International Conference on Very Large Databases, Cairo, Egypt (2000)
14. UCR Time Series datasets,  
[http://www.cs.ucr.edu/~eamonn/time\\_series\\_data/](http://www.cs.ucr.edu/~eamonn/time_series_data/)

# Which XML Storage for Knowledge and Ontology Systems?

Martin Bukatovič, Aleš Horák, and Adam Rambousek

Faculty of Informatics  
Masaryk University  
Botanická 68a, 602 00 Brno  
Czech Republic  
[{xbukatov,hales,xrambous}@fi.muni.cz](mailto:{xbukatov,hales,xrambous}@fi.muni.cz)

**Abstract.** New research concerning knowledge and ontology management systems in many cases need the versatility of native XML storage for manipulations with diverse and changing data structures. Within the DEB (Dictionary Editor and Browser) development platform, the efficiency of the background data storage for all kinds of structures and services including dictionaries, wordnet semantic networks, classical ontologies or lexical databases, tends to be a crucial property of the system.

In this paper, we describe a large set of tests that were run on four selected (out of twenty) available XML database systems, where the tests were run with the aim to recommend the best XML database for knowledge and ontology storage.

## 1 Introduction

The main advantages of storing data in the XML format is the data portability across systems and the versatility of the data structures – the storage systems, including query languages for manipulation, can handle arrays, hierarchies, texts, named substructures or links in one defined entity type with the possibility of automatic advanced schemata for syntactic checks of correctness. However, with such properties, the storage systems tend to be less efficient than standard relational databases, when it comes to processing large amounts of data, speaking of sizes in tens of megabytes or more [12].

After 5 years of development of the Dictionary Editor and Browser (DEB) platform that is designed to provide common useful features of dictionary writing systems, there are now more than ten actively used dictionary writing systems and lexicographic projects, which are based on the DEB platform. Two of them, DEBDict [3], general dictionary browser providing access to many dictionaries and lexical resources in several languages, and DEBVisDic [4], wordnet editor and browser used to build more than fifteen wordnets in different languages, are currently in use by more than 700 of registered users from all over the world. The freely available DEB server is currently installed in ten institutions from three continents, where it is used mostly as a XML-based data storage, presentation and manipulation system.

With the current deployment of the DEB platform, the current database storage is not able to efficiently process some kinds of search queries. Thus we have decided to analyze and compare available native XML database systems and provide a recommendation of the best performance for knowledge and ontology systems.

Database systems working with XML data (both native XML databases and XML enabled relational databases) are already widespread and used in many areas. Their performance was benchmarked by many projects using several benchmarks. In [5], a generally applicable benchmark XMach-1 is described and compared to other benchmarks. Results for several databases are presented, showing that native XML databases perform better than XML-enable relational databases. Unfortunately, no database is named, so the results are only general.

Nambiar et al. [6] use XOO7 benchmark to compare several XML-enabled and native XML databases. Their results suggest that XML enabled relational databases process data manipulation queries more efficiently. Native XML databases, on the other hand, are more efficient in navigational queries which rely on the document structure.

Extensive comparison experiments were conducted by Lu et al. [7]. Their results suggest that different XML benchmarks can show different weak and strong points of each database systems.

Differences in the results leads to the conclusion that customized XML benchmarks are needed in addition to a general XML benchmark to fully test the requirements of the application developed. For example for the business XML systems, Nicola, Kogan and Schiefer in [8] offer specialized benchmark, called “Transaction Processing over XML” (TPoX). This benchmark aims to provide good comparison of XML databases suitable for the business process modelling.

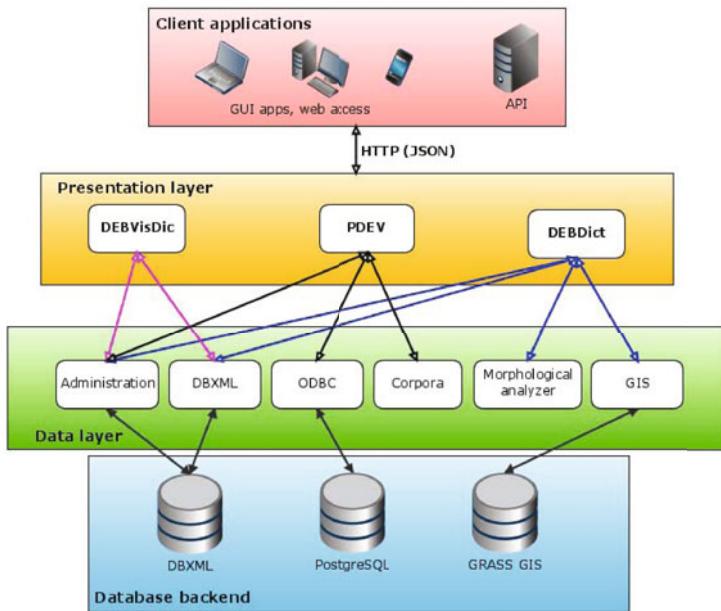
In the following sections, we present the results of comparing the XML enabled as well as native XML databases over data commonly used in dictionary writing systems.

## 2 The DEB Platform for Dictionary Writing Systems

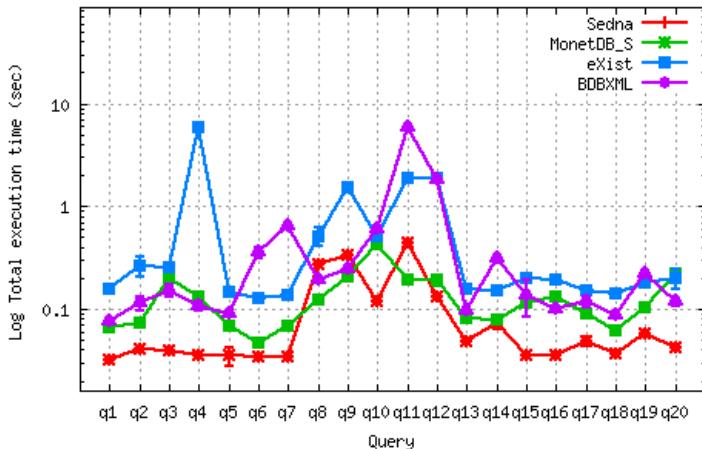
The DEB (Dictionary Editor and Browser, <http://deb.fi.muni.cz/>) is an open-source software platform for the development of applications for viewing, creating, editing and authoring of electronic and printed dictionaries. The platform is based on the client-server architecture (see the DEB platform schema in Figure 1). Most of the functionality is provided by the server side, and the client side offers (computationally simple) graphical interfaces to users. The client applications communicate with the server using the standard web HTTP protocol.

The server part is built from small, reusable parts, called servlets, which allow a modular composition of all services. Each servlet provides different functionality such as database access, dictionary search, morphological analysis or a connection to corpora.

The overall design of the DEB platform focusses on modularity. The data stored in a DEB server can use any kind of structural database and combine the



**Fig. 1.** The DEB platform schema



**Fig. 2.** Total execution time (in seconds) for a 1.8MB document

results in answers to user queries without the need to use specific query languages for each data source. The main data storage is currently provided by the Oracle Berkeley DB XML [1]. However, it is possible to switch to another database backend easily, without any changes to the client parts of the applications.

The main assets of the DEB development platform can be characterized by the following points:

- All the data are stored on the server and a considerable part of the functionality is also implemented on the server, while the client application can be very lightweight.
- Very good tools for team cooperation; data modifications are immediately seen by all the users. The server also provides authentication and authorization tools.
- Server may offer different interfaces using the same data structure. These interfaces can be reused by many client applications.
- Homogeneity of the data structure and presentation. If an administrator commits a change in the data presentation, this change will automatically appear in every instance of the client software.
- Integration with external applications.

### 3 Selected Databases

Although there are many native XML database, we have to select databases that correspond to the licence and technologies applied in the DEB platform. The most important features are the open source licence, active development and support of XML-related standards.

From more than 20 native XML or XML-enabled databases, we have chosen the following four systems according to the designated requirements.

#### 3.1 eXist

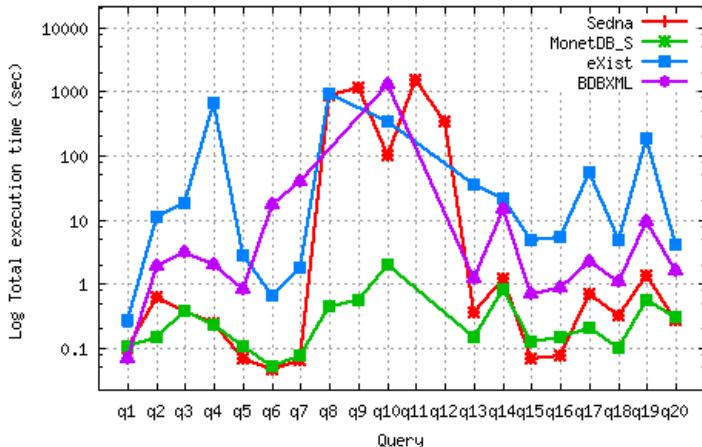
The eXist database [9] is developed in Java and licensed under LGPL, active since 2000 and currently developed by the group of independent developers. The database supports XQuery, XSLT and XUpdate standards for data manipulation, and DTD, XML Schema, RelaxNG and Schematron for validation.

Users are able to specify structural indexes (element and attribute structure in documents), range indexes (*contains*, *starts-with* and similar functions), and full-text indexes (Apache Lucene [10] is used for full-text indexing).

#### 3.2 MonetDB/XQuery

The MonetDB/XQuery database [11] is developed by CWI Amsterdam and several Linux distributions and MS Windows are officially supported. The database is licensed under a customised Mozilla Public License.

The main goal of MonetDB is to design a database for processing very large (in GBs) XML documents. The default database settings are optimized for document reading, offering indexing for quick query execution, although the indexes have to be rebuilt after every document update. Another option is an optimization for document updating, with simpler index structure and slower performance for search queries.



**Fig. 3.** Total execution time (in seconds) for a 114MB document

The database supports XQuery and partly XQuery Update [12]. It is also possible to use MonetDB internal query language. Indexing is automatized, without the possibility to alter settings in any way. The PF/Tijah [13] text search system is utilized for full-text searching.

### 3.3 Sedna

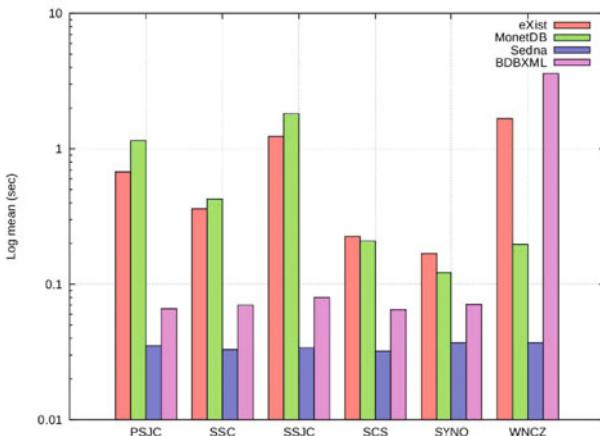
The Sedna database system [14] is developed by the Russian Academy of Sciences, and released under Apache Licence. Official packages for Windows, Linux, MacOS, FreeBSD and Solaris are available.

The database supports XQuery and custom variant of XQuery Update for data manipulation, and XML Schema for validation. Indexes have to be set manually and a special function must be used in the query to access the index. Full-text indexing is provided by external commercial tool dtSearch. Sedna offers several extensions, such as the capability of an SQL connection from XQuery, or the trigger support.

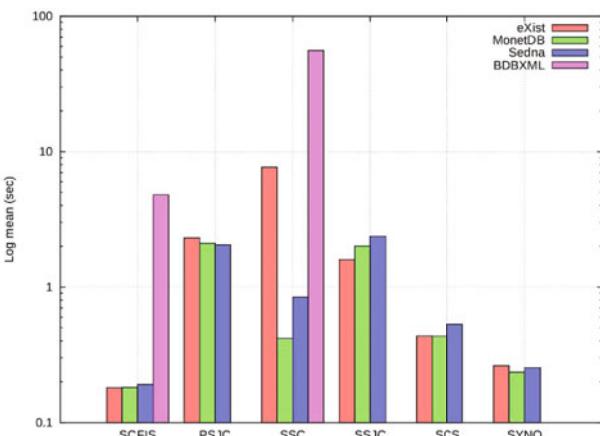
### 3.4 Oracle Berkeley DB XML

Berkeley DB XML [1] was created as an extension of Berkeley DB. Currently, the database is developed by Oracle and released for Windows and Linux. Users can choose between open source and commercial licence.

The underlying structure is still based on Berkeley DB and each document container is stored in a single file. The database supports XQuery and part of XQuery Update. The document validation according to a supplied XML Schema is checked only during document storage, later changes can render the document invalid. Users have to specify indexes manually, full-text indexing is also supported, although it is not possible to use regular expressions in queries.



**Fig. 4.** Average time (in seconds) for the *equality* query



**Fig. 5.** Average time (in seconds) for the *full-text* query

## 4 Database Comparison

Because of the special focus on dictionary writing systems, we have decided to run two different test suites. For the general database performance, we have evaluated and selected the XMark benchmark [15], and for the knowledge and ontology test, we have prepared a custom set of the most frequent queries and tasks.

In the tests, we have used the following database versions (preferring stable release over the development one): eXist 1.4.0, MonetDB/XQuery 2009-Aug-SP1, Sedna 3.2.91, and Oracle Berkeley DB XML 2.5.13.

#### 4.1 The XMark Tests

The XMark benchmark was developed in CWI Amsterdam with the aim to provide a benchmark suite for users and developers to choose the right XML database and to tune the database settings.

The benchmark includes the tool `xmlgen` to generate an XML document of a given size. The data and the structure are always the same, users are able to change just the document size. The test suite itself consists of 20 XQuery queries that model different operations with several collections of XML documents, ranging from simple search to complex linking and result generation.

We have tested the queries on documents of size from 1.8MB to 114MB. You can see the results for the smallest and the largest document in Figures 2 and 3. For the smallest document, all the queries were executed in less than a second, except for some queries in eXist and Berkeley DB XML. The problematic queries q11 and q12 are combining data from two collections and building very large result set. Although this is a complex task, it should not take so long on such a small document.

On the other hand, it is understandable in the case of the large document. While increasing the document size, the execution times are getting longer. For the 114MB document, much more queries are carried out in times above one second. MonetDB is providing the best results for large documents, and Sedna can be better for less complex queries on smaller documents.

Although the results of XMark tests can help users to pick the right database, real data and tasks should be taken into account, because the results vary significantly according to the document size and query complexity.

#### 4.2 Knowledge and Ontology Data Benchmark

Another step was to test the databases on real data and most frequent tasks of DEB applications. For the benchmark, the following lexicons and ontologies were used:

- The Dictionary of Literary Czech (SSJC), 180.000 entries,
- The Reference Dictionary of Literary Czech (PSJC), 200.000 entries,
- The Dictionary of Written Czech (SSC), 49.000 entries,
- The Dictionary of Words with Foreign Origin (SCS), 46.000 entries,
- The Dictionary of Czech Synonyms (SYNO), 23.000 entries,
- The Dictionary of Czech Phraseologisms and Idioms (SCFIS), 14.000 entries,
- The English WordNet (WNEN), 117.000 entries,
- The Czech WordNet (WNCZ), 28.000 entries.

We have analyzed the operations and have selected the most frequent query types as well as several queries requested by the users.

**Equality Query.** In the first run, XQuery was used to select entries with an element equal to a given value. In the second run, the query was rewritten as an XPath query. With this optimization, databases performed much better, significant improvement was seen for eXist and Berkeley DB XML. The results are shown in Figure 4.

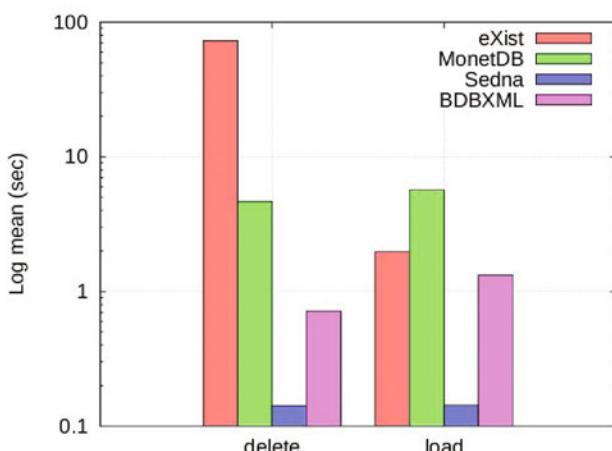
**Full-Text Search.** A more or less standard data set for full-text benchmarks is the INEX collection [16]. The current version of INEX collection 2009 contains 2.666.190 semantically annotated Wikipedia articles. The full-text search over the INEX database tests the database performance with a huge amount of data and complex-linked full-text structure. The tested databases have often problems with the kind of data structures used in INEX (e.g. Sedna was not able to build indexes for INEX at all, eXist did not return answers to many queries, MonetDB could not load the databases into 4 GB of memory). However, for the purpose of dictionary applications, the full-text search is usually used within short texts, such as definitions or examples. We thus offer the results of the comparison of full-text search over standard dictionary tags.

For the eXist database, the Lucene module was used for full-text search. We were unable to install PF/Tijah module on the testing server for MonetDB. And for Sedna, the commercial module was not tested. The results are shown in Figure 5.

Considering that full-text modules for MonetDB and Sedna were not used, it is surprising that these databases processed the queries in times comparable to eXist (sometimes even faster). Berkeley DB XML results are missing for most of the dictionaries, because several queries of the test suite were not completed in five minutes.

**Document Updates.** Because DEB platform applications are designed for editing the knowledge and ontology data, the documents are updated by teams of users. Another feature we needed to test is the performance during document deleting and saving. For this test, only the largest dictionary (PSJC) was used. The tests were run several times and each time five random documents were deleted and then saved again.

The average results are shown in Figure 6. Surprisingly, differences between the databases are quite significant.



**Fig. 6.** Average time (in seconds) for document update

## 5 Evaluation

According to the results of the tests, none of the available native XML databases can supersede the others for all kinds of operations needed for knowledge and ontology storage and manipulation. Berkeley DB XML cannot efficiently solve the queries involving multiple nodes and full-text queries. The eXist database contains the Lucene module for text search and supports many XML standards, so it can be recommended for deployment where these features are more important than the database performance. On the other hand the MonetDB database can be, according to its specific architecture, conveniently used for when working with very large amounts of XML data. For middle-size data collections, the Sedna database can provide the same performance as MonetDB, while offering richer set of features. The potential drawbacks of Sedna are the need to use special queries for the defined data indexes and the use of commercial tool for optimized full-text queries.<sup>1</sup>

## 6 Conclusions

Considering the results of XMark and the custom knowledge and ontology benchmark, the MonetDB/XQuery and the Sedna databases represent a good choice for the knowledge and ontology systems. MonetDB offers very good performance for very large documents, on the other hand, Sedna provides much more advanced features. Unfortunately, Sedna supports index usage only with its own special functions, so the queries need to be changed accordingly.

As a next step, both MonetDB and Sedna will be included in the DEB platform and compared in real operation.

**Acknowledgements.** This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536 and in the National Research Programme II project 2C06009 and by the Czech Science Foundation under the projects P401/10/0792 and 102/09/1842.

## References

1. Chaudhri, A.B., Rashid, A., Zicari, R. (eds.): XML Data Management: Native XML and XML-Enabled Database Systems. Addison Wesley Professional, Reading (2003)
2. Krishnamurthy, R., Kaushik, R., Naughton, J.: XML-to-SQL query translation literature: The state of the art and open problems. In: Bellahsène, Z., Chaudhri, A.B., Rahm, E., Rys, M., Unland, R. (eds.) XSym 2003. LNCS, vol. 2824, pp. 1–18. Springer, Heidelberg (2003)
3. Horák, A., Pala, K., Rambousek, A., Rychlý, P.: New clients for dictionary writing on the DEB platform. In: DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writings Systems, Italy, pp. 17–23. Lexical Computing Ltd, U.K (2006)

---

<sup>1</sup> However, the full-text queries without this optimization are already comparably fast.

4. Horák, A., Pala, K., Rambousek, A., Povolný, M.: First version of new client-server wordnet browsing and editing tool. In: Proceedings of the Third International WordNet Conference - GWC 2006, Jeju, South Korea, Masaryk University, Brno, pp. 325–328 (2006)
5. Böhme, T., Rahm, E.: Multi-user evaluation of XML data management systems with XMach-1. Efficiency and Effectiveness of XML Tools and Techniques and Data Integration over the Web, 148–159 (2008)
6. Nambiar, U., Lacroix, Z., Bressan, S., Lee, M., Li, Y.: Efficient XML data management: an analysis. E-Commerce and Web Technologies, 261–266 (2002)
7. Lu, H., Yu, J., Wang, G., Zheng, S., Jiang, H., Yu, G., Zhou, A.: What makes the differences: benchmarking XML database implementations. ACM Transactions on Internet Technology (TOIT) 5(1), 154–194 (2005)
8. Nicola, M., Kogan, I., Schiefer, B.: An XML transaction processing benchmark. In: Proceedings of the 2007 ACM SIGMOD international conference on Management of data, pp. 937–948. ACM, New York (2007)
9. Meier, W., et al.: eXist: An open source native XML database. LNCS, pp. 169–183. Springer, Heidelberg (2003)
10. Foundation, A.S.: Apache Lucene (2006), <http://lucene.apache.org>
11. Boncz, P., Grust, T., van Keulen, M., Manegold, S., Rittinger, J., Teubner, J.: MonetDB/XQuery: a fast XQuery processor powered by a relational engine. In: Proceedings of the 2006 ACM SIGMOD international conference on Management of data, p. 490. ACM, New York (2006)
12. W3C: XQuery Update Facility 1.0 (2009),  
<http://www.w3.org/TR/xquery-update-10>
13. Hiemstra, D., Rode, H., van Os, R., Flokstra, J.: PF/Tijah: text search in an XML database system. In: Proceedings of the 2nd International Workshop on Open Source Information Retrieval (OSIR), pp. 12–17 (2006)
14. Fomichev, A., Grinev, M., Kuznetsov, S.: Sedna: A Native XML DBMS. In: Wiedermann, J., Tel, G., Pokorný, J., Bieliková, M., Štuller, J. (eds.) SOFSEM 2006. LNCS, vol. 3831, p. 272. Springer, Heidelberg (2006)
15. CWI: XMark – An XML Benchmark Project (2009),  
<http://www.xml-benchmark.org>
16. Schenkel, R., Suchanek, F., Kasneci, G.: YAWN: A semantically annotated Wikipedia XML corpus. In: Datenbanksysteme in Business, Technologie und Web (BTW 2007), Aachen, Germany, Verlagshaus Mainz, pp. 277–291 (2007)

# Finding Temporal Patterns Using Constraints on (Partial) Absence, Presence and Duration

S. Peter and F. Höppner

Ostfalia University of Applied Sciences  
Robert Koch Platz 10-14, D-38440 Wolfsburg

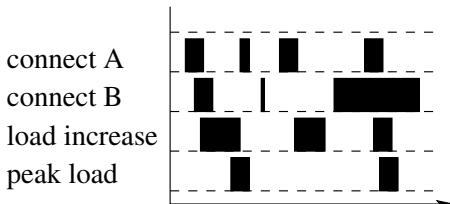
**Abstract.** When the evolution of variables over time is relevant to a classification task, established classifiers cannot be applied directly as the typical input format (data table) is not appropriate. We propose a new representation of temporal patterns that includes constraints on (partial) presence, (partial) absence as well as the duration of temporal predicates. A general-to-specific search-based algorithm is presented to derive classification rules. The approach evaluates promising on artificial and real data.

## 1 Introduction

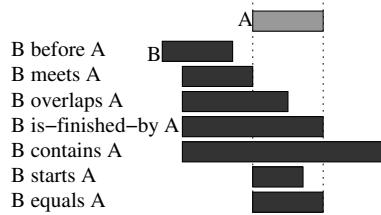
One important aspect of data mining is to identify dependencies and interrelationships that were unknown to the user before and deliver them in an easily understandable representation. Rule-based systems and decision trees, which fulfill such requirements, assume a databases of cases with characterizing features that held at the time of recording the case. In many areas, however, a *case* stretches over time, such as the traffic density over one day, the medication of a patient over the duration of illness, control variables of a production process over a production cycle, a workflow of a business process, etc. In such domains it is often not feasible to drop the temporal information, sometimes even the order of isolated events is not sufficient, but their temporal extent and contemporaneity is important. The discovery of complex relationships of the latter kind without compromising their interpretability requires a descriptive, graphical representation of the discovered patterns. In this paper, we propose a new temporal pattern representation and suggest an algorithm to learn classification rules automatically from temporal data.

## 2 Representing Evolving Data

We consider the history of a binary attribute as a temporal predicate. Denoting the temporal dimension by  $\mathbb{T}$ , a temporal predicate  $P_l$  is a function  $P : \mathbb{T} \rightarrow \mathbb{B}$ .  $l$  is called the label of the predicate  $P$ . We assume that all available data may be represented by such temporal predicates. The development of a nominal attribute with domain  $\{u, v, w\}$  may be represented by three predicates  $P_u$ ,  $P_v$  and  $P_w$ , denoting when each of the values held. A series of numerical values (time



**Fig. 1.** Representation of Evolving Data: the black rectangles denote the intervals when the predicate (labels to the left) holds



**Fig. 2.** Thirteen possible relationships between two intervals. The inverse relationships (after  $\leftrightarrow$  before) have been omitted.

series) can be represented by extracting different predicates such as  $P_{\text{increasing}}$  or  $P_{\text{high-valued}}$ . A set of such predicates (which we will call history  $H$ ) is often depicted by listing the various predicates against the temporal dimension (cf. Fig. ①). A suitable data representation is a (labelled) sequence of temporal intervals that indicate when the predicate did hold. Such a representation is used frequently, e.g. in the medical domain [7].

In the context of classification tasks the goal is to describe circumstances (prototypical histories) under which a certain target variable is likely to occur. Note that in contrast to stream mining approaches, where a single but potentially infinite stream of data is considered, we assume that multiple finite, labelled histories are available. Various ways to define such patterns in a stream of labeled intervals have been proposed in the literature, many of them relying on Allen's interval relationships [1] (cf. Fig. ②) or variants thereof. Some approaches define a history by specifying the exact relationship for every pair of intervals [3], others allow for a set of possible relationships [4]. The representation by sequences of chords [6] uses a partially ordered sequence of simultaneous (sub-) intervals to define a pattern. Some other proposals consider a different set of interval relationships or specify the next relationship only with respect to the union of the pattern discovered so far [5].

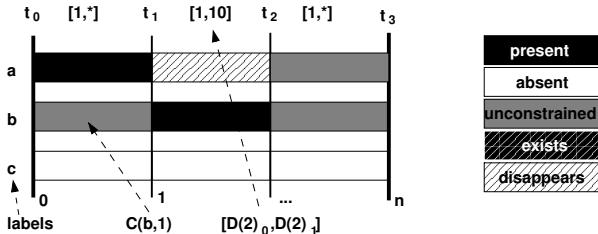
While these approaches have their individual strengths, they also have their weaknesses when it comes to represent certain simple situations. Thinking of predicting a certain state of some network server (breakdown, overload, attack, malfunction, etc.) on the history of, say, the last 24 hours, a situation as simple as “there was only one connection to server A” (during the last  $n$  hours) is usually impossible to discover for the approaches based on association rules [3,6], as they count occurrences of events only and rely on a quickly decreasing count of co-occurrences, such that an inclusion of *absent features* during counting undermines the initial assumptions of association mining. A situation like “there was a connection to B while the connection to A was lost” is impossible to represent for the approaches that rely on explicitly given interval relationship, as the exact position of B relative to A is not known [3]. Temporal constraints “the connection to A was lost for at least 4 hours” or “... at most 4 hours” are usually ignored completely or introduced in a postprocessing step.

We propose a new notion of a pattern, called *template history*, that shall be matched against an existing history later. To maximally support the understandability of the template, we keep the basic representation of Fig. 1, but relax the temporal alignment to allow for successful matching despite of dilational and translational effects. We do not care about a 1:1 mapping of time points (as in dynamic time warping) but concentrate on a few relevant points in time that are indicated by vertical *alignment lines*. This corresponds roughly to a discretization of the temporal axis, but we do not finally fix the position of the lines but adjust them at match-time. If  $m$  labels/predicates and  $n + 1$  adjustment lines are given, we obtain an  $m \times n$  matrix which allows us to pose different conditions on the predicates in each of the matrix cells. We distinguish four different conditions:

**Definition 1 (predicate constraint).** *Given a temporal interval  $T \subseteq \mathbb{T}$  and a predicate  $P$ , we say (a)  $P$  is present during  $T$  if  $\forall t \in T : P(t)$ , (b)  $P$  is absent during  $T$  if  $\forall t \in T : \neg P(t)$ , (c)  $P$  exists during  $T$  if  $\exists t \in T : P(t)$  and (d)  $P$  disappears during  $T$  if  $\exists t \in T : \neg P(t)$ . If no condition is posed, we say  $P$  is unconstrained during  $T$ . By  $\mathbb{C}$  we denote the set of constraints { present, absent, exists, disappears, unconstrained }.*

Every cell of the template matrix is now filled with one of the five constraints. Additionally a template may contain temporal constraints on the distance between the alignment lines (block duration).

**Definition 2 (template).** *A tuple  $T = (L, n, C, D)$  is called a template if  $L$  is a set of labels,  $n \in \mathbb{N}$ ,  $C : L \times \{1, \dots, n\} \rightarrow \mathbb{C}$  and  $D : \{1, \dots, n\} \rightarrow (\mathbb{T} \cup \{\infty\})^2$  with  $1 \leq D(i)_0 \leq D(i)_1$  for all  $1 \leq i \leq n$ .*



**Fig. 3.** Illustration of the template definition. The map  $D$  defines the block durations (time interval between alignment lines) and is shown on the top. The predicate constraints are coded by color.

Figure 3 shows an example template with  $n = 3$  blocks and thus four vertical alignment lines, where the leftmost and rightmost alignment line shall always represent the start and end of the history. The bottom row declares that a predicate  $P_c$  is absent in the whole history (2nd block),  $P_b$  is present ( $P_b$  may be present or not elsewhere).  $P_a$  is present from the very

beginning, but disappears while  $P_c$  is present in the 2nd block. The duration of the first block is arbitrary, the duration of the second block lies within  $[1, 10] = [D(2)_0, D(2)_1]$  time units, the last block may again have any (positive) duration.

Matching a template to a real history involves two steps: Firstly, the alignment lines need to be positioned appropriately such that, secondly, all temporal and predicate constraints hold.

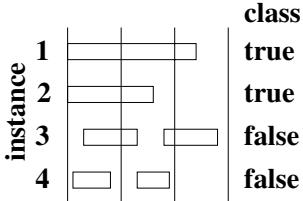
**Definition 3 (match).** Let  $T = (L, n, C, D)$  be a template and  $H$  be a history. Let  $[t_{\min}, t_{\max}]$  be the smallest interval subsuming  $\cup_{P \in H} \text{dom}(P)$ .  $T$  matches a history  $H$  if and only if (a) there is a predicate  $P_l \in H$  for every  $l \in L$ , (b) there are  $t_i \in \mathbb{T}$ ,  $0 \leq i \leq n$ , with  $t_0 = t_{\min}$ ,  $t_i \leq t_{i+1}$ ,  $t_n = t_{\max}$ , (c) for every  $l \in L$  and  $i \in \{1, \dots, n\}$  the constraint  $C(l, i)$  holds for  $P_l$  within  $[t_i, t_{i+1})$  and finally (d) for all  $1 \leq i \leq n$ :  $t_i - t_{i-1} \in \Delta_i$  with  $\Delta_i = [D(i)_0, D(i)_1]$ .

### 3 Finding Patterns

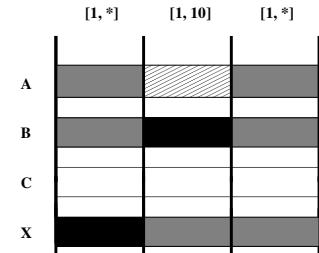
Next, we propose a method to explore the space of templates to discriminate differently labelled histories. The search algorithm implements a general-to-specific search: It begins with a pattern, which matches all instances, and tries to specialize it further to improve some measure of interestingness (we use the J-measure [8] as it balances the generality (applicability of the rule) and the interestingness (deviation from a priori knowledge)). An initial template that matches all histories must have at least one block, all matrix constraints are *unconstrained* and so are the temporal constraints  $[1, \infty]$ . While a propositional rule can only be specialized by an additional condition (like *outlook=sunny*), there are at least three ways to specialize a template: we may look at it in a finer resolution (by adding another alignment line), we may change or add a predicate constraint (for some label and block), or may introduce or change an existing temporal constraint. We thus have chosen three different specialization operators to address each of these aspects.

The general idea for all refinement operators is to search for specializations that improve the measure of interestingness, which basically requires that the specialized pattern still matches the positive instances but less negatives.

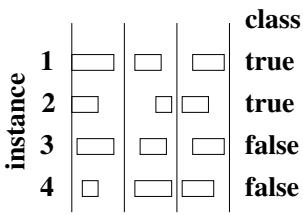
**Adding a predicate constraint.** Suppose we want to specialize the pattern in Fig. 3 by an additional constraint for label  $X$  (which is *unconstrained* so far). If four instances are given that match this template, we have to inspect the occurrences of  $X$  relative to the alignment lines of the template. This is shown (only for variable  $X$  and the four cases) in Fig. 4. We create confusion matrices for all possible refinements (each block and constraint combination) of the predicate  $X$  by counting how the instances will be classified by the considered specialization. For instance, the specialization shown in Fig. 5 would perfectly discriminate between the classes, as  $P_X$  is present in the first block only for the positive examples. Finally we apply the measure of interestingness on all confusion matrices to find the best refinement for  $X$ . We instantiate such a constraint refinement for every label to pick the best constraint specialization.



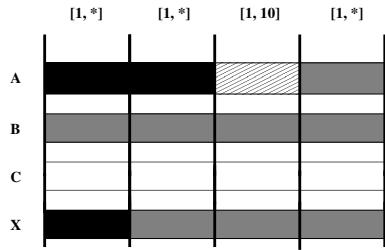
**Fig. 4.** Relative occurrence of the predicate  $X$  to the matches of the pattern shown in Fig. 3 in the instances



**Fig. 5.** Refined pattern for classify the instances correctly



**Fig. 6.** Relative occurrence of the predicate  $X$  to the matches of the pattern shown in Fig. 3 in the instances



**Fig. 7.** Refined pattern for classify the instances correctly

**Adding an alignment line.** The second operator adds new alignment lines to the template. Similar to Fig. 4, a different constellation of cases is shown in Fig. 6. Apparently an *exists*-constraint would match the positive classes (in any of the blocks), but unfortunately, it would also match the negative cases. However, we notice that all the positive cases start at the beginning of the first block, whereas the negative cases do not. By introducing a new alignment line that subdivides the first block, we may pose a *present*-constraint on the left part of the first block. The original template from Fig. 3 is extended by a new vertical line as shown in Fig. 7. As before, we again determine confusion matrices for different specializations and evaluate them for each block and combination of *absent/present*-constraints.

**Adding a temporal constraint.** Finally, the third operator tries to find a block length for a specific block, so that the new block length mostly holds by the positive examples but does not hold by the negative ones. Again, the best specialization is determined by the interestingness measure on the respective confusion matrix.

The search algorithm itself is relatively simple and based on a beam-search. It begins with a pattern which matches all instances. In every iteration the  $k$

best-evaluated patterns are further refined by the previously introduced operators. At the end of the iteration, only the  $k$  best specializations are preserved for the next loop. The search ends when no more improvements during the last  $p$  iterations were made. We do not stop immediately if the present iteration did improve the best rule ( $p = 1$ ) to avoid getting trapped in local minima due to the greedy nature of the approach. The best  $k$  templates are reported to the user.

The number of possible patterns grows quickly with the number of alignment lines and predicates, but the number of actually explored patterns is limited by the size of the beam. For each main iteration, the patterns in the beam are extended by the three presented operators. The necessary statistics to find the best specialization can be constructed in  $O(n \cdot m^2)$  time where  $n$  is the number of cases and  $m$  is the number of distinct durations considered as a temporal constraint. Rather than considering every possible duration constraint, the observed durations may be discretized beforehand, thereby limiting the number of choices and the overall runtime.

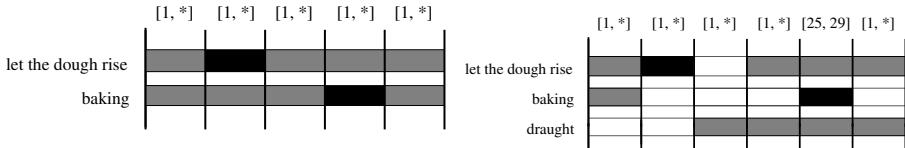
## 4 Experimental Evaluation

**Artificial example.** In order to evaluate the new representation we generated a synthetic dataset on the basis of a pizza recipe. In general the process of making a pizza consists of the following four steps: First we have to mix the ingredients to make the dough. Then we let the dough rise for 60 to 120 minutes in a place without (air) draught. Afterwards we roll out the dough and add the toppings. Finally we have to bake the pizza 25 to 29 minutes.

So the dataset consists of the labels: 'make dough', 'let the dough rise', 'roll out & coat the dough', 'draught' and 'baking'. Generally each instance is generated as follows: First a 'make dough' interval (5-20 minutes), followed by 'let the dough rise' (60-120 minutes), after 5-30 minutes the 'roll out & coat the dough' interval is present (5-10 minutes) and finally after 0-10 minutes there is a 'baking' interval (25-30 minutes). Furthermore it consists of the three classes: perfect pizza (the baking process fits the recipe), pizza burned (the baking process fits the recipe except the baking time is greater than 30 minutes) and dough not risen (the baking process fits the recipe except 'dough rise' is missing or during the time the dough rises there is a draught interval). The interval lengths were chosen randomly and depending on their fit into the abovementioned intervals the class label was selected appropriately.

The training set consists of 99 instances (33 per class) and the test set of 990 instances (330 per class). To compare the results of the new approach with the results of the old approaches we applied the algorithm described in Sect. 3 as well as a reduced version without time-, absent-, exists- and disappears-constraints to simulate the setting of previous approaches.

The best pattern found by the old approach for perfect pizza is shown in Fig. 8. As we can see it requires 'let the dough rise' followed by 'baking' but the pattern neither carries any information on how long the pizza should be



**Fig. 8.** Pattern for perfect pizza found by **Fig. 9.** Pattern for perfect pizza found by the limited pattern language (comparable new approach to earlier approaches)

backed to get not burned, nor is there a restriction of absent 'draught' during the proving process. The new approach found the pattern shown in Fig. 9, which requires that during the whole time the dough is rising (first two blocks, ends in third block due to absence constraint) there is no draught. Furthermore this pattern demands a baking time between 25 and 29 minutes due to the block 4 to 6, where blocks 4 and 6 forbid the presence of baking and the fifth block requires baking with a duration of 25-29 minutes. Similar patterns were derived for the other classes, too. Figure 10 summarizes how the patterns evaluated on the test set. As we can see the new approach performed considerably better than the old approach, mainly because the old approach lacks means to identify those cases where certain steps of the process were intentionally or incidentally absent.

$$\begin{aligned}
 \text{old approach: } & \begin{pmatrix} P & \neg P \\ \text{match} & 330 \ 490 \\ \neg\text{match} & 0 \ 170 \end{pmatrix} \begin{pmatrix} DnR & \neg DnR \\ \text{match} & 160 \ 0 \\ \neg\text{match} & 170 \ 660 \end{pmatrix} \begin{pmatrix} B & \neg B \\ \text{match} & 12 \ 170 \\ \neg\text{match} & 318 \ 490 \end{pmatrix} \\
 \text{new approach: } & \begin{pmatrix} P & \neg P \\ \text{match} & 330 \ 0 \\ \neg\text{match} & 0 \ 660 \end{pmatrix} \begin{pmatrix} DnR & \neg DnR \\ \text{match} & 0 \ 628 \\ \neg\text{match} & 330 \ 32 \end{pmatrix} \begin{pmatrix} B & \neg B \\ \text{match} & 330 \ 0 \\ \neg\text{match} & 0 \ 660 \end{pmatrix}
 \end{aligned}$$

**Fig. 10.** Confusion matrices for the best patterns found by old and new approach. P: perfect pizza, DnR: dough not rised, B: burned pizza.

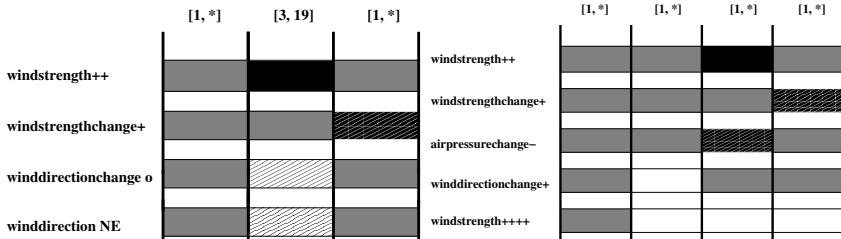
**Real data.** Furthermore we applied our approach to weather data collected by a weather station located on a small island in the northern sea (Helgoland). This station collected air-pressure, wind strength and wind direction hourly for several years. At first we had to preprocess these timeseries to a stream of intervals, where we used the following labels:

- air-pressure: very low (–), low (–), middle (o), high (+), very high (++).
- air-pressure slope and wind strength change: highly decreasing (–), decreasing (–), normal (o), increasing (+), highly increasing (++).
- wind strength: very low (+), low (++)+, high (+++), very high (++++)
- wind direction: N, NE, E, SE, S, SW, W, NW.

We tried to predict the occurrence of a strong wind (wind strength: high) and extracted all intervals within 72 hours before strong winds occurred as positive

examples. Negative examples (no strong winds to come) were extracted randomly from other time points. We do not expect to find new knowledge from the data, as it is already known that the actual value of the air pressure is irrelevant, but the change in the air pressure is a good indicator for strong winds, but we are again interested in finding evidence that the *new representation* is useful with real world scenarios.

An obvious pattern that simply requires quite strong winds followed by an increasing trend in wind strength leads to a rule with a J-Measure of 0.0583. Taking the newly introduced constraint types into account, this basic rule can be extended to increase the J-measure by more than 30%. Two of the patterns are shown in Fig. 11. We see that the ability to use the newly introduced constraints allows significant improvements also in real datasets. A much higher J-value is hard to obtain, because the J-measure is limited by the (relatively low) frequency of strong winds.



**Fig. 11.** Two different patterns found to predict an upcoming strong wind

Furthermore we applied our algorithm to the libras movement data set from the UCI repository [2]. It contains 15 different signs described by their characteristic hand movement over 45 frames, where the current x- and y-positions of the hand were recorded. There are 24 instances per sign, 360 in total. In the first step we have preprocessed the data in order to extract predicates that represent the hand movement. The extracted features address the speed of the hand movement in the x- and y-direction. Although one can easily think of more sophisticated features (rotation, absolute position, ...) and threshold extraction, we used a priori defined thresholds (quantiles) and the following labels only:

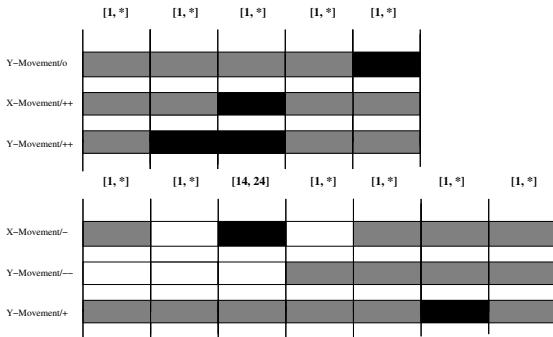
- X-Movement: fast left (--) , left (-), constant (o), right (+), fast right (++).
- Y-Movement: fast down (--) , down (-), constant (o), up (+), fast up (++).

For example, a fast hand movement to the upper left is recognized if predicates x-movement-- and y-movement++ hold at the same time.

For every sign we applied the algorithm with and without the newly introduced constraints. Figure 12 shows the the confusion matrices for the best pattern found for each of the two approaches.

As we can see the patterns with the newly introduced constraints perform considerable better, the number of false positives and false negatives have decreased. On average we have an increase of 3.6 true positives per pattern which

sign	id	old approach	new approach
	1	$\begin{pmatrix} 1 & -1 \\ \text{match } 22 & 3 \\ \neg\text{match } 2 & 333 \end{pmatrix}$	$\begin{pmatrix} 1 & -1 \\ \text{match } 23 & 0 \\ \neg\text{match } 1 & 336 \end{pmatrix}$
	7	$\begin{pmatrix} 7 & -7 \\ \text{match } 12 & 0 \\ \neg\text{match } 12 & 336 \end{pmatrix}$	$\begin{pmatrix} 7 & -7 \\ \text{match } 22 & 0 \\ \neg\text{match } 2 & 336 \end{pmatrix}$
	9	$\begin{pmatrix} 9 & -9 \\ \text{match } 0 & 256 \\ \neg\text{match } 24 & 80 \end{pmatrix}$	$\begin{pmatrix} P & \neg P \\ \text{match } 21 & 0 \\ \neg\text{match } 3 & 336 \end{pmatrix}$
	11	$\begin{pmatrix} 11 & -11 \\ \text{match } 16 & 6 \\ \neg\text{match } 8 & 330 \end{pmatrix}$	$\begin{pmatrix} 11 & -11 \\ \text{match } 19 & 0 \\ \neg\text{match } 5 & 360 \end{pmatrix}$
	12	$\begin{pmatrix} 12 & -12 \\ \text{match } 12 & 0 \\ \neg\text{match } 12 & 336 \end{pmatrix}$	$\begin{pmatrix} 12 & -12 \\ \text{match } 23 & 0 \\ \neg\text{match } 1 & 336 \end{pmatrix}$
	15	$\begin{pmatrix} 15 & -15 \\ \text{match } 5 & 270 \\ \neg\text{match } 19 & 66 \end{pmatrix}$	$\begin{pmatrix} 15 & -15 \\ \text{match } 15 & 0 \\ \neg\text{match } 9 & 336 \end{pmatrix}$

**Fig. 12.** Effect of newly introduced constraints on libras data [2]**Fig. 13.** The two best patterns found by the different approaches to describe sign four (top: old, bottom: new approach)**Fig. 14.** Instance of sign #4, plotted by connecting all hand positions

equals an increase of 15 percent. For the sign four, which represents a half circle swing (see Fig. 14), we show the identified patterns in Fig. 13. The first pattern was found by the old approach and tries to separate the instances by requiring certain movements – that may occur in other signs also. The second pattern, however, requires a movement to the left over a relatively long period of time (14-24 frames), during which no fast downward movement takes place. Afterwards, an upward movement is required, but the pattern contains no further information/constraints about the last frames, as they are not that helpful for

discriminating the signs. The key to the increased performance is the possibility of excluding certain co-occurring predicates.

## 5 Conclusions

We have proposed a new representation to define templates of evolving variables for classification tasks that overcome deficiencies of previous approaches. A general-to-specific approach to find useful templates has been presented. First experiments were promising and indicated the practical usefulness of the representation. The identified templates may also be used for feature generation when other classifiers shall be applied. Future work includes the application of the approach to business workflows.

## References

1. Allen, J.F.: Maintaining knowledge about temporal intervals. *Communications of the ACM* 26(11), 832–843 (1983)
2. Asuncion, A., Newman, D.: UCI machine learning repository (2007),  
<http://www.ics.uci.edu/~mlearn/MLRepository.html>
3. Höppner, F., Klawonn, F.: Finding informative rules in interval sequences. *Intelligent Data Analysis – An International Journal* 6(3), 237–256 (2002)
4. Höppner, F., Topp, A.: Classification based on the trace of variables over time. In: Yin, H., Tino, P., Corchado, E., Byrne, W., Yao, X. (eds.) IDEAL 2007. LNCS, vol. 4881, pp. 739–749. Springer, Heidelberg (2007)
5. Kam, P.-S., Fu, A.W.-C.: Discovering temporal patterns for interval-based events. In: Kambayashi, Y., Mohania, M., Tjoa, A.M. (eds.) DaWaK 2000. LNCS, vol. 1874, pp. 317–326. Springer, Heidelberg (2000)
6. Mörchen, F.: Time Series Knowledge Mining. PhD thesis, Philipps University Marburg (2006)
7. Shahar, Y., Musen, M.A.: RÉSUMÉ: A temporal abstraction system for patient monitoring. *Computers and Biomedical Research* 26, 155–273 (1993)
8. Smyth, P., Goodman, R.M.: An information theoretic approach to rule induction from databases. *IEEE Trans. Knowledge Discovery and Engineering* 4(4), 301–316 (1992)

# Clustering Based on Kolmogorov Information

Fouchal Said, Ahat Murat, Lavallée Ivan, Bui Marc, and Benamor Sofiane

Laboratoire d'Informatique et des Systèmes Complexes,

41, rue Gay Lussac, 75005 Paris

&

CNRS UMI ESS 3189

UCAD Dakar BP 5005

{said.fouchal,murat.ahat}@laisc.net,

ivan.lavallee@gmail.com,

{marc.bui,sofiane.benamor}@laisc.net

<http://laisc.net/>

**Abstract.** In this paper we show how to reduce the computational cost of *Clustering by Compression*, proposed by Cilibrasi & Vitányi, from  $O(n^4)$  to  $O(n^2)$ . To that end, we adopte the Weighted Paired Group Method using Averages (WPGMA) method to the same similarity measure, based on compression, used in *Clustering by Compression*. Consequently, our proposed approach has easily classified thousands of data, where Cilibrasi & Vitányi proposed algorithm shows its limits just for a hundred objects. We give also results of experiments.

**Keywords:** Kolmogorov complexity, Information theory, Clustering by compression, Classification.

## 1 Introduction

The goal of clustering is to organize objects into groups whose members are similar according, most often, to some proximity criteria defined by introducing distances [4].

The question that arises in this context is: which proximity criterion has more sense to form more homogenous groups than others?

There exists several proximity criteria, eg. euclidean, manhattan, ... etc. our work is focused on *Normalized Information Distance (NID)*, introduced by Cilibrasi & Vitányi, [2] [5] it is a distance based on Kolmogorov complexity (compression). We choose *NID* measure because it is based on “universal” information of Kolmogorov, it doesn’t need any background about data and can without changes be applied to different sorts of data [5].

*Normalized Information Distance* calculates quantitative proximity between two objects, it can be significant in different domain, notably in phylogenetic, music, ... [3] [5], but this distance is not made at its most. It is used with Hill climbing

algorithm in order to provide clusters. The clustering based on Hill climbing algorithm should be stopped by user in order to obtain clusters, therefore the results depend on user, they are not reliable. Additionally, its computational cost is  $O(n^4)$ , it is limited to clusters of only hundred objects [3] [5].

Our contribution consists in proposing a faster clustering algorithm which stopped automatically after finding all clusters, using the same distance, which allows to treat thousands of data, just in few hours. Our proposed clustering algorithm is the Weighted Paired Group Method using Averages [11], whose complexity is  $O(n^2)$  [1] [11] [16] [19] [20].

This paper is organized as following : In section 2 we introduce Kolmogorov information and its usage in clustering, specifically the *Clustering by Compression*. Our contribution and results are presented in section 3. Finally, in section 4 we give our conclusions.

## 2 Notations and Definitions

**Definition 1. Kolmogorov Complexity  $K(x)$**  : *Kolmogorov Complexity or descriptional complexity (also random complexity), is the size of the smallest universal calculator program which fully describes an object. Descriptional complexity defines the absolute information content of an object* [6] |  $x$  | [9] [13] [19].

*Remark 1.* The principal Kolmogorov Complexity property used in this article is universality. The Kolmogorov Complexity of an object depends on intrinsic information in an object [5] [13] [18].

**Definition 2. Conditional Kolmogorov Complexity  $K(x|y)$**  : *Kolmogorov Complexity  $K(s)$  provides the absolute information content of an object. Whereas, the Conditional  $K(s)$  Complexity deals with the common absolute information content between two different objects  $x$  and  $y$ , it is noted as  $K(x|y)$*  [13].

Kolmogorov Complexity definition is very simple to enunciate, but it is very complicated to acquire clear details about its real value, for any binary sequence. Kolmogorov Complexity is a non calculable function [8] [13] [18].

Kolmogorov Complexity is a theoretical object. Indeed, considering a word, it is impossible to define its Kolmogorov Complexity in reality.

To approach Kolmogorov Complexity in practice, we use a lossless compression algorithms. A such compression algorithm has the property to give a unique description of an object, the size of this description is less than or equal to original size.

**Definition 3.** *Data compression is the action used in order to reduce the physical size of an information block. Data compression is based on similarity research in the form and the pattern, in order to describe objects while removing duplications.*

---

<sup>1</sup> We mean by object here a binary sequence.

## 2.1 Clustering by Compression, Similarity Measure

We introduce here *Clustering by Compression*, it is an unsupervised clustering method, based on practical descriptional complexity (i.e. data compression), it is composed of two parts. The first part is calculating mutual proximities between all objects. The second part consist in creating, from a set of objects, an undefined number of clusters.

We define first a distance that measures similarity between two objects in a universal manner. We use Conditional Kolmogorov Complexity for this [3] [5]. The similarity measure<sup>2</sup> between two binary sequences  $A$  and  $B$  is defined by the shorter program to transform  $A$  to  $B$  and  $B$  to  $A$  ;

$$d(A, B) = \frac{\max(K(A|B), K(B|A))}{\max(K(A), K(B))}$$

Where,  $K(A|B)$  is the contained information in  $A$  related to  $B$ .

Since there is no algorithm to compute Kolmogorov Complexity, as seen above, the data compression is used to approach it. Lets consider a normal compressor  $C$  with the following properties:

- *idempotency* :  $C(x) = C(xx)$  ;
- *monotonicity* :  $C(xy) \geq C(x)$ ;
- *distributivity* :  $C(xy) + C(z) \leq C(xz) + C(yz)$ ;
- *symmetry* :  $C(xy) = C(yx)$ ;
- *subadditivity* :  $C(x | y) \leq C(x) + C(y)$ .

The similarity measure based on normal compressor  $C(A)$  is defined as following:  
If  $C(B) \leq C(A)$ , the distance between  $A$  and  $B$  equals :

$$d(A, B) = 1 - \left[ \frac{(c(A)+c(B))-c(AB)}{c(A)} \right];$$

If  $C(A) \leq C(B)$ , the distance equals :

$$d(A, B) = 1 - \left[ \frac{(c(A)+c(B))-c(AB)}{c(B)} \right].$$

$C(A)$  and  $C(B)$  are normalization coefficients, they intervene only when  $A$  and  $B$  have a different size. Clearly, we have  $d(A, B) \in [0, 1]$ .

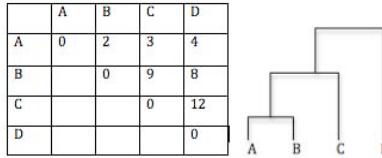
*Remark 2.* : The more common information content  $A$  and  $B$  have, the more they are close, hence the smaller  $d(A, B)$  is.

**Similarity Matrix.** Calculation of similarity measures, between objects, provides matrix of mutual distances [3] [7] [13], named similarity (or distance) matrix. Thus, the clustering leads to draw, from similarity matrix, a tree structure where data  $A$  and  $B$  are situated on neighboring leaves if and only if  $d(A, B)$  is the smallest distance compared to other objects (see Fig. 1.).

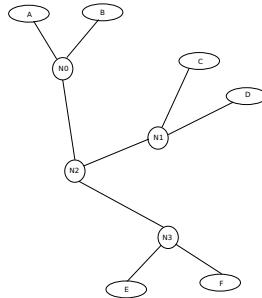
**Clustering Based on Hill Climbing Algorithm.** This clustering was tested with different kinds of data, notably to classify 24 species of mammals (using

---

<sup>2</sup> This is a metric measure [2] [5] [17], it has the symmetry, separation and triangular inequalities properties.



**Fig. 1.** Illustration of the similarity matrix with a dendrogram



**Fig. 2.** Dendrogram of data  $A, B, C, D, E$  and  $F$ , every internal node is connected to three edges

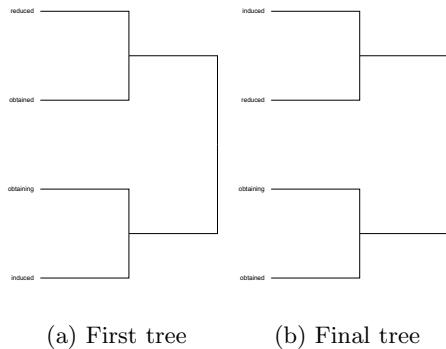
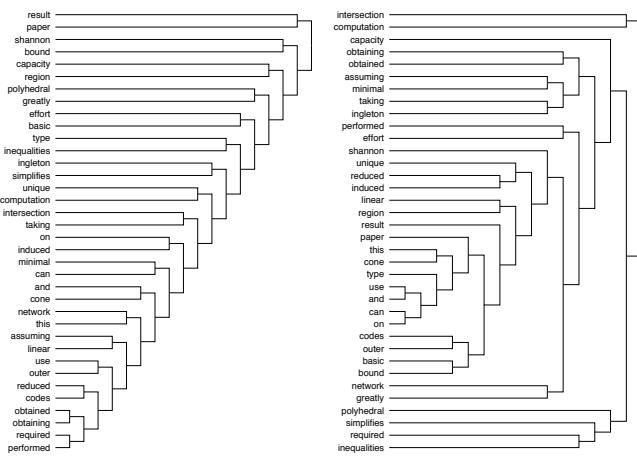
their mitochondrial DNA), the results are with few differences from those of paleontologists [3] [5] [9]. This similarity with expert results is due to the universality of the distance used, based on intrinsic similarities between objects. But, the major drawback of this method is that we can not classify more than hundred [3], due to the clustering algorithm with computational complexity of  $O(n^4)$  [3] [5].

The clustering algorithm used here is heuristic based on Hill-climbing algorithm, which begins with one initial solution (dendrogram, see Fig. 2.) to the problem at hand generated at random and then mutated. If the mutation results in higher fitness for the new solution than for the previous one, the new solution is kept ; otherwise, the current solution is retained. The algorithm is then repeated until no mutation can be found that causes an increase in the current solution's fitness, and this solution is returned as the result. We can find more details about this clustering algorithm in [3].

*Remark 3.* We observe that the best result is not obtained at the first step, we should repeat the Hill-climbing algorithm many and many time until obtaining the best possible result.

**Examples.** We have tested *Clustering by Compression* on different words, obtained from a scientific article abstract. The results shows that the best cost is not obtained in the first time. Some of the results are shown in Fig. 3, Fig.4. and Fig. 5.

	induced	obtaining	obtained	reduced
induced	0	0,529	0,500	0,200
obtaining	0,529	0	0,294	0,529
obtained	0,500	0,294	0	0,500
reduced	0,200	0,529	0,500	0

**Fig. 3.** Distances matrix for 4 data**Fig. 4.** Clustering by Compression with Hill-climbing of 4 data**Fig. 5.** Clustering by Compression with Hill-climbing of 36 data. We see at left the first generated dendrogram which is changed in every iteration, in order to improve its quality, until stopped by user. In our experiment, we stopped the algorithm when we obtained the same dendrogram (at right) as the *WPGMA* one (see Fig .7.), where similar words (have more letters in common) are closer to each other.

### 3 Contribution and Results

We present in this section our alternative clustering method, based on the universal similarity measure (used previously) and a faster clustering method (i.e. WPGMA), which has a computational complexity of  $O(n^2)$  [1] [2] [19], and also we discuss our experimental results.

The clustering algorithm which we use is Weighted Paired Group Method using Averages (*WPGMA*), it was developed by McQuitty in 1966, in order to build phylogenetic trees from similarities (or dissimilarities) matrix [1]. It works by merging (to cluster) at every iteration the nearest clusters (or leaves, cluster starts with leaves), until grouping all data in one cluster [1] [20].

#### 3.1 Algorithm

Consider:

- $D(i, j)$  is a distance based on Kolmogorov information between two objects  $i$  and  $j$ .
- $d_{i,j}$  is a distance between two clusters  $C_i$  and  $C_j$  it is equal to the average of Kolmogorov distance between two groups :

$$d_{i,j} = \frac{1}{2} \sum_{p \in C_i, q \in C_j} D(p, q).$$

If  $C_k = C_i \cup C_j$ , and  $C_l$  is a cluster , Then:  $d_{kl} = \frac{d_{il}|C_i| + d_{jl}|C_j|}{2}$

##### 1. Initialization:

Define  $n$  clusters, where every cluster  $C_i$  has a unique sequence number  $i$ ;  
Define a tree  $T$  with a set of  $n$  leaves, one leaf for every cluster and all leaves have the height 0.

##### 2. Iteration:

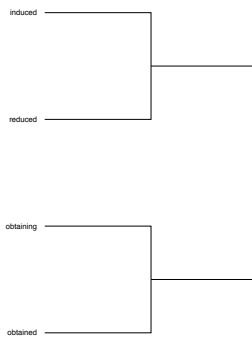
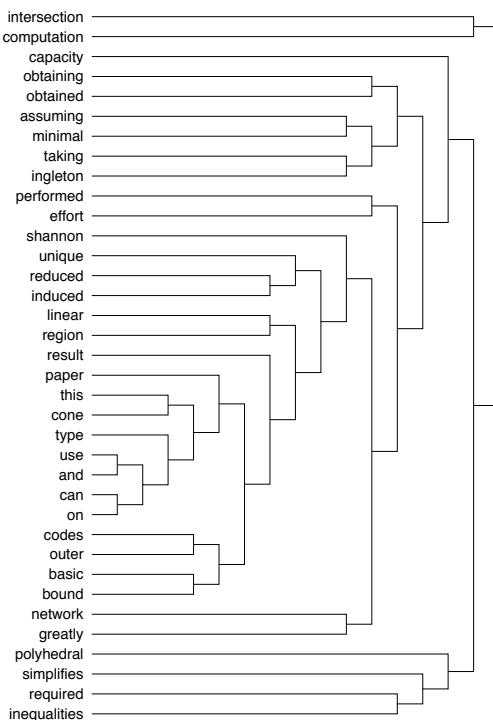
- Consider two clusters  $C_i, C_j$  as  $d_{i,j}$  is minimal;
- Define a new cluster  $C_k = C_i \cup C_j$ , with defining  $d_{kl}$ , for all  $l$  ;
- Define a new node  $k$  with sons  $i, j$ , and put it to the height  $\frac{d_{ij}}{2}$  ;
- Add  $C_k$  to the set of clusters, and remove  $C_i$  and  $C_j$ .

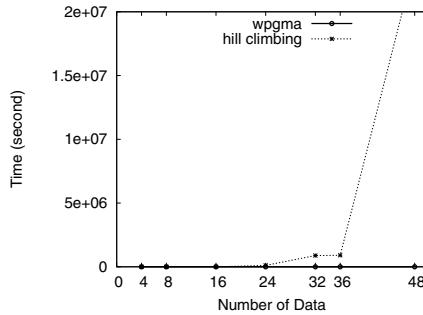
##### 3. When there are more than two groups $C_i, C_j$ , put the root in height $\frac{d_{ij}}{2}$ .

##### 4. End :

#### 3.2 Results

We did clustering with *WPGMA* with the same data used in Hill-climbing examples. The results shows that the best clustering (tree) is provided at the first time contrarily to the Hill climbing clustering. The results are shown in the figures Fig. 6 and Fig. 7.

**Fig. 6.** Clustering using WPGMA of 4 data**Fig. 7.** Clustering of 36 data by WPGMA This clustering provide only one (the first) dendrogram, where words are close to their similar ones



**Fig. 8.** Comparison of execution times of Hill climbing and WPGMA algorithms

The figure Fig. 8., shows the execution times of Hill climbing and WPGMA algorithms in the previous examples.

The comparison is represented by curves, continued for WPGMA and discontinued for Hill-climbing. On the  $x$ -axis is the number of data, represented by numbers (0-48). On the  $y$ -axis are the execution times, represented by seconds, of the two algorithms.

We observe that execution time of Hill climbing algorithm increase considerably, for 4 data it is executed in 0,0127 of a second and for 36 data it is executed 901868 seconds. On the contrary, we show that the execution time of our WPGMA algorithm version, which is fairly stable, increases slightly, for 4 data the execution time is 0,0002 of a second and for 36 data it is 0,01698 of a second.

## 4 Conclusion

We have introduced in this work *Clustering by Compression*, its advantages, notably the universality of the measure used, and its drawbacks specially in the case of Hill-climbing clustering. We proposed an amelioration to this clustering method by adapting the WPGMA clustering algorithm to the universal similarity measure. This adaptation decreases the complexity of the clustering by compression from  $O(n^4)$  to  $O(n^2)$ , and allows clustering of very large data rather than only hundred.

We have tried a clustering of a hundred objects with the hill-climbing based-on method, we did not obtain results after long time (more than a week). While clustering of five thousands of objects with our proposed WPGMA based method gave results in a few (four) Hours.

## References

1. Abrahams, J.: Code and parse trees for lossless source encoding. In: Proceedings of Compression and Complexity of Sequences, vol. 7.1, pp. 198–222 (1997)
2. Bennett, C.H., Gàcs, P., Li, M., Vitànyi, P.M.B., Zurek, W.: Information Distance. IEEE Transactions on Information Theory 44(4), 1407–1423 (1998)

3. Cilibrasi, R.: Statistical Inference Through Data Compression. Phd thesis, Amsterdam Universtity (2007)
4. Cilibrasi, R., Vitányi, P.M.B.: The Google Similarity Distance. IEEE Transactions on Knowledge and Data Engineering 19(3), 370–383 (2007)
5. Cilibrasi, R., Vitányi, P.M.B.: Clustering by compression. IEEE Transactions on Information Theory 51(4) (2005)
6. Cilibrasi, R., Vitányi, P.M.B.: A New Quartet Tree Heuristic for Hierarchical Clustering. In: IEEE/ACM Trans. Computat. Biol. Bioinf.; Presented at the EU- PASCAL Statistics and Optimization of Clustering Workshop, London, UK (2005) (submitted)
7. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley & Sons, Chichester (1991)
8. Delahaye, J.P., Zenil, H.: Towards a stable definition of Kolmogorov-Chaitin complexity. Fundamenta informaticae, 1–15 (2008)
9. Delahaye, J.P.: Complexités, Aux limites des mathématiques et de l'informatique. In: Belin, pour la science (2006)
10. Gronau, I., Moran, S.: Optimal implementations of UPGMA and other common clustering algorithms. Information Processing Letters 104(6), 205–210 (2007)
11. Guindon, S.: Méthodes et algorithmes pour l'approche statistique en phylogénie. Phd thesis, Université Montpellier II (2003)
12. Huffman, D.A.: A method for the construction of minimum redundancy codes. In: Proceeding of the IRE, pp. 1098–1101 (1951)
13. Lavallée, I.: Complexité et algorithmique avancée “une introduction”. In: 2 éme édition Hermann éditeurs (2008)
14. Levorato, V., Le, T.V., Lamure, M., Bui, M.: Classification prétopologique basée sur la complexité de Kolmogorov. Studia Informatica 7.1, 198–222 (2009)
15. Levorato, V.: Contributions à la Modélisation des Réseaux Complexes: Prétopologie et Applications. Phd thesis, Université de Paris 8, Paris (2008)
16. Loewenstein, Y., Portugaly, E., Former, M.L., Linial, M.: Efficient algorithms for accurate hierarchical clustering of huge datasets: tacking the entire protein space. Bioinformatics, 145–171 (2008)
17. Li, M., Chen, X., Li, X., Ma, B., Vitányi, P.M.B.: The similarity metric. IEEE Transactions on Information Theory 50(12) (2007)
18. Li, M., Vitányi, P.M.B.: An introduction to Kolmogorov Complexity and its applications, 2nd edn. Springer, Heidelberg (1997)
19. Murtagh, F.: Complexities of hierachic clustering algorithms: State of art. Computational Statistics Quarterly 1(2), 101–113 (1984)
20. Salemi, M., Vandamme, A.M.: The phylogenetic handbook: a practical approach to DNA and protein phylogeny. The Press Syndicate of the University of cambridge (2003)
21. Varré, J.S., Delahaye, J.P., Rivals, E.: Transformation distances: a family of dissimilarity measures based on movements of segments. Bioinformatics 15(3), 194–202 (1998)

# Rule Extraction from Support Vector Machine Using Modified Active Learning Based Approach: An Application to CRM

M.A.H. Farquad<sup>1,2</sup>, V. Ravi<sup>1,\*</sup>, and S. Bapi Raju<sup>2</sup>

<sup>1</sup> Institute for Development and Research in banking Technology, Castle Hills Road #1,  
Masab Tank, Hyderabad – 500 057 (AP) India

<sup>2</sup> Department of Computer & Information Sciences, University of Hyderabad,  
Hyderabad – 500 046 (AP) India  
farquadonline@gmail.com, rav\_padma@yahoo.com,  
bapics@uohyd.ernet.in

**Abstract.** Despite superior generalization performance Support vector machines (SVMs) generate *black box* models. The process of converting such opaque models into transparent model is often regarded as *rule extraction*. This paper presents a new approach for rule extraction from SVMs using modified active learning based approach (mALBA), to predict churn in bank credit cards. The dataset is obtained from Business Intelligence Cup 2004, which is highly unbalanced with 93% loyal and 7% churned customers' data. Since identifying churker is paramount from business perspective, therefore considering sensitivity alone, the empirical results suggest that the proposed rule extraction approach using mALBA yielded the best sensitivity compared to other classifiers.

**Keywords:** Support Vector Machines, Rule Extraction, modified active learning based approach, Customer Churn.

## 1 Introduction

Data mining involves the use of sophisticated data analysis algorithms to discover previously unknown, valid patterns and relationships in large datasets [1-4]. Data mining algorithms consists of statistics or machine learning based approaches, such as neural networks, decision trees etc. Similarly, the validity of the patterns discovered is dependent on how they compare to “*real world*” circumstances. Data mining has been effectively applied in wide range of applications, such as fraud detection [5] and scientific discovery [6] and manufacturing [7].

In recent past it is observed that, banks and the service industries has become more customer centric. The problem of customers shifting loyalties from one organization to another is called “*churn*”, and is common nowadays. Hence, there is a pressing need to develop algorithmic models that can predict which existing ‘*loyal*’ customer is going to churn out in near future [8]. Customer Relationship Management (CRM) is

---

\* Corresponding author. Ph: +91-40-2353 4981.

a process or methodology used to learn more about customers' needs and behaviors in order to develop stronger relationships with them. Research shows that, the customers with longer time relationship with the firm are more profitable [9, 10] than online bank customers [11]. Management should prepare an anti-churn strategy that is usually far less expensive than acquiring new customers [12, 13]. There are several types of CRM introduced for different purposes; *Operational CRM*, *Analytical CRM*, *Collaborative CRM* and *web-based CRM* [14-16].

Over a decade researchers have applied machine learning techniques for churn prediction problem; such as Multivariate Regression Analysis [17], Logistic Regression [9], Neural Networks [18], Random Forest [19], Decision Tree [20], FuzzyARTMAP [21], Support Vector Machines [22] and ensemble systems [8]. Yu Zhao et al., [23] concluded that using improved one-class SVM has shown best performance compared to other traditional methods like ANN, Decision Tree, and Naïve Bayes. SVM's generalization ability to deal with noisy data is reported using news paper subscription churn prediction data [24]. The efficiency of SVM has extended to predict the churn in Commercial Bank's VIP Customers [25]. Cao et al. reported that SVM-RFE extracts less key attributes and exhibits better satisfactory predictive effectiveness [26]. Further, recently SVMs efficiency is analyzed for customer churn prediction in landline telecommunications [27].

In this paper we present a modified active learning based approach for rule extraction from SVM using NBTree rule induction technique. The proposed approach is an extension to the approach presented by Martens et al., [41], where they used the training set with extra generated samples and using C4.5 rules were generated. During our proposed approach, support vectors are first obtained from SVM and using mALBA synthetic data instances are generated. The generated synthetic data is appended to support vectors and the target values are then replaced by the predictions of SVM. This modified data is fed to NBTree to generate rules.

This paper is structured as follows: In section 2 previous SVM rule extraction techniques are discussed. Section 3 describes the proposed approach. Next, in section 4, dataset description and experimental setup is detailed. Section 5 presents the results and discussions, and section 6 concludes this paper.

## 2 Rule Extraction from SVM

Gallant [28] initiated the work of rule extraction from a neural network that defines the knowledge learnt in the form of *if-then* rules. Even limited explanation can positively influence the system acceptance by the user [29]. A learning system might discover salient features in the input data whose importance was not previously recognized [30]. Rule extraction from opaque models improves generalization.

SVMs [31] have proved to be good alternative compared to other machine learning techniques specifically for classification problems [32]. Even though SVMs work well, it is completely non-intuitive to human experts, that they do not let us know the knowledge learnt by them during training in simple, comprehensible and transparent way.

Recently attempts have been made to extract rules from SVMs to represent the knowledge learnt by SVM during training. Extensive work was done towards devel-

oping rule extraction techniques for neural networks [33] but less work is reported towards rule extraction from SVM. SVM+Prototype [34], RulExtSVM [35], Extracting rules from trained support vector machines [36], Hyper rectangle Rules Extraction (HRE) [37], Fuzzy Rule Extraction (FREx) [38], Multiple Kernel-Support Vector Machine (MK-SVM) [39], SQReX-SVM [40], Active Learning-Based Approach (ALBA) [41], Hybrid rule extraction technique [42, 22, 43] and recently regression rule extraction technique [44], are some of the approaches proposed towards rule extraction from SVM.

### 3 Proposed Rule Extraction Approach

In this work, we propose a modified active learning based rule extraction procedure to extract rules from SVM using NBTree (Naive Bayes Tree) [45]. The proposed approach is applied to predict churn in bank credit cards. The dataset is obtained from Business Intelligence Cup 2004. The dataset is median scale and is highly unbalanced with 93% loyal and 7% churned customers' data. The proposed modified active learning based approach is described in Algorithm 1.

**Algorithm 1:** mALBA for Rule Extraction from SVM

Step 1: Train SVM and obtain the support vectors using training data [41].

# Calculate the average distance  $dist_k$  of training data to support vectors, in each dimension  $k$

Step 2: Calculate the  $dist_k$  between support vectors and training instances, in each dimension  $k$ .  $dist_k = dist_k + |d_k - sv_{j,k}|$

# Modified ALBA

Step 3: Randomly generate an extra data instance  $x_i$  close to support vectors

**For**  $i = 1$  to 500/1000 **do**

**For**  $k = 1$  to  $n$  **do**

$$x_{i,k} = sv(j, k) + \left[ (2 * rand - 1) \times \frac{dist_k}{2} \right]$$

**End for**

# Append the generated data to the support vectors

Step 4: Provide a class label  $y_i$  using the trained SVM as oracle.

**End for**

# Rule generation and evaluation

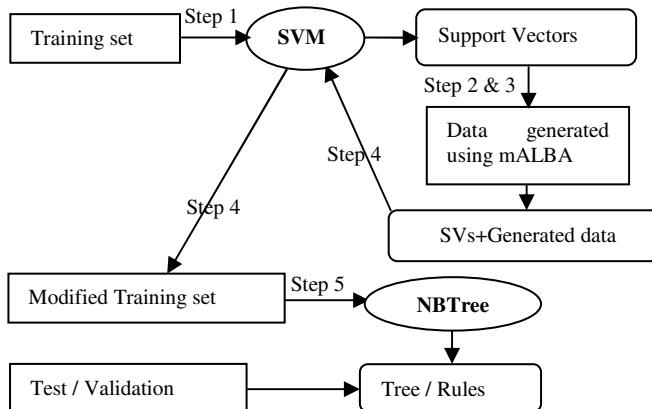
Step 5: Run rule induction algorithm on the modified data set and evaluate the performance of rules in terms of accuracy, fidelity and number of rules.

SVM model is first developed using training set under 10-fold cross validation and support vectors are extracted. We have chosen an RBF kernel for developing SVM model, as it is shown to achieve good overall performance [22]. Next, synthetic data is generated using mALBA. Before generation the data instances the distance  $dist$  between support vectors and the training set is calculated. Using the  $dist_k$ , the new data instances are generated which are near support vectors. 500 and 1000 data

instances are generated for empirical analysis. Generated data is then appended to the support vectors set and the predictions are obtained and the actual target values are then replaced by the predictions of SVM. This modified data is then fed to NBTree [45] to generate rules. The proposed approach is depicted in Fig. 1.

The current study in this paper is different from ALBA [41] approach in several ways, such as;

- They generated the instances using  $[(rand - 0.5) * dist_k / 2]$ , which generates the data near SVs that are between -0.5 to 0.5, whereas we are using  $[(2 * rand - 1) * dist_k / 2]$ , which generates the data near SVs that are between -1 to 1.
- They appended the generated data to training set whereas we are appending the generated data to the support vectors set.
- They employed C4.5 and RIPPER algorithm for rule generation, whereas we employed NBTree algorithm for rule generation.
- They evaluated their proposed approach for small scale and balanced problems whereas we analyzed medium scale and unbalanced problems in this study.



**Fig. 1.** Block diagram of the proposed rule extraction approach

## 4 Dataset Description and Experimental Setup

The dataset is from a Latin American bank that suffered from an increasing number of churns with respect to their credit card customers and decided to improve its retention system. The attribute information is tabulated in Table 1. The dataset consists of 14814 records, of which 13812 are loyal customers i.e. 93% and 1002 are churners i.e. 7%. Hence, the dataset is highly unbalanced in terms of the proportion of churners versus non-churners [46].

### 4.1 Experimental Setup

The available large scale unbalanced dataset is first divided into two parts of 80:20 ratios. 80% of the data is then used for training under 10 fold cross validation. 20% of

the data is named as validation set and stored for evaluating the efficiency of the rules generated under 10-FCV. The efficiency and validity of the rules generated during 10-FCV are then tested against the validation set, which is a subset of the original data.

To compare the performances with the original ALBA, we have applied the rule induction techniques

1. On ALBA.
2. On ALBA with support vectors set i.e. ALBA(SVs).
3. On mALBA.

**Table 1.** Feature description of churn prediction data set

Feature	Description	Value
Target	Target Variable	0-NonChurner 1-Churner
CRED_T	Credit in month T	Positive real number
CRED_T-1	Credit in month T-1	Positive real number
CRED_T-2	Credit in month T-2	Positive real number
NCC_T	Number of credit cards in months T	Positive integer value
NCC_T-1	Number of credit cards in months T-1	Positive integer value
NCC_T-2	Number of credit cards in months T-2	Positive integer value
INCOME	Customer's Income	Positive real number
N_EDUC	Customer's educational level	1 - University student 2 - Medium degree 3 - Technical degree 4 - University degree
AGE	Customer's age	Positive integer
SX	Customers sex	1 - male 0 - Female
E_CIV	Civilian status	1-Single 2-Married 3-Widow 4-Divorced
T_WEB_T	Number of web transaction in months T	Positive integer
T_WEB_T-1	Number of web transaction in months T-1	Positive integer
T_WEB_T-2	Number of web transaction in months T-2	Positive integer
MAR_T	Customer's margin for the company in months T	Real Number
MAR_T-1	Customer's margin for the company in months T-1	Real Number
MAR_T-2	Customer's margin for the company in months T-2	Real Number
MAR_T-3	Customer's margin for the company in months T-3	Real Number
MAR_T-4	Customer's margin for the company in months T-4	Real Number
MAR_T-5	Customer's margin for the company in months T-5	Real Number
MAR_T-6	Customer's margin for the company in months T-6	Real Number

## 5 Results and Discussions

Identifying potential churners correctly is the basic intension of many business decision makers. Hence, they place high emphasis on sensitivity alone which contributes towards the bottom-line of the fundamental CRM. Consequently in this paper, sensitivity is accorded top priority ahead of specificity and accuracy. We used the SVM library viz., LibSVM [47] for SVM. LibSVM is integrated software for support vector classification and is developed in MATLAB. RapidMiner4.5 community edition [48] is used for generating NBTree. The quantities employed to measure the quality of the classifiers are sensitivity, specificity and accuracy [49].

During our empirical study we generated 500 and 1000 extra instances separately, using the calculations described in step 3 of the proposed approach and using original ALBA calculations. The results obtained using NBTree and Decision Tree are presented in Table 2 and 3, respectively.

**Table 2.** Average Results obtained using NBTree

Extra Generated Data	Classifiers	10-Fold Cross validation				Validation		
		Sens*	Spec*	Acc*	t-test	Sens*	Spec*	Acc*
500	<b>SVM</b>	63.35	81.41	80.19	4.93	64.65	80.63	79.55
	<b>ALBA</b>	65.48	85.23	83.92	3.44	67.7	84.52	83.38
1000	<b>ALBA (SVs)</b>	74.93	83.05	82.5	0.82	76.55	82.74	82.32
	<b>mALBA</b>	<b>78.17</b>	<b>80.36</b>	<b>80.28</b>	-	<b>79.35</b>	<b>79.16</b>	<b>79.17</b>
	<b>ALBA</b>	65.35	85.88	84.46	2.2	68.05	84.75	83.68
	<b>ALBA (SVs)</b>	73.8	83.26	82.62	0.11	73.25	82.99	82.27
	<b>mALBA</b>	<b>74.3</b>	<b>82.84</b>	<b>82.31</b>	-	<b>75.9</b>	<b>83.3</b>	<b>82.87</b>

**Table 3.** Average Results obtained using Decision Tree

Extra Generated Data	Classifiers	10-Fold Cross validation				Validation		
		Sens*	Spec*	Acc*	t-test	Sens*	Spec*	Acc*
500	<b>ALBA</b>	60.22	83.56	81.97	4.377	65.2	82.65	81.48
	<b>ALBA (SVs)</b>	69.68	80.38	79.66	1.089	72.7	80.38	72.25
1000	<b>mALBA</b>	<b>73.81</b>	<b>76.91</b>	<b>76.7</b>	-	<b>75.05</b>	<b>75.99</b>	<b>76.43</b>
	<b>ALBA</b>	61.6	83.81	82.32	2.874	64.05	83	81.72
	<b>ALBA (SVs)</b>	67.69	80.55	79.68	0.658	71.65	79.7	79.17
	<b>mALBA</b>	<b>70.1</b>	<b>81.6</b>	<b>80.85</b>	-	<b>72.85</b>	<b>81.3</b>	<b>80.7</b>

It is observed that the rules extracted by the proposed rule extraction approach using mALBA with 500 extra instances yielded best average sensitivity of 78.17% under 10-FCV and the same set of rules yielded 79.35% sensitivity against validation set. Using ALBA, the sensitivity yielded under 10-FCV is 65.48% and against validation set the sensitivity obtained is 67.7%. It is observed from the results that the generated instances using mALBA are positively near the SVM boundary. And the extra generated data using ALBA when used with SVs set, the sensitivity yielded is 74.93% during 10-FCV and against validation it yielded 76.55% sensitivity.

The same set of experiments is carried out by generating 1000 extra instances, it is observed that mALBA yielded 74.3% sensitivity, whereas original ALBA yielded 65.35% sensitivity under 10-FCV. mALBA and ALBA with 1000 extra instances yielded 75.9% and 68.05% sensitivities against validation set, respectively. It is observed that the time taken by ALBA for rule extraction is more than mALBA, as the extra generated instances are appended to the training set in ALBA approach and the extra generated instances are appended to SVs set in mALBA. When the data is generated using ALBA calculations with SVs the sensitivity yielded under 10-FCV is 73.8% and against validation set the sensitivity yielded is 73.25%. It is observed that the generated samples with SVs yielded better sensitivity compared to ALBA. When the generated data is used with SVs, the complexity, time and rules are decreased. It is observed that instead of using all the training instances with the generated data for rule induction algorithm as [41], it is better to take SVs set with extra generated data

to reduce the complexity of the system and it also produces less number of rules without compromising the accuracy of the model.

For comparison purpose rules are also extracted using DT. It is observed that rules extracted using NBTree yielded best sensitivity compared to the rules extracted using DT. Furthermore, the average number of rules during 10-FCV using NBTree is 13 whereas the average number of rules extracted using DT is 85.5 for mALBA, 400 for ALBA and 98.5 for ALBA (SVs).

Using sensitivity, the classifiers are compared with t-test at  $n_1+n_2-2=10+10-2=18$  degrees of freedom at 10% level of significance. We tested if the difference in performances is statistically significant. The tabulated value of t-statistics for 18 degrees of freedom at 10% level of significance is 1.73. That means, if t-statistics value between two different classifiers is more than 1.73, we say that the difference between techniques is statistically significant otherwise not significant. The t-test values obtained between the sensitivities shows that mALBA is statistically significant to original ALBA but it is statistically insignificant to ALBA with SVs.

A rule set is considered to display a high level of *fidelity* if it can *mimic* the behavior of the machine learning technique from which it was extracted i.e. SVM in our study. The fidelity obtained using ALBA, ALBA with SVs and mALBA is presented in Table 4. It is observed that ALBA behaves 83.28% like SVM with 500 generated samples, whereas our proposed mALBA approach behaves 82.65% like SVM. The fidelity obtained using 1000 extra generated instances with ALBA, mALBA and ALBA with SVs is 81.64%, 79.1% and 79.03%, respectively. It is observed that ALBA mimics the behavior of SVM better than mALBA and ALBA with SVs.

**Table 4.** Average Fidelity Obtained using ALBA and Proposed mALBA

Extra Generated Data	Classifiers	NBTree	DecisionTree
500	<b>ALBA</b>	83.28	89.85
	<b>ALBA (SVs)</b>	80.88	86.56
	<b>mALBA</b>	82.65	85.64
	<b>ALBA</b>	81.64	90.54
	<b>ALBA (SVs)</b>	79.03	87.78
	<b>mALBA</b>	79.1	86.37

Table 5 presents the example rule set obtained using mALBA. The number of rules extracted using mALBA is very much less in number when compared to the rules extracted using ALBA [41]. It is observed that mALBA with 500 extra samples yielded the best sensitivity among other approaches tested and more generalized rules are obtained.

**Table 5.** Rules Obtained using mALBA

S. No	Antecedents	Consequent
1	If MAR_T <= 8.81 and MAR_T-5 <= -5.04	Non-Churner
2	If MAR_T <= 8.81 and MAR_T-5 > -5.04	Churner
3	If MAR_T > 8.81 and CRED_T < 594.145	Churner
4	If MAR_T > 8.81 and CRED_T > 594.145 and CRED_T-2 < 94.58	Churner
5	If MAR_T > 8.81 and CRED_T > 594.145 and CRED_T-2 > 94.58	Non-Churner
6	If CRED_T <= 598.1 and MAR_T_2 <= 14.045 and MAR_T_5 <= 973	Churner

## 6 Conclusions

In this paper, we present a modified active learning based approach for rule extraction from SVM to solve credit card customer churn prediction problem. The dataset is taken from Business Intelligence Cup organized by University of Chile in 2004. This is highly unbalanced data with 93% good customers and 7% churned customers. While solving the problems like churn prediction, sensitivity is accorded high priority. Accordingly, by considering sensitivity alone, it is observed that the proposed rule extraction approach using mALBA yielded the best sensitivity of 79.35%. It is also observed that when ALBA is used with SVs set obtained better accuracy than original ALBA [41]. It is observed that original ALBA using C4.5 for generating rules, yielded more number of rules, which may make the black box model a transparent model but the comprehensibility of the classifier is adversely affected. Efficiency of the feature selection using SVM-RFE can be analyzed in future works. Further, mALBA rule extraction approach can be employed to solve Insurance fraud detection problem.

## References

1. Usama, F., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence (1996)
2. Adriaans, P., Zantinge, D.: Data Mining. Addison-Wesley, New York (1996)
3. Osuna, E.E., Freund, R., Girosi, F.: Support vector Machines: Training and Applications. Technical Report, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, AI Memo No. 1602 (1997)
4. Osuna, E.E., Freund, R., Girosi, F.: Training support vector machines: an application to face detection. In: Proceedings of Computer Vision and Pattern Recognition, pp. 130–136 (1997)
5. Senator, T., Goldberg, H.G., Wooton, J., Cottini, M.A., Umarkhan, A.F., Klinger, C.D., Llamas, W.M., Marrone, M.P., Wong, R.W.H.: The Financial Crimes Enforcement Network AI System (FAIS): Identifying Potential Money Laundering from Reports of Large Cash Transactions. *AI Magazine* 16(4), 21–39 (1995)
6. Hall, J., Mani, G., Barr, D.: Applying Computational Intelligence to the Investment Process. In: Proceedings of CIFER 1996: Computational Intelligence in Financial Engineering, Washington, D.C. IEEE Computer Society Press, Los Alamitos (1996)
7. Ravi, V., Arul Shalom, S.A., Manickavel, A.: Sputter Process Variables Prediction via Data Mining. In: Proceedings of the 2004 IEEE Conference on Cybernetics and Intelligent Systems, Singapore (2004)
8. Kumar, D.A., Ravi, V.: Predicting credit card customer churn in banks using data mining. *International Journal for Data Analysis, Techniques and Strategies* 1(1), 4–28 (2008)
9. Bolton, R.N., Kannan, P.K., Bramlett, M.D.: Implications of loyalty program membership and service experiences for customer retention and value. *Journal of the Academy of Marketing Science* 28(1), 95–108 (2000)
10. Lejeune, M.A.P.M.: Measuring the impact of data mining on churn management. *Electronic Networking Applications and Policy* 11(5), 375–387 (2001)
11. Mols, N.P.: The Behavioral consequences of PC banking. *International Journal of Bank Marketing* 16(5), 195–201 (1998)
12. Chu, B.H., Tsai, M.-S., Ho, C.-S.: Toward a hybrid data mining model for customer retention. *Knowledge-Based Systems* 20(8), 703–718 (2007)

13. Ryals, L., Knox, S.: Cross-functional issues in the implementation of relationship marketing through customer relationship management. *European Management Journal* 19(5), 534–542 (2001)
14. Alajoutsijarvi, K., Mannermaaa, K., Tikkainen, H.: Customer relationships and the small software firm: a framework for understanding challenges faced in marketing. *Information and Management* 37, 153–159 (2000)
15. Laudon, K., Laudon, J.: *Management Information Systems: Managing the Digital Firm*, 7th edn. Prentice-Hall, Englewood Cliffs (2000)
16. Edwards, J.: Get It Together with Collaborative CRM. *insideCRM*. Tippit (2007), <http://www.insidecrm.com/features/collaborative-crm-112907/>
17. Bloemer, J., Ruyter, K.D., Peeters, P.: Investigating drivers of bank loyalty: the complex relationship between image, service quality and satisfaction. *International Journal of Bank Marketing* 16(7), 276–286 (1998)
18. Mozer, M.C., Wolniewicz, R., Grimes, D.B., Johnson, E., Kaushansky, H.: Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions of Neural Networks* 11(3), 690–696 (2000)
19. Larivière, B., Van den Poel, D.: Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: the case of financial services. *Expert Systems with Applications* 27(2), 277–285 (2004)
20. Hung, S.-Y., Yen, D.C., Wang, H.: Applying data mining to telecom churn management. *Expert Systems with Applications* 31(3), 515–524 (2006)
21. Naveen, N., Ravi, V., Kumar, D.A.: Application of fuzzyARTMAP for churn prediction in bank credit cards. *International Journal of Information and Decision Sciences* 1(4), 428–444 (2009)
22. Farquad, M.A.H., Ravi, V., Bapi, R.S.: Data Mining using Rules Extracted from SVM: an Application to Churn Prediction in Bank Credit Cards. In: Sakai, H., Chakraborty, M.K., Hassani, A.E., Ślęzak, D., Zhu, W. (eds.) RSFDGrC 2009. LNCS, vol. 5908, pp. 390–397. Springer, Heidelberg (2009)
23. Zhao, Y., Li, B., Li, X., Liu, W., Ren, S.: Customer Churn Prediction Using Improved One-Class Support Vector Machine. In: Li, X., Wang, S., Dong, Z.Y. (eds.) ADMA 2005. LNCS (LNAI), vol. 3584, pp. 300–306. Springer, Heidelberg (2005)
24. Coussement, K., Van den Poel, D.: Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications* 34(1), 313–327 (2008)
25. Zhaoa, J., Dang, X.-H.: Bank Customer Churn Prediction Based on Support Vector Machine: Taking a Commercial Bank's VIP Customer Churn as the Example. In: Proceedings of the 4th International Conference on Wireless Communications, Networking and Mobile Computing 2008 (WiCOM 2008), pp. 1–4 (2008)
26. Cao, K., Shao, P.-j.: Customer Churn Prediction Based on SVM-RFE. In: International Seminar on Business and Information Management ISBIM 2008, vol. 1, pp. 306–309 (2008)
27. Huang, B.Q., Kechadi, T.-M., Buckley, B., Kiernan, G., Keogh, E., Rashid, T.: A new feature set with new window techniques for customer churn prediction in land-line telecommunications. *Expert Systems with Applications* 37(5), 3657–3665 (2010)
28. Gallant, S.: Connectionist expert systems. *Communications of the ACM* 31(2), 152–169 (1988)
29. Davis, R., Buchanan, B.G., Shortliffe, E.: Production rules as a representation for a knowledge-based consultation program. *Artificial Intelligence* 8(1), 15–45 (1977)
30. Craven, M.W., Shavlik, J.W.: Using sampling and queries to extract rules from trained neural networks. In: The Proceedings of the Eleventh International Conference on Machine Learning, San Francisco, CA, USA (1994)
31. Vapnik, V.N.: *The Nature of Statistical Learning Theory*. Springer, New York (1995)

32. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.N.: Gene selection for cancer classification using support vector machines. *Machine Learning* 46(1-3), 389–422 (2002)
33. Tickle, A.B., Andrews, R., Golea, M., Diederich, J.: The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural network. *IEEE Transactions on Neural Networks* 9(6), 1057–1068 (1998)
34. Nunez, H., Angulo, C., Catata, A.: Rule extraction from support vector machines. In: European Symposium on Artificial Neural Networks Proceedings, pp. 107–112 (2002)
35. Fung, G., Sandilya, S., Bharat, R.R.: Rule extraction from linear support vector machines. In: Proceeding of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pp. 32–40. ACM Press, New York (2005)
36. Barakath, N., Diederich, J.: Eclectic rule-extraction from support vector machines. *International journal of Computer Intelligence* 2(1), 59–62 (2005)
37. Zhang, Y., Su, H., Jia, T., Chu, J.: Rule Extraction from Trained Support Vector Machines. In: Ho, T.-B., Cheung, D., Liu, H. (eds.) PAKDD 2005. LNCS (LNAI), vol. 3518, pp. 61–70. Springer, Heidelberg (2005)
38. Chaves, A.C.F., Vellasco, M.M.B.R., Tanscheit, R.: Fuzzy rule extraction from support vector machines. In: Fifth International Conference on Hybrid Intelligent Systems, Rio de Janeiro, Brazil, November 06-09 (2005)
39. Chen, Z., Li, J., Wei, L.: A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue. *Artificial Intelligence in Medicine* 41, 161–175 (2007)
40. Barakat, N.H., Bradley, A.P.: Rule Extraction from Support Vector Machines: A Sequential Covering Approach. *IEEE Transactions on Knowledge and Data Engineering* 19(6), 729–741 (2007)
41. Martens, D., Baesens, B., Gestel, T.V.: Decompositional Rule Extraction from Support Vector Machines by Active Learning. *IEEE Transactions on Knowledge and Data Engineering* 21(2), 178–191 (2009)
42. Farquad, M.A.H., Ravi, V., Bapi, R.S.: Rule Extraction using Support Vector Machine Based Hybrid Classifier. In: Presented in TENCON-2008, IEEE region 10 Conference, Hyderabad, India, November 19-21 (2008)
43. Farquad, M.A.H., Ravi, V., Bapi, R.S.: Support Vector Machine based Hybrid Classifiers and Rule Extraction Thereof: Application to Bankruptcy Prediction in Banks. In: Soria, E., Martín, J.D., Magdalena, R., Martínez, M., Serrano, A.J. (eds.) Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques, vol. 2, pp. 404–426. IGI Global, USA (2010)
44. Farquad, M.A.H., Ravi, V., Bapi, R.S.: Support vector regression based hybrid rule extraction methods for forecasting. *Expert Systems with Applications* (2010), doi:10.1016/j.eswa.2010.02.055
45. Ron, K.: Scaling Up the Accuracy of Naïve-Bayes Classifiers: a Decisoin-Tree Hybrid. In: Proceedings of KDD 1996, Portland, USA (1996)
46. Business Intelligence Cup - 2004: Organized by the University of Chile, [http://www.tis.cl/bicup\\_04/text-bicup/BICUP/202004/20public/20data.zip](http://www.tis.cl/bicup_04/text-bicup/BICUP/202004/20public/20data.zip)
47. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001), <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
48. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler, T.: YALE: Rapid Prototyping for Complex Data Mining Tasks. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2006 (2006)
49. Fawcett, T.: An introduction to ROC analysis. *Pattern Recognition Letters* 27, 861–874 (2006)

# Factorizing Three-Way Binary Data with Triadic Formal Concepts\*

Radim Belohlavek and Vilem Vychodil

Dept. Computer Science, Palacky University, Olomouc  
17. listopadu 12, CZ-771 46 Olomouc, Czech Republic  
[{radim.belohlavek,vilem.vychodil}@upol.cz](mailto:{radim.belohlavek,vilem.vychodil}@upol.cz)

**Abstract.** We present a problem of factor analysis of three-way binary data. Such data is described by a 3-dimensional binary matrix  $I$ , describing a relationship between objects, attributes, and conditions. The aim is to decompose  $I$  into three binary matrices, an object-factor matrix  $A$ , an attribute-factor matrix  $B$ , and a condition-factor matrix  $C$ , with a small number of factors. The difference from the various decomposition-based methods of analysis of three-way data consists in the composition operator and the constraint on  $A$ ,  $B$ , and  $C$  to be binary. We present a theoretical analysis of the decompositions and show that optimal factors for such decompositions are provided by triadic concepts developed in formal concept analysis. Moreover, we present an illustrative example, propose a greedy algorithm for computing the decompositions.

## 1 Introduction

### 1.1 Problem Description

Recently, there has been a growing interest in methods for analysis of three-way and generally  $N$ -way data that are based on various matrix decompositions. [10] provides an up-to-date survey with 244 references, see also [5]. An  $N$ -way data is represented by an  $N$ -dimensional matrix, called also  $N$ -dimensional array, or  $N$ -dimensional tensor. 2-dimensional matrices are the ordinary matrices whose entries are indexed by two indices (rows and column),  $N$ -dimensional matrices have  $N$ -indices. Decompositions of  $N$ -dimensional matrices go back as far as to 1920s and have been studied in psychometrics since the 1940s [10].

We are concerned with decompositions of three-way binary data, i.e. data represented by a 3-dimensional matrix which is denoted by  $I$  in this paper and whose entries, denoted  $I_{ijt}$ , are either 0 or 1. The matrix entries are interpreted as follows (clearly, other interpretations are possible):

$$I_{ijt} = \begin{cases} 1 & \text{if object } i \text{ has attribute } j \text{ under condition } t, \\ 0 & \text{if object } i \text{ does not have attribute } j \text{ under condition } t. \end{cases} \quad (1)$$

---

\* Supported by grant no. P202/10/0262 of the Czech Science Foundation and by grant no. MSM 6198959214.

For example (cf. Section 3), objects correspond to students, attributes to student qualities, and conditions to courses passed by the students.

Our aim is to decompose  $I$  in a way similar to the one employed in Boolean factor analysis, see e.g. [3][7]. Recall that in Boolean factor analysis, a decomposition  $I = A \circ B$ , defined by  $I_{ij} = \max_{l=1}^k A_{il} \cdot B_{lj}$ , of an object-attribute binary matrix  $I$  is sought into an object-factor matrix  $A$  and a factor-attribute matrix  $B$ , with  $k$  (number of factors) as small as possible.  $\circ$  is the well-known Boolean matrix multiplication. In our scenario, the goal is to decompose a 3-dimensional binary matrix  $I$  into a product  $\circ(A, B, C)$  of three binary matrices, an object-factor matrix  $A$ , an attribute-factor matrix  $B$ , and a condition-factor matrix  $C$  with the number of factors as small as possible. The operator  $\circ(\cdot, \cdot, \cdot)$ , defined in Section 2, is a 3-dimensional analogue of Boolean matrix multiplication.

As a main contribution of this paper, we show that optimal decompositions (those with the least number of factors) are attained by using so-called triadic concepts [13][7] as factors. In Sections 3 and 4, we present a detailed illustrative example, basic complexity considerations, and a greedy approximation algorithm to compute the decompositions. Section 5 presents issues for future work. Due to lack of space, proofs are omitted.

## 1.2 Related Work

Decompositions of (2-dimensional) matrices and the related methods of data analysis, such as factor analysis (FA), principal component analysis (PCA), independent component analysis (ICA), singular value decomposition (SVD), and others have been studied for a long time. Recently, there has been a growing interest in two topics. On one hand, there is a growing interest in the methods for decomposition of  $N$ -dimensional matrices, see [5] and in particular [10] for a survey. The reason behind is that  $N$ -way data naturally appear in many fields including psychometrics, chemometrics, signal processing, computer vision, neuroscience, numerical analysis, and others. On the other hand, there is an interest in the methods for decomposition of data which is constrained in some way. An example is the nonnegative matrix factorization [12]. Such constraints can be seen as semantic constraints which help us interpret the results of decompositions. Particularly relevant to our paper is the work on decompositions of binary data. Several methods, including modifications of the methods designed originally for real-valued data, have been developed, see [16] for an overview. A particular role among them have the methods which decompose a binary matrix into a Boolean product of binary matrices, see e.g. [3][7][14]. Namely, as reported in [14], Boolean matrix decompositions can be interpreted in a straightforward way. The present paper can be seen as an extension of [3] in which we described optimal decompositions of binary matrices, provided theoretical results on various aspects of Boolean decompositions, and an efficient approximation algorithm. In this paper, we seek to extend these results to three-way data. Such an extension is not obvious because several useful properties from the case of two-way data (such as a simple duality due to Galois connection induced by the input matrix) are no more available in the case of three-way data.

## 2 Decomposition and Factors

### 2.1 Decomposition

Consider an  $n \times m \times p$  binary matrix  $I$  with entries  $I_{ijt}$ . We are interested in decompositions of  $I$  into three binary matrices, an  $n \times k$  object-factor matrix  $A$  with entries  $A_{ik}$ , an  $m \times k$  attribute-factor matrix  $B$  with entries  $B_{jk}$ , an  $p \times k$  condition-factor matrix  $C$  with entries  $C_{tk}$ , with respect to a ternary composition  $\circ$  defined by

$$\circ(A, B, C)_{ijt} = \max_{l=1}^k A_{il} \cdot B_{jl} \cdot C_{tl}. \quad (2)$$

We look for  $I = \circ(A, B, C)$  with the smallest number  $k$  of factors.

*Remark 1.* (1) For  $p = 1$ , the problem becomes the problem of decomposition of a binary matrix into a Boolean product of binary matrices.

(2) Due to lack of space, we do not include observations on the various ways of possible compositions of 3- and lower-dimensional binary matrices. For real-valued matrices, see [10].

### 2.2 Factors for Decomposition

We are going to show the role of so-called triadic concepts for the decompositions. Triadic concepts were introduced in formal concept analysis (FCA). We provide the preliminaries and refer to [8] (ordinary, or dyadic, FCA) and [13, 17] (triadic FCA) for more information. Note that in FCA, one works with relations rather than binary matrices. Since the distinction between relations and binary matrices is only a formal one, we use  $I$  to denote both, an  $n \times m \times p$  binary matrix and a ternary relation between sets  $X$ ,  $Y$ , and  $Z$ , with  $|X| = n$ ,  $|Y| = m$ , and  $|Z| = p$ . The correspondence is:  $I_{ijt} = 1$  (matrix) iff  $\langle x_i, y_j, z_t \rangle \in I$  (relation).

*Preliminaries from Dyadic and Triadic FCA.* A *formal context* (or *dyadic context*) is a triplet  $\langle X, Y, I \rangle$  where  $X$  and  $Y$  are non-empty sets and  $I$  is a binary relation between  $X$  and  $Y$ , i.e.  $I \subseteq X \times Y$ .  $X$  and  $Y$  are interpreted as the sets of objects and attributes, respectively;  $I$  is interpreted as the incidence relation (“to have relation”). That is,  $\langle x, y \rangle \in I$  is interpreted as: object  $x$  has attribute  $y$ . A formal context  $\mathbf{K} = \langle X, Y, I \rangle$  induces a pair of operators  $\uparrow : 2^X \rightarrow 2^Y$  and  $\downarrow : 2^Y \rightarrow 2^X$  defined for  $C \subseteq X$  and  $D \subseteq Y$  by

$$\begin{aligned} C^\uparrow &= \{y \in Y \mid \text{for each } x \in C: \langle x, y \rangle \in I\}, \\ D^\downarrow &= \{x \in X \mid \text{for each } y \in D: \langle x, y \rangle \in I\}. \end{aligned}$$

These operators, called *concept-forming operators*, form a Galois connection [8] between  $X$  and  $Y$ . Usually, there is no danger of misunderstanding and both  $\uparrow$  and  $\downarrow$  may be denoted by the same symbol, e.g. one uses  $C'$  and  $D'$  instead of  $C^\uparrow$  and  $D^\downarrow$ . A *formal concept* (or *dyadic concept*) of  $\langle X, Y, I \rangle$  is a pair  $\langle C, D \rangle$  consisting of sets  $C \subseteq X$  and  $D \subseteq Y$  such that  $C^\uparrow = D$  and  $D^\downarrow = C$ ;  $C$  and  $D$

are called the *extent* and *intent* of  $\langle C, D \rangle$ . The collection of all formal concepts of  $\langle X, Y, I \rangle$  is denoted by  $\mathcal{B}(X, Y, I)$  and is called the *concept lattice* of  $\langle X, Y, I \rangle$ . That is,

$$\mathcal{B}(X, Y, I) = \{\langle C, D \rangle \mid C^\dagger = D, D^\dagger = C\}.$$

A concept lattice equipped with a partial order corresponding to a subconcept-superconcept hierarchy is indeed a complete lattice [8]. A formal context may be visualized by a binary matrix: rows and columns correspond to objects and attributes; an entry corresponding to  $x \in X$  and  $y \in Y$  equals 1 iff  $\langle x, y \rangle \in I$ . Formal concepts of  $\langle X, Y, I \rangle$  are just maximal rectangular areas in the corresponding binary matrix which are full of 1s [8].

A *triadic context* is a quadruple  $\langle X_1, X_2, X_3, I \rangle$  where  $X_1, X_2$ , and  $X_3$  are non-empty sets (interpreted as the sets of objects, attributes, and conditions, respectively), and  $I$  is a ternary relation between  $X_1, X_2$ , and  $X_3$ .  $I$  is interpreted as the incidence relation (“to have-under relation”). That is,  $\langle x, y, z \rangle \in I$  is interpreted as: object  $x$  has attribute  $y$  under condition  $z$ . For convenience, a triadic context is denoted by  $\langle X_1, X_2, X_3, I \rangle$ . A triadic context  $\mathbf{K} = \langle X_1, X_2, X_3, I \rangle$  induces the following dyadic contexts:  $\mathbf{K}^{(1)} = \langle X_1, X_2 \times X_3, I^{(1)} \rangle$ ,  $\mathbf{K}^{(2)} = \langle X_2, X_1 \times X_3, I^{(2)} \rangle$ ,  $\mathbf{K}^{(3)} = \langle X_3, X_1 \times X_2, I^{(3)} \rangle$ , with the binary relations  $I^{(1)}, I^{(2)}$ , and  $I^{(3)}$  defined by

$$\langle x_1, \langle x_2, x_3 \rangle \rangle \in I^{(1)} \text{ iff } \langle x_2, \langle x_1, x_3 \rangle \rangle \in I^{(2)} \text{ iff } \langle x_3, \langle x_1, x_2 \rangle \rangle \in I^{(3)} \text{ iff } \langle x_1, x_2, x_3 \rangle \in I.$$

for every  $x_1 \in X_1, x_2 \in X_2, x_3 \in X_3$ . The concept-forming operators induced by  $\mathbf{K}^{(i)}$  are denoted by  ${}^{(i)}$ . That is, for  $C \subseteq X_1$  and  $D \subseteq X_2 \times X_3$ , we have

$$\begin{aligned} C^{(1)} &= \{\langle x_2, x_3 \rangle \in X_2 \times X_3 \mid \text{for each } x_1 \in C: \langle x_1, x_2, x_3 \rangle \in I\}, \\ D^{(1)} &= \{x_1 \in X_1 \mid \text{for each } \langle x_2, x_3 \rangle \in D: \langle x_1, x_2, x_3 \rangle \in I\}, \end{aligned}$$

and similarly for  ${}^{(2)}$  and  ${}^{(3)}$ . A *triadic concept* of  $\langle X_1, X_2, X_3, I \rangle$  is a triplet  $\langle D_1, D_2, D_3 \rangle$  of  $D_1 \subseteq X_1, D_2 \subseteq X_2$ , and  $D_3 \subseteq X_3$ , such that for every  $\{i, j, k\} = \{1, 2, 3\}$  with  $j < k$  we have  $D_i = (D_j \times D_k)^{(i)}$ ;  $D_1, D_2$ , and  $D_3$  are called the *extent*, *intent*, and *modus* of  $\langle D_1, D_2, D_3 \rangle$ . The set of all triadic concepts of  $\langle X_1, X_2, X_3, I \rangle$  is denoted by  $\mathcal{T}(X_1, X_2, X_3, I)$  and is called the concept trilattice of  $\langle X_1, X_2, X_3, I \rangle$ ; the reader is referred to [17] to details on the notion of a trilattice and for the trilattice structure on  $\mathcal{T}(X_1, X_2, X_3, I)$ .

Triadic concepts can be represented by particular 3-dimensional binary matrices, namely cuboidal matrices. Formally,  $J$  is a cuboidal matrix (shortly, a cuboid) if there exist an  $n \times 1$  binary vector  $A$ , an  $m \times 1$  binary vector  $B$ , and a  $p \times 1$  binary vector  $C$ , such that  $J = \circ(A, B, C)$ .

The following theorem explains the role of cuboids for decompositions [2].

**Theorem 1.**  $I = \circ(A, B, C)$  for an  $n \times k$  matrix  $A$ ,  $m \times k$  matrix  $B$ , and  $p \times k$  matrix  $C$ , if and only if it is a max-superposition of  $k$  cuboids  $J_1, \dots, J_k$ , i.e.

$$I = J_1 \max \cdots \max J_k.$$

For each  $l = 1, \dots, k$ ,  $J_l = \circ(A_{\perp}, B_{\perp}, C_{\perp})$ , i.e. each  $J_l$  is the product of the  $l$ -th columns of  $A$ ,  $B$ , and  $C$ .

As a result, to decompose  $I$  using a small number of factors, one needs to find a small number of cuboids in  $I$  which are full of 1s and cover all the entries of  $I$  with 1s.

We say that a cuboid  $J$  is contained in  $I$  if  $J_{ijt} \leq I_{ijt}$  for all  $i, j, t$ . As the following theorem shows, triadic concepts of  $I$  correspond to maximal cuboids contained in  $I$ .

**Theorem 2.**  $\langle D_1, D_2, D_3 \rangle$  is a triadic concept of  $I$  if and only if  $J = \circ(c(D_1), c(D_2), c(D_3))$  is a maximal cuboid contained in  $I$  (i.e., any other cuboid which is contained in  $I$  is also contained in  $J$ ). Here,  $c(D_i)$  denotes the characteristic vector of  $D_i$ , i.e.  $c(D_i)(x) = 1$  iff  $x \in D_i$ .

We are going to use triadic concepts of  $I$  for decompositions of  $I$  the following way. For a set

$$\mathcal{F} = \{\langle D_{11}, D_{12}, D_{13} \rangle, \dots, \langle D_{k1}, D_{k2}, D_{k3} \rangle\}$$

of triadic concepts of  $I$ , we denote by  $A_{\mathcal{F}}$  the  $n \times k$  matrix in which the  $l$ -th column consists of the characteristic vector  $c(D_{l1})$  of the extent  $D_{l1}$  of  $\langle D_{l1}, D_{l2}, D_{l3} \rangle$ ,  $B_{\mathcal{F}}$  the  $m \times k$  matrix in which the  $l$ -th column consists of the characteristic vector  $c(D_{l2})$  of the intent  $D_{l2}$  of  $\langle D_{l1}, D_{l2}, D_{l3} \rangle$ ,  $C_{\mathcal{F}}$  the  $p \times k$  matrix in which the  $l$ -th column consists of the characteristic vector  $c(D_{l3})$  of the modus  $D_{l3}$  of  $\langle D_{l1}, D_{l2}, D_{l3} \rangle$ . That is,

$$(A_{\mathcal{F}})_{il} = \begin{cases} 1 & \text{if } i \in (D_{l1}), \\ 0 & \text{if } i \notin (D_{l1}), \end{cases} \quad (B_{\mathcal{F}})_{jl} = \begin{cases} 1 & \text{if } j \in (D_{l2}), \\ 0 & \text{if } j \notin (D_{l2}), \end{cases} \quad (C_{\mathcal{F}})_{tl} = \begin{cases} 1 & \text{if } t \in (D_{l3}), \\ 0 & \text{if } t \notin (D_{l3}). \end{cases}$$

If  $I = \circ(A_{\mathcal{F}}, B_{\mathcal{F}}, C_{\mathcal{F}})$ ,  $\mathcal{F}$  can be seen as a set of factors which fully explain the data. In such a case, we call the triadic concepts from  $\mathcal{F}$  *factor concepts*. Given  $I$ , our aim is to find a small set  $\mathcal{F}$  of factor concepts.

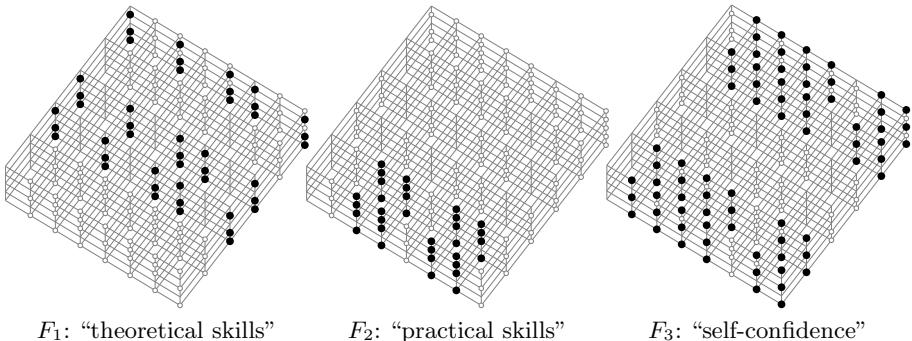
Using triadic concepts of  $I$  as factors is intuitively appealing because triadic concepts are simple models of human concepts according to traditional logic approach [13]. In fact, factors are often called “(hidden) concepts” in the ordinary factor analysis. In addition, the extents, intents, and modi of the concepts, i.e. columns of  $A_{\mathcal{F}}$ ,  $B_{\mathcal{F}}$ , and  $C_{\mathcal{F}}$ , have a straightforward interpretation: they represent the objects, attributes, and conditions to which the factor concept applies (see Section 3 for particular examples).

The next result says that triadic concepts of  $I$  are universal factors.

**Theorem 3 (universality).** For every  $I$  there is  $\mathcal{F} \subseteq \mathcal{T}(X, Y, I)$  such that  $I = \circ(A_{\mathcal{F}}, B_{\mathcal{F}}, C_{\mathcal{F}})$ .

The following theorem may be considered the main result. It says that, as far as exact decompositions of  $I$  are concerned, triadic concepts are optimal factors in that they provide us with decompositions of  $I$  with the least number  $k$  of factors.

**Theorem 4 (optimality).** If  $I = \circ(A, B, C)$  for  $n \times k$ ,  $m \times k$ , and  $p \times k$  binary matrices  $A$ ,  $B$ , and  $C$ , there exists a set  $\mathcal{F} \subseteq \mathcal{T}(X, Y, I)$  of triadic concepts of  $I$  with  $|\mathcal{F}| \leq k$  such that for the  $n \times |\mathcal{F}|$ ,  $m \times |\mathcal{F}|$ , and  $p \times |\mathcal{F}|$  matrices  $A_{\mathcal{F}}$ ,  $B_{\mathcal{F}}$ , and  $C_{\mathcal{F}}$  we have  $I = \circ(A_{\mathcal{F}}, B_{\mathcal{F}}, C_{\mathcal{F}})$ .



**Fig. 1.** Geometric meaning of factors as maximal cuboids

This means that when looking for decompositions of  $I$ , one can restrict the search to the set of triadic concepts instead of the set of all possible decompositions.

### 3 Illustrative Example

In this section, we present an illustrative example of factorization. We consider input data containing information about students and their performance in various courses. Such data is frequently obtained from student evaluation and recommendation systems that are used during the process of admission to universities. Factor analysis of this type of data can help reveal important factors describing skills of students under various conditions.

Our model data is represented by a triadic context  $\langle X, Y, Z, I \rangle$  where  $X = \{a, b, \dots, h\}$  is a set of students (objects);  $Y = \{\text{co, cr, di, fo, in, mo}\}$  is a set of student qualities (attributes): communicative, creative, diligent, focused, independent, motivated; and  $Z = \{\text{AL, CA, CI, DA, NE}\}$  is a set of courses passed by the students (conditions): algorithms, calculus, circuits, databases, and networking. The fact that  $x$  is related with  $y$  under  $z$  is interpreted so that "student  $x$  showed quality  $y$  in course  $z$ ". We consider  $I$  given by the following table:

	AL	CA	CI	DA	NE
	0 0 0 0 0 0	0 0 0 0 0 0	0 0 0 0 0 0	0 0 0 0 0 0	0 0 0 0 0 0
a	1 1 1 1 1 1	0 0 1 1 0 1	1 1 0 0 1 1	0 0 1 1 0 1	1 1 0 0 1 1
b	1 1 0 0 1 1	0 0 0 0 0 0	1 1 0 0 1 1	1 1 0 0 0 0	1 1 0 0 1 1
c	1 1 1 0 0 1	0 0 1 1 1 1	1 1 0 0 0 0	1 1 1 1 0 1	1 1 0 0 0 0
d	1 1 1 1 1 1	0 0 1 1 0 1	1 1 0 0 1 1	0 0 1 1 0 1	1 1 0 0 1 1
e	1 1 0 0 1 1	0 0 0 0 0 0	1 1 0 0 1 1	1 1 0 0 0 0	1 1 0 0 1 1
f	1 1 1 1 1 1	0 0 1 1 0 1	1 1 0 0 1 1	1 1 1 1 1 0 1	1 1 0 0 1 1
g	1 1 0 0 1 1	0 0 0 0 0 0	1 1 0 0 1 1	0 0 0 0 0 0	1 1 0 0 1 1
h	0 0 1 1 0 1	0 0 1 1 0 1	0 0 0 0 0 0	0 0 1 1 0 1	0 0 0 0 0 0

The rows of the table correspond to students, the columns correspond to attributes under the various conditions (courses). The triadic context  $\langle X, Y, Z, I \rangle$  contains 14 triadic concepts:

$$\begin{aligned}
D_1 &= \langle \emptyset, \{\text{co, cr, di, fo, in, mo}\}, \{\text{AL, CA, CI, DA, NE}\} \rangle, \\
D_2 &= \langle \{\text{f}\}, \{\text{co, cr, mo}\}, \{\text{AL, CI, DA, NE}\} \rangle, \\
D_3 &= \langle \{\text{c, f}\}, \{\text{co, cr, di, fo, mo}\}, \{\text{AL, DA}\} \rangle, \\
D_4 &= \langle \{\text{b, c, e, f}\}, \{\text{co, cr}\}, \{\text{AL, CI, DA, NE}\} \rangle, \\
D_5 &= \langle \{\text{a, d, f}\}, \{\text{mo}\}, \{\text{AL, CA, CI, DA, NE}\} \rangle, \\
D_6 &= \langle \{\text{a, d, f}\}, \{\text{co, cr, di, fo, in, mo}\}, \{\text{AL}\} \rangle, \\
D_7 &= \langle \{\text{a, c, d, f}\}, \{\text{co, cr, di, fo, mo}\}, \{\text{AL}\} \rangle, \\
D_8 &= \langle \{\text{a, c, d, f, h}\}, \{\text{di, fo, mo}\}, \{\text{AL, CA, DA}\} \rangle, \\
D_9 &= \langle \{\text{a, b, d, e, f, g}\}, \{\text{co, cr, in, mo}\}, \{\text{AL, CI, NE}\} \rangle, \\
D_{10} &= \langle \{\text{a, b, c, d, e, f, g}\}, \{\text{co, cr}\}, \{\text{AL, CI, NE}\} \rangle, \\
D_{11} &= \langle \{\text{a, b, c, d, e, f, g}\}, \{\text{co, cr, mo}\}, \{\text{AL}\} \rangle, \\
D_{12} &= \langle \{\text{a, b, c, d, e, f, g, h}\}, \emptyset, \{\text{AL, CA, CI, DA, NE}\} \rangle, \\
D_{13} &= \langle \{\text{a, b, c, d, e, f, g, h}\}, \{\text{mo}\}, \{\text{AL}\} \rangle, \\
D_{14} &= \langle \{\text{a, b, c, d, e, f, g, h}\}, \{\text{co, cr, di, fo, in, mo}\}, \emptyset \rangle.
\end{aligned}$$

Following the observations from Section 2, it suffices to take  $\mathcal{F} = \{D_1, \dots, D_{14}\}$  as the set of factor concepts which then yields a factorization of  $I$  into an  $8 \times 14$  object-factor matrix  $A_{\mathcal{F}}$ , a  $6 \times 14$  attribute-factor matrix  $B_{\mathcal{F}}$ , and a  $5 \times 14$  conditions-factor matrix  $C_{\mathcal{F}}$ . However, there exists a smaller set  $\mathcal{F}$  of factor concepts consisting of

$$\begin{aligned}
F_1 &= D_8 = \langle \{\text{a, c, d, f, h}\}, \{\text{di, fo, mo}\}, \{\text{AL, CA, DA}\} \rangle, \\
F_2 &= D_4 = \langle \{\text{b, c, e, f}\}, \{\text{co, cr}\}, \{\text{AL, CI, DA, NE}\} \rangle, \\
F_3 &= D_9 = \langle \{\text{a, b, d, e, f, g}\}, \{\text{co, cr, in, mo}\}, \{\text{AL, CI, NE}\} \rangle.
\end{aligned}$$

If we fix the order of objects, attributes, and conditions in sets  $X$ ,  $Y$ , and  $Z$ , respectively, we can uniquely represent subsets of object, attributes, and conditions by characteristic vectors. For instance, we let  $\text{a} < \text{b} < \dots < \text{h}$ , i.e.,  $\text{a}$  has index 1,  $\text{b}$  has index 2, etc. Similarly, we assume  $\text{co} < \text{cr} < \dots < \text{mo}$  and  $\text{AL} < \text{CA} < \dots < \text{NE}$ . As a consequence, extents, intents, and modi of  $F_1, F_2, F_3$  can be represented by characteristic vectors as follows:

$$\begin{aligned}
F_1 &= \langle 10110101, 001101, 11010 \rangle, & F_2 &= \langle 01101100, 110000, 10111 \rangle, \\
F_3 &= \langle 11011110, 110011, 10101 \rangle.
\end{aligned}$$

Using  $\mathcal{F} = \{F_1, F_2, F_3\}$ , we obtain the following  $8 \times 3$  object-factor matrix  $A_{\mathcal{F}}$ ,  $6 \times 3$  attribute-factor matrix  $B_{\mathcal{F}}$ , and  $5 \times 3$  conditions-factor matrix  $C_{\mathcal{F}}$ :

$$A_{\mathcal{F}} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \quad B_{\mathcal{F}} = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}, \quad C_{\mathcal{F}} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}.$$

One can check that  $I = \circ(A_{\mathcal{F}}, B_{\mathcal{F}}, C_{\mathcal{F}})$ , i.e.,  $I$  decomposes into three (two-dimensional) matrices using three factors. Note that the meaning of the factors can be seen from the extents, intents, and modi of the factor concepts. For instance,  $F_1$  applies to students  $a, c, d, f, h$  who are diligent, focused, and motivated in algorithms, calculus, and databases. This suggests that  $F_1$  can be interpreted as “having good background in theory / formal methods”. In addition,  $F_2$  applies to students who are communicative and creative in algorithms, circuits, databases, and networking. This may indicate interests and skills in “practical subjects”. Finally,  $F_3$  can be interpreted as a factor close to “self-confidence” because it is manifested by being communicative, creative, independent, and motivated. As a result, by finding the factors set  $\mathcal{F} = \{F_1, F_2, F_3\}$ , we have explained the structure of the input data set  $I$  using three factors which describe the abilities of student applicants in terms of their skills in various subjects.

Let us recall that the factor concepts  $\mathcal{F} = \{F_1, F_2, F_3\}$  can be seen as maximal cuboids in  $I$ . Indeed,  $I$  itself can be depicted as three-dimensional box where the axes correspond to students, their qualities, and courses. Figure 1 shows the three factors depicted as cuboids. White and black circlets in Figure 1 correspond to elements in  $I$ . Namely, a circlet is present on the intersection of  $x \in X$ ,  $y \in Y$ , and  $z \in Z$  in the diagram iff  $\langle x, y, z \rangle \in I$ . Furthermore, the circlet is black iff  $\langle x, y, z \rangle \in I$  belongs to the factor  $F_i$  which is iff  $x$  belongs to the extent of  $F_i$ ,  $y$  belongs to the intent of  $F_i$ , and  $z$  belongs to the modus of  $F_i$ .

## 4 Algorithm

We now outline an algorithm for computing the matrix decompositions described in Section 2. Since the problem of finding a minimal decomposition of  $\langle X, Y, Z, I \rangle$  is reduced to a problem of finding a minimal subset  $\mathcal{F} \subseteq \mathcal{T}(X, Y, Z, I)$  of formal concepts which cover the whole set  $I$ , we can reduce the problem of finding a matrix decomposition to the set-covering problem. The universe  $U$  that should be covered corresponds to  $I \subseteq X \times Y \times Z$ . The family  $\mathcal{S}$  of subsets of the universe  $U$  that is used for finding a cover is, in fact, the set of all triadic concepts  $\mathcal{T}(X, Y, Z, I)$ . More precisely,  $\mathcal{S} = \{A \times B \times C \mid \langle A, B, C \rangle \in \mathcal{T}(X, Y, Z, I)\}$ . In this setting, we are looking for  $\mathcal{C} \subseteq \mathcal{S}$  as small as possible such that  $\bigcup \mathcal{C} = U$ . Thus, finding factor concepts is indeed an instance of the set-covering problem. The set covering optimization problem is NP-hard and the corresponding decision problem is NP-complete. However, there exists an efficient greedy approximation algorithm for the set covering optimization problem which achieves an approximation ratio  $\leq \ln(|U|) + 1$ , see [6].

Algorithm 1, implementing the above-mentioned greedy approach in our setting, computes a set of factor concepts by first computing the set of all triadic concepts which are stored in  $\mathcal{S}$ , see lines 1–8, and then iteratively selecting formal concepts from  $\mathcal{S}$ , maximizing their overlap with the remaining tuples in  $U$ , see lines 9–17. Notice that the triadic concepts are computed by a reduction to the dyadic case [9]. In line 2, we iterate over all concepts in  $\mathcal{B}(X, Y \times Z, I^X)$

**Algorithm 1.** COMPUTEFACTORS( $X, Y, Z, I$ )

---

```

1  /* compute a set  $\mathcal{S}$  of all triadic concepts */
2  set  $\mathcal{S}$  to  $\emptyset$ ;
3  foreach  $\langle A, J \rangle \in \mathcal{B}(X, Y \times Z, I^X)$  do
4    |  foreach  $\langle B, C \rangle \in \mathcal{B}(Y, Z, J)$  do
5      |    if  $A = (B \times C)^{(X)}$  then
6        |      add  $\langle A, B, C \rangle$  to  $\mathcal{S}$ ;
7      |    end
8    |  end
9  end
10 /* compute a set  $\mathcal{F}$  of factor concepts */
11 set  $\mathcal{F}$  to  $\emptyset$ ;
12 set  $U$  to  $I$ ;
13 while  $U \neq \emptyset$  do
14   |  select  $\langle A, B, C \rangle \in \mathcal{S}$  which maximizes  $|U \cap (A \times B \times C)|$ ;
15   |  add  $\langle A, B, C \rangle$  to  $\mathcal{F}$ ;
16   |  set  $U$  to  $U \setminus (A \times B \times C)$ ;
17   |  remove  $\langle A, B, C \rangle$  from  $\mathcal{S}$ ;
18 end
19 return  $\mathcal{F}$ 

```

---

where  $I^X = \{\langle x, \langle y, z \rangle \rangle \mid \langle x, y, z \rangle \in I\}$ , cf. **K<sup>(1)</sup>** in Section 2. In line 3, we iterate over all concepts in  $\mathcal{B}(Y, Z, J)$  where  $J$  was obtained as an intent in the previous line. The condition in line 4 is needed to check whether  $A$  is maximal, i.e., whether  $\langle A, B, C \rangle$  is a triadic concept. Notice that **9** computes triadic concepts by an analogous reduction which utilizes two nested NEXTCLOSURE algorithms, however, arbitrary algorithm for computing dyadic formal concepts can do the job (e.g., CbO, Lindig's algorithm), see **III** for a survey and comparison.

We have observed by experiments that Algorithm **1** computes nearly optimal sets of factor concepts in terms of their sizes. Because of the limited scope of the paper, we postpone detailed performance evaluation of the algorithm to a full version of the paper.

## 5 Further Issues

Future work will include the following topics: Algorithms and experiments (we proposed a greedy algorithm, based on the idea from our **3**; the algorithm need not compute all triadic concepts, instead it computes good factor concepts one by one directly from the data, resulting in a high speed-up; a paper is in preparation); approximate factorization (decompositions for which  $\circ(A, B, C)$  approximately equals  $I$ ); complexity and approximability of the problem of finding decompositions; extension to ordinal data (see **1,2,4** for the case of two-way data).

## References

1. Belohlavek, R.: Optimal decompositions of matrices with grades. In: IEEE IS 2008, pp. 15-2–15-7 (2008)
2. Belohlavek, R.: Optimal triangular decompositions of matrices with entries from residuated lattices. *Int. J. of Approximate Reasoning* 50(8), 1250–1258 (2009)
3. Belohlavek, R., Vychodil, V.: Discovery of optimal factors in binary data via a novel method of matrix decomposition. *J. Computer and System Sci.* 76(1), 3–20 (2010)
4. Belohlavek, R., Vychodil, V.: Factor analysis of incidence data via novel decomposition of matrices. LNCS (LNAI), vol. 5548, pp. 83–97. Springer, Heidelberg (2009)
5. Cichocki, A., et al.: Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation. J. Wiley, Chichester (2009)
6. Cormen, T.H., et al.: Introduction to Algorithms, 2nd edn. MIT Press, Cambridge (2001)
7. Frolov, A.A., et al.: Boolean factor analysis by Hopfield-like autoassociative memory. *IEEE Trans. Neural Networks* 18(3), 698–707 (2007)
8. Ganter, B., Wille, R.: Formal Concept Analysis. Mathematical Foundations. Springer, Berlin (1999)
9. Jäschke, R., Hotho, A., Schmitz, C., Ganter, B., Stumme, G.: TRIAS – An Algorithm for Mining Iceberg Tri-Lattices. In: Proc. ICDM 2006, pp. 907–911 (2006)
10. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. *SIAM Review* 51(3), 455–500 (2009)
11. Kuznetsov, S., Obiedkov, S.: Comparing performance of algorithms for generating concept lattices. *J. Exp. Theor. Artif. Intelligence* 14(2-3), 189–216 (2002)
12. Lee, D., Seung, H.: Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–791 (1999)
13. Lehmann, F., Wille, R.: A triadic approach to formal concept analysis. In: Ellis, G., Rich, W., Levinson, R., Sowa, J.F. (eds.) ICCS 1995. LNCS, vol. 954, pp. 32–43. Springer, Heidelberg (1995)
14. Miettinen, P., Mieličáinen, T., Gionis, A., Das, G., Mannila, H.: The Discrete Basis Problem. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) PKDD 2006. LNCS (LNAI), vol. 4213, pp. 335–346. Springer, Heidelberg (2006)
15. Stockmeyer, L.J.: The set basis problem is NP-complete. IBM Research Report RC5431, Yorktown Heights, NY (1975)
16. Tatti, N., Mieličáinen, T., Gionis, A., Mannila, H.: What is the dimension of your binary data? In: Perner, P. (ed.) ICDM 2006. LNCS (LNAI), vol. 4065, pp. 603–612. Springer, Heidelberg (2006)
17. Wille, R.: The basic theorem of triadic concept analysis. *Order* 12, 149–158 (1995)

# Application of Ontological Engineering in Customs Domain

Panagiotis Loukakos and Rossitza Setchi

Cardiff School of Engineering, Cardiff University, Queen's Building, Cardiff, CF24 3AA  
{loukakosp, setchi}@cf.ac.uk

**Abstract.** Customs is considered a very complex business domain in terms of concepts, processes and rules. The complexity is increased because a number of actors are involved in various activities such as National Administrations and Economic Operators. All these should share a common understanding about various concepts. The paper presents a research work which examines the use of ontological engineering as a tool to represent and share various terms and knowledge in the area of Customs. The developed ontology is expressed in OWL syntax. Some illustrative examples are provided demonstrating how the complex relationships between various Customs concepts could be modelled.

**Keywords:** Ontology, Customs, Risk Analysis, OWL.

## 1 Introduction

Customs plays a vital role in the economy, environment and security. The main role of EU Customs is to facilitate trade and at the same time to protect the interests of the European Union and its citizens. Currently twenty-seven national administrations implement the community customs code ([1], [2]). The Customs Union is an important element for the operation of the single market, which requires common rules to be applied at its external borders. Those common rules cover aspects such as common tariff, health and environment controls, protection of economic interests, etc. However, nowadays, Customs are facing new challenges and in order to achieve these demands, modernization of customs procedures and controls is required as well as cooperation of various services [3]. In the context of Customs modernization, the electronic customs initiative started with the aim of creating a paperless environment [3]. The information exchange between customs IT systems is considered an essential element so as the EU's economy to continue to compete in a global context [3].

Customs is a very complex business domain due to the number and the nature of its processes, the number and complexity of business rules that govern the processes, the number of the actors involved, as well as the number of terms and the concepts that are used in the various procedures. Normally, a customs procedure is performed in more than one location, which means that the processes are distributive. A strong collaboration and common understanding of procedures and business rules is required for the performance of customs business since a number of actors such as Economic Operators and Customs Authorities are involved in the completion of customs procedures. In the context of electronic customs, Information is exchanged among those

actors and hence it is believed that everyone should share the same understanding about the various concepts. In addition, a number of entities (e.g. declarations, authorizations and guarantees) is required for the completion of a customs procedure. A large number of business rules has to be satisfied in order to proceed with the trading of goods. Thus, the representation of knowledge for that domain is considered a challenging task.

According to [4], *Ontology is the term used to refer to the shared understanding of some domain of interest which may be used as a unifying framework to solve the above problems in the above described manner.* Ontologies have various usages [4], [5]. One of the most important is the *Communication* between people, groups or roles with different needs so as ensuring that all parties share the same understanding. In reality, ontologies enable consistent knowledge representation as well as avoiding any ambiguity. Finally, it enables all parties to have at the end the same perception for various concepts.

The purpose of developing an ontology for EU Customs is to represent the Customs concepts with formal representation and to model the complex relationships that exist in this domain. This activity also helps to examine the use of ontological engineering as a tool to represent and share various terms and knowledge in the area of Customs. It is believed that such knowledge representation could be of vital importance considering the complexity of the customs domain as well as the requirement for strong collaboration between the National Administrations and Economic Operators. It can facilitate the knowledge and experience exchange at various levels. This ontology will be used in the context of research work for risk assessment in Customs. Hence, it is currently focused more on concepts related to risk analysis. It is foreseen to be used as an upper layer (more generic) of a more specialised ontology for risk assessment in Customs domain. However, the ontology can be elaborated to fully depict the various concepts of Customs business.

The rest of the paper is organised as follows. Section 2 contains a brief overview of related work in ontologies development for Customs domain. Section 3 describes the ontological engineering approach followed in this study and section 4 presents some illustrative examples of modelled knowledge from the developed ontology for EU Customs demonstrating how the ontology could be used for common understanding and knowledge sharing. Finally, conclusions are presented in the last section of the paper.

## 2 Related Work

Two related works have been found on domain ontologies for Customs [6] and [7]. In the first research, domain ontology for import and export has been developed to acquire Harmonised System (HS) codes for given products. In fact, the ontology is used for reasoning and particularly for specifying intelligently the HS code of a given product based on its product name. According to the authors of this work, the ontology is intended to be used by the Customs and quarantine departments in order to automate and improve their inspections processes since the HS code can be used to identify the applied policies to the product. Hence, accurate assignment of an HS code to a product implies more efficient and effective inspection. It is believed that the

work as currently presented in [6] concerns only some Tariff aspects and particularly the HS Nomenclature in Customs business.

The second work [7] presents an ontology which includes the Customs Domain Concepts and the Risk Assessment Ontology. This ontology developed for RACWeb project, which co-funded by the European Commission under the “Information Society Technology” Programme, Framework Programme 6. In this work, two main ontologies developed. The first one is the “Customs Ontology” and the second is the “Risk Assessment Ontology”. The “Customs Ontology” consists of three layers. The first layer defines some general concepts, the second layer specifies some Customs-specific concepts and finally the third layer extends the second layer with some very specific concepts for Inward Processing and Export customs procedures. The ontologies of third layer were further specialised with some national-specific customs ontologies. Finally, the “Risk Assessment” Ontology has also a layered architecture. The first layer is a general “Risk Assessment” Ontology and the second layer contains a more specialised ontology of “Customs Risk Assessment”. According to the authors of this work, the purpose of this ontology was to store information and then to use this information for performing risk assessment with some rules expressed in Semantic Web Rule Language (SWRL). It is believed that this work [7] focuses more on data modelling of customs-related information. This can also been seen by examining the various relationships between concepts, which have been defined from a data-modelling perspective. Moreover, the Customs ontology is limited to the inward processing and to export customs procedures. It is worth noting that OWL-DL was also used for expressing the Customs domain specific ontology (as the ontology presented in this paper).

The ontology developed by the authors of this paper attempts to cover more concepts of customs for a more generic purpose, although currently it mainly focuses more on the risk analysis area. It is envisaged to be enriched with both data properties presenting attributes of concepts and object properties modelling complex relationships between concepts, which cannot be easily identified by reading documentation such as legislations. Moreover, the ontology architecture presented in [7] is similar to the one will be followed in this research meaning that layered approach will be followed.

### 3 Ontological Engineering Approach

The approach towards ontological engineering proposed by Uschold and Gruninger [4] has been used in this work. The purpose and the scope of this ontology for EU Customs have been expressed in section 1. However, it is worth noting that the scope of this activity is not to develop a complete ontology for EU Customs. It can be considered as a first step for a future work. Continuous refinement is very important for the ontology’s effectiveness. Regarding the building of the ontology, the following aspects should be mentioned:

- **Ontology Capture:** The ontology was developed using various sources of information (libraries, online resources, existing knowledge, etc.) such as [1-2, 8-15]. During this step, a number of concepts were defined, a hierarchy of concepts was built, descriptions were added for concepts and finally the relationships among them were defined. As it is mentioned at the beginning, the ontology focuses on risk analysis.

- **Ontology Coding:** It is very important to choose the appropriate representation language for expressing the ontology of a specific domain [4] and [16]. The Ontology Web Language (OWL) has been used for building the ontology being discussed in this paper. OWL facilitates greater machine interpretability of the content and therefore enables the processing of the content from computers as well as the performance of reasoning tasks [17]. The OWL provides the following three sub-languages: OWL-Lite, OWL-DL and OWL-Full. Each sub-language has different expressiveness. The OWL-Lite is the least expressing sub-language while the OWL-Full is the most expressive sub-language. On the other hand, the OWL-DL sub-language is more expressive than OWL-Lite and less expressive than OWL-Full.

The OWL-DL sub-language has been used for the purpose of this ontology. The complex nature of Customs domain requires quite expressive language for describing the various concepts and the relationships among them. Therefore, the OWL-Lite could not be used for the ontology for EU Customs due to its simplicity in terms of expressiveness. On the other hand, the OWL-Full is very expressive sub-language; however, one of its main disadvantages is the low decidability due to its power and hence, no efficient reasoning can be currently performed [18]. As a conclusion, the OWL-DL has been selected for producing the ontology. Although, the ontology is not used in this research for automated processing, the motivation is to represent the knowledge with formal language for future use. Therefore, this version of the ontology could be considered as starting point for further work. The selection of OWL-DL enables the use of this ontology in the future for automated reasoning and machine processing.

The ontology coding was performed using Protégé v4.0.2 (build 115). The generated OWL code is compliant to OWL 2.0 [19] and has been generated with OWL API (version 2.2.1.1138) provided with Protégé tool. Moreover, the FACT++ reasoner provided by Protégé tool has been used for some classification.

- **Integrating existing ontologies:** No integration was performed with other technologies.

The development of ontology is an iterative process. Therefore, the ontology is continuously updated and verified. According to [20], Ontology can be evaluated by the development team, by other development teams, and by end users or experts. Each actor validates it from different perspective. Normally, the development team focus the evaluation on the technical properties of the concepts whereas end users evaluate the actual value and correctness of defined concepts within a given organization or domain. Some evaluation of the specific ontology has been performed by the development team. The FACT++ reasoner has also been used for the validation of the ontology. However, it is worth noting that the ontology has not been validated from any Customs Authority or any other Customs organisation. It has been developed based on the sources mentioned in *Ontology Capture* bullet above hence this is the frame of reference for the technical evaluation. Moreover, it is envisaged this ontology to be further enriched which implies that the continuous evaluation of the ontology is required. The outcome of this work is envisioned to be used in a broader research for risk analysis in Customs.

## 4 Ontology for EU Customs

The ontology for EU Customs has taken advantage of all OWL components for representing various concepts of Customs business. Three annotation types have been used as attributes for the definition of classes. These are the *label*, *comment* and *source* attributes. The *comment* annotation has been used to provide a small description about the specific class and hence the user of the ontology to be able to understand the various business concepts. An example is the *comment* annotation '*means the act whereby a person indicates in the prescribed form and manner a wish to place goods under a given customs procedure, with an indication, where appropriate, of any specific arrangements to be applied*' for the "Customs Declaration" class. The *source* annotation mainly indicates the knowledge source from which this class captured. For the specific example, the '*Article 4(10) of Modernised Customs Code No 450/2008*' was the source for the "Customs Declaration" class. Some illustrative examples from the developed ontology are presented and discussed in the following paragraphs.

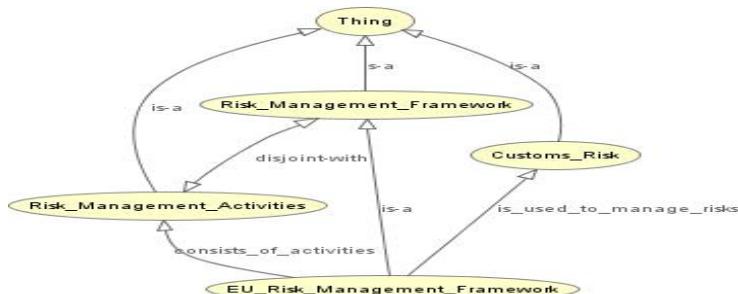
The ontology can be used to present the hierarchical structure of various entities. Focusing on the "Risk Management Framework" and specifically the "EU Risk Management Framework" class, it can be realised that the "EU Risk Management Framework" consists of some activities and that manages the Customs Risk by using the OWL component *object property*. The object properties are used to express the various relationships between the concepts/classes of the ontology. In this case, the "EU Risk Management Framework" class has the object properties '*consists\_of\_activities*' and '*is\_used\_to\_manage\_risks*'. The first one specifies that the "EU Risk Management Framework" consist of a number of activities, which are defined by the entity "Risk Management activities". The second object property defines that the "EU Risk Management Framework" is used to manage the "Customs Risk" entity (see Fig. 1).

The "Risk Management Framework" is a disjoint class with "Risk Management Activities" class. The latter class aims to represent the various activities that shall be performed in the context of a risk management (see Fig. 1). As it is shown in Fig. 2, the ontology models the main activities of risk management as defined in [8]. Hence, the "Risk Analysis" activity consists of the "Identify Risk Data", "Analyse Risks" and "Weigh Risks" sub-activities. The "Weigh Risks" sub-class has been defined as a value partition of "High Risk", "Medium Risk" and "Low Risk" aiming to model the classification of customs risk (see Fig. 3). The value partition is considered as a design pattern and it is not part of OWL [21]. This pattern has been used to restrict the values of "Weigh Risks" and indicate that it has equivalent class the "High Risk or Low Risk or Medium Risk".

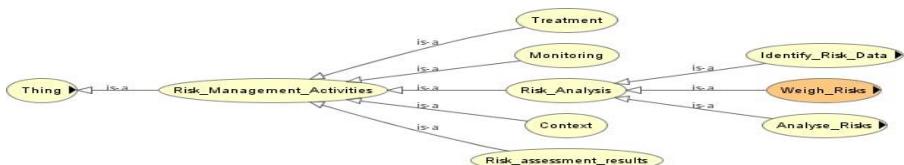
Moreover, the hierarchy of "Analyse Risks" class depicts that two types of analysis shall be performed; analysis on proven risks and on potential risks (see Fig. 4). This is also reflected by looking at the "Customs Risk" concept, which classifies the risks into potential and proven risks [8], [15] (see Fig. 5). The ontology presents the fact that the "Analyse Proven Risks" and "Analyse Potential Risks" activities are used to analyse the "Proven Risks" and "Potential Risks" respectively through OWL object properties. In this particular case, the object property '*is\_used\_to\_analyse\_potential\_risk*' relates the "Analyse Potential Risks" class with the "Potential Risks" class. The same applies for the object property '*is\_used\_to\_analyse\_proven\_risks*', which links the "Analyse Proven Risks" class with the "Proven Risks" class.

The equivalent class feature has also been used in the ontology to indicate whether one entity is the same with another because it is often two different names to be used for the same concept. The “Analyse Risk of consignment” class denotes the activity, which is performed for analyzing the risk of a consignment using the declaration data [8], [9]. This is modelled in the ontology with two object properties. The first one (‘analyse\_risk\_of\_consignment’) relates the “Analyse Risk of consignment” class with the “Consignment” class while the second one (‘is\_based\_for\_analysing\_the\_risk\_of\_consignment\_on’) relates the “Analyse Risk of consignment” class to “Customs Declaration” or “Summary Declaration” classes. Finally, the object property ‘is\_considered\_for\_the\_control\_decision’ between the “Analyse Risk of consignment” class and the “Control consignment” class verifies the fact that outcome of risk analysis of consignment will be one of the criteria for selecting to perform movement inspection [2]. In any case, the “Control consignment” (consignment inspection) activity is considered as part of the Risk Management activity and particular of the Treatment activity [8], [15]. This has been modelled in the ontology with the object property ‘is\_part\_of\_treatment’ between the “Control consignment” class and the “Treatment” class (Fig. 6).

Finally, the representation of complex structure and relationships was achieved with the development of the ontology. Apart from the inheritance, which can be shown through classes/sub-classes, a number of relationships could be defined between entities. An example is the “Person established in the Community” class, which has relationships between the entities “Natural Person”, “Legal Person” and “Association of Persons”. This happens because a Person established in the Community can be a natural person, legal person or association of persons considering that also satisfy the conditions defined in the legislation [2] (Fig. 7).



**Fig. 1.** Relationships of EU Risk Management Framework class with other concepts



**Fig. 2.** Risk Management Activities class hierarchy with two levels of children



Fig. 3. Weigh Risks class hierarchy

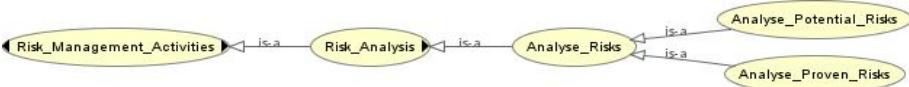


Fig. 4. Analyse Risks class hierarchy

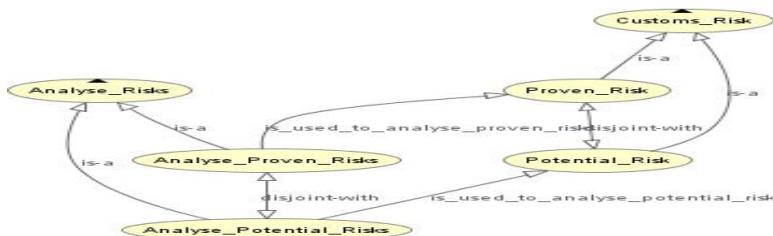


Fig. 5. Relationships between Analyse Risks and Customs Risk classes

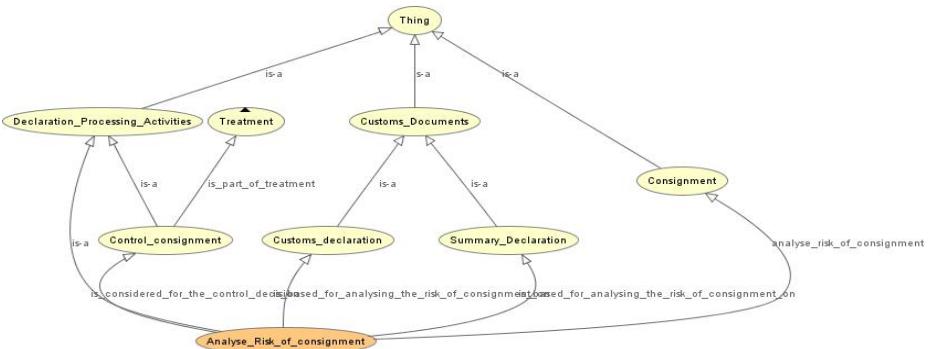
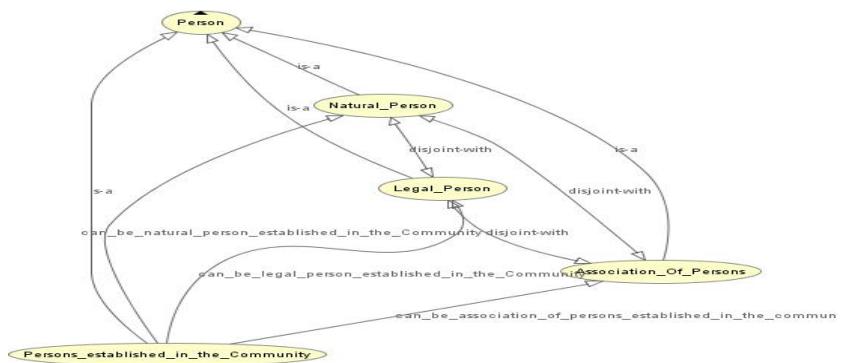


Fig. 6. Relationships of Analyse Risk of Consignment class with other concepts

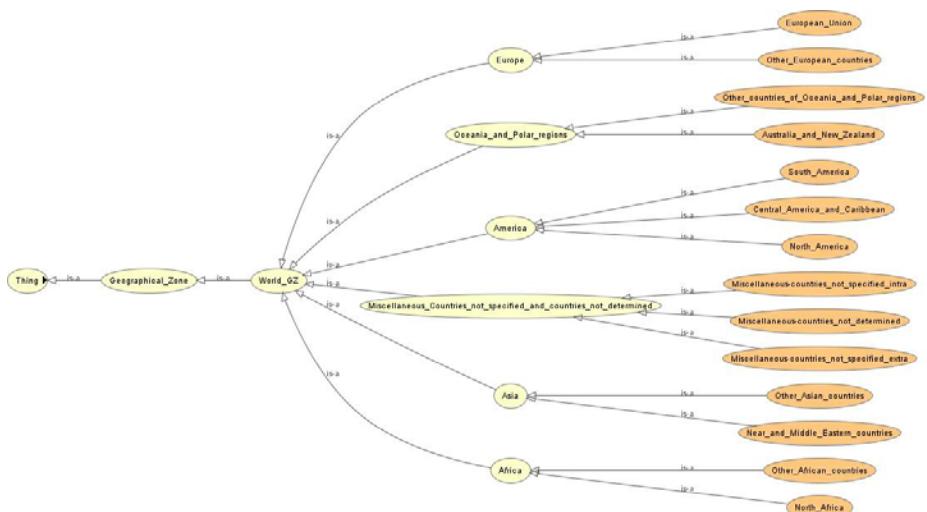
The ontology has also defined information about the Economic zones and Geographical zones [11] (Fig. 8). Such information could be useful in Ontology for Customs domain. In order to achieve the modelling of the aforementioned information in the ontology, the Geographical zones and Economic zones were represented as *enumerated* classes. This implies that all countries should be modelled as individuals. These individuals are members of the "Country" class, which also have three Data properties (*EU\_Code\_Name*, *ISO-3166\_Code\_Name* and *EU\_Geonomencalature\_Code*). Hence,

for each individual (specific country) of type “Country”, the specific data properties were also defined. Using the FACT++ reasoner integrated with Protégé tool, a new classification was inferred showing the relationship between Geographical zones and Economic zones. This relationship inferred by examining the enumeration of various classes. Some equivalent classes were inferred (same enumerations) or new classifications created (classes/subclasses) based on whether the enumeration of one class (a set of individuals-countries) was subset of another.

Finally, as it is mentioned above, all individuals are initially defined with type “Country” class. Following the reasoning, new types per individual was inferred based on their membership in the various enumerated classes (e.g. a specific Geographical zone).



**Fig. 7.** Relationships between sub-classes of Person concept



**Fig. 8.** Geographical zone class hierarchy

## 5 Conclusions and Future Work

This paper presents the ontology for EU Customs being developed under a research work for Risk Analysis in Customs domain. It demonstrates how the ontology can be used for communication between the various actors involved in the Customs business as well as in Customs' Information Systems. It enables the representation and sharing of various concepts in the area of Customs. In particular, the ontology shows clearly the classification of various Customs entities, the complex relationships that exist among the various entities. Moreover, it is shown how some classes can be inferred based on their definition (characteristics). The current version of the ontology for EU Customs is considered as a first attempt to model the knowledge of such a complex domain. It must be enriched with more concepts and relationships. It requires a significant effort in order to fully represent the knowledge of Customs and consider the ontology complete. Generally, the Knowledge Acquisition phase is a challenging task for such complex business domains. Moreover, the applicability of rules could be considered as a future work. Finally, formal validation of the ontology and continuous refinement is very important for the ontology's effectiveness.

## References

1. EEC: Council Regulation (EEC) No. 2913/92 of October 12, 1992 establishing the Community Customs Code. Official Journal, 88 (1992)
2. EEC: REGULATION (EC) No. 450/2008 of The European Parliament and oF The Council of April 23, 2008 laying down the Community Customs Code (Modernised Customs Code). Official Journal, 64 (2008)
3. DGTAXUD: Policy Issues,  
[http://ec.europa.eu/taxation\\_customs/customs/policy\\_issues/index\\_en.htm](http://ec.europa.eu/taxation_customs/customs/policy_issues/index_en.htm)
4. Uschold, M., Gruninger, M.: Ontologies: Principles, methods and applications. *Knowledge Engineering Review* 11, 93–136 (1996)
5. Studer, R., Benjamins, V.R., Fensel, D.: Knowledge engineering: Principles and methods. *Data & Knowledge Engineering* 25, 161–197 (1998)
6. Zang, B., Li, Y., Xie, W., Chen, Z., Tsai, C.-F., Laing, C.: An ontological engineering approach for automating inspection and quarantine at airports. *J. Comput. Sci.* 74, 196–210 (2008)
7. Dimakopoulos, T., Kassis, K.: Domain specific ontology - Risk Assessment for Customs in Western Balkans. RACWeB - IST PROJECT 045101 (2008)
8. DGTAXUD: Standarised Framework for Risk Management in the Customs Administrations of the EU. EUROPA- Taxation and Customs Union (2003),  
[http://ec.europa.eu/taxation\\_customs/resources/documents/framework\\_doc.pdf](http://ec.europa.eu/taxation_customs/resources/documents/framework_doc.pdf)
9. EEC: Commission Regulation (EEC) No. 2454/93 of July 2, 1993 laying down provisions for the implementation of Council Regulation (EEC) No. 2913/92 establishing the Community Customs Code, p. 779 (1993)
10. EEC: COMMISSION REGULATION (EC) No. 1192/2008 of November 17, 2008 amending Regulation (EEC) No. 2454/93 laying down provisions for the implementation of Council Regulation (EEC) No. 2913/92 establishing the Community Customs Code. Official Journal, 51 (2008)

11. EUROSTAT: Geonomenclature 108 (2005)
12. EEC: Commission Regulation (EC) No 1833/2006 of December 13, 2006 on the nomenclature of countries and territories for the external trade statistics of the Community and statistics of trade between Member States, 19–28 (2006)
13. ISO: ISO 3166-1-alpha-2 code In: (ISO), I.O.f.S. (ed.), Vol. 2008 (2006)
14. DGTAXUD: Customs Glossary,  
[http://ec.europa.eu/taxation\\_customs/common/glossary/customs\\_index\\_en.htm](http://ec.europa.eu/taxation_customs/common/glossary/customs_index_en.htm)
15. DGTAXUD: Risk Management for Customs in the EU,  
[http://ec.europa.eu/taxation\\_customs/customs/customs\\_control\\_s/risk\\_management/customs\\_eu/index\\_en.htm](http://ec.europa.eu/taxation_customs/customs/customs_control_s/risk_management/customs_eu/index_en.htm)
16. Shanks, G., Tansley, E., Weber, R.: Using ontology to validate conceptual models. Commun. ACM 46, 85–89 (2003)
17. Smith, M.K., Welty, C., McGuinness, D.L. (eds.): OWL Web Ontology Language Guide. W3C Recommendation (February 10, 2004)
18. Antoniou, G., Harmelen, F.v.: A Semantic Web Primer. MIT press, London (2004)
19. W3C\_OWL\_Working\_Group (ed.): OWL 2 Web Ontology Language Document Overview. W3C Recommendation (October 27, 2009)
20. Gomez-Perez, A.: Some ideas and examples to evaluate ontologies. In: Proceedings of the 11th Conference on Artificial Intelligence for Applications. IEEE Computer Society, Los Alamitos (1995)
21. Jupp, S., Moulton, G., Rector, A., Stevens, R., Wroe, C.: A Practical Guide To Building OWL Ontologies Using Protege 4 and CO-ODE Tools. The University Of Manchester (2007)

# Classification and Prediction of Academic Talent Using Data Mining Techniques

Hamidah Jantan<sup>1</sup>, Abdul Razak Hamdan<sup>2</sup>, and Zulaiha Ali Othman<sup>2</sup>

<sup>1</sup> Faculty of Computer and Mathematical Sciences  
Universiti Teknologi MARA (UiTM) Terengganu,  
23000 Dungun, Terengganu, Malaysia

hamidahjtn@tganu.uitm.edu.my

<sup>2</sup> Faculty of Information Science and Technology  
Universiti Kebangsaan Malaysia (UKM)  
43600 Bangi, Selangor, Malaysia  
{arh, zao}@ftsm.ukm.my

**Abstract.** In talent management, process to identify a potential talent is among the crucial tasks and need highly attentions from human resource professionals. Nowadays, data mining (DM) classification and prediction techniques are widely used in various fields. However, this approach has not attracted much interest from people in human resource. In this article, we attempt to determine the potential classification techniques for academic talent forecasting in higher education institutions. Academic talents are considered as valuable human capital which is the required talents can be classified by using past experience knowledge discovered from related databases. As a result, the classification model will be used for academic talent forecasting. In the experimental phase, we have used selected DM classification techniques. The potential technique is suggested based on the accuracy of classification model generated by that technique. Finally, the results illustrate there are some issues and challenges rise in this study, especially to acquire a good classification model.

**Keywords:** Classification, Prediction, Data mining (DM), Academic talent.

## 1 Introduction

Classification in data mining(DM) is among the popular machine learning technique for knowledge discovery and future prediction. This method is categorized as supervised learning, where the class level or classification goal is already known. Afterward, classification model made by this techniques will be used for related future prediction [1, 2]. Nowadays, there are many areas adapting this approach such as in finance, medical, marketing, stock, telecommunication, manufacturing, health care, education, customer relationship and many others [3, 4]. Nevertheless, this technique has not attracted much attention of Human Resource (HR) people [5]. The valuable data in HR databases can provide a rich resource for knowledge discovery and for decision support tool development.

Recently, among the challenges of HR professionals are managing an organization talent. Talent is defined as an outcome to ensure the right person is in the right job at

the right time [6, 7]. This task involves a lot of human decisions; which sometimes are very uncertain and difficult. Human decisions depend on various factors like human experience, knowledge, preference and judgment. In fact, the process to identify an existing talent in organization highly requires human decisions and this task is among the prominent talent management issues and challenges [7, 8]. Academicians in higher education institution are evaluated based on their performance in academic position, and the position represents the academic talent that he/she has. For that reason, in this study, we attempt to use classification techniques by dealing with academic talent position as a class level. In the experimental phase, academic talent data was collected from selected higher education institutions, and the data will be used as training and test datasets. At the end, the purpose of this paper is to suggest the possible classification techniques for academic talent forecasting through some experiments using the selected classification algorithms.

This paper is organized as follows. The second section describes the related work on DM in HR especially for talent management, classification and prediction in DM; and some of the possible classification techniques. The third section discusses on the experiment setup in this study. Section 4 shows some experiments results and analysis. Finally, the paper ends at Section 5 with the concluding remarks and suggestion for future research direction.

## 2 Related Work

### 2.1 DM in Human Resource

DM approach has evolved various techniques such as database oriented techniques, statistics, machine learning, pattern recognition, knowledge discovery and others. DM tasks for knowledge discovery are generally categorized as clustering, association, classification and prediction [1-3, 5]. In fact, DM technique has been applied in many fields, but its application in HR is very rare [3, 5]. Recently, there are some researches that show interest in solving HR problems using this approach [3, 5, 9-11]. Table 1 lists some of the applications in HR that uses DM techniques which have been discussed very limitedly in HR field. In HR, DM technique is quite popular in some personnel selection problems such as to choose a right candidate for a job, assigning project, employee training needs and etc. However, prediction applications in HR are infrequent, and there are some examples of applications such as to predict the length of service, sales premium, persistence indices of insurance agents and analyze miss-operation behaviors of operators and others [3]. Nevertheless, there are few discussions on talent management such as talent forecasting, project assignment, career path development and talent recruitment using DM approach [3, 10, 11]. Due to these reasons, this study attempts to use DM techniques for talent forecasting. The selected classification techniques will be identified through some experiments. At the end, the generated model or classification rules from selected classifier will be used to predict the potential talent which is based on past experience knowledge. This method will help decision maker consistently in selecting or evaluating proses especially to choose the right person for a job.

**Table 1.** Some of HR Tasks using DM Techniques

<b>HR Task</b>	<b>DM Technique</b>
Project Assignment	<i>Fuzzy Data Mining and Fuzzy Artificial Neural Network</i> [10]
Employee selection and Job Attitudes Study	<i>Decision tree</i> [3] <i>Fuzzy Data Mining</i> [12]
Training	<i>Association rule mining</i> [13]
Recruit and Retain Talent	<i>Rough Set Theory</i> [11]

## 2.2 Talent Management and DM

In any organization, talent management has become an increasingly crucial approach to determine the success of HR functions. Talent is considered as the capability of any individual to make a significant difference to the current and future performance of the organization [7, 14]. However, managing talent involves human resource planning that emphasizes processes for managing people in organization. Talent management can be defined as a process to ensure leadership continuity in key positions and encourages individual advancement; and decision to manage supply, demand and flow of talent through human capital engine [6]. Nowadays, talent management is very crucial and needs direct attention from HR professionals. TP Track Research Report has found that among the top current and future talent management challenges are developing existing talent; forecasting talent needs; attracting and retaining the right leadership talent; engaging talent; identifying existing talent; attracting and retaining the right leadership and key contributor; deploying existing talent; lack of leadership capability at senior levels and ensuring a diverse talent pool [8]. The talent management process consists of recognizing the key talent areas in the organization, identifying the people in the organization who constitute key talent, and conducting development activities for the talent pool to retain and engage them and also have them ready to move into more significant roles [6]. These processes involve HR activities that need to be integrated into an effective system [15].

Recently, with the new demand and increased visibility, HR seeks a more strategic role by turning to DM methods [5]. In that case, DM techniques can be used to discover useful knowledge from patterns identification from the existing data in HR databases. Thus, this study concentrates on identifying the patterns that relate to the talent management. The talent patterns can be generated by using some of the major DM techniques such as clustering to list the employees with similar characteristics, to group the performances and etc. From the association technique, patterns that are discovered can be used to associate the employee's profile for the most appropriate program/job, employee's attitude to performance and etc. In prediction and classification, the pattern can be used to predict the percentage of accuracy in employee's performance, behavior, and attitudes, predict the performance progress throughout the performance period, and also identify the best profile for different employee and etc. [9, 16]. The matching of DM problems and talent management needs is very crucial. Therefore, it is very important to determine the suitable DM techniques [9]. Due to

that reason, this study focuses on one of the talent management challenges i.e. to identify existing talent regarding the key talent in an organization by predicting their talent. In this case, we use the past data from the employee databases to implement the classification process. The classification rules generated from selected classifier will be used to forecast the potential talent.

### **2.3 Classification and Prediction in DM**

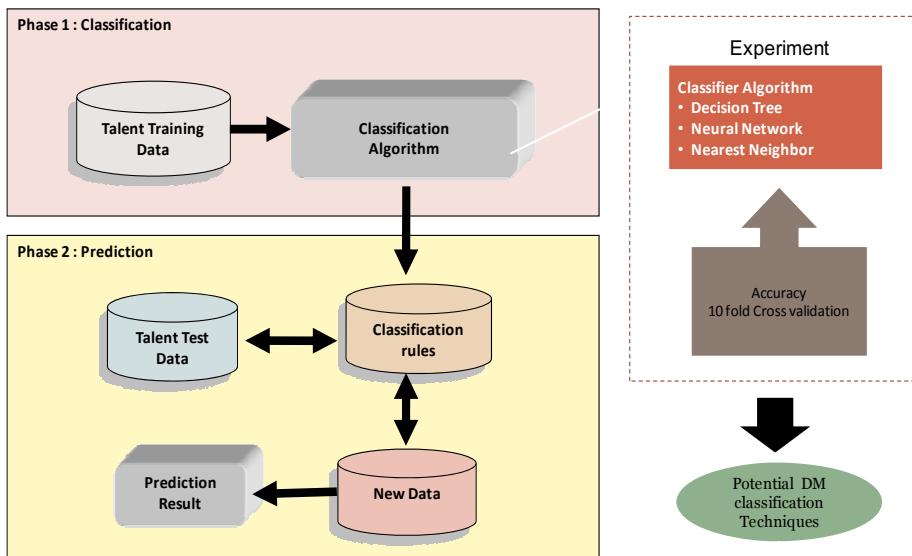
Database or data warehouse is rich with hidden information that can be used to provide intelligent decision making. Intelligent decision refers to the ability to make automated decision that is closely similar to human decision [2]. Classification and prediction abilities are among the methods that can produce intelligent decision. Currently, many classification and prediction methods have been proposed by researchers in machine learning, pattern recognition, and statistics. We focus our discussion on classification and prediction in machine learning. Classification and prediction in DM are two forms of data analysis that can be used to extract models to describe important data classes or to predict future data trends [1]. The classification process has two phases; the first phase is learning process whereby training data are analyzed by classification algorithm. Learned model or classifier is represented in the form of classification rules. The second phase is classification, and test data are used to estimate the accuracy of classification rules. If the accuracy is considered acceptable, the rules can be applied to the classification of new data.

In most cases, techniques used for data classification are decision tree, Bayesian methods, Bayesian network, rule-based algorithms, neural network, support vector machine, association rule mining, k-nearest-neighbor, case-based reasoning, genetic algorithms, rough sets and fuzzy logic. In this study, our discussion focuses on three classification techniques i.e. decision tree, neural network and k-nearest-neighbor. However, decision tree and neural network are found useful in developing predictive models in many fields [17]. The advantage of decision tree technique is that it does not require any domain knowledge or parameter setting, and is appropriate for exploratory knowledge discovery. The second technique is neural-network which has high tolerance of noisy data as well as the ability to classify pattern on which they have not been trained. It can be used when we have little knowledge of the relationship between attributes and classes. Next, the K-nearest-neighbor technique is an instance-based learning using distance metric to measure the similarity of instances. All these three classification techniques have their own advantages and disadvantages [2]. In this study, we attempt to use all these major classification techniques for academic talent data in order to propose the suitable techniques for talent forecasting.

## **3 Experiment Setup**

The purpose of this experiment is to identify talent patterns from related HR databases for academic talent forecasting. There are several classification techniques used in order to handle this issue. The fundamental process of classification and prediction in DM and our experiment setting are shown in Fig 1. In this experiment, the selected classification techniques are based on the common techniques for classification and

prediction especially in DM. The first classification technique chosen is neural network which is quite popular in DM community and used as pattern classification technique [2, 12, 17]. The second technique is decision tree known as ‘divide-and-conquer’ approach from a set of independent instances and the third technique is nearest neighbor that is based on the distance metric. In this study, we attempt to use C4.5/J4.8 and Random Forest(RF) for decision tree; Multilayer Perceptron (MLP) and Radial Basic Function Network (RBFC) for neural network; and K-Star for the nearest neighbor technique which are summarized in Table 2.



**Fig. 1.** Classification and Prediction in DM

The input variables in the classification process are academic talent attributes; and the outcome of this process is academic position. All the components and attributes involved in this experiment are shown in Fig 2. In this study, the performance attributes are extracted from the yearly performance appraisal records, previous knowledge and expertise records. In this experiment, we have five training datasets which contain 17 attributes for related performance components. The data are collected from academic Curriculum Vitae (CV) of Information Technology lecturers from 20 public universities in Malaysia. The descriptions about the datasets are demonstrated in Table 3.

However, due to the confidentiality and security of data, for the exploratory purposes, we simulate 199 performance data that are based on the academic talent factors as first dataset. The second, third and forth datasets are from actual data which are collected from academic CV of the respective universities. In this study, we categorize university according to the status of the university either Research University (RU) or Non Research University (NRU). This status is given to the university based on their experiences and contributions in higher learning education. The classification process used is based on 10 fold cross validation for training and test dataset [2].

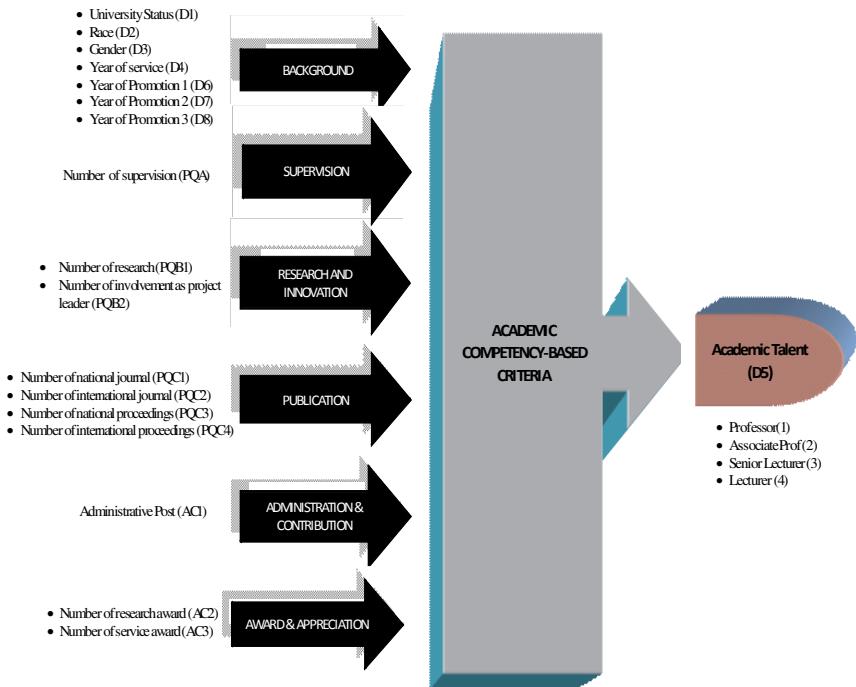
**Table 2.** Potential Classification Techniques

<i>DM Techniques</i>	<i>Classification Algorithm</i>
<i>Decision Tree</i>	<ul style="list-style-type: none"> <li>• <b>C4.5</b> (Decision tree induction – the target is nominal and the inputs may be nominal or interval. Sometimes the size of the induced trees is significantly reduced when a different pruning strategy is adopted).</li> <li>• <b>Random forest</b> (Choose a test based on a given number of random features at each node, performing no pruning. Random forest constructs random forest by bagging ensembles of random trees).</li> </ul>
<i>Neural Network</i>	<ul style="list-style-type: none"> <li>• <b>Multi Layer Perceptron</b> (An accurate predictor for underlying classification problem. Given a fixed network structure, we must determine appropriate weights for the connections in the network).</li> <li>• <b>Radial Basic Function Network</b> (Another popular type of feed forward network, which has two layers, not counting the input layer, and differs from a multilayer perceptron in the way that the hidden units perform computations).</li> </ul>
<i>Nearest Neighbor</i>	<ul style="list-style-type: none"> <li>• <b>K*Star</b> (An instance-based learning using distance metric to measure the similarity of instances and generalized distance function based on transformation)</li> </ul>

**Table 3.** Dataset Description

<i>Data</i>	<i>Description</i>	<i>Number of data</i>
<i>Dataset1</i>	Simulated Data	199
<i>Dataset2</i>	All academic CV data	295
<i>Dataset3</i>	RU academic CV	135
<i>Dataset4</i>	NRU academic CV	160
<i>Dataset5</i>	All academic CV – Academic contribution by Ratio (Year of Service)	295

In this experiment, the DM tools used are WEKA and ROSETTA toolkit. Besides that, this experiment attempts to identify the possible techniques using selected classifier algorithm for full attributes of datasets. In this case, we use all attributes that we have defined previously for full dataset. This experiment focuses on the accuracy of the classifier in order to identify the suitable classifier algorithm for the datasets. The accuracy of classifiers is based on the percentage of test set samples which are correctly classified.



**Fig. 2.** Selected Attribute for Academic Talent

## 4 Results and Discussions

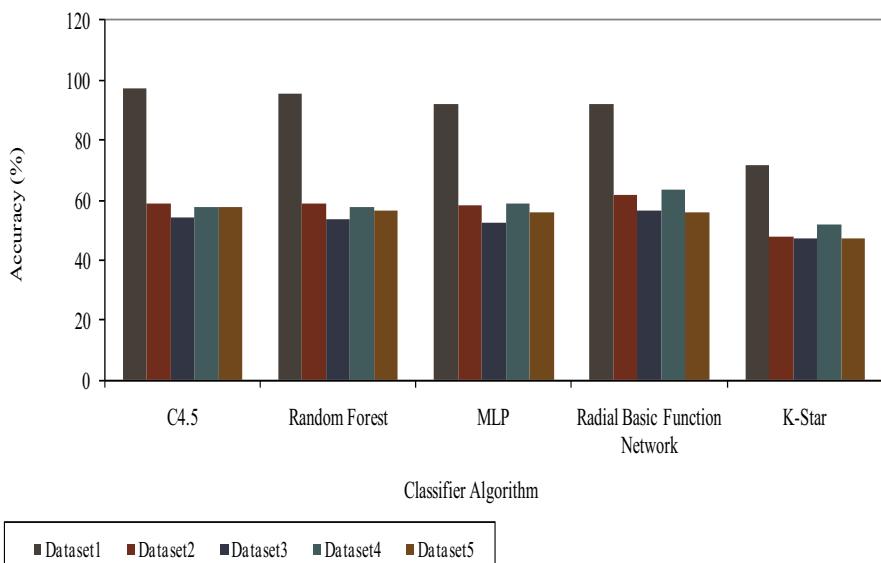
In this experiment, the accuracy of classification techniques is determined based on the 10 fold cross validation for selected classifier algorithm. As a result, the accuracy of generated model for all datasets is shown in Table 4. Besides that, the trend of accuracy for each classifier is shown in Figure 3. The result illustrated the highest accuracy of the model is the first dataset which contains simulated data. On the other hand, the accuracy of real data for datasets is slightly low compared to the simulated data, which is most of them are less than 60%. In this experiment, the accuracy for each classifiers is not far different for each dataset. In DM classification process, especially in model construction, the accuracy of model should be higher or acceptable enough in order to produce a good model for forecasting. In classifier analysis, the C4.5/J4.8 classifier has the highest accuracy for *dataset1*, *dataset3* and *dataset5*; and the RBFN classifier has the highest accuracy for *dataset2* and *dataset4*. However, the difference of accuracy among them is quite minimal for each dataset.

This experiment results demonstrate some issues and challenges raised when the classification process implemented to the actual data. There are some possible reasons regarding this results. Firstly, the importance of whole attributes should be taken into consideration, especially in attribute selection because in some way this factor will affect the accuracy of model. In this experiment, the training dataset is for full attribute without looking at attribute reduction. The process to determine the importance of

attribute can be implemented by using attribute reduction techniques such as Boolean reasoning, Genetic algorithm and others. Secondly, the distribution of data and the proses of handling missing value in data preprocessing need to be re-evaluated in order to get good datasets. These factors should be considered in future work in order to improve the accuracy of model. In addition, from this experiment, we observe the great potential to use C4.5 classification technique in the next stage of DM process which is known as prediction. Besides that, the result also indicates the suitability of the decision tree classifier for the datasets.

**Table 4.** The Accuracy of Model for Datasets

<i>Classifier Algorithm</i>	<i>Dataset</i>				
	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>
C4.5 /J4.8	97.44%	58.95%	54.43%	57.52%	57.90%
Random forest	95.31%	58.95%	53.79%	58.03%	56.88%
Multi Layer Perceptron(MLP)	92.28%	58.24%	52.36%	59.00%	56.16%
Radial Basis Function Network	92.30%	61.91%	54.40%	63.31%	56.23%
K-Star	71.81%	48.00%	47.31%	51.73%	47.24%



**Fig. 3.** Accuracy of Model for Each Classifier

## 5 Conclusions and Future Work

This article has clarified the significance of the study on selected DM classification techniques for academic talent forecasting. However, other DM techniques such as Support Vector Machine (SVM), Fuzzy logic and Artificial Immune System (AIS) should also be considered for future work using the same dataset. The attribute relevancy is considered as a factor on the accuracy of the classifier algorithm. In future work, the attribute reduction process should be implemented using any reduction techniques to validate these findings. The important attribute can be determined using some techniques such as C4.5 decision tree. The technique is used to identify the important attribute by using the concept of information gain. Thus, the entire result will indicate whether the number of attributes will affect the accuracy of the classifier or not.

The C4.5 classifier has the highest accuracy in the experiment; in next experiment, other decision tree techniques should be evaluated using the same method. Besides that, the significance or hypothesis testing for the results is needed, especially to justify the findings. In this experiment, C4.5 classifier algorithm is the potential classifier for future work. Thus, this technique will be used to construct the classification rules and the rules can be used to predict the potential academic talent. In conclusion, the ability to continuously change and obtain new understanding of the classification and prediction in HR researches has thus, lead to as imperative contribution to DM in HR.

## Acknowledgement

This research was conducted as a part of the eScience project funded by MOSTI (Ministry of Science, Technology and Innovation), Malaysia (01-01-01-SF0236).

## References

1. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publisher, San Francisco (2006)
2. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publisher, San Francisco (2005)
3. Chien, C.F., Chen, L.F.: Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems and Applications* 34, 380–390 (2008)
4. Wang, H., Wang, S.: A knowledge management approach to data mining process for business intelligence. *Industrial Management & Data Systems* 108, 622–634 (2008)
5. Ranjan, J.: Data Mining Techniques for better decisions in Human Resource Management Systems. *International Journal of Business Information Systems* 3, 464–481 (2008)
6. Cubbingham, I.: Talent Management: Making it real. *Development and Learning in Organizations* 21, 4–6 (2007)
7. Cappelli, P.: Talent Management for the Twenty-First Century, <http://www.hbr.org>
8. A TP Track Research Report Talent Management: A State of the Art. Tower Perrin HR Services (2005)

9. Jantan, H., Hamdan, A.R., Othman, Z.A.: Knowledge Discovery Techniques for Talent Forecasting in Human Resource Application. In: World Academy of Science, Engineering and Technology, Penang, Malaysia, pp. 803–811 (2009)
10. Huang, M.J., Tsou, Y.L., Lee, S.C.: Integrating fuzzy data mining and fuzzy artificial neural networks for discovering implicit knowledge. *Knowledge-Based Systems* 19, 396–403 (2006)
11. Chien, C.F., Chen, L.F.: Using Rough Set Theory to Recruit and Retain High-Potential Talents for Semiconductor Manufacturing. *IEEE Transactions on Semiconductor Manufacturing* 20, 528–541 (2007)
12. Tung, K.Y., Huang, I.C., Chen, S.L., Shih, C.T.: Mining the Generation Xer's job attitudes by artificial neural network and decision tree-empirical evidence in Taiwan. *Expert Systems and Applications* 29, 783–794 (2005)
13. Chen, K.K., Chen, M.Y., Wu, H.J., Lee, Y.L.: Constructing a Web-based Employee Training Expert System with Data Mining Approach. In: The 9th IEEE International Conference on E-Commerce Technology and The 4th IEEE International Conference on Enterprise Computing, E-Commerce and E-Services, CEC-EEE 2007 (2007)
14. Lynne, M.: Talent Management Value Imperatives: Strategies for Execution. In: The Conference Board (2005)
15. CHINA UPDATE. HR News for Your Organization: The Tower Perrin Asia Talent Management Study, <http://www.towersperrin.com>
16. Jantan, H., Hamdan, A.R., Othman, Z.A.: Data Mining Techniques for Performance Prediction in Human Resource Application. In: 1st Seminar on Data Mining and Optimization, pp. 41–49. CAIT UKM, Selangor (2008)
17. Tso, G.K.F., Yau, K.K.W.: Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy* 32, 1761–1768 (2007)

# Test-Cost Sensitive Classification on Data with Missing Values in the Limited Time

Chang Wan

School of Information Science and Technology  
SUN YAT-SEN UNIVERSITY, China  
[wanchang@mail2.sysu.edu.cn](mailto:wanchang@mail2.sysu.edu.cn)

**Abstract.** Much work [1] [2] has been done to deal with the test-cost sensitive learning on data with missing values. Most of the previous works only focus on the cost while ignore the importance of time. In this paper, we address how to choose the unknown attributes to be tested in the limited time in order to minimize the total cost. We propose a multi-batch strategy applying on test-cost sensitive Naïve Bayes classifier and evaluate its performance on several data sets. We build graphs from attributes and it includes the vertices cost and set cost. Then we use randomized algorithm to select the unknown attributes in each testing cycle. From the results of the experiments, our algorithms significantly outperforms previous algorithms[3][4].

## 1 Introduction

In many real-world applications, since different misclassification errors can cost quite differently, cost-sensitive learning has been proposed to try to minimize the total cost other than lower the error rate. Among the whole range of costs in cost-sensitive learning, two types of costs are typically considered to be the most important: misclassification costs and test costs. For example, in a binary classification task, there are costs of false positive(FP) and costs of false negative(FN). Usually, these two types of costs are not equivalent. What is more, in many cases the data is not complete. As a result, It may need time and more cost to get the missing values of the attributes to improve the quality of classification work. Unfortunately, obtaining the missing values may cost much time but in some cases time is very important. After all, it is very useful to know how to minimize the sum of misclassification costs and test costs in the limited time.

Tasks involving both misclassification costs and test costs in the limited time are abundant in real-world applications. In medical diagnosis(see Table 1), medical tests are like attributes in machine learning whose values may be obtained at cost(test costs), and misdiagnoses are like the misclassifications which may also bear a cost(misclassification costs). In some cases, a patient has a sudden hurt and must have the operation within a time(limited time). We must do some tests to get more information about the hurt in order to make the strategy safer for the operation but the tests need time and different tests need different time. If we test all the missing values at the same time, it will takes much money and

**Table 1.** An Example of a New Case Containing Missing Values and Their Associated Costs and required time for Getting a Value

Tests:	Blood Test	X-ray Test	CAT Scan	Ultrasound	MRI Scan	Diagnosis
Known:	?	Normal	?	?	Abnormal	To be decided
Costs:	\$500		\$200	\$300		MC + TC
Required Time:	20min		30min	60min		<=LT

MC is the misclassification cost, TC is the test cost and LT is the limited time.

some patients may not afford it. A solution is testing a missing value which seems to be the most useful and using the result of this test with previous information to decide the next value to be tested or stopping testing. However, a factor we cannot ignore is the time. Testing the value one by one requires lots of time. The patient may not be able to wait. So there is a confliction of time and money. In this case, we can have tests several times within the limited time and a batch of values will be tested each time. So we must decide the combination of missing values to be tested and the sets of the values to be tested together in a time to minimize the sum of test costs and misclassification costs in the limited time.

## 2 Related Work

Much work has been done in machine learning on minimize the total cost of classification. In particular, two types of costs are considered [5]:

1. Misclassification costs: these are the costs incurred by classification errors. Works such as [6][7][8] considered classification problems with non-uniform misclassification costs.
2. Test costs: these are the costs incurred for obtaining attribute values. Previous work such as [9][10] considered the test costs alone without incorporating misclassification cost. As pointed out in [11], it is obviously an oversight.

As far as we know, [3] has proposed a unified framework of test-cost sensitive classification. They applied both sequential and batch strategy to test-cost sensitive decision tree and test-cost sensitive Naïve Bayes. The former strategy is aimed at deciding testing a most useful attribute a time or stopping testing. The latter one is aimed at deciding testing some of the attributes at one time. But one factor they did not think appropriately is the time. They just thought that tests should be done one by one but this is not always true in real-world. We can choose attributes that may be tested at the same time in order to minimize the used time. So here, we proposed a strategy to try to minimize the total cost in the limited time. We emphasize that as long as the required time is shorter than the limited time, the lower the total cost( misclassification costs and the test costs) is, the better the solution is. Our study is aimed at the case which is urgent and does not have much time but there are lots of missing values.

In [12], Andrew Arnt et al. took the time cost into consideration. Andrew Arnt et al. proposed a decision theoretic approach based on a Markov Decision

Process(MDP), where they try to minimize the expected value of a cost function reflecting the quality of strategy. Two major differences of our work from theirs have to be mentioned. First, in our study, the instances are time independent and there is no effects on arriving time of instances while in [12], instances are considered to be a sequence of time-sensitive ones. Besides, we emphasize that as long as the classification is done in the given limited time, the lower the sum of misclassification cost and test cost is, the better the strategy is, whereas Andrew Arnt et al. proposed a cost function to incorporate all the three type of costs and used this function to reflecting the quality of the strategy.

### 3 Classifier

#### 3.1 The Standard Naïve Bayes Classifier

Naïve Bayes classifier is shown to perform well in practice to minimize classification errors, even in many domains containing clear attribute dependence [11]. For classification, the standard Naïve Bayes algorithm computes the posterior probability  $P(c_j|x)$  of sample  $x$  belonging to class  $c_j$  according to the Bayes rule:

$$P(c_j|x) = \frac{P(x|c_j)P(c_j)}{P(x)} \quad (1)$$

$x$  is predicted to belong to the class  $c_{j^*}$ , where  $j^* = \arg \max_j P(c_{j^*}|x)$ . When there exist missing values in sample  $x$ , the corresponding attributes are simply left out in likelihood computation and the posterior probability is computed only based on the known attributes.

However, classification errors are not the only criteria in evaluating a learned model. In practice, costs involved during classification are even more important in deciding whether the model is effective in making correct decisions. Therefore, the standard Naïve Bayes classifier should be extended to be cost-sensitive.

#### 3.2 Test-Cost Sensitive Naïve Bayes Classifier

In this paper, we consider two types of costs in Naïve Bayes classification: misclassification costs and test costs. Misclassification costs is considered when there are different types of classification errors and the costs they bring are different. The standard Naïve Bayes algorithm (NB) can be extended to take the misclassification costs into account. Suppose that  $C_{ij}$  is the cost of predicting a sample of class  $c_i$  as belonging to class  $c_j$ . The expected misclassification costs of predicting sample  $x$  as class  $c_j$  is also known as the conditional risk [4], which is defined as:

$$R(c_j|x) = \sum_i C_{ij} * P(c_i|x) \quad (2)$$

$P(c_i|x)$  is the posterior probability given by the standard NB classifier. Sample  $x$  is then predicted to belong to class  $c_{j^*}$  which has the minimal conditional risk:

$$R(c_{j^*}|x) = \min_j R(c_j|x). \quad (3)$$

## 4 Test Strategies

### 4.1 Previous Test Strategies

Two test strategies are proposed in [3]: sequential strategy and batch strategy.

Sequential test strategy is as follows. During the process of classification, based on the results of previous tests, decisions are made sequentially on whether a further test on an unknown attribute should be performed, and if so, which attribute to select. More specifically, the selection of a next unknown attribute to test is not only dependent on all the values of initially known attributes, but also dependent on the values of those unknown attributes previously tested. We noticed that they did not take the time limit into account. So we extended this strategy a little in our study in order to make it obey our rules. The change is that the time for stopping testing is that there is no useful attribute or no sufficient time for testing useful attributes. The detail of the sequential strategy is offered in [3].

Sequential strategy seems good because it chooses the predicted most useful attribute to test and each time it can use both the beginning known attributes and the previous tested attributes. But its shortage is obvious. It need much time. In some cases, because of the time limit, the sequential strategy is not suitable. As a result, some researchers proposed the batch strategy which let people make the decision that some attributes may be tested at one time. The detail of it can be obtained in [3]. Also, we extended this strategy a little in our study for the comparison. By using this strategy, we decide to test all the seemed useful attributes at the first time and do all the testing at the same time. This is totally different from the sequential strategy. Its advantage is apparent. It saves time but it is not so accurate to know whether the attributes should be tested since at first we have little information because of missing values.

### 4.2 Proposed Test Strategy

In order to make use of the two previous strategies' advantages, we propose a new strategy: multi-batch strategy. It means that we have tests several times within the time limit. A batch of unknown attributes will be tested each time. To decide which attributes should be tested each time, we do as follows. First, we build a graph  $G(V, E)$  where each vertices stands for a attribute. The graph includes two cost: vertices cost and set cost.  $Cost_i$  stands for the vertices cost for attribute  $i(A_i)$ . It can be obtained from the following equation:

$$Cost_i = \lambda / Util_i \quad (4)$$

Where  $\lambda$  is the smoothing factor. In this study we set it to 1000.  $Util_i$  can be got as follows:

$$Util_i = Gain(A, i) - C_{test}(i); \quad (5)$$

$$Gain(A, i) = C_{mc}(A) - C_{mc}(A^*); \quad (6)$$

$C_{test}(i)$  means the test cost of  $A_i$ ,  $A$  represents all the known attributes at present.  $A^* = A \bigcup A_i$ .  $C_{mc}(A)$  stands for the misclassification cost of  $A$  which can be easily got from (3). However, it is not trivial as the calculation of  $C_{mc}(A^*)$  since the value of unknown  $A_i$  is not revealed until the test is performed. We calculate it by taking expectation over all possible values of  $A_i$  as follows:

$$C_{mc}(A^*) = E[min_j(R(c_j|A \bigcup A_i)|A)] \quad (7)$$

$$= \sum_{k=1}^{\|A_i\|} (P(A_i = V_{i,k}|A) * min_j R(c_j|A, A_i = V_{i,k}));$$

$V_{i,k}$  is a value of  $A_i$ . More details about  $Util_i$  are included in [3]. Overall, by using (4~7), we can get the vertices cost. Set cost stands for the association of attributes. So, we can apply association rule to get it. In this study, we use fp-tree algorithm [13] because of its high efficiency. We simply set the support to 50%.  $Cost_{ij}$  represents the original set cost between  $A_i$  and  $A_j$  which can be got from fp-tree algorithm.

$$Cost_{ij} = \sum_m^{\|A_i\|} \sum_n^{\|A_j\|} Num(V_{i,m}, V_{j,n}); Num(V_{i,m}), Num(V_{j,n}) >= 0.5Num \quad (8)$$

$$Cost_{ij}^* = \alpha \frac{Cost_{ij}}{Num} \quad (9)$$

$Num$  is the number of instances in data set.  $V_{i,k}$  is a value of  $A_i$ . We use  $Cost_{ij}^*$  as the final set cost.  $\alpha$  is a smoothing factor we use to balance the vertices cost and set cost. In our study, we set  $\alpha$  to 800.

Each time, a subset of vertices is selected and its corresponding attributes are tested. We tried to minimize the sum of vertices costs and set costs brought by the subset in the selecting procedure. It is easy to prove that such a combinatorial optimization problem can not be solved in polynomial time. Therefore, a simple randomized algorithm is employed to provide approximations. As a result, after the graph is build, we induce randomized algorithm to initially select which attributes should be tested each time and whether the test should be stopped. The algorithm is shown in algorithm 1. Algorithm 2 shows how to use randomized algorithm in deciding the final tested attributes or stopping testing.

## 5 Experiments

### 5.1 Procedure

The experiments were in three phases: Building the model followed by using test strategies followed by comparison between 5 type strategies.

In this study, we use the training data to build the test cost sensitive Naïve Bayes classification. For comparison, we use other four methods besides multi-batch strategy, namely lazy Naïve Bayes(LNB), extracting Naïve Bayes(ENB),

---

**Algorithm 1.** Randomized algorithm(graph G(V,E),nselect,attribute)

---

**Input:**

graph G(V,E)– a graph having vertices cost and set cost from the attributes  
 nselect–number of attributes to be selected  
 attribute–an array to indicate the present known attributes.

**Output:**

bestsubset–a subset of vertices indicating which attributes are selected

---

**Steps:**

- 1: bestsubset  $\leftarrow \emptyset$ , bestres  $\leftarrow +\infty$
  - 2: **for**  $i \leftarrow 1$  to 1000
  - 3: Randomly generate  $V' \in V$ ,  $\|V'\| = nselect$   
     the corresponding attributes of  $V'$  are unknown according to attribute
  - 4: calculate the total vertices cost  $C_1$
  - 5: calculate the total set cost  $C_2$
  - 6: **if** besres  $> C_1 + C_2$
  - 7:     **then** besres  $\leftarrow C_1 + C_2$
  - 8:     bestsubset  $\leftarrow V'$
  - 9: return bestsubset
- 

---

**Algorithm 2.** Tested Attributes Decision( $RA, n\_a, util, missing\_rate, RT, LT$ )

---

**Input:**

$RA$ –randomized algorithm

$n\_a$ –number of attributes(no include class)

$util$ –a vector includes every unknown attribute's  $util$

$missing\_rate$ –a proportion of missing attributes in total attributes

$RT$ –a attributes' tests required time vector

$LT$ –limited time

---

## Steps:

- 1: Times  $\leftarrow 1$
  - 2: **while** true
  - 3: max  $\leftarrow$  max number of  $util$
  - 4: **if** max  $\leq 0$
  - 5:     **then** algorithm terminate
  - 6: nselect  $\leftarrow$  number of points which  $util > 0$
  - 7: **if** Times == 1
  - 8:     **then if** nselect  $> \frac{3}{20}n\_a * missing\_rate$
  - 9:         **then** nselect  $\leftarrow \frac{3}{20}n\_a * missing\_rate$
  - 10:     **else if** Times == 2
  - 11:         **then if** nselect  $> \frac{3}{16}n\_a * missing\_rate$
  - 12:             **then** nselect  $\leftarrow \frac{3}{16}n\_a * missing\_rate$
  - 13:         **else if** nselect  $> \frac{1}{4}n\_a * missing\_rate$
  - 14:             **then** nselect  $\leftarrow \frac{1}{4}n\_a * missing\_rate$
  - 15: Times  $\leftarrow$  Times + 1
  - 16: use  $RA$  get result
  - 17: **if** every selected attributes'  $RT[i] > LT$
  - 18:     **then** algorithm terminate
  - 19: test selected attributes which  $RT[i] \leq LT$
  - 20:      $LT \leftarrow LT - maxRT[i]$  of tested attributes
-

cost-sensitive Naïve Bayes with sequential strategy(NBS) and cost-sensitive Naïve Bayes with batch strategy(NBB). Our method is cost-sensitive Naïve Bayes with multi-bath strategy(NBMB). LNB given in [4] just predicts class labels based on the present known attributes and requires no further tests on unknown attributes. On the other hand, ENB tests all the unknown attributes and predicts without missing values. NBS and NBB are introduced before in Sec.(3.2) and more details of them can be got in [3].

Several experiments were carried out on data sets from the UCI ML repository[14]. The eight data sets are listed in Table 2. We choose these data sets because the number of their attributes, instances and class are different. So, we can prove that our algorithms can be used in many kinds of situations. The numerical attributes in data sets were discretized to 5 intervals with equal length. For convenience, we try to use the data sets without missing values except “Dermatology”. In our study, we just ignore the instances with missing values since “Dermatology” just has little instances incomplete.

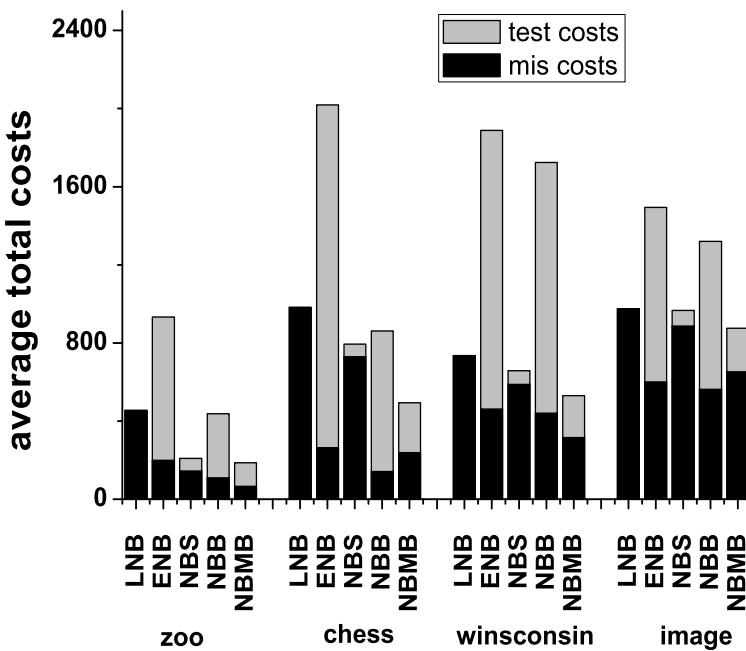
**Table 2.** Data Sets Used in the Experiments

name of data sets	number of instances	number of attributes	number of class labels
Zoo	101	17	7
Chess	3196	37	2
Breast	569	32	2
Image	2310	20	7
Dermatology	358	35	6
Wine	178	14	3
Waveform	5000	22	3
Statlog	690	14	2

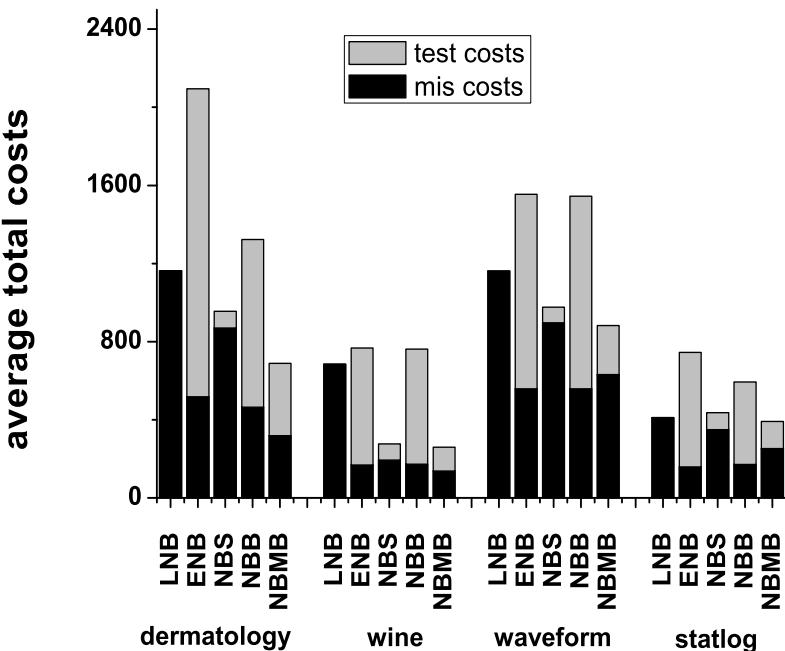
We ran a three-fold cross validation on the data sets. For the testing samples, a certain percentage (missing rate) of attributes are randomly selected and marked as unknown. If we decided to test a unknown attribute, then the real value of it was to be revealed. The performance of the algorithms is measured by the total cost, the sum of test cost and misclassification cost. The test cost of every attribute is randomly between 40~60. The required time of every attribute is randomly set between 80~120.  $LT = 2.5 \frac{\sum_{i=a}^{n-a} RT[i]}{n-a}$  since we focus on the urgent case. The missing rate is 97% which means that we hardly have any information about the case we have to predict. We simply set all the misclassification cost  $C_{MC} = 70n_a$ .

## 5.2 Results

The results are shown in Fig.1 and Fig.2. Each group of 5 bars represents the runs of 5 algorithms on one particular data set. The height of a bar represents the average total cost and, therefore, the lower the better. Each bar consists of two parts: The lower dark portion stands for the misclassification costs while the upper light portion stands for the test costs. We can see the large different of total cost from the figures.



**Fig. 1.** Experiments on first 4 data sets



**Fig. 2.** Experiments on other 4 data sets

### 5.3 Discussion

There are some very interesting observations from these experiments. First, sometimes even we known all the values of missing attributes, we can not get the minimum misclassification cost. It makes sense because from the feature selection prospective, some features are useless. Knowing all the attributes may have negative effects on classification accuracy.

Second, from the results, we can see that our algorithm NBMB performs best among the 5 algorithms. It is not unexpected for the following reasons. In our study, the case have high missing rate and short limited time. LNB offers little information, ENB costs much on testing, NBS can not take sufficient tests because the time is limited, NBB can not select more useful attributes to test because its selection is based on the initial situation. As a result, our NBMB is the most suitable algorithm among these 5 algorithms.

## 6 Conclusion and Future Work

In this paper, we proposed a strategy NBMB to deal with test-cost sensitive classification on data with missing values in the circumstances where time is limited. In this study, we focus on the case which has lots of missing values and is very urgent. Our study is very different from the previous study because we take the time limit into account. This leads our study is more proper in real world. Experiments show that our method outperforms other competing algorithms. In the future, we plan to study the missing rate, number of attributes and limited time's impact on the algorithm. For example, we may learn that with the limited time become longer, maybe the NBS shows better results than NBMB and the change of the performance of these algorithms with the mentioned factors' changes.

## Acknowledgment

I am really grateful to Mr. Ce Guo, an undergraduate student from School of Information Science and Technology at SUN YET-SEN UNIVERSITY, for giving me so many recommendations and advices about this study.

## References

1. Ling, C.X., Sheng, V.S., Yang, Q.: Test strategies for cost-sensitive decision trees. *IEEE Transactions on Knowledge and Data Engineering* 18(8) (2006)
2. Chai, X., Deng, L., Yang, Q., Ling, C.X.: Test-cost sensitive naive bayes classification. In: Fourth IEEE International Conference on Data Mining (2004)
3. Yang, Q., Ling, C., Chai, X., Pan, R.: Test-cost sensitive classification on data with missing values. *IEEE Transactions On Knowledge And Data Engineering* (5) (2006)

4. Duda, R.O., Hart, P.E., Stork, D.G.: *Pattern Classification*. Wiley and Sons, Inc., Chichester (2001)
5. Turney, P.D.: Types of cost in inductive concept learning. In: Workshop Cost-Sensitive Learning at the 17th Int'l Conf. Machine Learning (2000)
6. Elkan, C.: The foundations of cost-sensitive learning. In: 17th Int'l Joint Conf. Artificial Intelligence, pp. 973–978 (2001)
7. Domingos, P.: metacost: A general method for making classifiers cost-sensitive. *Knowledge Discovery and Data Mining*, 155–164 (1999)
8. Kai, M.T.: Inducing cost-sensitive trees via instance weighting. In: Źytkow, J.M. (ed.) PKDD 1998. LNCS, vol. 1510, pp. 139–147. Springer, Heidelberg (1998)
9. Nunez, M.: The use of background knowledge in decision tree induction. *Machine Learning*, 231–250 (1991)
10. Tan, M.: Cost-sensitive learning of classification knowledge and its applications in robotics. *Machine Learning J.*, 7–33 (1993)
11. Domingos, P., Pazzani, M.: On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 103–130 (1997)
12. Arnt, A., Zilberstein, S.: Learning policies for sequential time and cost sensitive classification. In: Proceedings of the 1st International Workshop on Utility-Based Data Mining (UBDM 2005) held with KDD 2005 (2005)
13. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: A frequent pattern tree approach. *Data Mining and Knowledge Discovery* (2004)
14. Blake, C.L., Merz, C.J.: Uci repository of machine learning databases (1998),  
<http://www.ics.uci.edu/~mlearn/MLRepository.html>

# Modified K-Means Clustering for Travel Time Prediction Based on Historical Traffic Data

Rudra Pratap Deb Nath<sup>1</sup>, Hyun-Jo Lee<sup>2</sup>, Nihad Karim Chowdhury<sup>1</sup>,  
and Jae-Woo Chang<sup>2</sup>

<sup>1</sup> Department of Computer Science & Engineering, University of Chittagong, Bangladesh  
[prataprudracsecu@gmail.com](mailto:prataprudracsecu@gmail.com), [nihad@cu.ac.bd](mailto:nihad@cu.ac.bd)

<sup>2</sup> Department of Computer Engineering, Chonbuk National University, South Korea  
[{o2near,jwchang}@chonbuk.ac.kr](mailto:{o2near,jwchang}@chonbuk.ac.kr)

**Abstract.** Prediction of travel time has major concern in the research domain of Intelligent Transportation Systems (ITS). Clustering strategy can be used as a powerful tool of discovering hidden knowledge that can easily be applied on historical traffic data to predict accurate travel time. In our Modified K-means Clustering (MKC) approach, a set of historical data is portioned into a group of meaningful sub-classes (also known as clusters) based on travel time, frequency of travel time and velocity for a specific road segment and time group. With the use of same set of historical travel time estimates, comparison is also made to the forecasting results of other three methods: Successive Moving Average (SMA), Chain Average (CA) and Naïve Bayesian Classification (NBC) method. The results suggest that the travel times for the study periods could be predicted by the proposed method with the minimum Mean Absolute Relative Error (MARE).

**Keywords:** Intelligent Transportation System (ITS), K-means Clustering, Successive Moving Average (SMA), Chain Average (CA), Naïve Bayesian Classification (NBC).

## 1 Introduction

Recently, accurate estimation of travel times has been central for traffic data analysis to various Advanced Travelers Information System (ATIS) and ITS applications such as trip planning, vehicular navigation systems and dynamic route guidance systems. Moreover, Travel time prediction is also becoming increasingly important with the development of ATIS [1]. In addition, Travel time forecasting provides information that helps travelers to decide whether they should change their routes, travel mode, starting time or even cancel their trip [2]. So, the reliable and accurate travel time prediction on road network plays an important role in any kind of dynamic route guidance systems to fulfill the users' desires. On top of that, the importance of travel time information is also indispensable to find the fastest path (i.e. shortest path according to travel time) that connects the origin and destination. Besides, accurate travel time information also helps delivery industries to progress their service quality by delivering on time.

Travel time prediction is based on vehicle speed, traffic flow and occupancy which are extremely sensitive to external event like weather condition and traffic incident [3]. Addressing the uncertainty on the road network is also a crucial issue in the research domain. Prediction on uncertain situation is very complex, so it is important to reach optimal accuracy. Yet, the structure of the traffic flow of a specific road network fluctuates based on daily, weekly and occasional events. For example, the traffic condition of weekend may differ from that of weekday. So, time-varying feature of traffic flow is one of the major issues to estimate accurate travel time [12].

In this study, we focus a new method that is able to predict travel time reliably and accurately. Generally this effort is the extension of our previous works. In this research, we have tried to combine the advantages of our previous methods namely NBC [12], SMA and CA [13] by eliminating the shortcomings of those methods. Proposed MKC method is able to address the arbitrary route on road networks that is given by user. Furthermore proposed method flushes a functional relationship between traffic data as input variables and predicted travel time as the output variables. According to the experimental result, our method exhibits satisfactory performance in terms of prediction accuracy. At the same time, the result is considered to be superior rather than other prediction methods like NBC, SMA and CA.

The format of the remaining portions of this paper is depicted as follows: Section 2 introduces some related research in this field. An outline of MKC method with example is illustrated in section 3. Section 4 exhibits a concise experimental evaluation. Finally, the conclusion statement and direction of future research is discussed in section 5.

## 2 Literature Review and Motivation

Numerous researchers have paid their attention on the accurate travel time prediction as it is one of the major issues for effective dynamic route guidance systems. Various methodologies have been investigated till date for computation and prediction of travel times with varying degree of successes. In this section, a historical background on the topic of travel time prediction is discussed briefly.

Park et al [5], [6] proposed Artificial Neural Network (ANN) models for forecasting freeway corridor travel time rather than link travel time. One model used a Kohonen Self Organizing Feature Map (SOFM) whereas the other utilized a fuzzy c-means clustering technique for traffic pattern classification. Lint et al [7], [8] proposed a state-space neural network based approach to provide robust travel time predictions in the presence of gaps in traffic data.

Kwon et al [9] focused their research on linear regression method. A linear predictor consisting of a linear combination of the current times and the historical means of the travel times was proposed by Rice et al [10]. They proposed a method to predict the time that would be needed to traverse a given time in the future. Wu et al [3] applied support vector regression (SVR) for travel time predictions and compared its results to other baseline travel-time prediction methods using real highway traffic data. Most recent research in this field has been proposed by Erick et al [11]. They investigated a switching model consisting of two linear predictors for travel time prediction.

An efficient method for predicting travel time by using NBC was proposed by Lee et al [12] which had also been scalable to road networks with arbitrary travel routes. The main idea of NBC was that it would give probable velocity level for any road segment based on historical traffic data. It was shown from experiments that NBC could reduce MARE significantly rather than the other predictors.

In our previous research, we formulated two completely new methods namely SMA and CA that were based on moving average. In that research, we eliminated the drawbacks of conventional moving average approach such as unwanted fluctuation in data set. These methods were also scalable to large network with arbitrary travel routes. Moreover, both methods were less expensive in terms of computational time. Consequently, it was revealed that these proposed methods can reduce error significantly, compared with existing methods [13].

Travel time prediction forms an integral part of any ATIS. The grouping style of whole day is efficiently and effectively done by NBC. But a significant problem will arise when we calculate velocity level for a particular route. Moreover, this method emphasize on those data whose probabilities are higher i.e. it does not concern with all data. Although NBC is capable to predict more accurately ordinary, usually it doesn't give significant result in the uncertain situation. Nevertheless, we have to compute velocity class prior probability and velocity class posterior probability. So, it takes more computational time. On the other hand, SMA and CA compute all data which are not based on probability theory. Though they provide an almost accurate travel time, they are also failed to find out uncertain data from the available traffic data. There are various knowledge based techniques that can be used in the traffic data. Clustering is one of the leading tools for discovering hidden knowledge that can be applied in the large historical traffic data set. To address the uncertain situation and predict the travel time more accurately, we propose MKC method. In this study, we try to eliminate the shortcomings of traditional K-means Clustering approach as well as NBC, SMA and CA. The key challenges of this research are to reduce prediction error as well as to predict the uncertain situation. At the same time, proposed method can also be scalable to large network with arbitrary travel routes. To motivate the presentation of proposed method, the following sections will explore complete scenario of our proposed method.

### 3 Proposed Travel Time Prediction Methods

In this section, a new method for foretelling travel time from historical traffic data using MKC method is depicted. Cluster analysis or clustering is an assignment of separating the set of observations into subset. A cluster is therefore a collection of objects which are similar between themselves and are dissimilar to the objects belonging to other clusters. K-means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well known clustering problems. The main advantages of K-means algorithm are its simplicity and speed which allows it to run on large datasets. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume K clusters) fixed a priori. The main concept is to define K centroids, one for each cluster. These centroids should be positioned in a cunning way because location verification causes result verification. So,

the better choice is to place them as much as possible far away from each other. In our MKC method, we incorporate a technique so that centroids of different clusters maintain a sufficient difference. The main disadvantage of K-means clustering is that it doesn't yield the same result with each run, since the resulting clusters depend on the initial random assignments. We have formulated MKC method in a cunning way such that we can eliminate this kind of shortcoming of well-defined K-means algorithm.

Initially, an origin with start time and destination is initialized by user. A route may consist of several road segments from origin to destination. First of all, we apply our MKC method on the data set of the first road segment to calculate the end time of first road segment which in turn becomes the start time of the next road segment. Finally, applying successive repetition approximate travel time from origin to destination can be measured.

Again, the road environment of the same road network for running vehicles on the different time periods of a day is different. So, for our convenience, the whole day time is separated into several groups according to the time as shown in Table 1.

**Table 1.** Time Group definition

Start_time_range	Time_group	Start_time_range	Time_group
06:01~10:00	1	16:01~18:00	6
10:01~11:00	2	18:01~22:00	7
11:01~12:00	3	22:01~00:00	8
12:01~14:00	4	00:01~06:00	9
14:01~16:00	5		

**Table 2.** Sample historical traffic data

Vehicle_ID	Road_ID	Time_group	Start_time	End_time	Travel_time (min)	Velocity (km/min)
1	1	6	16:50	16:57	7	1.8725
2	1	6	17:20	17:31	11	1.1916
3	1	6	17:43	17:56	13	1.0082
4	1	6	16:02	16:11	9	1.456
5	1	6	16:16	16:32	16	0.8192
6	1	6	16:05	16:18	13	1.0082
7	1	6	17:03	17:10	7	1.8725
8	1	6	17:11	17:18	7	1.8725
9	1	6	17:35	17:46	11	1.1916
10	1	6	16:09	16:16	7	1.8725

For example: if a vehicle starts from any road segment between 16:01 and 18:00, its *Time\_group* will be 6. Table 2 illustrates the sample snapshot of historical traffic data for any road segment. Each record of the table contains seven attributes. The value of *Time\_group* is calculated from the *Start\_time*. *Travel\_time* is the difference from *End\_time* to *Start\_time*. Dividing length of road segment by *Travel\_time*, *Velocity* is measured.

To calculate approximate travel time for any road segment, we introduce MKC method in the following section with appropriate example.

### 3.1 Modified K-Means Clustering Method

Step by step, MKC procedure is given below to predict travel time.

#### PROCEDURE

**Step 1:** Frequency for each travel time is measured by counting the repetition of that travel time in different records.

**Step 2:** Define Prediction relation that contains three attributes namely *Frequency*, *Travel\_time* and *Velocity*. Each tuple of Prediction relation must contain distinct travel time.

**Step 3:** Find the greatest value from the *Frequency* attribute ( $f_{\max}$ ). If two or more tuples contain the greatest value then find the greatest *Travel\_time* for available highest frequencies. A tuple  $P(x_p, y_p, z_p)$  is chosen as a centroid of *Cluster1*, where  $x_p$  is the maximum frequency,  $y_p$  is the corresponding maximum travel\_time associated with  $x_p$  and  $z_p$  is the velocity associated with travel\_time  $y_p$ .

**Step 4:** Compare each tuple  $T_i(x_i, y_i, z_i)$  of relation *Prediction* with the centroid  $P(x_p, y_p, z_p)$  of *Cluster1* by using the following formula:

$$COST(P, T_i) = |x_p - x_i| + |y_p - y_i| + |z_p - z_i| \quad (1)$$

Choose tuple  $Q(x_q, y_q, z_q)$  as the centroid of *Cluster2*, where  $COST(P, Q)$  is maximum.

**Step 5:** Build two clusters where the centroid of *Cluster1* is tuple  $P(x_p, y_p, z_p)$  and that of *Cluster2* is tuple  $Q(x_q, y_q, z_q)$ .

**Step 6:** Define the cluster memberships of tuples by assigning them to the nearest cluster representative tuple. The cost is given by Eq.1.

**Step 7:** Re-estimate the cluster centre (we consider arithmetic mean) by assuming the memberships found above are correct.

**Step 8:** Step 6 and Step 7 are repeated until no change in clusters

**Step 9:** After complete preparation of clusters, desired predicted time is calculated separately for each cluster by using the following formula:

$$\tau_i = \frac{\sum_{i=1}^N f_i * t_i}{\sum_{i=1}^N f_i} \quad (2)$$

Where,  $\tau_i$  is the travel time obtained from  $i$ -th cluster,  $N$  is the total number of tuple in associated cluster,  $f_i$  is the *Frequency* of the  $i$ -th tuple, and  $t_i$  is the *Travel\_time* of the  $i$ -th tuple.

**Step 10:** If  $\tau_1$  and  $\tau_2$  are desired travel times calculated from *Cluster1* and *Cluster2* respectively, then the final predicted approximate travel time,  $T$  for the road segment in the specific time group is the arithmetic mean of  $\tau_1$  and  $\tau_2$ .

$$\text{i.e. } T = (\tau_1 + \tau_2) / 2 \quad (3)$$

### 3.2 Explanation of MKC Method with Example

Considering the sample historical traffic data of Table 2 that contains data Road\_id =1 and Time\_group=6, the steps of MKC procedure are explained below:

**Step 1:** There are 10 records in Table 2 where *Road\_id* and *Time\_group* are common. First step of MKC reveals to find the frequency of each distinct travel time. If we observe Table 2, then we find that the frequency of *Travel\_time* 7 is four (4) because the number of repetition of *Travel\_time* 7 in different records is four. Similarly, frequencies of *Travel\_time* 16,9,13, and 11 are 1, 1, 2, and 2 respectively.

**Step 2:** *Prediction* relation is illustrated in Table 3. Each tuple in relation has three attributes namely *Frequency*, *Travel\_time* and *Velocity*. The relation also reveals that it contains only those tuples that have distinct travel time.

**Table 3.** *Prediction* relation of Table 2

Frequency	Travel_time(min)	Velocity (km/min)	Frequency	Travel_time(min)	Velocity (km/min)
1	16	0.8192	2	11	1.1916
1	9	1.456	4	7	1.8725
2	13	1.0082			

**Step 3:** The Frequency column of relation *Prediction* represents that the maximum value of it is 4. No more than one tuple contain the highest frequency. So, the centroid for *Cluster1* is the tuple  $P(x_p, y_p, z_p) = (4, 7, 1.8725)$ .

**Step 4:** Table 4 calculated the cost of each tuple  $T_i(x_i, y_i, z_i)$  from the seed of *Cluster1* by using Eq. 1.

**Table 4.** Comparison of each tuple with the centroid of *Cluster1*

Frequency	Travel_time (min)	Velocity (km/min)	Distance from (4,7,1.8725)
1	16	0.8192	$ 4-1  +  7-16  +  1.8725 - 0.8192  = 3 + 9 + 1.0533 = \mathbf{13.0533}$
1	9	1.456	$3 + 2 + 0.4165 = 5.4165$
2	13	1.0082	$2 + 6 + 0.8643 = 8.8643$
2	11	1.1916	$2 + 4 + 0.6809 = 6.6809$
4	7	1.8725	0

The maximum cost (**13.0533**) from centroid of *Cluster1* is marked as block in the Distance column of Table 4. So, the tuple  $Q(x_q, y_q, z_q) = (1, 16, 0.8192)$  is selected as the centroid of *Cluster2*.

**Step 5:** Two clusters are built where the centroid of *Cluster1* is the tuple  $P(x_p, y_p, z_p) = (4, 7, 1.8725)$  and that of *Cluster2* is the tuple  $Q(x_q, y_q, z_q) = (1, 16, 0.8192)$ .

**Step 6:** Table 5 decides the cluster memberships of tuples by assigning them to the nearest cluster representative tuple. The numbers marked as block indicate the lowest cost comparison to other. Eq. 1 is also used to find cost. 1<sup>st</sup> scenario of both clusters is shown in Table 6.

**Table 5.** Deciding cluster memberships

Frequency	Travel_time(min)	Velocity(km/min)	Distance from Cluster1 centroid (4,7,1.8725)	Distance from Cluster2 centroid (1,16,0.8192)
1	16	0.8192	$3+9+1.0533=13.053$	<b>0</b>
1	9	1.456	$3+2+0.4165=5.4165$	$0+7+0.6368=7.6368$
2	13	1.0082	$2+6+0.8643=8.8643$	$1+3+0.189=4.189$
2	11	1.1916	$2+4+0.6809=6.6809$	$1+5+0.3724=6.3724$
4	7	1.8725	<b>0</b>	$3+9+1.0533=13.0533$

**Table 6.** 1<sup>st</sup> scenario of both clusters with their members

	Frequency	Travel_time(min)	Velocity(km/min)
Cluster1	4	7	1.8725
	1	9	1.456
Cluster2	1	16	0.8192
	2	13	1.0082
	2	11	1.1916

**Step 7:** Re-estimating of new centroid for each cluster.

New centroid for Cluster1

$$\begin{aligned} P_1(x_p, y_p, z_p) &= ((4+1)/2, (7+9)/2, (1.8725+1.456)/2) \\ &= (5/2, 16/2, 3.3285/2) = (2.5, 8, 1.664). \end{aligned}$$

New centroid for Cluster2

$$\begin{aligned} Q_1(x_q, y_q, z_q) &= ((1+2+2)/3, (16+13+11)/3, (0.8192+1.0082+1.1926)/3) \\ &= (5/3, 40/3, 3.02/3) = (1.6, 13.3, 1.006). \end{aligned}$$

**Step 8:** Repetition of Step 6 with new centroids of both clusters. Blocking numbers indicate lowest cost comparing to other. Detail description illustrates in Table 7.

**Table 7.** Deciding cluster memberships with new centroids

Frequency	Travel_time(min)	Velocity(km/min)	Distance from Cluster1 new centroid (2.5,8,1.664)	Distance from Cluster2 new centroid (1.6,13.3,1.006)
1	16	0.8192	$1.5+8+0.84=10.34$	$0.6+2.7+0.189=3.489$
1	9	1.456	$1.5+1+0.208=2.708$	$0.6+4.3+0.45=5.35$
2	13	1.0082	$0.5+5+0.655=6.155$	$0.4+0.3+0.0022=0.7022$
2	11	1.1916	$0.5+3+0.4724=3.9724$	$0.4+2.3+0.1856=2.8856$
4	7	1.8725	$1.5+1+0.2085=2.7085$	$2.4+6.3+0.8665=9.5665$

Re-estimating the cluster memberships from Table 7, 2<sup>nd</sup> scenario of both clusters has been represented in Table 8.

After repetition of step 7 we get that the most recent centroids of Cluster1  $P_2(x_p, y_p, z_p)$  and Cluster2  $Q_2(x_q, y_q, z_q)$  are (2.5, 8, 1.664) and (1.6, 13.3, 1.006) respectively. The most recent centroids of both clusters are similar to the 2<sup>nd</sup> most recent centroids. So, the need of repetition of step 6 and step 7 again and again are unnecessary. Table 8 shows the final clusters.

**Table 8.** 2<sup>nd</sup> scenario of both clusters with new centroids

	Frequency	Travel_time(min)	Velocity(km/min)
Cluster1	4	7	1.8725
	1	9	1.456
Cluster2	1	16	0.8192
	2	13	1.0082
	2	11	1.1916

**Step 9:** By using Eq. 2, desired travel time from *Cluster1* and *Cluster2* can be measured

Expected Travel Time from Cluster1      Expected Travel Time from Cluster2

Here, N=2                  Here, N=3

$$\begin{aligned} \text{So, } \tau_1 &= (4*7+1*9) / (4+1) \\ &= (28+9) / 5 = 7.4 \end{aligned}$$

$$\begin{aligned} \text{So, } \tau_2 &= (1*16+2*13+2*11) / (1+2+2) \\ &= (16+26+22) / 5 = 12.8 \end{aligned}$$

So, expected travel time from *Cluster1*,  $\tau_1 = 7$  min (applying round operation) and expected travel time from *Cluster2*,  $\tau_2 = 13$  min (applying round operation)

**Step 10:** The final approximate travel time, T (for *Road\_id*=1 and *Time\_group*= 6) is predicted by using Eq. 3 such as the simple arithmetic mean of  $\tau_1$  and  $\tau_2$ . So, the final approximate travel time is  $T = ((7+13)/2)$  min =10 min.

## 4 Simulation Results

We used Pusan National University (PNU) generator [12] to measure the performance of different predictors. This generator is based on real traffic situation in Pusan city, South Korea. Global Positioning System (GPS) sensor is used to collect real traffic delay for building this well-organized PNU generator. Traffic pattern of Pusan city was extracted from this data. According to this traffic pattern, generator simulates and generates trajectory data which is almost same as real data. By using this generator, 167,669 trajectories are generated. Every trajectory may be composed of several road segments. This data organization format sufficiently reflects real traffic situations. For computing easily and efficiently and accurate evaluation of performance of the algorithms, data is divided into two categories, namely training data and test data sets. 365 days traffic data are used as training data set and 30 days traffic data are used as testing data set. Testing data sets are chronologically after 365 days data used for training. Data from 365 training days are used for fitting the model. However, 30 days test data are used to measure prediction performance for all methods.

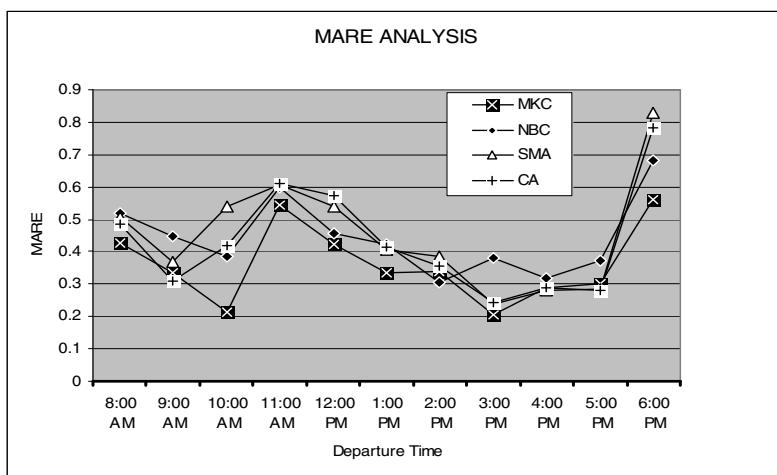
The prediction error indices, MARE are used to compare the accuracy among all prediction methods. MARE is the simplest and well-known method for measuring overall error in travel time prediction. MARE measures the magnitude of the relative error over the desired time range. The MARE is measured by the following formula:

$$MARE = \frac{1}{N} \sum_{t=1}^N \frac{|x(t) - x^*(t)|}{x(t)} \quad (4)$$

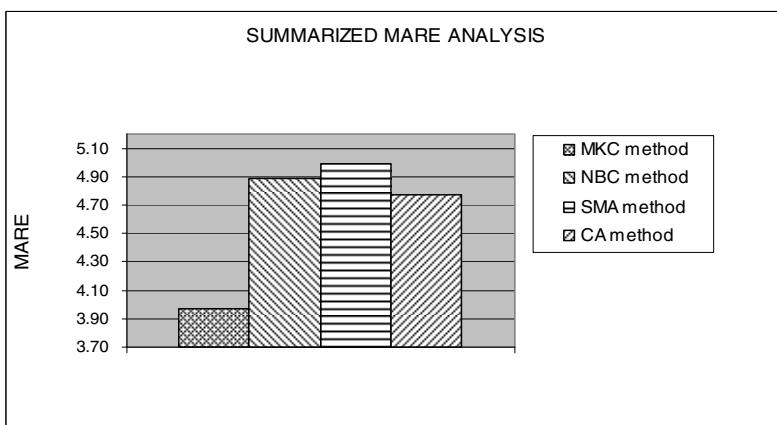
where,  $x(t)$  is the observation value,  $x^*(t)$  is the predicted value and  $N$  is the number of samples. In experimental evaluation, proposed methods are tested against other predictors like NBC, SMA and CA. Prediction errors of all predictors from 8 AM to 6 PM are examined. There are 11 test cases evaluated between 8 AM and 6 PM.

The line chart shown in Fig.1 illustrates relative performance of all travel time predictors. From the overall point of view, proposed method performs much better than NBC, SMA and CA method. In case of MKC method, it is shown that seven test cases exhibit errors less than 0.40. At 10.00 AM, 3.00 PM and 6.00 PM our method MKC predicted more accurately than others and datasets of those period included uncertain data. By contrary, NBC, SMA and CA outperform our method in one, two and one cases respectively but that are slight differences.

Summarized result of MARE for different travel time predictors are shown in Fig. 2. MARE of MKC, NBC, SMA and CA are 3.96, 4.891, 4.99 and 4.76 respectively. Thus, our proposed method reduces MARE from NBC, SMA and CA method by 19%, 20% and 17% respectively.



**Fig. 1.** MARE of each method during different time interval



**Fig. 2.** Summarized MARE of each prediction method

## 5 Conclusion

In our research, an efficient and scalable method for predicting travel time with arbitrary routes in road network is focused. The K-means algorithm is one of the simplest clustering techniques. One of the significant disadvantages of K-means clustering is that it doesn't yield the same result with each run, since the resulting clusters depend on the initial random assignments. Fixing two centroids by a selective procedure our MKC method eliminates this drawback of traditional K-means clustering. Moreover, the centroids are placed in a cunning way so that they maintain as much as possible far way from each other. In our clustering approach, our first cluster amalgamates those data whose possibilities are higher and second cluster congregates those data that probabilities are not significant but help to address the uncertain situation. Another shortcoming of standard clustering methods is that they ignore measurement errors or uncertainty associated with data. If these errors are available, they can play a significant role in improving the clustering decision. If we take output only from the first cluster then the results are very good in most cases but in uncertain situations the results are worse. So, we take the predicted travel time by analyzing both clusters such that the algorithm can be able to address uncertain situation. However, our method is able to predict in uncertain situations more accurately comparing with other methods. Moreover performance analysis portion of this research reveals that our proposed method outperforms other methods in most cases. The superiority of MKC is that the more the historical traffic data set increases the more the predictor is able to predict accurately. As our future, we will extend our clustering approach considering not only day time but also week days and seasonal patterns. This may help us to address uncertain situations more efficiently. We will also pay attention in the relationship between the length of roadways and accuracy of the prediction. We also try to improve our algorithm by addressing which data in the historical data are associated with uncertain situation.

## References

1. Chen, M., Chien, S.: Dynamic freeway travel time prediction using probe vehicle data: Link-based vs. Path-based. *J. of Transportation Research Record*, TRB Paper No. 01-2887, Washington, D.C. (2001)
2. Wei, C.H., Lee, Y.: Development of Freeway Travel Time Forecasting Models by Integrating Different Sources of Traffic Data. *IEEE Transactions on Vehicular Technology* 56 (2007)
3. Chun-Hsin, W., Chia-Chen, W., Da-Chun, S., Ming-Hua, C., Jan-Ming, H.: Travel Time Prediction with Support Vector Regression. In: *IEEE Intelligent Transportation Systems Conference* (2003)
4. Kwon, J., Petty, K.: A travel time prediction algorithm scalable to freeway networks with many nodes with arbitrary travel routes. In: *Transportation Research Board 84th Annual Meeting*, Washington, D.C. (2005)
5. Park, D., Rilett, L.: Forecasting multiple-period freeway link travel times using modular neural networks. *J. of Transportation Research Record* 1617, 163–170 (1998)
6. Park, D., Rilett, L.: Spectral basis neural networks for real-time travel time forecasting. *J. of Transport Engineering* 125(6), 515–523 (1999)

7. Lint, J.W.C.V., Hoogenoorn, S.P., Zuylen, H.J.v.: Towards a Robust Framework for Freeway Travel Time Prediction: Experiments with Simple Imputation and State-Space Neural Networks. In: Presented at 82 Annual Meeting of the Transportation Research Board, Washington, D.C (2003)
8. Lint, J.W.C.V., Hoogenoorn, S.P., Zuylen, H.J.v.: Freeway Travel Time Prediction with State-Space Neural Networks: Modeling State-Space Dynamics with Recurrent Neural Networks. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1811, TRB, National Research Council, Washington, D.C, 30–39 (2002)
9. Kwon, J., Coifman, B., Bickel, P.J.: Day-to-day travel time trends and travel time prediction from loop detector data. *J. of Transportation Research Record*, No. 1717, TRB, National Research Council, Washington, D.C., 120–129 (2000)
10. Rice, J., Van Zwet, E.: A simple and effective method for predicting travel times on freeways. *IEEE Trans. Intelligent Transport Systems* 5(3), 200–207 (2004)
11. Schmitt Erick, J., Jula, H.: On the Limitations of Linear Models in Predicting Travel Times. In: *IEEE Intelligent Transportation Systems Conference* (2007)
12. Lee, H., Chowdhury, N.K., Chang, J.: A New Travel Time Prediction Method for Intelligent Transportation System. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) *KES 2008, Part I. LNCS (LNAI)*, vol. 5177, pp. 473–483. Springer, Heidelberg (2008)
13. Chowdhury, N.K., Nath, R.P.D., Lee, H., Chang, J.: Development of an Effective Travel Time Prediction Method using Modified Moving Average Approach. In: Velásquez, J.D., Ríos, S.A., Howlett, R.J., Jain, L.C. (eds.) *KES 2009, Part I. LNCS*, vol. 5711, pp. 130–138. Springer, Heidelberg (2009)

# An Ontology-Based Approach for Autonomous Systems' Description and Engineering

## The OASys Framework

Julita Bermejo-Alonso, Ricardo Sanz,  
Manuel Rodríguez, and Carlos Hernández

Autonomous Systems Laboratory (ASLab), Universidad Politécnica de Madrid  
[jbermejo@etsii.upm.es](mailto:jbermejo@etsii.upm.es),  
[{ricardo.sanz,manuel.rodriguezh,carlos.hernandez}@upm.es](mailto:{ricardo.sanz,manuel.rodriguezh,carlos.hernandez}@upm.es)

**Abstract.** Ontologies provide a common conceptualisation that can be shared by all stakeholders involved in an engineering development process. They provide a good means to analyse the knowledge domain, allowing to separate the descriptive and the problem-solving knowledge. They can also be as generic as needed allowing its reuse and easy extension. These features made ontologies useful for representing the knowledge of software engineering techniques applied to autonomous systems. This work describes an ontology-based framework consisting of two intertwined elements: a domain ontology for autonomous systems (OASys) to capture any autonomous system's structure, function, and behaviour; and an ontology-based engineering methodology that generates models for autonomous systems, based on the knowledge contained in OASys and other domain ontologies. Both elements have been used in a case study to assess the suitability of the developed framework.

**Keywords:** Ontology, ontology-based methodology, knowledge-based engineering, autonomous systems.

## 1 Introduction

Autonomous systems refer to systems capable of operating in a real-world environment without any form of external control for extended periods of time. Reasons to provide systems with autonomy range from cost reduction to improved performance and dependability. Moreover, managing the increasing complexity of these systems has turned into delegating their configuration, optimisation, and repair to the systems themselves [1]. Autonomy is understood as a mongrel property that requires a combination of different capabilities:

- To obtain data and information (either from the environment or the system itself)
- To transform or refine information in a way that can be used by the decision-making elements
- To handle, to understand and to generate concepts

- To leverage cognitive aspects, meaning by such the capacity to reason, to infer and to learn
- To make decisions to achieve the system's goal, based on data provided at a sufficient level of detail
- To disseminate the decision taken to the appropriate execution or action elements

The former aspects require new engineering paradigms to cater for the idiosyncrasy of such systems: the perception process as major input for the system's knowledge, a more precise definition of system's goals to allow some decision-making to be moved from the designer to the system itself, the potential reconfiguration of the system as the goals, capabilities or the environment changes, and finally the means to assess the adequacy of the current configuration against the current mission. We are carrying out a long-term research programme, considering a wide range of autonomous systems. The strategy to increase a system's autonomy will be by exploiting cognitive control loops based on knowledge in the form of different models: of the system, of the environment, and of the task the system must fulfil in a particular environment. We have followed an ontological approach to capture the concepts of our research programme, as a knowledge-representation and software support for autonomous system's engineering.

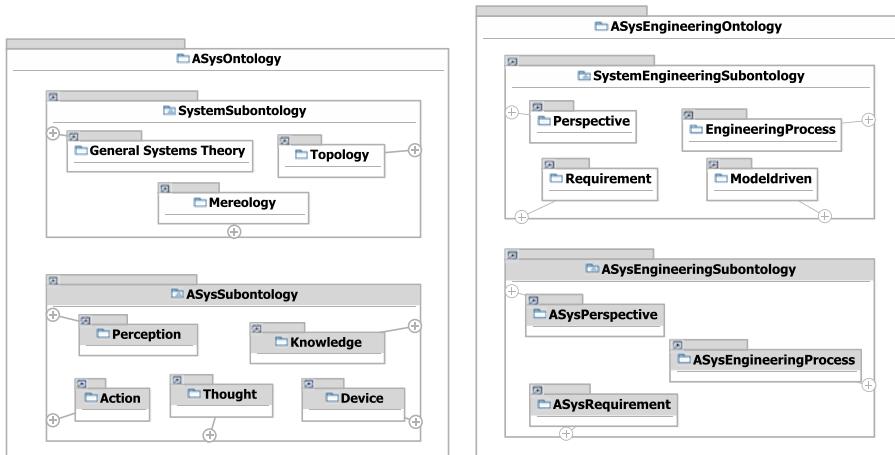
This paper describes our efforts and conclusions on applying such approach, with the following structure. Section 2 briefly describes our ontology-based framework, consisting of a domain ontology for autonomous systems (OASys), and a related methodology to exploit OASys to generate models based on the knowledge it contains. Next, Section 3 exemplifies the usage of the framework in a research testbed. Section 4 reviews previous research on the domain, and compares it with our approach. Finally, Section 5 provides some concluding remarks, and suggests further work.

## 2 OASys Framework for Autonomous Systems

The OASys Framework captures and exploits the concepts to support the description and the engineering process of any software-intensive autonomous system. This has been done by developing two different elements: an autonomous systems domain ontology (OASys), and an OASys-based engineering methodology.

### 2.1 Ontology for Autonomous Systems (OASys)

OASys is a domain-ontology consisting of two ontologies (Fig. 1): the ASys Ontology with the ontological elements to describe an autonomous system structure, behaviour and function, and the ASys Engineering Ontology to provide the concepts for the engineering process of an autonomous system. Each one is organised using Subontologies, and Packages. Subontologies address autonomous systems' description and engineering at different levels of abstraction, whereas packages are organisational elements used to gather concepts semantically related to a specific aspect within a subontology. The different packages have been



**Fig. 1.** OASys structure: ontologies, subontologies and packages

formalised using UML class diagrams to show the ontological elements in terms of concepts, attributes and relationships.

**System Subontology** contains the elements necessary to define any system, organised in three different packages: the *General Systems Theory Package* gathers concepts to characterise any kind of system's structure and behaviour based on the General Systems Theory; the *Mereology Package* collects taxonomies of the whole–part concepts and relationships based on existing mereological theories; the *Topology Package* provides concepts for topological connections based on topological concepts.

**ASys Subontology** specialises the previous concepts for autonomous systems, containing several packages to address its structure, behaviour and function: the *Device Package* gathers concepts to describe the different aspects of devices; the *Perception Package* to formalise the perceptive and sensing processes; the *Knowledge Package* to conceptualise the different kinds of knowledge used, such as the types of goals in an autonomous system; the *Thought Package* to describe the goal-oriented processes concepts; the *Action Package* summarises the concepts about the operations and performing actors.

**System Engineering Subontology** gathers the concepts related to an engineering process as general as possible, based on different metamodels, specifications and glossaries used in software-based developments. The subontology is organised as follows: the *Requirement Package* conceptualises system's requirements; the *Perspective Package* addresses viewpoints in a system development; the *Engineering Process Package* describes an engineering development process in terms of phases, tasks, and the obtained workproducts; the *Model-driven Package* provides the elements based on model-driven engineering.

**ASys Engineering Subontology** contains the specialisation and additional ontological elements to describe an autonomous system's generic engineering process, organised as different packages: the *ASys Requirement Package* to provide those concepts to characterise process and system quality requirements; the *ASys Perspective Package* to describe an autonomous system from different aspects; the *ASys Engineering Process Package* to describe an autonomous system generic engineering process.

## 2.2 The OASys-Based Methodology

The OASys-based methodology provides some guidelines and exemplifies the application of the OASys ontological elements to obtain the autonomous system's models during the phases of an engineering process. The ontological elements

**Table 1.** OASys-based Methodology: ASys Requirement phase

PHASE	TASK	SUBTASK	WORK PRODUCT		RELATED OASYS PACKAGE
ASys Requirement	System UseCase	UseCase Modelling	UseCase Model	System UseCase Model Subsystem UseCase Model	System Engineering Subontology: Requirement Package
		Use Case Detailing	UseCase Specification	UseCase Description	ASys Engineering Subontology: ASys Requirement Package
	Requirement Characterisation	Non-functional Requirement Functional Requirement	Requirement Specification	Process Characterisation System Characterisation	

**Table 2.** OASys-based Methodology: ASys Analysis phase

PHASE	TASK	SUBTASK	WORK PRODUCT		RELATED OASYS PACKAGE
ASys Analysis	Structural Analysis	System Modelling	Structural Model	Structure Model Topology Model	System Subontology: General Systems Theory Package, Mereology Package, Topology Package
		Knowledge Modelling	Knowledge Model	GoalStructure Model Procedure Model Quantity Model Ontology Model	ASys Subontology: Knowledge Package
	Behavioural Analysis	Behaviour Modelling	Behavioural Model	Behaviour Model	System Subontology: GST Package
	Functional Analysis	Function Modelling	Functional Model	Agent Model Operation Model Responsibility Model	ASys Subontology: Action Package ASys Subontology: Thought Package ASys Subontology: Perception Package

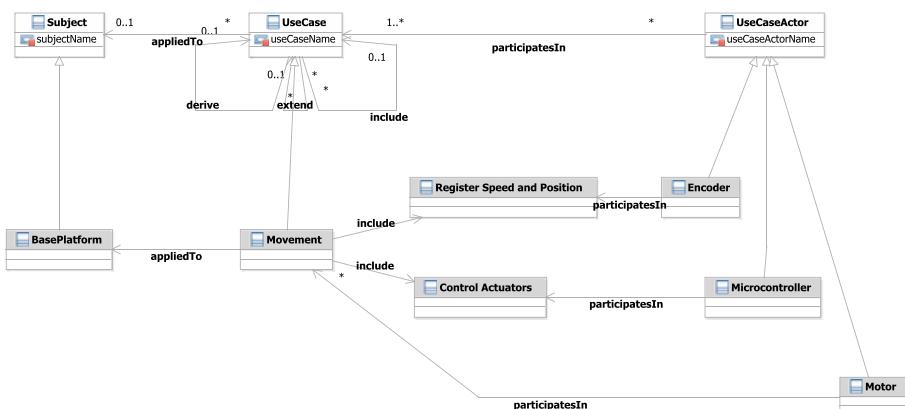
defined in the System Engineering and ASys Engineering subontologies serve as semantic guide for the phases, tasks, and work products names and usage. The methodology focuses on two main phases: the ASys Requirement phase to identify the autonomous system's requirements (Table 1), and the ASys Analysis phase to consider the autonomous system's analysis of its structure, behaviour and function (Table 2).

### 3 A Case Study: The Robot Control Testbed

The OASys Framework is being applied in the description and engineering of a Robot Control Testbed (RCT), which is a robotic-based application, consisting of a base platform and different interconnected subsystems to cover a wide range of functionalities. Not all the developed models are described in this section, only some to exemplify the usage of the ontology and the methodology.

The RCT ASys Requirement Phase identified stakeholders' requirements, using traditional engineering techniques. An Use Case Model as WorkProduct was obtained, considering the elements in the Requirement Package: the concepts of *Subject*, *UseCase*, and *UseCaseActor* and the relations of *include* and *appliedTo*. Figure 2 shows how these elements are instantiated into the case study ones. The top white UML classes represents the original concepts and relationships in the Requirement Package. The lower shaded classes, the instantiated elements. Each instantiated class is related to its corresponding one using a UML generalisation relation to express the is-a relationship between them. The instantiated relation receives the same name as the original relation in the package, known as construct overloading. However, the specialised relation is so with a different range and domain, and thus with different semantics.

As part of the ASys Analysis Phase, the Structural Analysis task analyses the autonomous system's structure using the structural concepts General Systems Theory Package, and the mereotopological relationships in the Mereology



**Fig. 2.** Use Case Modelling: example of instantiation for the Base Platform of the RCT

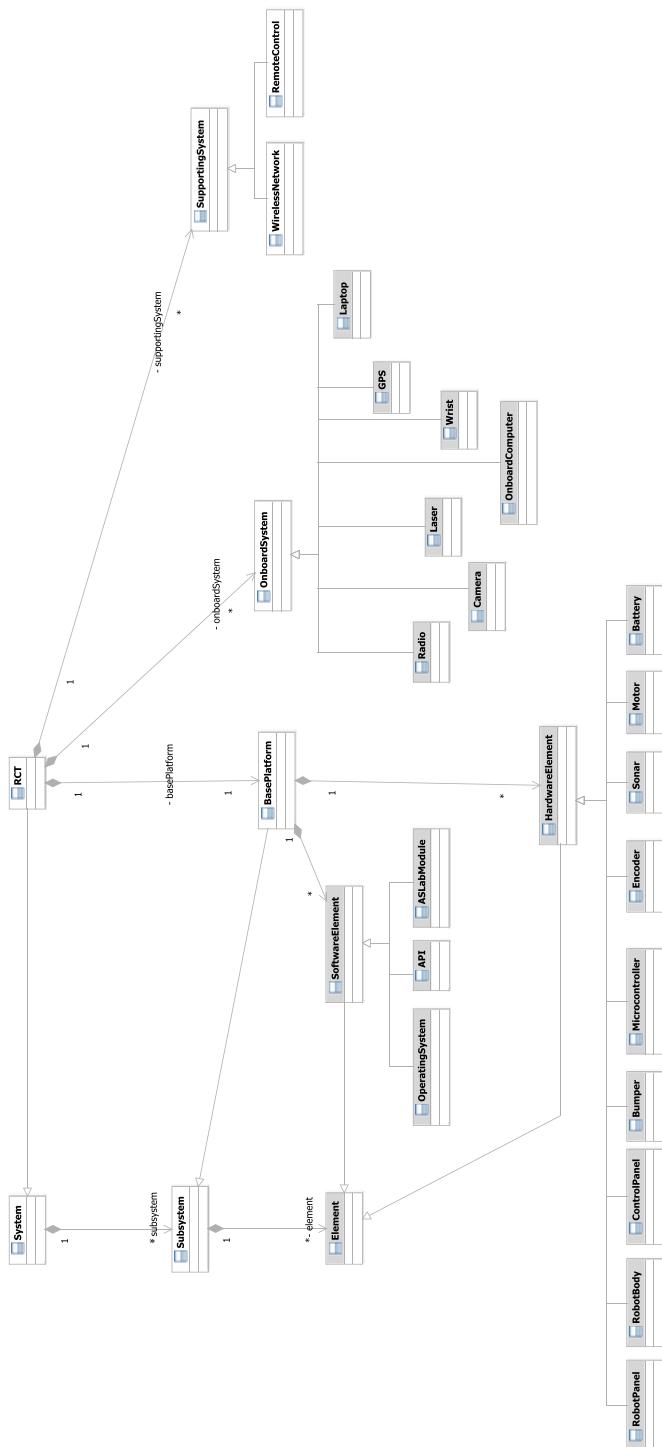


Fig. 3. System Modelling subtask: the RCT Structure Model

and Topology Packages. As WorkProduct, a Structural Model for the RCT is obtained instantiating these ontological elements, consisting of a RCT Structure Model and a RCT Topology Model. Figure 3 shows an example of the Structure Model for the RCT. On the left, the original UML classes that formalise the concepts of *System*, *Subsystem*, and *Element*. On the right, their instantiation into RCT actual elements. To point out that part–whole relations defined explicitly in the Mereology Package, such as *isPartOf* or *isExclusivelyPartOf*, can be instantiated using the UML aggregation and composition relations. The UML composition relation has been used to model that the RCT System is composed of a Base Platform, several Onboard Systems and different Supporting System. The original UML relation cardinalities on the left side of the figure have been particularised into the required ones in the RCT. For example, the RCT only contains one Base Platform, which is expressed by the cardinality 1 close to the UML class *BasePlatform*. Subsystems can contain different subsystems of elements, hence they can be in turn decomposed. In the RCT System Model, the Base Platform is decomposed into the software and hardware elements, using respectively the *SoftwareElement* and *HardwareElement* classes. UML Generalisation relations are used to classify additional Elements within the two former categories.

## 4 Discussion and Related Work

Our ontology-based framework merges two different aspects. Firstly, it is used to describe autonomous systems. Secondly, it should provide support for conceptual modelling and software engineering for such systems. Regarding autonomous systems, ontologies have been used with an approach closer to the viewpoint of knowledge representation based on a specification of a conceptualization, as opposed to defining the meaning of terms or to understanding the world. Ontologies are used as representation-based mechanisms based on a computational language, to describe the different entities participating in the design and operation of the autonomous systems: different domains, the environment, the objects the systems interact with, the possible actions to be taken, resources to be considered, etc. In a similar way, OASys has conceptualised how the engineers developing autonomous systems characterise them, describing the different elements taking part in their operation. However, our approach differs in considering these elements not only in terms of environment objects, resources or actions, but on the main aspects we consider fundamental for autonomy: the perception process, the knowledge to be used, the system’s goals, the thought process as functional decomposition, and the actors to carry out the system’s actions.

Moreover, we have considered the autonomous systems domain with a global view. Previous ontologies developed for this kind of systems focused on a particular application type, such as mobile robots or agent based systems. Our approach has been to define the different elements to describe any autonomous system, in a way general enough to be reused among different applications. OASys can be further complemented with subdomain, task or application ontologies, without loosing its reusability and generality features.

OASys has provided similar benefits to those referred to in the literature on ontologies for autonomous systems [10]: clarifying the structure of knowledge, knowledge sharing and reuse, and easing the interoperability among heterogeneous agents. Additional advantages have been a common conceptualisation not only of the autonomous systems domain but also of the engineering process for this kind of systems, making easier sharing the terminology among different developers. It remains to explore the reasoning capabilities based on the ontological relationships and commitments defined in OASys.

For software engineering, ontologies have played a twofold role [7]: as conceptual basis for the definition of software components in the software engineering process, and to describe the terminology of the software engineering domain itself. OASys and its methodology fall within the first role, providing a conceptual basis to define the early stages of an engineering process specific for autonomous systems according to our research view. Ontologies in this domain provide a series of benefits [9], such as a specialised representation vocabulary, transference of knowledge in software projects, same conceptualisation for different software applications, and conceptual mismatches reduction among users. Similar benefits have been obtained during the testbeds development using OASys: specific vocabulary for our research programme, same conceptualisation for software applications to be developed, and common meaning of elements throughout the engineering process.

## 5 Concluding Remarks and Further Work

We learned some lessons whilst developing the models for the case study. The model development benefited from the underlying ontological commitments in OASys, avoiding meaning and conceptual mismatches. However, the ontology development and the construction of the models was not exempt of challenges.

The decision to formalise OASys using a software engineering language such as UML was based on two facts. Firstly, the review of software engineering techniques and ontologies made as part of the research showed most of them as being UML-centered. Secondly, OASys was designed to support the model-driven engineering process in the research programme. UML is not an ontology development language, hence it has been considered to have some drawbacks for this task [5]: lightweight ontologies, incomplete semantics, software heritage, and lack of inference mechanisms. It has as advantage its widespread use among engineering practitioners, the support of a graphical representation, and the existence of UML-based tools. Some of the shortcomings have been addressed in the Ontology UML Profile [4], and concretise in the Ontology Definition Metamodel (ODM) [8] that defines metamodels, mappings and profiles to allow the interaction between UML and ontological languages such as OWL and RDF.

An additional issue was to define the role of OASys for metamodeling. The case study models were obtained by instantiating, i.e., creating an entity that conforms to the definition. During the instantiation in the case study, we encountered the two forms of metamodeling described in [3], [2]: linguistic instance-of

relationships used to express that the concept has been made from a specification concept (e.g., to describe specific autonomous system's objects), and ontological instance-of relationships (is-a) used to express its similarity to an ontological element of a higher level ontology (e.g. to instantiate the packages ontological elements in the case study elements). The existence of the two different meta-modelling approaches leads to some metamodeling problems [6]. To address the ontological and linguistic instantiations, OASys currently plays the role of an ontological model expressed with UML modelling constructs, inheriting its features from the UML Metamodel itself.

The models in this paper explicitly show the ontological is-a relationships, as the UML generalisation relations between the OASys elements and the actual RCT ones. This is time consuming and clutters the models. This semantic information is not shown when obtaining the models using an UML based tool, where the relation is inherited but not explicitly shown. We are evaluating the use of roles in the UML relations to express the original element one class is instantiated from. UML stereotypes are also considered to address this point.

Finally, the construction of the conceptual models was made ad-hoc, i.e., specifically for the case study. A proper methodology based on ontological and software patterns with the aid of conceptual modelling tools is a further step to enrich these conceptual models. This methodology, to be defined in the next stage of our research programme, will define how a concept is selected, how to integrate a concept into a pattern, how to establish and to import its relationships with other concepts, and how to detail or to add its attributes as part of the development of a concrete model.

## References

1. Kephart, J.O., Chess, D.M.: The vision of autonomic computing. Computer 36(1), 41–50 (2003)
2. Assmann, U., Zschaler, S., Wagner, G.: Ontologies, meta-models, and the model-driven paradigm. In: Calero, C., Ruiz, F., Piattini, M. (eds.) *Ontologies for Software Engineering and Software Technology*, pp. 249–273. Springer, Heidelberg (2006)
3. Atkinson, C., Kuhne, T.: Model-driven development: a metamodeling foundation. IEEE Software 20(5), 36–41 (2003)
4. Gasevic, D., Djuric, D., Devedzic, V.: *Model Driven Architecture and Ontology Development*. Springer, Heidelberg (2006)
5. Gómez Pérez, A., Fernández López, M., Corcho, O.: Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. In: *Advanced Information and Knowledge Processing*. Springer, Heidelberg (2004)
6. Henderson-Sellers, B., Gonzalez-Peres, C.: *Metamodelling for Software Engineering*. John Wiley and Sons Ltd., Chichester (2008)
7. Hesse, W.: Ontologies in the software engineering process. In: *Tagungsband Workshop on Enterprise Application Integration (EAI 2005)*, Berlin, Germany. GITOV-Verlag (2005)

8. Object Management Group. Ontology Definition Metamodel Version 1.0 (May 2009)
9. Ruiz, F., Hilera, J.: Using ontologies in software engineering and technology. In: Calero, C., Ruiz, F., Piattini, M. (eds.) *Ontologies for Software Engineering and Software Technology*, pp. 49–95. Springer, Heidelberg (2006)
10. Stojanovic, L., Schneider, J., Maedche, A., Libischer, S., Studer, R., Lumpp, T., Abecker, A., Breiter, G., Dinger, J.: The role of ontologies in autonomic computing systems. *IBM Systems Journal* 43(3), 598–616 (2004)

# Search Space Reduction for an Efficient Handling of Empty Answers in Database Flexible Querying

Mohamed Ali Ben Hassine, Chaker Abidi Nasri, and Habib Ounelli

Department of Computer Sciences, Faculty of Sciences of Tunis,  
El Manar 1, Campus Universitaire, 2092, Tunis, Tunisia  
{mohamedali.benhassine,habib.ounelli}@fst.rnu.tn,  
chaker.abidi@gmail.com

**Abstract.** The developed topic in this paper deals with a proposal of an approach based on Formal Concept Analysis to solve the problem of empty answers in the case of fuzzy conjunctive queries. Unlike our previous approach [1], which handles the whole context, our new one handles smaller contexts, determined after a pretreatment step. Thus, we suggest two algorithms. The first one detects the failure reasons of a flexible query whereas the second one provides neighbor answers. The neighbor answers are always in the first context, so we do not need to check the entire context. Similarly, to detect the failure reasons, we find frequent cases which can happen in one of the two contexts. Hence, we can reduce the search space and ensure the improvement of the response time of our previous algorithms.

**Keywords:** Flexible Queries, FCA, Cooperative Systems, Empty Answers.

## 1 Introduction

The querying of DataBases (DB), by the use of the boolean conditions, is rigid for certain applications. On the one hand, DB systems are often hard to use because they do not explicitly attempt to cooperate with their users. They provide literal answers to queries posed to them. Such answers may not always be the best ones. They can be correct, but may not contain the information the user really wants or can be empty. The user needs more information, or might even need different information, than the query requests. To remedy the latter restriction, extending the classical notion of query processing to cooperative answering has been explored [2]. On the other hand, the traditional queries in relational DB are unable to satisfy the user needs' to deal with imprecise data. Therefore, flexible querying systems are introduced to allow imprecise specification of queries and to reduce the risk of empty answers [3]. Several works have been proposed to deal with empty answers [1][2][4][5][6]. We are interested in our previous work [1][7], which is based on Formal Concept Analysis (FCA) [8] and fuzzy logic [9]. We proposed an approach to solve the problem of empty answers in fuzzy

queries. When a query fails, our approach detects its failure reasons and proposes, to the user, the possible neighbor answers accompanied by their satisfaction degrees. Although the experimentations done in some DB have demonstrated the efficiency of our approach, when the DB is large or when the query contains too many conditions, the algorithms attain their limits.

In this paper, we try to solve the limits of our previous work. We carry out a pretreatment on the global query context and, split it into small ones. According to this pretreatment, the neighbor answers are always in the first context, so we do not need to check the entire context. Similarly, to detect the failure reasons, we find frequent cases which can happen in one of the two contexts. Therefore, we reduce the search space and improve the response time of our previous algorithms.

The next section makes an overview of previous approaches of cooperative answers in DB systems, with an interest to the approach of Hachani et al. [1]. Section 3 presents the big lines of our approach. Then, Section 4 deals with the problem of empty answers and Section 5 illustrates our work with an example. Finally, Section 6 gives some concluding remarks directions for the future.

## 2 Related Work

A number of researchers in the area of DB have recognized the practical need for cooperative answering behavior in standard, widely available information systems. This need has led people to consider how to adapt and develop cooperative techniques especially for DB. It is well-known that the problem often approached in this field is the "empty answer problem", that is, the problem of providing the user with some alternative data when there is no data fitting his query. Several approaches have been proposed to deal with this issue [2]. Some of them are based on a relaxation mechanism that expands the scope of the query [5][10][11][12]. The main objective of these approaches is to modify a failing user query into a relaxed query whose answer is non-empty, or at least to identify the cause of the failure. In the context of flexible queries, similar problems could still arise. The empty answer problem is defined in the same way as in the Boolean case. The well known works concerned with the cooperative systems and empty answers in the context of fuzzy queries are those of Andreason and Pivert [13], Motro [6], Voglozin et al. [14], Bosc et al. [4][5] and, Hachani et al. [1][7].

We are interested in this last work, in which, we proposed a cooperative approach of flexible querying based on FCA. In fact, FCA was introduced by Wille [15] with the aim of identifying similar groups of objects from a collection G of objects described by a set of attributes M. The paradigm occurred within the lattice theory [8]. [1] gives more details about the basic concepts of FCA used in the rest of this paper. This approach attempts to explain the reasons of the query failure and, proposes the nearest approximate (neighbor) answers. To deal with this problem, two algorithms have been proposed. These algorithms, which handle the whole query context, give good results specially when the query does

not contain too many conditions. This does not lead to say that the running time is far from the real needs of the users. For example, with 12 fuzzy conditions applied on a German or Diabetes datasets [16], the first algorithm detects the failure reasons of the query in about 1 second.

### 3 Proposed Approach

We attempt, in our approach, to keep the main goal of the previous one [1], i.e. handling empty answers, while improving its performance.

#### 3.1 Definitions

Let us remind ourselves of some important definitions useful to understand clearly our approach [1]:

**Definition 1.** A fuzzy query  $FQ$  is an extended SQL query applied to the relation  $r$  with linguistic terms and fulfillment degree through the following form:

$SELECT < attributes > FROM < r > WHERE < fuzzy condition_1 > threshold \tau_1 [ AND ... < fuzzy condition_n > threshold \tau_n];$

The fuzzy condition is defined as an attribute with a linguistic term such as "price is low". The fulfillment threshold  $\tau$  (default is 0) indicates that the condition must be satisfied with minimum degree  $\tau \in [0, 1]$  to be considered. FQ is transformed to a conceptual query (CQ).

**Definition 2.** A conceptual query  $CQ = (R, \{A_1, A_2, \dots, A_n\})$  is a formal concept where:

- $R$  is the extent of  $CQ$  ( $g(\{A_i\})$ ) and represents the set of the expected answers (initially  $R = \emptyset$ ).
- $\{A_i\}$  is the intent of  $CQ$ .  $A_i$  is an attribute in the generated context, denoted as  $Cr$ , from the table  $r$  ( $A_i \in Cr$ ) and represents also a fuzzy condition included in  $FQ$  described as follows:  $A_i = \text{attribute } IS \text{ linguistic-term}$  where attribute  $\in r$  and linguistic-term is a value  $\in$  values of attribute.

Example: Consider the query  $FQ$ :

$SELECT name FROM Employee WHERE age is young and salary is average;$   
The translation of  $FQ$  is as follows:  $CQ := (\emptyset, \{\text{age is young}, \text{salary is average}\}).$

**Definition 3.** A fuzzy context of a query is the result of a projection operator applied on the original context (fuzzy) for the attributes appearing in the  $FQ$  condition(s) and having a degree greater than its associated threshold  $\tau$ :

$$\text{fuzzy\_context\_query} := \Pi_{\{A_i\}}(\sigma_{A_i \geq \tau_i} \text{Original context})$$

**Definition 4.** We call  $CQ' = (R', \{A_{s1}, A_{s2}, \dots, A_{sk}\})$  a subquery of  $CQ$  iff  $\{s1, s2, \dots, sk\} \subseteq \{1, \dots, n\}$  and  $R'$  is the set of the objects (attended answers) satisfying the set of conditions  $\{A_{s1}, A_{s2}, \dots, A_{sk}\}$  ( $R' \supseteq R$ ).

Given a conceptual query  $CQ$  and its associated concept lattice  $L$ ,  $CQ' = (R', \{A_{s1}, A_{s2}, \dots, A_{sk}\})$  is called a conceptual subquery of  $CQ$  iff  $\exists C(X, Y) \in L \setminus Y = \{s1, s2, \dots, sk\}$ .

**Definition 5.** Let  $CQ = (R, \{A_i\})$  be a conceptual query and  $L$  its associated lattice ( $L$  is the set of all the generated concepts  $C(X, Y)$ ). The set of failure reasons  $FR$  is defined as follows:  $FR = \{fr\}$

$$fr : \{A_{sk}\} \subseteq \{A_i\} / \forall C(X, Y) \in L / \infimum(L), \{A_{sk}\} \not\subseteq Y.$$

**Definition 6.** A Minimal failure reason ( $Mfr$ ) is a failure reason that does not include any other failure reason ( $fr$ ):  $Mfr \in FR$  and  $\#fr \in FR / fr \subset Mfr$ .

### 3.2 Functioning Description

The architecture and the environment of work used in our previous approach are still unchanged. This architecture is implemented under a module called IDFQ [17]. The flexible querying processing, depicted in Figure 1, keeps the big lines of the steps used in our previous work [1] with few differences. These differences concern the building of the concept lattice (we use Galicia [18]), the addition of a pretreatment step and, the rewriting of the previous algorithms in a new way.

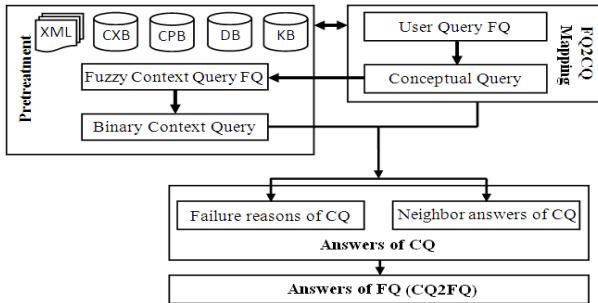


Fig. 1. Empty answers processing

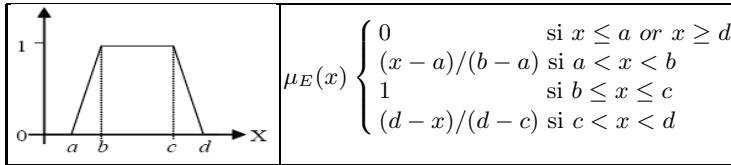
### 3.3 Empty Answers Processing

Given a failed FQ, the processing of empty answers consists in:

1. Translating the query FQ to the concept format (CQ).  
FQ:  $\text{SELECT } <\text{attributes}> \text{ FROM } <\text{relation}> \text{ WHERE primary key IN CQ}$
2. Building the fuzzy context of FQ (Definition 5) while extracting, from the KB, the scales of the fuzzy criteria used in the clause WHERE of FQ. To compute these scales, we used the trapezoid membership function depicted in Table 1. The resulting fuzzy context will be transformed into a query binary context based on  $\alpha$ -cut [19]. In this work,  $\alpha = 0$ . At this level, the pretreatment step is added.
3. Providing approximate answers for the failed query FQ and looking for its failure reasons. This task is detailed in the next sections.

### 3.4 Pretreatment on CQ

The main improvement of our approach starts with a pretreatment at the level of the binary context of the query. This pretreatment consists in:

**Table 1.** Trapezoidal fuzzy set and its corresponding function

- a) Deleting the unused objects and attributes: This step consists in removing, from the binary context, the unused attributes (respectively the unused objects), i.e. the attributes which are not possessed by any object (respectively the objects which do not posses any attribute of CQ). This is explained by the fact that the deleted objects will not appear in the answer. Similarly, each deleted attribute constitutes a minimal failure reasons of size 1. Therefore, we omit this useless treatment and we look for more information: the hidden reasons. In this manner, we can reduce the size of the binary context.
- b) Sorting the binary context: In this step, the objects which have more frequent attributes (more 1 in their columns) are placed before, i.e. at the top of the context, and those which have few attributes (more 0 in their columns) are placed after, i.e. toward the bottom of the context.
- c) Decomposing horizontally the binary context query: This step consists in splitting, horizontally, the binary context already sorted into two contexts: the first one contains the same number of frequent attributes in his lines whereas the second one contains the remaining objects. In this cutting, the binary relation between these objects and their attributes is still kept.

These steps are illustrated, in Algorithm 1, through the following functions:

- *Card(S)* returns the number of elements of the set S.
- *Delete(BCxt)* removes from the binary context (BCxt) the unused attributes and objects. It returns the set of the deleted attributes.
- *Sort\_context (BCxt)* sorts the binary context (BCxt) as described above.
- *Decompose\_context (BCxt, BCxt1, BCxt2, nbCxt)* decomposes BCxt into two (nbCxt<sup>1</sup>) binary contexts (BCxt1 and BCxt2) as described above.

---

**Algorithm 1.** Binary Context Pretreatment

---

**input** : FC: set of Fuzzy Conditions of FQ; BCxt: Binary Context source

**output**: DFC : set of Deleted Fuzzy Conditions

nbCxt: number of decomposed binary contexts

BCxt1: First binary context; BCxt2: Second binary context

**begin**

```

DFC ← Delete(BCxt);
if Card (DFC ) = Card (FC ) then nbCxt ← 0;
else
  Sort_context (BCxt) ;
  Decompose_context (BCxt, BCxt1, BCxt2, nbCxt );

```

---

<sup>1</sup> nbCxt can be 1 if the context is not decomposed.

## 4 Handling of Empty Answers

The handling of empty answers constitutes the aim of our approach. Empty answers are uninformative to the user. Thus, when a query fails, it is often a surprise to the user. We think that is interesting to identify the causes of the query failure and present some approximate answers. For this reason, we propose two algorithms: the first one detects the minimal failure reasons whereas the second generates the neighbor answers. The proposed algorithms focus in some parts of the original context determined at the pretreatment step (i.e the first one or the second one, and in few situations both of them).

### 4.1 Detection of Query Failure Reasons

The detection of query failure reasons starts by finding the minimal failure reasons of size 1 (i.e. the deleted attributes at the pretreatment step) if they exist. Then, this process continues by looking for the minimal failure reasons of size superior to 1 and inferior or equal to the number of deleted attributes. Two situations are present:

1. **Detection of evident failure reasons of size one:** This case happens when, at the pretreatment step, we have one or more deleted attributes and all the remaining ones are owned by at least one object. In this situation, we do not need to check the whole context, neither to build its concept lattice. The deleted attributes represent the only failure reasons and, are of size one.
2. **Detection of hidden failure reasons:** In this scenario, there is not any object of the context (after deleting unused attributes) which owns all the attributes together<sup>2</sup> (line 3 of Algorithm 2).

Algorithm 2 describes this processus through the following functions<sup>3</sup>:

- *Find(Cxt)* returns a boolean value: 1 if there is at least an object which owns all the attributes of the context *Cxt* and 0 otherwise.
- *Count1(Cxt)* returns the number of the value 1 owned by the first object of the context *Cxt*.
- *latticebuilding(Cxt)* build the concept lattice of the context *Cxt*.
- *ParentsIntents (X, L)* returns, from the set of concepts *L*, the intents of the concepts located directly above the concept *X*.
- *Compare (PFR, n)* returns 1 if PFR contains only elements having the size equal or superior to *n* and 0 otherwise.
- *Failure\_reasons (F, C, n)* returns the set of failure reasons from the set *F*, which are not equal or not included in any elements of the set *C*. This function tests only the elements of *F* whose size is inferior or equal to *n*.
- *Minimal (PFR)* removes from PFR the elements which include others ones of PFR so that only the minimal elements that remain.

<sup>2</sup> Also, each object cannot own all the attributes together minus one, otherwise we return to the previous case.

<sup>3</sup> The function *Card* is already defined in Algorithm 1.

**Algorithm 2.** Detection of minimal failure reasons

---

```

input : FC : set of Fuzzy conditions of FQ
        DFC : set of Deleted Fuzzy Conditions
        BCxt1 : First binary context; BCxt2 : Second binary context
        nbCxt : number of decomposed binary contexts
output: MFR : set of minimal failure reasons
1 begin
2   MFR ← DFC ;
3   if (( $0 < Card(DFC) < Card(FC)-1$ ) and  $Find(BCxt1) = 0$ ) or ( $Card(DFC)=0$ )
    then
4     RFC ← FC - DFC ; // Remaining Fuzzy Conditions
5     N1 ← Count1(BCxt1);
6     CA ← Set of combination of RFC from size 2 to (N1 + 1);
7     if  $N1=1$  then
8       MFR ← MFR ∪ CA ;
9       else if  $N1>1$  then
10      L1 ← latticebuilding(BCxt1);
11      C1 ← ParentsIntents (Infimum(L1),L1);
12      PFR ← Failure_reasons (CA, C1, N1);
        // Possible Failure Reasons
13      if  $Card(PFR)=1$  then
14        MFR ← MFR ∪ PFR;
15      else if ( $Card(PFR)>1$  and  $Compare(PFR,N1=1)$  or
16       $nbCxt=1$  then
17        MFR ← MFR ∪ Minimal (PFR);
18      else if  $nbCxt = 2$  and  $Count1(BCxt2) > 1$  then
19        L2 ← latticebuilding(BCxt2);
20        C2 ← ParentsIntents (Infimum(L2),L2);
21        CFR ← Failure_reasons (PFR, C2, Count1(BCxt2)) ;
          // Commun Failure Reasons
        MFR ← MFR ∪ Minimal (CFR);

```

---

Algorithm 2 is more efficient than the one presented on our previous approach and, this depends on the query and the DB. In fact, we find frequent cases, in which, we are limited only to the first context, and consequently the building and the navigation in a small lattice. In other cases, we even do not need to build the lattice. Accordingly, the running time decreases noteworthy. If we are in a situation where the two contexts have to be checked, our approach does not exceed much the previous one.

## 4.2 Generation of Neighbor Answers

Generating approximates answers in the neighborhood of the query criteria is not a simple task. In fact, we must provide the user with answers which satisfy the maximum his query criteria. To minimize the search space, we start from the first

binary context (BCxt1) defined at the pretreatment step because it contains all the information that we look for. After generating concepts from this binary context, Algorithm 3 try to retrieve the extents of the concepts having, in their intents, the maximum of CQ intents<sup>4</sup>. In some cases, we can restrict ourselves to the binary context without building the concept lattice. After some experimentations, Algorithm 3 has shown that it is more efficient than the one of our previous approach. The evaluation of the failed query CQ consists in evaluating:

- The WHERE clause: We look for the concepts (Parents) linked directly to the infimum of the concept lattice of L1. These concepts represent the neighbor subqueries (Definition 4). Their extents regroup their answers. Algorithm 3 associates also to each answer its satisfaction degree, which will be used in the global answers sorting. It is important to note that this degree is computed using the min operator [9].
- The SELECT clause: for each returned objects from CQ, a selection operator is applied to find answers to FQ. These answers are presented to the user with their satisfaction degrees.

Algorithm 3 is defined by the following functions<sup>4</sup>:

- *Extent (X)* returns the extent of concept X.
- *Intent (X)* returns the intent of concept X.
- *Parents (X,L)* returns, from the set of concepts (lattice) L, those located directly above the concept X.
- *Degrees (O, A, FCxt)* returns a set of satisfaction degrees of an object O for each attribute A in the Fuzzy Context(FCxt).
- *Min (S)* returns the minimum value of a set S.
- *Ranking (TD)* returns a set TD ranked by its satisfaction degrees.
- *Object(Cxt)* returns the set of objects of the context (Cxt).

## 5 Illustrative Example

Let *Employee* be a relation described by the attributes: Ssn, Name, Age, Salary, Number of years service (Nbyearser), Child number (Chnb) and Height. Let us consider the following query: "Retrieve the employees with high age, low salary, average number of years service, low number of child and average height".

The fuzzy query FQ is written as follow:

FQ: SELECT name, Id FROM Employee

WHERE age is *high* and number of years service is *average*

and salary is *low* and number of child is *low* and height is *average*;

FQ is translated to the conceptual query CQ (Definition 1):

CQ=( $\emptyset$ , {Hage, Lsalary, Anbyearser, Lchnb, Aheight})

To answer this this query:

1. We build the fuzzy context of the query CQ (Table 3(a)).
2. We transform this fuzzy context into a binary one (Table 3(b)).

---

<sup>4</sup> The functions *Card* and *Find* are already defined in the previous algorithms.

**Algorithm 3.** Neighbour Subqueries

---

**input** : L1: concept lattice of BCxt1  
 FCxt: Fuzzy Context  
 FC: set of Fuzzy conditions of FQ  
 DFC: set of Deleted Fuzzy Conditions  
 BCxt1: First binary context

**output:** R: Neighbor subqueries with their satisfaction degrees

**begin**

```

if ((0 < Card (DFC) < Card (FC)-1) and Find(BCxt1)= 1) or (Card
(DFC) = Card (FC) - 1) then
  RFC ← FC - DFC ; // Remaining Fuzzy Conditions
  T ← Object(BCxt1);
  for j ← 1 to Card(T) do
    DegTj ← Min (degrees (T(j), RFC, FCxt));
    TD ← TD ∪ (T(j), DegTj) ; // TD is a set of pairs composed
      of an object and its satisfaction degree
  TR ← Ranking (TD);
  R ← (RFC, TR) ;
else if (Card (DFC) = 0) or ((0 < Card (DFC) < Card (FC) -1) and
Find(BCxt1) = 0) then
  cmax ← Parents(Infimum(L1),L1);
  K ← Card(cmax);
  for i ← 1 to K do
    T ← Extent(cmaxi);
    TD ← Ø ;
    for j ← 1 to Card(T) do
      DegTj ← Min (Degrees (T(j), Intent(cmaxi), FCxt));
      TD ← TD ∪ (T(j), DegTj) ;
    TR ← Ranking (TD);
    R ← R ∪ (Intent(cmaxi), TR) ;
  
```

return (R);

---

**Table 2.** Extension of the table "Employee"

Ssn	Name	Age	Salary	Nbyearser	Chnb	Height
1	Ali	30	257	11	3	160
2	Mohamed	32	233	12	3	155
3	Hanene	45	144	20	4	160
4	Sameh	56	377	30	5	155
5	Bassem	46	257	17	5	175
6	Hassen	48	562	27	4	185
7	Amal	34	456	17	3	170
8	Ahmed	38	388	13	4	175
9	Farah	59	644	32	5	174
10	Ikram	30	277	12	3	163

**Table 3.** Fuzzy (a) and Binary (b) context of the query

	Hage	Lsalary	Anbyearser	Lchnb	Aheight		Hage	Lsalary	Anbyearser	Lchnb	Aheight
1	0.0	0.86	0	0	0.0		1		x		
2	0.0	1.00	0	0	0.0		2		x		
3	0.0	0.88	1	0	0.0		3		x		
4	1.0	0.00	0	0	0.0		4	x		x	
5	0.0	0.86	1	0	0.0		5		x	x	
6	0.0	0.00	0	0	0.0		6				x
7	0.0	0.00	1	0	1.0		7			x	
8	0.0	0.00	0	0	0.0		8				x
9	0.5	0.00	0	0	0.2		9	x	x		
10	0.0	0.46	0	0	0.6		10				x

(a)

(b)

**Table 4.** The first (a) and the second (b) binary context after the pretreatment step

	Hage	Lsalary	Anbyearser	Aheight		Hage	Lsalary	Anbyearser	Aheight
3		x	x			1		x	
5		x	x			2		x	
7			x	x		4	x		
9	x	x							
10		x							

(a)

(b)

- We carry out the pretreatment (deletion, sorting and splitting) on the binary context of the query. In our example, we delete the objects 6 and 8, and the attribute "Lchnb". In result, we have two binary contexts: the first one contains the objects which have the same maximum number of attributes (Table 4 (a)), whereas the second one contains the rest of the objects with their associated attributes (Table 4 (b)).
- To handle empty answers, we begin from the attribute "Lchnb" which represents the only minimal failure reason of size 1. Then, Algorithm 2 generates the concept lattice of the first binary context to detect the others minimal failure reasons of size superior to 1 and inferior to 3 ( $3 = [count1(BCxt1)+1]$  where  $BCxt1$  is the first binary context). At the end of the Algorithm 2, we obtain the list of the minimal failure reasons (see Table 5). These reasons give, to the user, an idea about incompatible fuzzy conditions.

**Table 5.** Minimal Failure Reasons

Low child number
High age and average number of years service
High age and low salary
Low salary, average number of years service and average height

**Table 6.** Neighbour Answers with Their Satisfaction Degrees

Id(Concept Extent)	Conditions( Concept Intent)	Answers	Sat_deg
7	Average number of years service and average height	Amal	1.00
3	Average number of years service and low salary	Hanene	0.88
5		Bassem	0.86
10	Low Salary and average height	Ikram	0.46
9	High Age and average height	Farah	0.20

5. The last step consists in applying Algorithm 3 to obtain the neighbor answers. Table 6 presents these answers, in the original format (FQ), ranked by their satisfaction degrees.

## 6 Conclusion

This paper presents an optimization of our previous approach [1] which deals with the problem of empty answers in cooperative systems. Instead of handling the whole context, our new approach handles small contexts, having similar features and, determined after a pretreatment step. Thus, we suggest two algorithms. The first one detects the failure reasons of a flexible query whereas the second one provides neighbor answers. The neighbor answers are always in the first context, so we do not need to check the entire context. Similarly, to detect the failure reasons, we find frequent cases which can happen in one of the two contexts. Hence, we reduce the search space and ensure the improvement of the response time of our previous algorithms. In future work, the vertical decomposition of the original context will be also discussed. We think that functional dependency between attributes will help us to do it.

## References

1. Hachani, N., Ben Hassine, M.A., Chettaoui, H., Ounelli, H.: Cooperative answering of fuzzy queries. *J. Comput. Sci. Technol.* 24(4), 675–686 (2009)
2. Gaasterland, T., Godfrey, P., Minker, J.: An overview of cooperative answering. *Journal of Intelligent Information Systems* 1(2), 123–157 (1992)
3. Bosc, P., Pivert, O.: Some approaches for relational databases flexible querying. *J. Intell. Inf. Syst.* 1(3/4), 323–354 (1992)
4. Bosc, P., HadjAli, A., Pivert, O.: Cooperative answering to flexible queries via a tolerance relation. In: An, A., Matwin, S., Raś, Z.W., Ślezak, D. (eds.) *ISMIS 2008. LNCS (LNAI)*, vol. 4994, pp. 288–297. Springer, Heidelberg (2008)
5. Bosc, P., HadjAli, A., Pivert, O.: Incremental controlled relaxation of failing flexible queries. *J. Intell. Inf. Syst.* 33(3), 261–283 (2009)
6. Motro, A.: Cooperative database systems. *International Journal of Intelligent Systems* 11(10), 717–731 (1996)
7. Chettaoui, H., Ben Hassine, M.A., Hachani, N., Ounelli, H.: Using fca to answer fuzzy queries in cooperative systems. In: *FSKD*, vol. 3, pp. 14–20 (2008)
8. Ganter, B., Wille, R.: *Formal Concept Analysis: Mathematical foundations*. Springer, New York (1998) (Translator-Franzke, C.)
9. Zadeh, L.A.: Fuzzy sets. *Information and Control* 8(3), 338–353 (1965)
10. Gaasterland, T.: Cooperative answering through controlled query relaxation 12(5), 48–59 (September 1997)
11. Godfrey, P.: Minimization in cooperative response to failing database queries. *Int. J. Cooperative Inf. Syst.* 6(2), 95–149 (1997)
12. Muslea, I.: Machine learning for online query relaxation. In: *KDD 2004: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 246–255. ACM, New York (2004)

13. Andreassen, T., Pivert, O.: On the weakening of fuzzy relational queries. In: Raś, Z.W., Zemankova, M. (eds.) ISMIS 1994. LNCS, vol. 869, pp. 144–153. Springer, Heidelberg (1994)
14. Voglozin, W.A., Raschia, G., Ughetto, L., Mouaddib, N.: Querying the saintetiq summaries – dealing with null answers. In: Proc. of the 14th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2005), Reno, Nevada, USA, May 2005, vol. 1, pp. 585–590 (2005)
15. Wille, R.: Restructuring lattice theory: an approach based on hierarchies of concepts. In: Rival, I. (ed.) Ordered sets, pp. 445–470. Reidel, Dordrecht (1982)
16. Asuncion, A., Newman, D.: UCI machine learning repository (2007)
17. Ben Hassine, M.A., Ounelli, H.: Idfq: An interface for database flexible querying. In: Atzeni, P., Caplinskas, A., Jaakkola, H. (eds.) ADBIS 2008. LNCS, vol. 5207, pp. 112–126. Springer, Heidelberg (2008)
18. Valtchev, P., Grosser, D., Roume, C., Hacene, M.R.: Galicia: An open platform for lattices. In: Using Conceptual Structures: Contributions to the 11th Intl. Conference on Conceptual Structures (ICCS 2003), pp. 241–254. Shaker, Aachen (2003)
19. Wolff, K.E.: Concepts in fuzzy scaling theory: order and granularity. *Fuzzy Sets Systems* 132(1), 63–75 (2002)

# Using Association Rules to Discover Color-Emotion Relationships Based on Social Tagging

Haifeng Feng, Marie-Jeanne Lesot, and Marcin Detyniecki

Universite Pierre et Marie Curie - Paris 6, CNRS, UMR7606, LIP6, France  
`{haifeng.feng,marie-jeanne.lesot,marcin.detyniecki}@lip6.fr`

**Abstract.** Relationships between colors and emotions have been studied for a long time in several domains, such as psychology and artistic theories. In this paper, we extract such relations appearing in social tagging systems, in which users can freely choose the images they upload and annotate, as well as the annotation tags. We first study two color representations that can be used to encode the chromatic contents of such images and select the most appropriate one for discovering color-emotion relationships, based on their performance for a classification task. We then extract, from this image corpus and based on the selected encoding, association rules characterizing relations between colors and emotions. We use the Apriori algorithm with a particular focus on the implications of color presence and absence on the emotion presences, commenting and discussing the obtained results.

**Keywords:** affective computing, color-emotion relationships, social tagging system, association rules.

## 1 Introduction

Color-emotion relationships have been studied for a long time in multiple fields, as artistic theories [1] or psychology [2,3,4]. Besides, in the machine learning domain, a large variety of methods has been proposed to study or extract such relationships, e.g. artificial neural networks [5], SVM [6,7] or fuzzy c-means [8]. Yet these studies are based on small databases, containing less than 2000 images.

In this paper, we use machine learning approaches to explore color-emotion relationships, considering a very large image set, acquired from the social tagging system Flickr ([www.flickr.com](http://www.flickr.com)): we aim at studying the associations that can be derived from such systems that allow users to upload and share images, as well as to tag them using any linguistic terms as labels. No constraints (others than legal constraints) restrain the user, neither for the image selection nor for the tagging step. Thus, querying such systems offers the possibility to build very large image sets that are representative of users. This is the main difference with image sets used in psychological studies, where images are selected by the researchers.

More precisely we propose to study the relationships between colors and emotions by retrieving images simultaneously tagged by one emotional term and at least one chromatic term. Such queries select from the Flickr images those that are considered as, first, triggering emotions and, second, having a significant chromatic content, according to the human evaluation provided by the linguistic tags. Thus such queries can retrieve images relevant for the study of color-emotion relationships. In particular, the chromatic tags make it possible to exclude images whose emotional content would be related to the semantic content (e.g. a smiling face being associated to joy or a spider to fear), like in the IAPS image base [9] or in the study carried out by [10].

For this image set, we first compare two representations of colors: on one hand an objective and numeric one, based on HSV histograms, and on the other hand a subjective and linguistic one, based on the color tags associated to the images. After determining the representation that appears the most appropriate for the relationships between colors and emotions by comparing their performance in a classification task, we apply the Apriori algorithm [11] to discover color-emotion relationships in the form of association rules.

The paper is organized as follows: in the next section, we describe the constitution of the image set from Flickr. In Section 3, we describe the two color representations and the experiments carried out to compare them. In Section 4, we present the extraction of rules associating colors and emotions. Lastly, in Section 5, we draw the conclusions and describe future works.

## 2 Data Acquisition

### 2.1 Social Tagging Systems

With the development of the web and the amount of available data, many social tagging systems appeared for sharing and retrieving resources. These resources include bookmarks in the case of the social bookmarking web service del.icio.us ([www.delicious.com](http://www.delicious.com)), or more frequently photos and videos, as is the case of Flickr ([www.flickr.com](http://www.flickr.com)). In these systems, users are allowed to share resources and to tag them, enabling efficient mining in the huge resulting data sets. No indexing rules are imposed, leading to free and authentic resource annotation.

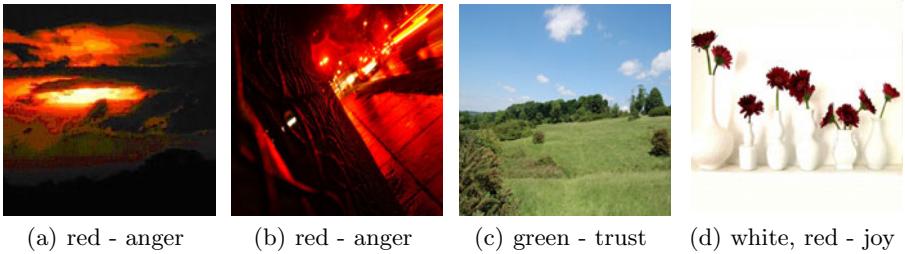
The Flickr system, considered in this paper, is a highly popular system, commonly used worldwide for storing, sharing and tagging photos within family and friends. Therefore it provides a source to build a reliable, authentic and very large corpus of images to study color-emotion relationships.

### 2.2 Collection of Images

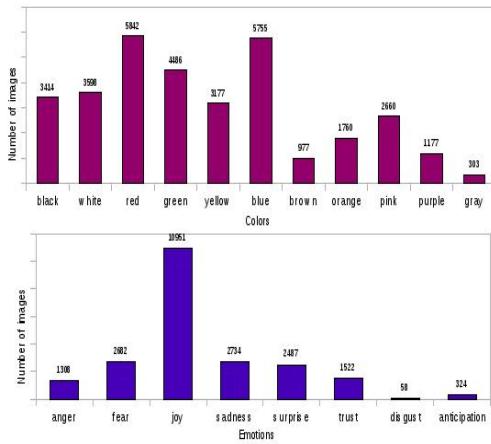
The image set considered throughout the paper is obtained by retrieving images from Flickr using queries of the following form:

**Emotions AND Colors [\*upload start time \*upload end time]**

Indeed, such queries collect images simultaneously tagged with both chromatic and emotional terms, i.e. images for which the users indicate a relation between



**Fig. 1.** Examples of collected images, with their emotional and chromatic tags



**Fig. 2.** Color and emotion distributions: number of images for each tag

color and emotions. The optional time part of the query makes it possible to restrict the images according to their upload date.

For the selection of the emotional tags used in the queries, we consider the psychological emotion model proposed by Plutchik [12] that contains 8 basic emotions declined on three intensity levels. In this paper, the intermediate level is considered, i.e. the emotions *sadness*, *anger*, *joy*, *disgust*, *surprise*, *fear*, *anticipation* and *trust*. The emotional part of the query is a disjunction of these terms, in the form (*sadness OR anger OR ... OR trust*).

Regarding the color part of the query, we consider the chromatic terms defined in the reference work on color naming of Berlin and Kay [13], i.e. *black*, *white*, *red*, *green*, *purple*, *pink*, *brown*, *gray*, *blue*, *orange* and *yellow* using again a disjunction of these terms in the query.

In the image set returned by the above mentioned query, some images are labeled with several emotions, constituting 2.2% of the retrieved images. In order to focus on non ambiguous relations between colors and emotions, we remove such images. Moreover, we remove images whose only color tags are *black* and *white* because they are considered as black-and-white images that would not be

useful in the considered framework. As a result, we obtain a total set of 22 066 images, Figure 2 illustrates some examples.

Figure 2 illustrates some statistical properties of the tag distribution in the constituted image set. It can be observed that the emotion distribution is highly unbalanced: around half the images are tagged as *joy*; on the contrary, *disgust* and *anticipation* are very under-represented, containing only 58 and 324 images respectively. We thus remove these two emotions from the study in the following.

Regarding the color distribution, it can be seen that it is unbalanced as well, although to a lesser extent: the left 6 colors are much more frequent than the others. Besides, it appears that 30% of the images are tagged with several color terms; in 70% cases, users decide that a single color is sufficient to characterize the image color content.

### 3 Comparison of Two Color Encodings

#### 3.1 Color Representations

In this section, we compare two color representations to select the most appropriate one to discover color-emotion relationships. On one hand, we use HSV histograms to numerically encode the chromatic content of the images, using 36 non-uniform bins [14].

On the other hand, we consider a linguistic description based on the chromatic tags associated with the images: we associate each image to a binary vector in  $\{0, 1\}^{11}$ , where each component indicates the absence or presence of each of the 11 considered chromatic terms. It must be underlined that this description offers  $2^{11} = 2048$  different encoding possibilities and that actually only 412 are encountered in the image set established in the previous section.

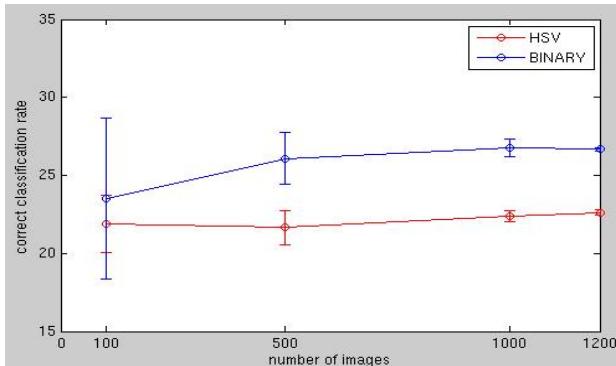
#### 3.2 Experimental Protocols

In order to compare these two color encodings, we perform classification experiments using C4.5 [15] integrated in WEKA [16] to predict the 6 emotions (*sadness*, *anger*, *joy*, *surprise*, *fear* and *trust*) from the image chromatic content.

As C4.5 is sensitive to unbalanced data distributions, we apply random samplings without replacements, constituting data sets respectively containing a constant value of 100, 500, 1000 and 1500 images for each emotion. For each data set, we compute the correct classification rate, averaged on 6 runs. To evaluate the result quality, we compare it to a baseline classifier that randomly predicts the 6 emotions, whose correct classification rate is thus 16%.

#### 3.3 Results and Discussions

Figure 3 shows the correct classification rates of all experiments. They all are well above the baseline random classifier, and are stable for both encodings, insofar as it does not vary significantly with the size of the training set.



**Fig. 3.** Comparison of the 2 image encodings, for different sizes of the training corpus

Moreover the curves show that the binary linguistic encoding is significantly superior to the numerical HSV encoding. This is all the more noteworthy since, as mentioned before, it suffers from a low number of distinctive possible codes. The reason that can be proposed to explain this result is related to subjectivity of the linguistic encoding: it results from a human interpretation of the images, and indicates the colors that are judged important by the users. The latter may significantly differ from the most frequent color, or more generally from the color distribution as encoded in the HSV histograms: the human tagger can focus on a rare color that would be a minor color in the histograms, and nevertheless conveys the main impression from the image.

This can be illustrated by Figure 1(a) and 1(d): the image 1(a) is tagged with *red* whereas the most frequent color, as deduced from the histogram is *black*. Likewise, image 1(d) is tagged with *red* although this color is not the dominant color in image and can even be considered as rare from the histogram.

As a result, the binary linguistic encoding appears to be the most appropriate color representation and is used in the following.

## 4 Color-Emotion Relationships Discovery with Association Rules

The classification task in the previous section was not aimed at predicting emotions from colors, but at selecting the most appropriate color encoding. Indeed, we are not interested in a global model to perform classification, but in identifying relevant regularities relating colors and emotions. The decision trees built by C4.5 offer some interesting rules, such as "in images not tagged with *gray*, about 99% are tagged with *joy*". In order to discover more such rules and exploiting the fact that the linguistic encoding is binary, we turn to discover such relationships through association rules.

#### 4.1 Experimental Protocol

We are interested in maximal rules of the form  $A \Rightarrow B$  where  $A$  denotes a conjunction of presence of chromatic terms and  $B$  represents the presence of one emotion. To identify these rules, we apply the Apriori algorithm with a support threshold of 0.01. This low value comes from the fact that the numbers of images for some emotions are very low.

There exist many interestingness measures for association rules [17][18]. In the following, we focus on the *relative risk* measure, defined for a rule  $A \Rightarrow B$  as

$$RR = \frac{N_{ab}}{N_a} \times \frac{N_{\bar{a}}}{N_{\bar{a}b}}$$

where  $N_a$  represents the number of images whose tags contain the terms in  $A$ ,  $N_{ab}$  the number of images presenting both  $A$  and  $B$ .  $N_{\bar{a}}$  and  $N_{\bar{a}b}$  respectively represent the number of images not tagged with  $A$  and images tagged with  $B$  but not  $A$ . A relative risk value higher than 1 indicates that  $B$  is significantly more present in image tags when  $A$  also is present, than when  $A$  is not. This measure presents the advantage of being normalized with respect to the global frequencies of  $A$  and  $B$ , which is necessary for the data we consider in which the number of images for each emotion and for each color varies significantly.

Therefore, we extract rules for which relative risk is higher than 1.4, meaning that the frequency of emotion  $B$  in the color conjunction  $A$  is 40% superior to the frequency of emotion  $B$  when the color conjunction is not in the image tags. This indicates significant color presences that characterize the emotions.

Moreover as the relative risk of  $A \Rightarrow B$  is inverse of the relative risk of  $\bar{A} \Rightarrow B$ , we also extract rules for which relative risk is lower than  $1/1.4 = 0.71$ : such rules characterize an emotion by the absence of colors and are also highly relevant. We do not exhaustively look for absence rules, in the form "the absence of *several* colors  $A \Rightarrow$  emotion  $B$ ", because all colors and emotions are predominantly absent of all data, which would result in too many rules. Focusing on presence rules for which  $RR < 0.71$  provides the means to extract only relevant such rules. The lower the relative risk, the more significant the absence of the colors.

Lastly, to ease the interpretation of the obtained results, we also indicate for each rule the values of *support* ( $S = N_{ab}/N$ ), *confidence* ( $C = N_{ab}/N_a$ ) and *lift* ( $L = (N_{ab}/N_a)/(N_b/N)$ ), where  $N$  denotes the total number of images.

#### 4.2 Results and Discussions

**Emotion Characterization.** Table I presents the obtained rules for which the relative risk is higher than 1.4 or lower than 0.71 and that are maximal in terms of premise length: for instance the rule *green blue*  $\Rightarrow$  *joy* is removed because it is included in the rule *red green yellow blue*  $\Rightarrow$  *joy* (rule n°3). As a result 18 color-presence rules and 6 color-absence rules, characterizing the presence of an emotion by the presence (resp. the absence) of color combinations, are identified.

For the presence rules, it appears that all have low support, due to the low number of co-occurrences of any color-emotion combination. This is also the

**Table 1.** Maximal rules with relative risk (RR) higher than 1.4 or lower than 0.71, with their support (S), confidence (C) and lift (L), sorted by the relative risk value

	No.	Rule	S	RR	C	L
color presence	1	<i>black</i> $\Rightarrow$ <i>fear</i>	0.037	<b>2.46</b>	0.24	2
	2	<i>red</i> $\Rightarrow$ <i>anger</i>	0.025	<b>2.00</b>	0.09	1.58
	3	<i>red green yellow blue</i> $\Rightarrow$ <i>joy</i>	0.010	<b>1.72</b>	0.85	1.71
	4	<i>yellow purple</i> $\Rightarrow$ <i>joy</i>	0.010	<b>1.66</b>	0.81	1.64
	5	<i>red yellow orange</i> $\Rightarrow$ <i>joy</i>	0.011	<b>1.63</b>	0.80	1.61
	6	<i>yellow pink</i> $\Rightarrow$ <i>joy</i>	0.013	<b>1.62</b>	0.79	1.60
	7	<i>black</i> $\Rightarrow$ <i>anger</i>	0.013	<b>1.61</b>	0.08	1.47
	8	<i>black</i> $\Rightarrow$ <i>sadness</i>	0.027	<b>1.57</b>	0.17	1.44
	9	<i>blue pink</i> $\Rightarrow$ <i>joy</i>	0.016	<b>1.56</b>	0.76	1.54
	10	<i>red purple</i> $\Rightarrow$ <i>joy</i>	0.010	<b>1.53</b>	0.75	1.51
	11	<i>green orange</i> $\Rightarrow$ <i>joy</i>	0.012	<b>1.51</b>	0.74	1.50
	12	<i>green</i> $\Rightarrow$ <i>trust</i>	0.019	<b>1.51</b>	0.09	1.37
	13	<i>blue purple</i> $\Rightarrow$ <i>joy</i>	0.010	<b>1.50</b>	0.74	1.49
	14	<i>green purple</i> $\Rightarrow$ <i>joy</i>	0.010	<b>1.50</b>	0.74	1.49
	15	<i>pink</i> $\Rightarrow$ <i>surprise</i>	0.018	<b>1.47</b>	0.15	1.39
	16	<i>red pink</i> $\Rightarrow$ <i>joy</i>	0.015	<b>1.47</b>	0.72	1.46
	17	<i>green pink</i> $\Rightarrow$ <i>joy</i>	0.015	<b>1.47</b>	0.72	1.45
	18	<i>blue orange</i> $\Rightarrow$ <i>joy</i>	0.012	<b>1.44</b>	0.71	1.43
color absence	1	<i>yellow</i> $\Rightarrow$ <i>fear</i>	0.013	<b>0.71</b>	0.09	0.74
	2	<i>black</i> $\Rightarrow$ <i>surprise</i>	0.012	<b>0.69</b>	0.08	0.72
	3	<i>green</i> $\Rightarrow$ <i>sadness</i>	0.017	<b>0.65</b>	0.08	0.69
	4	<i>black</i> $\Rightarrow$ <i>joy</i>	0.052	<b>0.64</b>	0.33	0.68
	5	<i>blue</i> $\Rightarrow$ <i>anger</i>	0.011	<b>0.64</b>	0.04	0.70
	6	<i>yellow</i> $\Rightarrow$ <i>sadness</i>	0.012	<b>0.63</b>	0.08	0.66

reason why *joy* is the only emotion for which rules involving several colors are obtained: for all other emotions, the support of the itemsets is below the used threshold. For the same reason, confidence also tends to be low, except for the *joy* rules for which it achieves values higher than 0.7. On the other hand, it can be observed that for all the obtained rules, the *lift* values are higher than 1, indicating that the proportion of images tagged with the involved emotion among those presenting the considered color combination is significantly higher than the global proportion. Regarding the absence rules, for which only the relative risk value is significant, it appears that the lowest value remains high (0.63). The number of absence rules is thus small.

The rule with maximal relative risk characterizes the presence of *fear* by that of *black*, with a very high score: *fear* is 2.5 times more present in images tagged with *black* than in images not tagged with *black*. Moreover, this relation is all the stronger as it is the only rule that concludes to *fear* in the presence rules. The absence rules indicate that *fear* can be derived from the absence of *yellow*. It can be underlined that this characterization corresponds to a commonly accepted

and intuitive representation: it illustrates the fact that the data gathered from the social tagging system reflect widely spread cultural characteristics.

The second best presence rule establishes a link between *red* and *anger* associated to high values both for relative risk (2) and lift (1.5). This means that *red* can be interpreted as having a priority to characterize *anger*. Still, it must be underlined too that, from rule *n°3*, *red* combined with other colors implies *joy*. Besides, it appears that *anger* is also related to *black*, to a lesser extent but still with a significant relative risk value higher than 1.5, which can be seen as a less obvious result.

The following rules involve *joy*, with many different colors. As already mentioned, these results show that the proposed approach still has a bias to the overall frequency: the emotion distribution is so unbalanced that the support threshold of 0.01, although very low, still leads to many rules associated to the most frequent emotion, even if the considered interestingness measure is not sensitive to the frequencies. As a consequence, it appears that the color characterization of *joy* is more difficult than that of rarer emotions: many different colors are attached to it. Still, it appears that most colors are bright ones, even if purple also appears to play an important role.

*Sadness* appears to be characterized by the presence of *black*, and *trust* by that of *green*, which again constitutes results compatible with common representations.

Lastly, it can be underlined that dark colors, and in particular *black*, are associated to all three negative emotions (*fear*, *anger* and *sadness*), and that its absence is considered as characterizing the most positive emotion, *joy*. On the contrary, the absence of the brightest color, *yellow*, is associated to both *fear* and *sadness*.

**Color Characterization.** In this section, we turn to exploiting the identified rules to characterize the color use, instead of analysing them emotion after emotion. To that aim, Table 2 shows the rules as a double entry matrix, indicating for each color-emotion pair whether the color characterizes the emotion by its

**Table 2.** Significant colors for the emotions with the relative risk (P: significant presence, A: significant absence, in parentheses, RR value)

Color	sadness	anger	fear	surprise	trust	joy
red	–	P (2.0)	–	–	–	–
green	A (0.65)	–	–	–	P (1.51)	–
yellow	A (0.63)	–	A (0.7)	–	–	–
black	P (1.57)	P (1.61)	P (2.46)	A (0.69)	–	A (0.64)
blue	–	A (0.64)	–	–	–	–
pink	–	–	–	P (1.47)	–	–
red green yellow blue	–	–	–	–	–	P (1.72)

presence (P), its absence (A) or does not characterize it (-). Moreover, it recalls, in parentheses, the associated relative risk value.

The table highlights the fact that *red* only characterizes the *anger* emotion, underlining the strength of this relationship. On the contrary it also shows that *black* cannot be associated to a single emotion, but to three of them, even if *fear* obtains the highest score. These three emotions are the three negative emotions in our study. *Green* appears to be a color highly typical for *trust*, as it only characterizes the presence of this emotion. As reciprocally, no other representative color appears for *trust*, *green* is very appropriate to characterize it. The presence of *yellow* does not appear to characterize any emotion, but its absence is related to *sadness* and *fear*. *Blue* only characterizes an emotion by its absence, namely *anger*, which means it is not a very significant color.

Lastly, it is interesting to see that several colors do not appear in any rules: this holds for the not so frequent colors, *orange*, *purple*, *brown* and *gray*, but also for *white*. Now the latter has a significant overall frequency: its absence indicates that this color may convey many different emotions, and cannot be used directly.

## 5 Conclusion

In this paper, we collect a set of images with emotional content, for which the emotional impression is related to the chromatic content, querying the social tagging system Flickr. We propose two image encodings and study the color-emotion relationships through existing classification and association rule algorithms. The classification experiments make possible to select the most appropriate color representation: they show that the linguistic encoding is more significant and relevant than the numeric one to extract color-emotion relations. This is due to the subjectivity of the linguistic coding, that results from a human interpretation of the chromatic content of the images. Relationships between colors and emotions can be extracted using the Apriori algorithm, in the form of association rules.

The identified relations show interesting results, that are often compatible with common representations about the emotions conveyed by colors: they highlight the widely spread cultural impressions shown by the social tagging system. Some other rules are less expected and encourage to refine the obtained results.

In particular, future works will include the introduction of fuzzy chromatic description based on HSV relations between colors so as to possibly enhance rare chromatic tags. Another perspective aims at developing a more refined interpretation of absence rules, and at studying their relevance in the framework of freely chosen labels. Indeed, the absence of tagging constraints and indexing rules, that leads to natural and authentic image annotation, also leads to question the relevance of tags that may depend on the personal subjectivity of the annotator.

## Acknowledgments

This work was supported by the project GENIUS funded by ANR, n°07 TLOG 005.

## References

1. Itten, J.: The art of color: the subjective experience and objective rationale of color. Wiley, Chichester (1997) (translated from the German version published in 1961)
2. Hemphill, M.: A note on adults' color-emotion associations. *Journal of Genetic Psychology* 54, 275–281 (1996)
3. Ou, L.-C., Luo, M.R., Woodcock, A., Wright, A.: A study of colour emotion and colour preference. *Color Research and Application* 29, 232–240 (2004)
4. Clarke, T., Costall, A.: The emotional connotations of color: a qualitative investigation. *Color Research and Application* 33, 406–410 (2008)
5. Hayashi, T., Hagiwara, M.: Image query by impression words - the IQI system. *IEEE Transactions on Consumer Electronics* 44, 347–352 (1998)
6. Wang, W., Yu, Y., Jiang, S.: Image Retrieval by Emotional Semantics: A Study of Emotional Space and Feature Extraction. *IEEE Systems, Man and Cybernetics (SMC)* 4(8-11), 3534–3539 (2006)
7. Wu, Q., Zhou, C., Wang, C.: Content-Based Affective Image Classification and Retrieval Using Support Vector Machines. *Affective Computing and Intelligent Interaction*, 239–247 (2005)
8. Wei, K., He, B., Zhang, T., He, W.: Image Emotional Classification Based on Color Semantic Description. In: Tang, C., Ling, C.X., Zhou, X., Cercone, N.J., Li, X. (eds.) ADMA 2008. LNCS (LNAI), vol. 5139, pp. 485–491. Springer, Heidelberg (2008)
9. Lang, P.J., Bradley, M.M., Cuthbert, B.N.: International affective picture system (IAPS): Technical manual and affective ratings, University of Florida, Center for Research in Psychophysiology, Gainesville (1999)
10. Schmidt, S., Stock, W.G.: Collective Indexing of Emotions Images. A Study in Emotional Information Retrieval (EmIR). *Journal of the American Society for Information Science and Technology* 60, 863–876 (2009)
11. Xie, Y., Li, Y., Wang, C., Lu, M.: The Optimization and Improvement of the Apriori Algorithm. In: Int. Symp. on Intelligent Information Technology Application Workshops, pp. 1101–1103 (2008)
12. Plutchik, R., Kellerman, H.: Emotion: Theory, Research and Experience. Academic Press, San Diego (1990)
13. Berlin, B., Kay, P.: Basic color terms: their universality and evolution. University of California Press, Berkeley (1969)
14. Zhang, L., Lin, F., Zhang, B.: A CBIR Method Based on Color-Spatial Feature. In: Proc. IEEE Region 10 Annual Int. Conference, TENCON 1999, Cheju, Korea (1999)
15. Kohavi, R., Quinlan, R.: Decision Tree Discovery. In: Klosgen, Zytkow (eds.) Handbook of Data Mining and Knowledge Discovery, pp. 267–276. Oxford University Press, Oxford (2002)
16. WEKA, Machine Learning Platform (2008),  
<http://www.cs.waikato.ac.nz/ml/index.html>
17. Zhang, Y., Zhang, L., Nie, G., Shi, Y.: A Survey of Interestingness Measures for Association Rules, Business Intelligence and Financial Engineering (BIFE). In: Proc. Int. Conference on Business Intelligence and Financial Engineering, pp. 460–463 (2009)
18. Lenca, P., Vaillant, B., Meyer, P., Lallich, S.: Association Rule Interestingness Measures: Experimental and Theoretical Studies. In: Guillet, F., Hamilton, H. (eds.) Quality Measures in Data Mining, pp. 51–76. Springer, Heidelberg (2007)

# A Conceptual Framework for Role-Based Knowledge Profiling Using Semiotics Approach

Nazmona Mat Ali and Kecheng Liu

Informatics Research Centre, University of Reading,  
Reading, RG6 6WB, United Kingdom  
[{n.matali,k.liu}@reading.ac.uk](mailto:{n.matali,k.liu}@reading.ac.uk)

**Abstract.** The difficulty in finding knowledge related to individual's roles as well as an appropriate mechanism for representing organisational knowledge are seen as a crucial process nowadays. Knowledge profiling is therefore designed to overcome these issues by providing a systematic set of phases or steps that carry a collection and documentation of individual and organisational knowledge. This paper shows the use of two methods of semiotics approach towards information systems (i.e. organisational morphology and norms analysis) to identify and analyse knowledge based on semiotics point of view whereby knowledge is often referred as norms. This means, knowledge and hence norms have ability in directing, coordinating and controlling actions in acceptable manners by members within a community or society. The outcome of a successful knowledge profiling is a profile that consists of an inventory of existing and needed knowledge requirements as well. In order to illustrate the execution of knowledge profiling, this study has used the roles of academic members at higher education institutions as a case study.

**Keywords:** knowledge profiling, semiotics, knowledge, norms, organisational morphology, norm analysis method (NAM).

## 1 Introduction

The emergence of knowledge profiling theory lately is seen as a great advantage for organisations to be more productive and competitive in the business world. Its benefits are not only for industrial organisations but also intended for higher education organisations. This is because most universities nowadays are preparing their academic members to face many complex demands. Members of academic are being appointed for performing an increasingly diverse range of roles which of course involve many tasks. These roles and responsibilities however are often ambiguous and thus difficult for them to act appropriately. At the same time, this situation contributes to another issue where academic members are often burdened with repetitive questions and requests for help from their colleagues particularly from new members. Furthermore, the implementation of the concept of organisational knowledge is essentially required for an organisation which integrates all knowledge at the organisational level from individual knowledge of its members [1]. However, it is important for this knowledge to be represented in a formal way that people can easily understand and

follow, and later to be automated into computer-based systems. This knowledge consequently can be shared and accessed by other members.

The study has taken a step further by using semiotics approach which is mainly concerned with signs that must stand for something in the real world and fulfil someone's purpose in the same time [2]. Therefore, semiotics has introduced the notion of knowledge as norms in order to provide guidance for people to act on an agreed moral and acceptable within the social context. Norms are generally triggered by signs and signs can be anything whether in the form of documents, oral communication or behaviour. A norm moreover can in turn be treated as a sign in another context since executing norms will lead to more signs being produced [3]. Semiotics approach is also seen to coincide with the nature of the study since it is closely related with the organisational action which certainly involves individuals who have many tasks and possess knowledge, and also context that must be taken into consideration. Thus, an effective approach to derive a conceptual framework for role-based knowledge profiling is semiotics theory, which is elaborated in this paper.

The rest of the paper is organised as follows: Section 2 reviews existing literature on knowledge profiling framework and discuss their constraints particularly if they are being implemented in a given case study. Section 3 introduces a concept of semiotics and its relationship between signs, norms and knowledge. This followed by brief descriptions of two semiotics' methods that are essentially employed in this study. Section 4, illustrates the use of those methods using a case study. Finally, section 5 summarises the study and outlines future research directions.

## 2 Existing Knowledge Profiling Frameworks and Their Constraints

The significance of knowledge profiling for individuals and organisations is recently recognised due to its capability in documenting and presenting knowledge gained from individuals or groups. Until now however, there has been little research in this field and thus offers a lot of exploration and investigation in order to practically prove in reality context. Therefore, in order to achieve an inclusive knowledge profiling, the current practices and any barriers in existing knowledge profiling techniques are examined. Currently, there are three techniques of knowledge profiling that have been introduced by [4], [5] and [6]. These three frameworks have shown different purposes and consequently involve different phases. The review of these techniques is illustrated in Table 1.

Though these three knowledge profiling techniques have brought benefits to the context where it is applied, there are still ample spaces for enhancement mainly in representing individual knowledge which is related to roles and tasks assigned by an organisation. For example, most knowledge is demonstrated in the form of narratives and it is difficult to elicit the essence of its actions. This structure of knowledge is also referred as descriptive knowledge, knowledge which is usually expressed in declarative sentences instead of directing people on how to act in appropriate manners (e.g. how best to perform some task). In the contrary, the implementation of individuals' roles and subsequently tasks are basically related with the way how individuals tend to behave and think within their context i.e. an community or society. Therefore, the

representation of knowledge in norms can facilitate the use of knowledge itself which essentially provides guidance for people to act in certain ways. Most importantly, those actions are accepted and agreed by members within an organisation. Furthermore, this representation of knowledge hence norms, is compliance with automation purpose, that is for developing computer-based systems.

**Table 1.** Three frameworks of existing knowledge profiling (KP)

Author(s)	Aim	Phases	Contexts where KP is applied
Thellefsen [4]	To create realistic representation of knowledge organisation which is mainly for sharpening the terminology, removing redundant and misleading connotations.	Involves six main steps (for further details see [4])	Epistemology (e.g. concept or knowledge domain)
Engel, Huppert and Cleveringa [5]	To document and record knowledge that has been obtained through practical experience and based on lessons learned from stakeholders point of view.	Consists eleven main steps (for further details see [5])	Project or programme where practical experience is highly used
Edwards and Gibson [6]	To help researchers from all disciplines and background to bring diverse sources of knowledge together to address common issues as well as assemble the right research team.	Involves four main steps (for further details see [6])	Community-based research project

### 3 Semiotics Approach for Conceptualisation of Knowledge Profiling

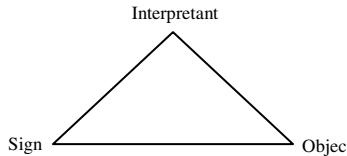
The use of semiotics approach for executing a process of knowledge profiling is seen as a great advantage for organisations to understand two main aspects in information systems: technical and human aspect. Therefore, many research projects currently have focused their attention on using semiotics approach to solve the problems (e.g. [7-9]).

Therefore, understanding of fundamental concepts of semiotics is important in order to embed this theory of sign into the case study. This includes semiotics process called semiosis and its relationship between signs and norms. Two methods of semiotic that have been identified as suitable techniques to classify and analyse norms are also shown in this paper.

#### 3.1 Relationship between Signs and Norms

Semiotics differs with other theories, which leads us to a very different view. This theory is concerned with human aspects without ignoring technical aspects at the same time. Basically, there are many philosophers who contribute in semiotics and one of them is Charles Sanders Peirce (1839-1914). Pierce's well known contribution is on the interaction between a sign, object and interpretant, called as a semiosis. These elements are strongly correlated and interdependent each other as shown in Fig. 1.

- *The sign*: the form which the sign takes (not necessarily material)
- *An interpretant*: not an interpreter but rather the sense made of the sign or the concept it represents.
- *An object*: to which the sign refers or what the sign ‘stand for’

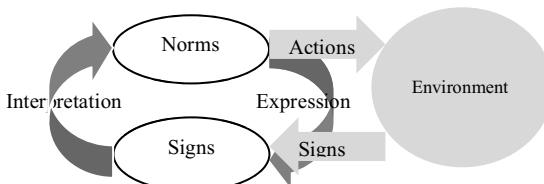


**Fig. 1.** A version of Peirce's semiotic triangle

The concept of interpretant however needs a human role for the sign makes sense. In other word, there is always someone involved, that is, to whom the signification makes sense [10]. The role of the interpreter therefore must be accounted for either within the formal model of the sign or as an essential part of the process of semiosis [11]. At the same time, it is important to note that the interpreter must have knowledge or norm in order to associate the sign and the object. Norms and signs are inseparable and its relationship has been represented by Stamper [12] in his model in Fig. 2. and quoted as the following:

*“To recognise when a norm should be triggered, the subject needs information (signs) relating to the condition. The resulting attitude may not produce an immediate outcome but sooner or later will be revealed in words or comportment, or sometimes translated eventually into action. In either case the result will be more signs.”*

Fig. 2. shows both the signs and norms as complementary to each other since norms can be expressed in all kinds of signs (e.g. documents, oral communication or behaviour) which it can be turn be treated as a sign in another process of semiosis. Concurrently, signs can be interpreted into meaningful interpretation when they trigger norms in the mind of interpreters which directly affects to whom the sign makes sense.



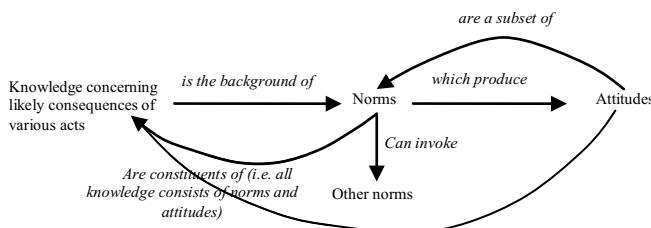
**Fig. 2.** The relationship between signs and norms [2]

### 3.2 The Notion of Knowledge as Norms

The notion of knowledge as norms has given huge advantages to organisations to understand the way of people doing their works and simultaneously showing how

they act in particular conditions. The essences of norm are it can be shared and importantly all actions being taken are on an agreed moral or acceptable within the social context [10, 13]. This clearly depicts the desirability of semiotics theory is concerned with formality aspects without ignoring the human and social aspects. These norms further can be integrated at the organisation level and thus can be documented systematically through knowledge profiling process. The study primarily refers to two main sources by Braf [14] and Stamper [15] that highly contribute in making relationship between norm and knowledge. For Braf [14], '*a norm is knowledge concerning value standards for action and action and action result and governs human behaviour*'. Norms are believed have a crucial role in the change of work and later influencing a culture within an organisation. This is how norms play their functions in an organisational aspects by assisting people to understand how an organisation works, who and what governs the activities being performed.

In the point of view of Stamper [15], he argues that norms together with attitudes constitute knowledge as illustrated in Fig. 3. A norm is more like a field of force that makes the members of the community tend to behave or think in a certain way [2]. All knowledge can then be classified into four main categories; perceptual, evaluative, cognitive and behavioural. Perceptual norm refers knowledge about what exists, cognitive norm reflects knowledge about how the world functions, evaluative norm means knowledge about what is good or bad, safe or dangerous etc and behavioural norm refers knowledge about how what one should do.

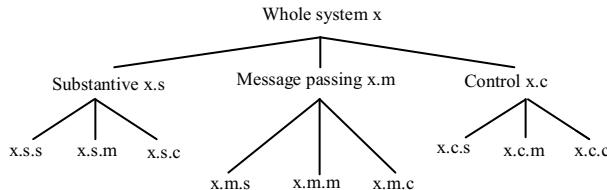


**Fig. 3.** Illustration of Stamper's [15] relation between knowledge, norms and attitudes

### 3.3 Organisational Morphology: A Functional View of Norms

Basically, the terminology of morphology is initiated from the biology discipline which deals with the structure of animals and plants. This concept is then adapted into organisational theory and unlike the conventional view of organisational structure, semiotics approach has taken a step further by introducing the functional view of norms known as organisational morphology. It is usefully used to identify tasks and afterwards classify them into three types of tasks. All tasks are directed by norms and hence three types of norms exist: substantive, communication and control norms. Therefore, norms can be classified according to the types of objects that the norms are applied to [10]. Each type of tasks can be further divided at a more detailed level, for example substantive message task (x.m.s), messages about message (x.m.m) and control of messages (x.m.c) as illustrated in Fig. 4. Consequently, the concept of recursion is introduced where communication or control tasks in their turn can be

treated as substantive tasks. The level of detail however depends on the context where it is applied and on what is adequate for the purpose of analysis [16]. All these tasks should be found in any business processes of an organisation.



**Fig. 4.** Organisational morphology [16]

Substantive tasks are concerned with the essential tasks of an organisation which conforming norms that associated with these tasks will directly contribute to the achievement of the organisation's goals. This type of task usually requires some physical actions such as giving a lecture as a part of teaching role of academic members in higher education institutions. Communication tasks which are also known as message passing emphasise on how to inform relevant people about relevant facts, work procedures and what actions are to be taken as well as when and by whom. In current practices, this type of task is usually carried out through activities such as announcements of events, sending letters for meetings, and emails and telephone calls for notifying people of any activities. Unlike substantive and communication tasks, control tasks act as a force to monitor and evaluate substantive and communication tasks. Normally, these tasks are executed through implementation of rewards and punishments which indirectly guide members within an organisation to what they are supposed to do.

### 3.4 Norm Analysis Method (NAM)

Norm analysis is a method for eliciting the detail of norms [17] which consists four main stages as follows:

#### i. Responsibility analysis

Generally, each task has at least two norms: startNorm and finishNorm. The task however could have other norms called as operationalnorms. StartNorm is concerned with a norm that make a particular task to begin meanwhile finishNorm refers to a norm that make that task has ended. OperationalNorm reflects to a norm to keep in existence of a particular task [18]. These norms need to be analysed in terms of who is responsible for executing those norms. It is important to note that *who* is here referred to an agent that might be a person, group, organisation, software or physical artefact.

#### ii. Information identification

The main purpose of this stage is to identify relevant information which is essentially required for the responsible agent to make decisions corresponding to the particular

task. Note here that the responsible agents have been determined in the responsibility analysis phase.

### iii. Trigger analysis

In this stage, any determiners or triggers that cause the particular task to start, sustain or finish are identified. There are two types of triggers namely pre-condition and post-condition. Pre-condition refers the conditions for invoking the norm and post-condition is concerned with the state of affair or action that can be performed after all identified pre-conditions are fulfilled. Note that for a particular task there are at least two post-conditions i.e. for startNorm and finishNorm.

### iv. Norm specification

It is important for each norm to be explicitly expressed in a formal representation thereby people can easily follow and can subsequently be automated into computer-based systems. Based on the nature of the case study, most norms fall into the category of behavioural norms that corresponds to what has to be done (i.e. obligatory, permitted or forbidden). These norms can thus be represented using the following form:

**whenever <context> if <condition> then <agent> is <deontic operator> to <action>**

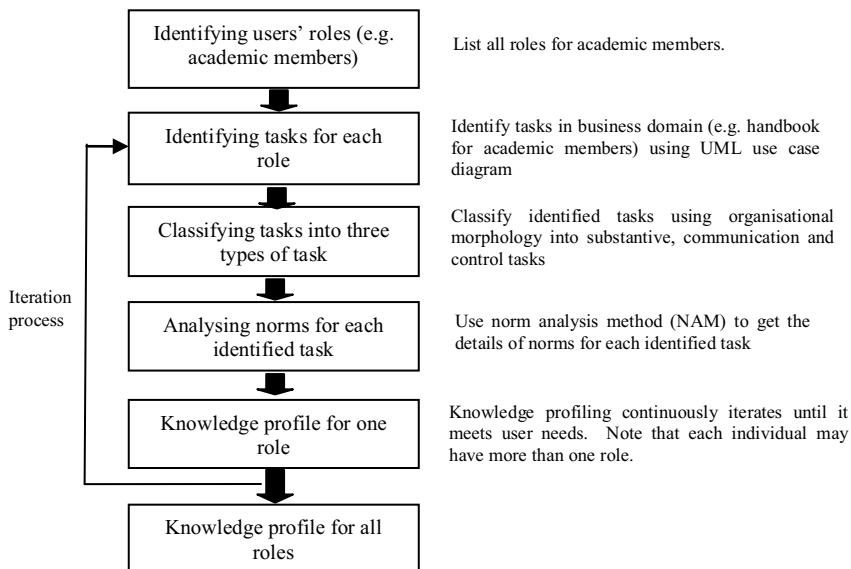
Here, the context and condition are defined from the determined pre-conditions in the previous stage, trigger analysis.

- *Context* is certainly related to the situation in which the responsible agent exists or roles that the agent plays. It often describes who and what.
- *Condition* is corresponded with any circumstances that need to be met in order to relevant actions can be executed.
- *Agent* refers to who will execute the action. Here, an agent can be a person, group, organisation, software or physical artefact.
- *Deontic operator* specifies what kind of action will be executed whether obliged, permitted or prohibited.
- *Action* reflects what act to be performed based on the conditions that have been determined in the trigger analysis phase.

## 4 Case Study: A Role-Based Knowledge Profiling for Academic Members

It is well known that academic members are the main resource for any universities in the world. They play important roles in providing human capital for industries as well as governments. They also lead in growing the countries' economy through their findings of research activities. More than that, in recent years they are being appointed for performing an increasingly diverse range of roles such as academic administrative, student development and society services. Many universities however do not provide a comprehensive and well-documented profile that governs particularly to new academic members on what they are supposed to do when entering an academic world. In fact, the existing profiles do not fulfil academic members' requirements

when there is unclear separation between essential tasks and supporting tasks. Those profiles also do not provide the details of tasks such as when the task should be executed, in what condition and who is responsible to execute that task. These ambiguous tasks will essentially give a huge impact to the whole of organisation's operation. The conceptual framework of role-based knowledge profiling is therefore designed to support academic members in collecting and documenting individual knowledge mainly related to roles that have been assigned by their organisations as shown in Fig. 5.



**Fig. 5.** A conceptual framework for role-based knowledge profiling

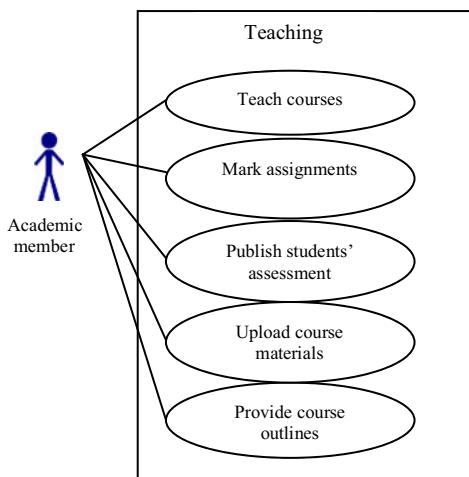
### i. Identifying users' roles

The users here are referred to academic members who are involve in diverse roles assigned by an organisation (i.e. a university). The identification of roles can be easily recognised if the execution of an organisation's current activities is well understood. The investigation can thus be initiated by analysing relevant documents such as academic handbooks, university's website and teaching portfolio. At the same time, interview sessions with potential users can also be carried out to obtain confirmation and additional information. In general, most academic members involve in teaching, research, academic-related administrative, student development, and consultation services.

### ii. Identifying tasks for each role

The main purpose of the second stage is to identify tasks and its norms for each role. Similar to the first stage above, the identification of tasks can be essentially done by analysing academic handbooks and teaching portfolios as well. These two main

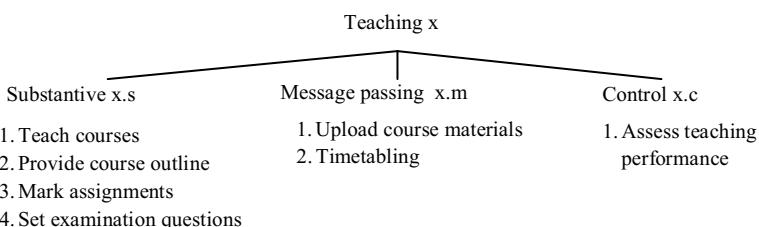
resources usually describe the basic activities that academic members are supposed to do. In this stage, we employ a UML (Unified Modeling Language) use case diagram to depict roles of academic members as shown in Fig. 6. A member of academic is represented in term of actor and drawn as a stick figure. Tasks for each role are illustrated by horizontal ellipses. Note that there are possibilities for discovering additional information in the later stages in order to provide a comprehensive knowledge profile.



**Fig. 6.** UML use cases for a role of teaching

### iii. Classifying tasks into three types of task

At this stage, the tasks that have been identified in the previous stage will be classified into three types of tasks: substantive, communication and control tasks. In the meantime, the relationship between these three types of tasks should also be analysed since they tend to complement each other in order to achieve organisation's objectives. In the example of task *teach courses*, there is another task is needed to monitor and evaluate that task (e.g. *assess teaching performance of academic members*) as illustrated in Fig. 7. Though this control task is performed by students, it is important for academic members to be informed who evaluate their teaching performance.



**Fig. 7.** The classification of tasks into three types of tasks

#### iv. Analysing norms for each identified task

The fourth stage is concerned with the details of norms where norms are often executed whether formally, informally or by technical means. Norm analysis method (NAM) is therefore applied to identify responsible agents as well as specify conditions to be met in order to some actions that correspond to obligatory, permitted and forbidden can be performed. For example of the task *provide course outline*, three norms have been identified as demonstrated in Table 2-4. It is important to note that each task at least have two norms, startNorm and finishNorm. In order to keep in existence some tasks might need another norm known as operationalNorm. The result from analysing those tasks can afterwards be used to determine the way how those norms will be implemented formally, informally or technically. In this case study, for instance the finish norm can ideally be automated in computer-based system which helps to circulate course outlines easily and effectively.

#### v. Knowledge profile for one role

Knowledge profiling is an iterative process which continues until individuals' needs are met. The main outcome from this process therefore is a profile which acts as an inventory to document existing as well as needed knowledge requirements. This knowledge profile is also importantly used to represent individual knowledge and organisational knowledge whenever any changes are required to improve particular

**Table 2.** The details of startNorm for the task *provide course outline*

Stages	Tasks	Outcome
1	Responsibility analysis	Coordinator
2	Information identification	Course code, staff id
3	Triggers analysis	Pre-condition <i>a member of academic staff is a coordinator for the particular of course, the semester is begin</i> Post-condition <i>the course outline is needed to be prepared</i>
4	Norm specification	<b>whenever</b> <i>a member of academic staff is a coordinator for the particular of course</i> <b>if</b> <i>the semester is begin</i> <b>then</b> <i>the coordinator is obligated to prepare a course outline</i>

**Table 3.** The details of operationalNorm for the task *provide course outline*

Stages	Tasks	Outcome
1	Responsibility analysis	head of department
2	Information identification	Course code, staff id
3	Triggers analysis	Pre-condition <i>a member of academic staff is a head of department, the semester is begin</i> Post-condition <i>the course outline is needed to be checked and approved</i>
4	Norm specification	<b>whenever</b> <i>a member of academic staff is a head of department</i> <b>if</b> <i>the semester is begin</i> <b>then</b> <i>the head of department is obligated to approve a course outline</i>

**Table 4.** The details of finishNorm for the task *provide course outline*

Stages	Tasks	Outcome
1	Responsibility analysis	Coordinator
2	Information identification	Course code, staff id
3	Triggers analysis	Pre-condition <i>a member of academic staff is a coordinator for the particular of course, the semester is begin</i> Post-condition <i>the course outline is needed to be circulated to students</i>
4	Norm specification	<b>whenever</b> <i>a member of academic staff is a coordinator for the particular of course</i> <b>if</b> <i>the semester is begin</i> <b>then</b> <i>the coordinator is obligated to circulate a course outline</i>

tasks. This kind of situation usually happens in business process reengineering. At the same time, these knowledge requirements should be stored in any form of database or knowledge base if an organisation intends to automate some of the tasks to achieve efficiency.

## 5 Conclusions

In this paper, we have highlighted the critical needs for developing knowledge profiling for documenting role-related knowledge of academic members. Our review of the literature has suggested that the representation of knowledge in norms is seen as great advantages in order to facilitate the use of knowledge which directly governs how people behave, think, make judgements and perceive the world. Furthermore, the utilisation of organisational morphology to identify tasks and then classify them based on three types of norms: substantive, communication and control norms, essentially assists individuals and organisations in devoting their resources to essential activities rather than supporting activities. As a result, based upon the theory of sign and its methods, we have developed a conceptual framework of role-based knowledge profiling that enables individuals and organisations collect and document their knowledge hence norms in an organised and meaningful manner. This study however still requires further works in order to provide a comprehensive knowledge profile which includes; identifying an appropriate mechanism for norm-based workflow, refining and validating the conceptual framework, and evaluating the content of knowledge profile.

## References

1. Dima, A.M., Stancov, V.: Taxonomies of Organisational Knowledge. Revista Informatica Economica (2008)
2. Stamper, R., Liu, K., Hafkamp, M., Ades, Y.: Understanding the Roles of Signs and Norms in Organisation. Journal of Behaviour and Information Technology 19(1), 15–27 (2000)

3. Salter, A.M.: A Normative Approach to Modelling Action and Communication in Organisational Processes. School of Computing, Staffordshire University (2003)
4. Thellefsen, T.: Knowledge Profiling: The Basis for Knowledge Organisation. *Library Trends* 52(3), 507–514 (2004)
5. Engel, A., Huppert, W., Cleveringa, R.: Knowledge Profiling: Promoting Easy Access to Knowledge and Experience Generated in Projects and Programmes. In: International Fund for Agricultural Development (IFAD), Namibia (December 2007)
6. Edwards, K.E., Gibson, N.L.: Knowledge Profiling as Emergent Theory in Community-Based Participatory Research. *Progress in Community Health Partnerships: Research, Education and Action* 2.1, 73–79 (2008)
7. Jorna, R.J.: Introduction: Organizational Semiotics and Social Simulation. *Semiotica* 2009(175), 311–316 (2009)
8. Liu, K., Sun, L., Rong, W.: Semiotic Modelling for Complex Enterprise Systems. In: Proceeding of International Conference on Informatics and Semiotics in Organisations, Beijing, China (2009)
9. Nobre, A.L.: Semiotic Learning: A Conceptual Framework for Facilitating Learning in Knowledge-intensive Organisations. In: Department of Management, p. 236. University of Evora, Evora (2009)
10. Liu, K.: Semiotics in Information Systems Engineering. Cambridge University Press, Cambridge (2000)
11. Chandler, D.: Semiotics: The Basic, 2nd edn. Routledge, New York (2007)
12. Stamper, R.: Information Systems as a Social Science. In: Proceeding of The Ifip TC8/WG8.1 International Conference on Information System Concepts: An Integrated Discipline Emerging, Deventer. Kluwer, B.V, The Netherlands (1998)
13. Smith, L.: Norms in Human Development: Introduction. In: Smith, L., Voneche, J. (eds.) *Norms in Human Development*, pp. 1–31. Cambridge University Press, New York (2006)
14. Braf, E.: Multiple Meanings of Norms. In: Gazendam, H.W.M., Jorna, R.J., Cijssouw, R.S. (eds.) *Dynamics and Change in Organisations: Studies in Organisational Semiotics*, pp. 165–181. Kluwer Academic Publishers, Netherlands (2003)
15. Stamper, R.K.: Organisational Semiotics: Informatics without the Computer? In: Liu, K., Clarke, R.J., Andersen, P.B., Stamper, R.K. (eds.) *Information, Organisation and Technology: Studies in Organisational Semiotics*. Kluwer Academic Publisher, Boston (2001)
16. Stamper, R., Liu, K., Huang, K.: Organisational Morphology in Re-engineering. In: Proceeding of The 2nd European Conference on Information Systems, Nijenrode University, Breukelen, The Netherlands (1994)
17. Salter, A., Liu, K.: Using Semantic Analysis and Norm Analysis to Model Organisation. In: Proceeding of 4th International Conference on Enterprise Information Systems. Universidad de Castilla-La Mancha, Ciudad Real - Spain (2002)
18. Ousmanou, K.: A Method for the Articulation of Users Requirements for Personalised Information Provision. In: School of Systems Engineering, University of Reading, Reading (2007)

# Using Biased Discriminant Analysis for Email Filtering

Juan Carlos Gomez<sup>1</sup> and Marie-Francine Moens<sup>2</sup>

<sup>1</sup> ITESM, Eugenio Garza Sada 2501, Monterrey NL 64849, Mexico

[juancarlos.gomez@invitados.itesm.mx](mailto:juancarlos.gomez@invitados.itesm.mx)

<sup>2</sup> Katholieke Universiteit Leuven, Celestijnenlaan 200A, B-3001 Heverlee, Belgium  
[sien.moens@cs.kuleuven.be](mailto:sien.moens@cs.kuleuven.be)

**Abstract.** This paper reports on email filtering based on content features. We test the validity of a novel statistical feature extraction method, which relies on dimensionality reduction to retain the most informative and discriminative features from messages. The approach, named Biased Discriminant Analysis (BDA), aims at finding a feature space transformation that closely clusters positive examples while pushing away the negative ones. This method is an extension of Linear Discriminant Analysis (LDA), but introduces a different transformation to improve the separation between classes and it has up till now not been applied for text mining tasks.

We successfully test BDA under two schemas. The first one is a traditional classification scenario using a 10-fold cross validation for four ground truth standard corpora: LingSpam, SpamAssassin, Phishing corpus and a subset of the TREC 2007 spam corpus. In the second schema we test the anticipatory properties of the statistical features with the TREC 2007 spam corpus.

The contributions of this work is the evidence that BDA offers better discriminative features for email filtering, gives stable classification results notwithstanding the amount of features chosen, and robustly retains their discriminative value over time.

## 1 Introduction

In data mining the goal is to find previously unknown patterns and relations in large databases [12], and a common task is automatic classification. Here, given a set of instances with known values of their attributes and their classes, the aim of this task is to predict automatically the class of a new instance, when only the values of its features are known.

When classifying email messages, commonly the data contained in them are very complex, multidimensional or represented by a large number of features. Since when using many features, we need a corresponding increase in the number of annotated examples to train from to ensure a correct mapping between the features and the classes [1][5], the use of any kind of dimensionality reduction method is useful in the classification task.

In this paper, we want to discriminate between two classes of email messages (spam or phishing from ham) with the purpose of detecting potentially dangerous emails. In order to do it, we advocate an approach which is an extension of the traditional *Linear Discriminant Analysis* (LDA), named *Biased Discriminant Analysis* (BDA) [3][15]. This method is especially suited to find a feature space transformation that closely clusters the positive examples while pushing away the negative ones. The method has never been used in text mining tasks, but seems promising in binary classification such as spam and phishing mail filtering.

We test this algorithm under two scenarios. First, we try to obtain the smallest number of features to have a good performance in classification. In that case, these features can be understood as the core profile of a data set, which allows a fast training of classifiers. Second, we test the capability of these core profiles of persisting over time, in order to anticipate new dangerous messages, we do this by ordering emails by date and by training our method on older emails and testing on more recent ones. Both schemata are evaluated based on standard datasets for spam and phishing mail filtering.

We compare the BDA results with the basic LDA method to see the improvements included with the more recent algorithm. Additionally, for comparison we present the results for a classifier trained with the complete vocabulary composed by all the unique terms in each data set.

The contribution of our work is the evidence that the technique of Biased Discriminant Analysis offers excellent discriminative features for the filtering, that gives stable results notwithstanding the amount of features chosen, and that robustly retains their discriminative value over time. Our findings contribute to the development of more advanced email filters and open new opportunities for text classification in general.

The remainder of this paper is organised as follows. Section 2 overviews related work on dimensionality reduction. Section 3 introduces the model for dimensionality reduction using BDA. Section 4 discusses our experimental evaluation of the method for email classification. Section 5 concludes this work with lessons learnt and future research directions.

## 2 Related Work

Several methods have been proposed for email filtering [13], and one of the most promising approaches is the use of content-based filters [19], which use the text content of the messages rather than black lists, header or sender information. In this sense, machine learning and data mining methods are especially attractive for this task, since they are capable of adapting to the evolving features of spam and phishing messages' content. There is plenty of work devoted to email filtering [18], including some seminal papers for spam classification using traditional Bayesian filters like [2]. There are also interesting works on phishing detection like [11], describing a set of features to distinguish phishing emails. Nevertheless most of the methods devoted to email filtering use bag-of-words as features and Bayesian methods to perform the classification. Recently, works like [7] and

[16] where the authors use compression models and n-grams to produce more robust features and more sophisticated classifiers like Support Vector Machines, are starting to emerge. Our work pretends to contribute with a new approach in the task of email classification specifically focusing on feature extraction.

Dimensionality reduction has been popular since the early 90s in text processing tasks, like, for example, the technique of Latent Semantic Analysis (LSA) [10]. LSA is an application of principal component analysis where a document is represented along its semantic axes or topics. In a text categorization task, documents are represented by a LSA vector model both when training and testing the categorization system. However, these models do not exploit class information.

Probabilistic topic models such as probabilistic Latent Semantic Analysis (pLSA) [14] and Latent Dirichlet Allocation (LDrA) [6] are currently popular as topic representation models. Documents are represented as a mixture of topic distributions and topics as a mixture of word distributions. However, in text categorization, information about the text categories is not taken into account. Very recently, the LDrA model has been used for spam classification [4].

*Linear Discriminant Analysis* (LDA) uses class information in order to separate well the classes. Recently, the computer vision community has successfully proposed several variants of LDA that artificially pull apart the positive and the negative examples. One of these is *Biased Discriminant Analysis* (BDA), used by Huang et al. [15] for learning a better ranking function based on relevance feedback in image search. To our knowledge, this LDA variant is only recently developed and has not been used for text classification or email filtering.

BDA is an eigenvalue based method. An eigenvalue is a number indicating the weight of a particular pattern or cluster expressed by the corresponding eigenvector. The larger the eigenvalue the more important the pattern is.

### 3 Biased Discriminant Analysis

The final goal of this work is to classify email messages in a priori defined mutual exclusive classes. Our concrete aim is to transform the original feature space of emails into a less dimensional space. This space would be expressed in terms of statistical similarities and differences between messages. The new space is intended to be easier to deal with because of its size, carrying the most important part (core profiles) of the information needed to filter emails.

Let  $\{(\mathbf{x}_1, c_1), (\mathbf{x}_2, c_2), \dots, (\mathbf{x}_n, c_n)\}$  be a set of email messages with their corresponding classes, where  $\mathbf{x}_i \in \mathbb{R}^d$  is the  $i$ -th email, represented by a  $d$  dimensional row vector, and  $c_i \in \mathbf{C}$  is the class of  $\mathbf{x}_i$ . In this work we have  $\mathbf{C} = \{-1, +1\}$ , where -1 refers to the negative class  $N$  (ham messages) and +1 to the positive class  $P$  (spam or phishing).

The goal of the data dimensionality reduction is to learn a  $d \times l$  projection matrix  $\mathbf{W}$ , which can project to:

$$\mathbf{z}_i = \mathbf{x}_i \mathbf{W} \quad (1)$$

where  $\mathbf{z}_i \in \mathbb{R}^l$  is the projected data with  $l \ll d$ , such that in the projected space the data from different classes can be effectively discriminated.

As was mentioned before, BDA [15] is a variant of LDA, where BDA seeks to transform the feature space so that the positive examples cluster together and each negative instance is pushed away as far as possible from this positive cluster, resulting in the centroids of both the negative and positive examples being moved. BDA aims at maximizing the following function:

$$\mathbf{W}^* = \arg \max_{\mathbf{W}} \frac{|\mathbf{W}^T \mathbf{S}_{PN} \mathbf{W}|}{|\mathbf{W}^T \mathbf{S}_P \mathbf{W}|} \quad (2)$$

The inter-class (positive-negative) scatter matrix  $\mathbf{S}_{PN}$ , is computed as follows, where  $\mu_P$  is the mean of the examples in the positive class:

$$\mathbf{S}_{PN} = \Sigma_{\mathbf{y} \in N} (\mathbf{y} - \mu_P)^T (\mathbf{y} - \mu_P) \quad (3)$$

and the intra-class matrix (positive) scatter matrix  $\mathbf{S}_P$ , is computed as follows:

$$\mathbf{S}_P = \Sigma_{\mathbf{x} \in P} (\mathbf{x} - \mu_P)^T (\mathbf{x} - \mu_P) \quad (4)$$

We then perform an eigenvalue decomposition on  $\mathbf{S}_P^{-1} - \mathbf{S}_{PN}$ , and construct the  $d \times l$  matrix  $\mathbf{W}$  whose columns are composed by the eigenvectors of  $\mathbf{S}_P^{-1} - \mathbf{S}_{PN}$  corresponding to its largest eigenvalues.

The goal of BDA is to transform the data set  $\mathbf{X}$  into a new data set  $\mathbf{Z}$  using the projection matrix  $\mathbf{W}$ , with  $\mathbf{Z} = \mathbf{X}\mathbf{W}$  in such a way the examples inside the new data set are well separated by class. In this case,  $\mathbf{Z}$  represents the training data. The test examples, for which we do not know the class, are projected using the matrix  $\mathbf{W}$  to be represented in the new BDA space. Then, if  $\mathbf{q}$  is a test example, its projection using BDA is  $\mathbf{u} = \mathbf{q}\mathbf{W}$ .

## 4 Experimental Results

The four public email corpora we use for performing our tests are: Ling-Spam (*LS*)<sup>1</sup> [2], SpamAssassin (*SA*)<sup>2</sup>, TREC 2007 spam corpus (*TREC*)<sup>3</sup>[9] and a subset of Phishing Corpus (*PC*), created by randomly selecting 1,250 phishing messages from the Nazario's corpus<sup>4</sup> and 1,250 ham messages from the TREC corpus. The number of emails in each corpus is listed in table II.

For this work, before performing BDA, the emails have to be transformed into vectors. First we remove the structure information, i.e. the header and the HTML tags, to retain only the text content, as seen in figure II. We focus in this article on content based email filtering and realize that important discriminative information in headers of the email is ignored (e.g., address of the sender). Second, we build a vocabulary by removing stop words and words that are evenly distributed over the classes, the latter using a mutual information statistic, obtaining 5000 initial features. Finally, we weight the remaining words in each document by a TF-IDF schema while representing each message as a vector.

<sup>1</sup> Available at: <http://nlp.cs.aueb.gr/software.html>

<sup>2</sup> Available at: <http://spamassassin.apache.org/publiccorpus/>

<sup>3</sup> Available at: <http://plg.uwaterloo.ca/gvcormac/treccorpus07/>

<sup>4</sup> Available at: <http://monkey.org/jose/wiki/doku.php?id=PhishingCorpus>

It is going to be HUGE. Target sym: CDYV, Price (current): \$0.089, 5 Day Target price: \$0.425, Action: Strong Buy/Hold.. Here comes the REAL BIG ONE!.. See the news, catchall, call your broker!!

**Fig. 1.** Typical text content of a spam email from TREC spam corpus. We use only the words inside the messages to perform the BDA.

**Table 1.** Number of messages per corpus

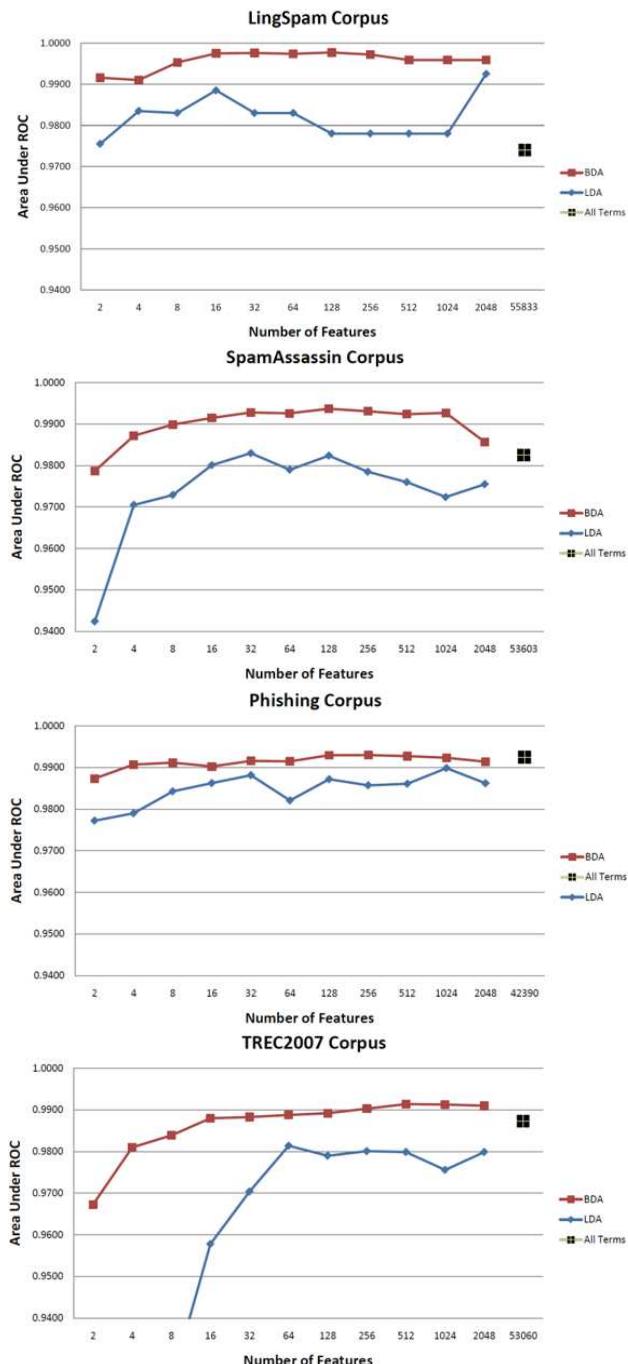
Corpus	Spam	Phishing	Ham	Total
LS	481		2,412	2,893
SA	1,897		4,150	6,047
TREC	50,199		25,220	75,419
PC		1,250	1,250	2,500
Total	52,576	1,250	33,032	

The training model is constructed by applying the BDA to the message vectors of the training set. Then, a classifier is trained based on this data and tested using the new messages expressed also as vectors in the BDA space. Experimentally we decided to use the bagging ensemble classifier [8], using as single classifier the C4.5 decision tree [17]. The rationale for this is that C4.5 is fast and easy to train and has good performance with small number of features, and that bagging presents a general well behavior by weighting the results of the trees and by reducing the variance of the data set and the overfitting.

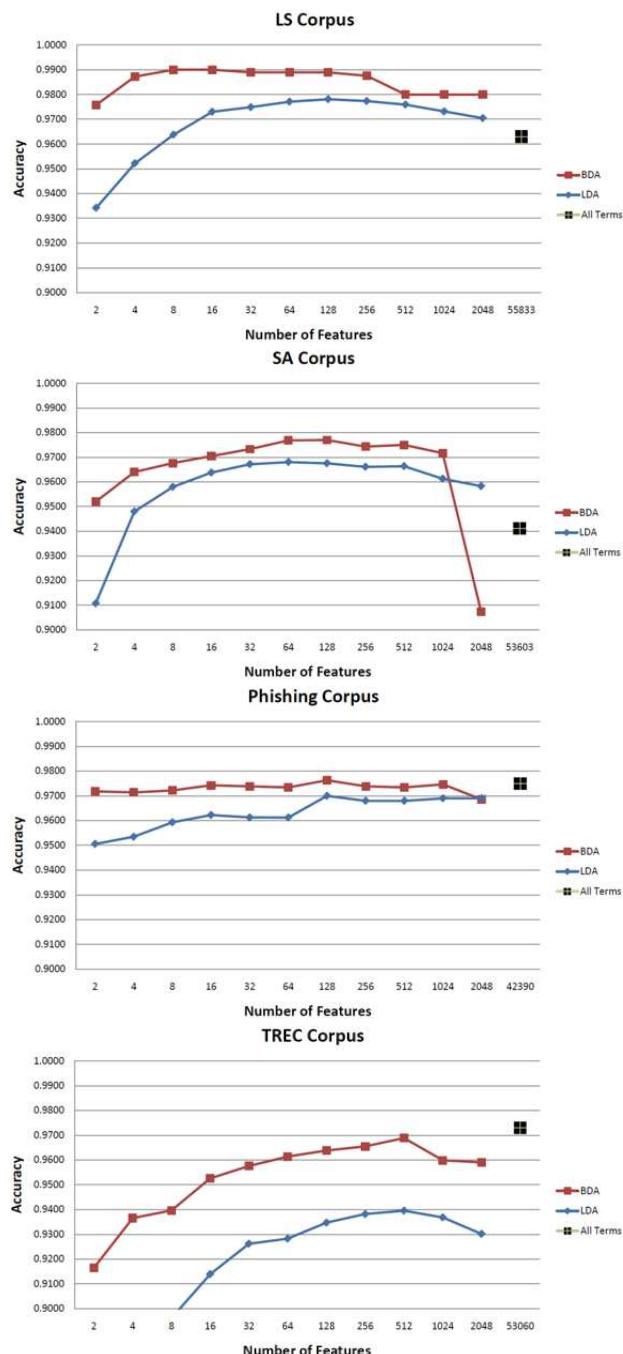
Because the feature extraction results in a ranked list of features, we perform experiments by considering the 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, and 2048 highest ranked features for each method BDA and LDA and then we compare all the results. Additionally, we compare with a baseline method (AUT) which uses all unique terms from each training set in the classification.

We present results using the area under the ROC, which aims at a high true positive rate and a low false positive rate. The ROC metric is very important for commercial settings, where the cost for misclassifying a legitimate email as illegal (false positive) is really high. In addition we provide results in terms of overall accuracy of the classification for better understanding the behavior of the algorithms.

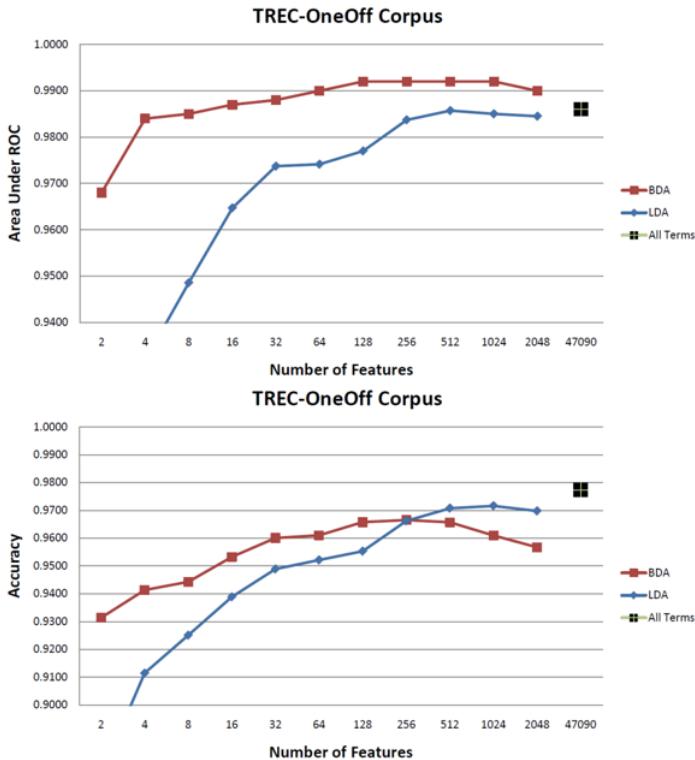
Figures 2 and 3 show the results for area under ROC and accuracy for the application of BDA, LDA and the baseline AUT to the four public corpora, performing a normal classification using a 10-fold cross validation for spam and phishing filtering. From these results we can observe that the statistical features extracted by BDA are well suited to discriminate the spam or phishing messages from the ham messages in these corpora. BDA is able to reach good performance almost independently of the number of features used. There are several published works where some of these corpora are used, like [2], [7] and [19]. Nevertheless in these works the authors do not present the complete information about the performance of their methods, and a direct comparison with our results is not possible.



**Fig. 2.** Performance of algorithms for area under ROC for LS, SA, PC and TREC corpora using 10-fold cross validation



**Fig. 3.** Performance of algorithms for accuracy for LS, SA, PC and TREC corpora using 10-fold cross validation



**Fig. 4.** Performance of algorithms for area under ROC (above) and accuracy (below) for TREC corpus, performing a one-off experiment

Using 10-fold cross validation BDA performs better than LDA and the baseline AUT in LS and SA corpora for both, area under ROC and accuracy. In the case of PC and TREC corpora, AUT reaches good results, but the number of features used makes it impractical to implement a filter like that. On the other hand, with a much smaller number of features BDA is able to reach competitive (or even better) results with the corresponding minor cost of training and especially of testing a filter.

In figure 4 we see the performance of the methods for the one-off experiment for TREC corpus. In this case we want to test the persistency and robustness of the features extracted; then we select a small subset of 9,020 messages, corresponding to the first week of the data set, for training, and the rest 66,399 messages, corresponding to (almost) 11 weeks in the future, for testing. Similar to the previous experiment, the statistical features extracted by BDA reach a good performance with only a small number of features. This means BDA performs quite well when creating robust predictive profiles from past training data. It is possible to observe that the number of active features is generally low, even if the corpus presents a big variation of topics. The testing of these profiles

with completely new data containing unseen messages prove that the statistical features from BDA are able to generalize over a bigger variation inside a data set.

The main aims of this work are the filtering of spam and phishing messages, while performing this filtering with just a few features representing a small set of core profiles and proving the persistency of the core features over time (robustness). As is confirmed by the results presented in this section, we have accomplished these goals.

## 5 Conclusions

In this paper we performed content-based email filtering using a statistical feature extraction method. This method, named Biased Discriminant Analysis (BDA), is an approach understood as a dimensionality reduction technique, which especially aims at better discriminating positive from negative examples. The obtained essential statistical features carry very useful information (core profiles of the data set), which is highly discriminative for email classification and robust to persist over time.

The results show a very good classification performance when using BDA for filtering emails in standard benchmarking data sets using 10-fold cross validation and in an anticipatory scenario, where the task was to separate spam and phishing from ham messages. In this sense, BDA is effective for classifying emails and robust when predicting the type of email when trained on older data. Overall, BDA performs excellently and better than standard LDA and (most of the times) than the baseline bag-of-words method, measured in terms of the area under the ROC and accuracy, even when emails are described with very few features. The results obtained with BDA are not very dependent on the number of features chosen, which is an advantage in a text classification task.

Spam filters in practical settings often rely on a small set of signature rules that form a profile. The proposed technique perfectly fits this scenario, by yielding excellent classification results based on a limited number of features. The focus of our work was on content-based filtering. It would be interesting to investigate how our method can be integrated with non-content based features for email filtering such as the provenance of the emails. In the future we want to apply our method to other text classification tasks, possibly customized with multilinear (tensor) methods for dimensionality reduction, that have the possibility to include, for instance, content and non-content features in email filtering.

## Acknowledgment

We thank the EU FP6-027600 Antiphish consortium (<http://www.antiphishresearch.org/>) and in particular Christina Lioma, Gerhard Paaf, Andre Bergholz, Patrick Horkan, Brian Witten, Marc Dacier and Domenico Dato.

## References

1. Aha, D.W., Kibler, D.F., Albert, M.K.: Instance-based learning algorithms. *Machine Learning* 6, 37–66 (1991)
2. Androuloutsopoulos, I., Koutsias, J., Chandrinou, K.V., Ch, K.V., Palioras, G., Spyropoulos, C.D.: An evaluation of naive bayesian anti-spam filtering. In: Lopez de Mantaras, R., Plaza, E. (eds.) ECML 2000. LNCS (LNAI), vol. 1810, pp. 9–17. Springer, Heidelberg (2000)
3. Baudat, G., Anouar, F.: Generalized discriminant analysis using a kernel approach. *Neural Computation* 12(10), 2385–2404 (2000)
4. István, B., Jácint, S., Benczúr, A.A.: Latent dirichlet allocation in web spam filtering. In: AIRWeb 2008: Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web, pp. 29–32 (2008)
5. Bishop, C.: Neural Networks for Pattern Recognition. Clarendon Press, Oxford (1995)
6. Blei, D.M., Ng, A.Y., Jordan, M.I., Lafferty, J.: Latent dirichlet allocation. *Journal of Machine Learning Research* 3, 2003 (2003)
7. Bratko, A., Cormack, G., Filipic, B., Lynam, T., Zupan, B.: Spam filtering using statistical data compression models. *Journal of Machine Learning Research* 7, 2673–2698 (2006)
8. Breiman, L.: Bagging predictors. In: *Machine Learning*, pp. 123–140 (1996)
9. Cormack, G.V.: Spam track overview. In: TREC-2007: Sixteenth Text REtrieval Conference (2007)
10. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41, 391–407 (1990)
11. Fette, I., Sadeh, N., Tomasic, A.: Learning to detect phishing emails. In: WWW 2007: Proceedings of the 16th International Conference on World Wide Web, pp. 649–656. ACM, New York (2007)
12. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press, London (1990)
13. Goodman, J., Heckerman, D., Rounthwaite, R.: Stopping spam. *Scientific American* 292(4), 42–88 (2005)
14. Hofmann, T.: Probabilistic latent semantic indexing. In: *Uncertainty in Artificial Intelligence*, pp. 50–57 (1999)
15. Huang, T.S., Dagli, C.K., Rajaram, S., Chang, E.Y., Mandel, M.I., Poliner, G.E., Ellis, D.P.W.: Active learning for interactive multimedia retrieval. *Proceedings of the IEEE* 96(4), 648–667 (2008)
16. Kanaris, I., Kanaris, K., Houvardas, I., Stamatatos, E.: Words vs. character n-grams for anti-spam filtering. *Int. Journal on Artificial Intelligence Tools* 16(6), 1047–1067 (2007)
17. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo (1993)
18. Guzella, T.S., Caminhas, W.M.: A review of machine learning approaches to spam filtering. *Expert Systems with Applications* 36, 10206–10222 (2009)
19. Yu, B., Xu, Z.-b.: A comparative study for content-based dynamic spam classification using four machine learning algorithms. *Knowledge-Based Systems* 21(4), 355–362 (2008)

# Use of Geospatial Analyses for Semantic Reasoning

Ashish Karmacharya<sup>1,2</sup>, Christophe Cruz<sup>2</sup>, Frank Boochs<sup>1</sup>, and Franck Marzani<sup>2</sup>

<sup>1</sup> Institut i3mainz, am Fachbereich 1 - Geoinformatik und Vermessung, Fachhochschule Mainz,  
Lucy-Hillebrand 2, 55128 Mainz

{ashish,boochs}@geoinform.fh-mainz.de

<sup>2</sup> Laboratoire Le2i, UMR-5158 CNRS, UFR Sciences et Techniques, Université de Bourgogne,  
B.P. 47870, 21078 Dijon Cedex, France

{christophe.cruz, franck.marzani}@u-bourgogne.fr

**Abstract.** This work focuses on the integration of the spatial analyses for semantic reasoning in order to compute new axioms of an existing OWL ontology. To make it concrete, we have defined Spatial Built-ins, an extension of existing Built-ins of the SWRL rule language. It permits to run deductive rules with the help of a translation rule engine. Thus, the Spatial SWRL rules are translated to standard SWRL rules. Once the spatial functions of the Spatial SWRL rules are computed with the help of a spatial database system, the resulting translated rules are computed with a reasoning engine such as Racer, Jess or Pellet.

**Keywords:** OWL, SWRL, Spatial functions, GIS system, Built-ins, Spatial Knowledge Reasoning.

## 1 Introduction

This paper discusses a method to integrate the spatial technologies and Semantic Web technologies. This is undertaken by using rules. Actually, they have always played an important role for knowledge-based systems [9]. In the semantic Web context, rules are defined with the help of the Rule Markup Language. The derived language “Semantic Web Rule Language” (SWRL) combines the RuleML and the OWL-DL [1]. The method consists in extending the SWRL language with spatial built-ins. The Spatial SWRL API, part of the project ArchaeoKM [5], [6], [7], provides an authoring environment for the definition of rules and allows the execution of these rules. The results of this work are applied to the domain of archaeology and the project ArchaeoKM. The main concept behind ArchaeoKM is to use knowledge from archaeologists to manage the excavated information. The application ArchaeoKM facilitates archaeologists to manage the information and the knowledge concerning the findings and objects collected on the site. This is done by defining the geo-localization of objects, the enrichment and the population of an ontology of domain. Presently, it concerns the domain of industrial archaeology. This project has already been presented at CAA 2009 [6].

The Open Geospatial Consortium (OGC) plays a major role in developing a consensus among different stakeholders on various aspects of geospatial technology. The

OGC is concerned with the data interoperability and has developed different standards for the efficient interoperability. In addition, groups like Geospatial Incubator have taken the works of OGC to formulate steps in updating the W3C geo vocabulary and preparing the groundwork to develop comprehensive geospatial ontology [11]. The domain of archaeology benefits from this work and could surely be of benefits for lots of other domains. As a proof of concept, we present an example of what is possible to compute with our method. For instance, it is possible to identify possible flooding zones according to river bank bursts due to excessive water during rainy season. This is a very common exercise for a flood management system in hydrology and it provides interesting clues for industrial archaeology. The SWRL rule provided below (rule 1) incorporates the spatial built-ins within to support the hydrology example provided above.

$$\text{River}(\text{?x}) \wedge \text{Building}(\text{?y}) \wedge \text{spatialswrlb:Buffer}(\text{?x}, 50, \text{?z}) \wedge \text{spatialswrlb:Intersection}(\text{?z}, \text{?y}, \text{?res}) \rightarrow \text{isLiableToFloodingBy}(\text{?y}, \text{?x}) \quad (1)$$

Section 2 is about the existing research projects focusing on the spatial components and their lacking areas which have motivated us to carry out this research work. Section 3 presents the cutting edge technologies covering knowledge representation and the reasoning process through Semantic Web technology and its components like the Web Ontology Language (OWL) and the Semantic Web Rule Language (SWRL). Section 4 deals with the spatial representation in GIS systems. This section includes the presentation of the spatial relationship functions and the spatial processing functions. Section 5 presents the ontology adjustment process which is necessary to do before the processing of spatial rules. Section 6 gives a description of the Spatial Built-ins related the spatial functions. The last section concludes the papers.

## 2 Related Works and Motivation

The existing GIS systems do not use semantic explicitly. They primarily focus on geometry and store them in their native formats. These systems perform spatial analyses through their spatial queries based on spatial functions and operations. These spatial functions and operations are more or less similar in all the systems. The database systems have taken steps to integrate the spatial functionalities in their systems. With the advancement Spatial Database Management Systems (SDBMS), it is now possible to store the geometry in those database systems and do not have to rely on the GIS tools to store or retrieve the geometries. It has even become possible to perform spatial operations within those database systems. Research projects like GIS DILAS [12] or 3D MURALE [13] take advantages of those features of current database systems to carry out spatial operations within their systems.

The inclusion of semantic into any information system adds the efficient on the system as a whole [14]. There have been few research works to include the semantic layers within GIS but they are not as many as in some other research areas. In addition the current research works mostly focus on the use of semantic for semantic interoperability of the GIS data so that the GIS data could be exchanged over broader and heterogeneous platforms [15]. The ontology is also being used data mapping in

order to have comprehensive data integration. This has been discussed in the research works [16, 17, 18].

Though there are several research works, common consensus foundation ontology has not yet been agreed upon. Open Geospatial Consortium (OGC) is playing a major role to develop a consensus among different stakeholder on various aspect of geospatial technology. Data interoperability is a major area in which OGC is concerned upon and it has developed different standards for this. Groups like Geospatial Incubator have taken the works of OGC to formulate steps in updating the W3C geo vocabulary and preparing the groundwork to develop comprehensive geospatial ontology. In the process it has been reviewing different spatial ontologies that exist in the web [19].

It can be clearly notices that their does exist clear lack of research work in this area especially in the field of spatial analyses through knowledge modeling. The research this paper discusses is use of knowledge through rules in performing these analyses. We believe activities within any spatial analysis are execution of spatial rules to come out to an analytical result.

## 2.1 ArchaeoKM

ArchaeoKM is a web base tool to support the archaeologists to manage their information during their excavations. As already been mentioned it is based on Semantic Web technology and knowledge management. It provides supports archaeologists to manage their data and document collected during excavation through simple but efficient mechanism of annotations. The data and documents are stored in their proprietary format and they are annotated to the relevant objects through semantic annotations. It provides a base for data integration. The objects are identified and tagged within the orthophoto. Those identified objects get populated into the domain ontology which is basically a graphical representation of the excavation site. The simplified structure of domain ontology is given in figure 1. in the appendix. This ontological structure is modified to adjust the spatial analysis which is given in figure 2 in the appendix. The details on the adjustment will be discussed in section 5.

ArchaeoKM is a rule based system. It uses the advancement in rule engines through rule languages to manage the knowledge. Once the objects are identified and tagged within the domain ontology, a knowledge base is created which reflects the knowledge of the archaeologist who has tagged the objects. Now this knowledge base could be used to manage the knowledge. ArchaeoKM uses rule languages of rule engines (primarily SWRL and Jena Rule) to manage them. This paper highlights the process of inclusion of spatial rules within ArchaeoKM through built-ins of SWRL.

## 3 The Cutting Edge Technologies

This section deals with a short introduction to the main Semantic Web technologies. The OWL language that allows the definition of ontologies of domain, SWRL that allows the definition of rules on ontologies and SWRL built-ins that allow to compute advanced processes.

### 3.1 OWL

OWL is a knowledge representation language and a standard (W3C recommendation) for expressing ontologies in the Semantic Web. The OWL language facilitates greater machine understandability of Web resources by providing additional constructors for building class, property descriptions and new axioms, along with a formal semantics. Concepts are sets of classes of individual objects. Classes provide an abstraction mechanism for grouping resources with similar characteristics [4]. In any graphical representation of knowledge classes are represented through the nodes. Descriptions on OWL classes are discussed in details in [4]. A property restriction is an unnamed class containing all individuals that satisfy the restriction. Properties are binary relationships between two objects. In general they are the relationships between two classes which apply to the individuals of those classes. They are known as roles in description logic and are represented through links in the graphical representation. OWL provides two main categories of properties: Object properties – relationships between concepts and consequently instances of the concepts and Data properties – relation of an instance to the data value.

### 3.2 SWRL

Semantic Web Rule Language (SWRL) [1] is a rule language based on the combination of the OWL-DL (SHOIN(D)) with Unary/Binary Datalog RuleML which is a sublanguage of the Rule Markup Language. One restriction on SWRL called DL-safe rules were designed in order to keep the decidability of deduction algorithms. This restriction is not about the components of the language but on its interaction. SWRL includes a high-level abstract syntax for Horn-like rules.

The SWRL as the form, *antecedent*  $\rightarrow$  *consequent*, where both antecedent and consequent are conjunctions of atoms written  $a_1 \wedge \dots \wedge a_n$ . Atoms in rules can be of the form  $C(x)$ ,  $P(x,y)$ ,  $Q(x,z)$ , *sameAs*( $x,y$ ), *differentFrom*( $x,y$ ), or *builtIn*( $pred, z_1, \dots, z_n$ ), where  $C$  is an OWL description,  $P$  is an OWL individual-valued property,  $Q$  is an OWL data-valued property,  $pred$  is a datatype predicate URIref,  $x$  and  $y$  are either individual-valued variables or OWL individuals, and  $z, z_1, \dots, z_n$  are either data-valued variables or OWL data literals. An OWL data literal is either a typed literal or a plain literal [2]. Variables are indicated by using the standard convention of prefixing them with a question mark (e.g.,  $?x$ ). URI references (URIrefs) are used to identify ontology elements such as classes, individual-valued properties and data-valued properties. For instance, the following rule asserts that one's parents' brothers are one's uncles where parent, brother and uncle are all individual-valued properties. This could be executed with the SWRL presented in rule (2).

$$\text{parent}(\text{?x}, \text{?p}) \wedge \text{brother}(\text{?p}, \text{?u}) \rightarrow \text{uncle}(\text{?x}, \text{?u}) \quad (2)$$

### 3.3 SWRL Built-Ins

The set of built-ins for SWRL is motivated by a modular approach that will allow further extensions in future releases within a (hierarchical) taxonomy. SWRL's built-ins

approach is also based on the reuse of existing built-ins in XQuery and XPath, which are themselves based on XML Schema by using the datatypes. This system of built-ins should also help in the interoperability of SWRL with other Web formalisms by providing an extensible, modular built-ins infrastructure for Semantic Web Languages, Web Services, and Web applications. Many built-ins are defined and a non exhaustive list can be found below.

- *Comparisons*
- *Math Built-Ins*
- *Built-Ins for Boolean Values*
- *Built-Ins for Strings, etc.*

The next SWRL rule is an example using the Math built-in “`swrlb:greaterThan`”. If the result of the built-in is true for a Person `?p` then this Person `?p` is a member of the concept Adult.

$$\text{Person}(\text{?p}) \wedge \text{hasAge}(\text{?p}, \text{?age}) \wedge \text{swrlb:greaterThan}(\text{?age}, 18) \rightarrow \text{Adult}(\text{?p}) \quad (3)$$

## 4 Spatial Components

This section discusses the spatial components within GIS technology and the database system. It is important to evaluate the spatial features within the existing technologies in order to take the advantage from their developments. Additionally, the spatial functions of database system are utilized to execute spatial rules within spatial built-ins.

Today most database systems provide support to the spatial extension. This paper uses *PostGIS* a spatial extension *PostgreSQL* for the arguments but same could be applied in other database system too. PostGIS supports the storage of *point*, *line*, *polygon*, *multipoint*, *multiline*, *multipolygon*, and *geometrycollections*. It follows the specification provided by OGC for the *simple features* to store these objects. Those are specified in the Open GIS Well Known Text (WKT) or Well Known Binary (WKB) Formats. It stores 3Dimensional coordinates as Extended Well Known Text (EWKT) and Extended Well Known Binary (EWKB) – the extensions it defined. They are different from Simple Feature Specification by OGC as they embed Spatial Reference Identifier (SRID) within them. Besides providing functionalities for storing the geometries and exporting/importing geometries from/into the database, *PostgreSQL* with its spatial extension *PostGIS* provides a range of spatial functions which are spatial relationship functions and spatial processing functions.

- The **spatial relationship functions** are generally binary functions. These functions return a Boolean value. However, when they are used with a proper SQL statement, these functions can be used to identify the objects with which they are related to. The functions are used as SQL statement. The examples of spatial functions under this category are *touch*, *disjoint*, *overlap*, *within* and are used through *st\_touch*, *st\_disjoint*, *st\_overlap*, *st\_within* respectively in PostGIS.

- The **spatial processing functions** provided in this section allow the processing of the object geometries. The results themselves are sets of geometries. The spatial built-ins *Buffer* and *Intersection* discussed in (1) belong to this category. Besides *buffer* and *intersection* there are functions like *Difference*, *Union* under this category. Those functions are executed through *st\_buffer*, *st\_intersection*, *st\_difference* and *st\_union* respectively in PostGIS.

## 5 Ontological Adjustment

This section highlights our approach to integrate the spatial functions in the ontology and make it possible to store the results of the spatial functions in an existing ontology. This paper uses the domain ontology described in ArchaeoKM [5], [6], [7] and consists in adding new axioms (concepts, relations, attributes, etc.) for our purpose. The initial and adjusted ontological structures are shown in figure 1 and 2 in the appendix. Once the Spatial SWRL rules are executed, the results of these rules will generate information that have to be stored in the enriched part of the ontology. The main process of enriching the ontology schema consists in adding the concept *feat:siteFeature*. All the objects, that define a domain concept and have a geometrical definition in the spatial database, requires to be instances of the concept *feat:siteFeature*. This concept is important as it allows the definition of links between the adjusted domain ontology and the spatial functions. These spatial analysis properties are specializations the relationship *sa:hasSpatialRelAnalysis*. The concept *sa:spatialAnalysis* refers to the spatial functions as its specialized concepts and are defined through its inheritance. In addition, the links between the ontology and the database are defined using the link *feat:hasAnnotation*. The *shape:feature* relates to the geometrical definition of excavated objects and the *an:tag* refers to the same geometrical definition but stored into the database. Details on how *an:tag* or *feat:Annotation* functions can be read in [5], [6], [7].

### 5.1 Spatial Relationship Functions

Spatial relation functions return the true or false about the relationship of the objects. Most of these relationship functions operate in same way so all of these functions are adjusted within the ontology in similar fashion. We are discussing the four spatial functions to highlight our point but the same approach is applied to others. The following four sub-relations of the relationship *sa:hasSpatialRelAnalysis* define spatial relationships between two objects. The result of a spatial function process between two objects of the kind of the concept *feat:siteFeature* can be a new link between them. This new link is of kind of e.g. Table 1.

**Table 1.** Ontology adjustment concerning the Spatial Relation Functions

Spatial Relationship Functions	ObjectProperties
Disjoint	<i>sa:hasDisjoint(x,y)</i>
Touches	<i>sa:hasTouch(x,y)</i>
Within	<i>sa:hasWithin(x,y)</i>
Overlaps	<i>sa:hasOverlaps(x,y)</i>

The variables x and y are of the type of the concept *feat:siteFeature*. It means that it could be an object or the result of a spatial processing function.

## 5.2 Spatial Processing Functions

The current database systems provide spatial processing functions to process the geometry of an object for spatial analysis. These functions return geometries themselves. However we focus on four distinct functions. These spatial processing functions are Buffer, Union, Intersection and Difference. The returned geometries are also stored in the spatial database in order to be computed by future spatial functions.

**Table 2.** Ontology adjustment concerning the Spatial Processing Functions

Spatial Processing Functions	Concept	Object Property
Buffer	<i>feat:sp_Buffer</i>	<i>sa:hasBuffer(x,y)</i>
Union	<i>feat:sp_Union</i>	<i>sa:hasUnion(x,y)</i>
Intersection	<i>feat:sp_Intersection</i>	<i>sa:hasIntersection(x,y)</i>
Difference	<i>feat:sp_Difference</i>	<i>sa:hasDifference(x,y)</i>

As a solution, we define four new concepts called *feat:sp\_buffer*, *feat:sp\_union*, *feat:sp\_intersection* and *feat:sp\_difference* which are of kind of *feat:siteFeature*. By inheritance, these four concepts have a spatial definition in the spatial database which are defined with the help of the relationship *feat:hasAnnotation* like any other finding objects. There is also four *sa:hasSpatialRelAnalysis* defined corresponding to each spatial processing function (*sa:hasBuffer*, *sa:hasUnion*, *sa:hasIntersection*, *sa:hasDifference*). They are used to keep a link between the first spatial geometry of the spatial function and the results of this spatial function (e.g. Table 2). The variables x and y are of the type of the concept “*feat:siteFeature*”. It means that it could be an object or the result of a spatial processing function.

## 6 Definition of the Spatial SWRL Built-Ins

The previous section discussed on how different spatial functions are integrated with the ontology. From this adjustment, the Spatial SWRL Built-ins can be defined for each spatial function. Before the definition of these Built-ins, it is necessary to explain the way the engine works beforehand and how the engine translates Spatial SWRL rules into standard SWRL rules. The example given in rule 1 and 4 uses five axioms. The axioms River and Building is of the kind of the concept “*feat:siteFeature*”. It means that they have both a spatial geometry stored in the database. The axiom “*isLiableToFloodingBy*” is a relationship that links two object of the kind of the concept “*feat:siteFeature*”. It means that a building “?y” can be liable to flooding by a river “?x” if all the axioms of the antecedent are true. This rule is computed for every river and building that is present in the ontology. The axiom “*spatialswrlb:Buffer*” is to compute a buffer for the feature “?x”, and the axiom “*saptialswrlb:Intersection*” is used to compute the intersection of the second feature “?y” with the result of the buffer operation. If there is a result “?res” of the intersection function, then a new relation is created.

The role of the translation engine consists in:

1. *interpreting the Spatial SWRL rules*
2. *computing the spatial functions within spatial database*
3. *updating the ontology and the spatial database with the results of the spatial functions*
4. *translating the spatial SWRL rules into standard SWRL rules*
5. *running the rules with the help of a standard rule engine as Racer, Jess or Pellet*

The two next sections explain how the spatial built-ins are translated into SWRL rules. The computing of the spatial functions is out of the scope of this paper. However, it uses SQL statements.

## 6.1 Spatial Relationship Built-Ins

Concerning these built-ins, the translation engine computes the spatial function in the database within all the instances of the built-in parameters. For instance, the built-in *spatialswrlb:Disjoint(?x, ?y)* is interpreted by the translation engine and compute all the instances of the kind of the variables *?x* and *?y*. If the result is true for any couple of instances, then a new relationship *sa:hasDisjoint* is created in the ontology between the couple of instances. After what, the axiom *spatialswrlb:Disjoint(?x, ?y)* is replaced in the rule by the axiom *sa:hasDisjoint(?x, ?y)*. Consequently, the rule is now a standard rule.

## 6.2 Spatial Processing Built-Ins

Concerning these built-ins, the translation is a bit more complex. Actually, the translation engine has to interpret the spatial built-ins and to compute the new geometry for each built-in. The resulting geometries are stored in the spatial database and a new individual of the kind of the *feat:sp\_Buffer*, for example, is created in order to keep a link with the database. In addition, a link of the kind of the relationship *sa:hasBuffer*, for example, is created in order to keep a relationship between the first individual parameter of the built-in and the new individual *feat:sp\_Buffer*. Once the ontology is updated, the axiom *spatialswrlb:Buffer(?x, ?value, ?res)*, for instance, is replaced by the following two axioms *sa:hasBuffer(?x, ?res) ^ feat:sp\_Buffer(?res, bufDistance(?value))*. The parameter *?res* is to refer the resultant instances of *feat:sp\_Buffer*, for instance. Similarly *bufDistance(?value)* defines the buffering distance. It is a data property but is an important factor defining a buffer zone. Due to a lack of space, the complete translation table is not given.

The example (1) is a Spatial SWRL rule and the example (4) is its translation into a standard SWRL rule done by the translation engine. Meanwhile, the translation engine has computed the necessary geometries and has updated the domain ontology with individuals and relationships allowing the run of the translated rule by a reasoning engine. Thus, a spatial reasoning is executed on the adjusted domain ontology.

$$\begin{aligned} \text{River(?x)} \wedge \text{Building(?y)} \wedge \text{sa:hasBuffer(?x, ?z)} \wedge \text{feat:sp_Buffer(?z)} \wedge \\ \text{sa:hasIntersection(?z, ?res)} \wedge \text{sa:hasIntersection(?y, ?res)} \wedge \\ \text{feat:sp_Intersection(?res)} \rightarrow \text{isLiableToFloodingBy (?y, ?x)} \end{aligned} \quad (4)$$

## 7 Conclusion

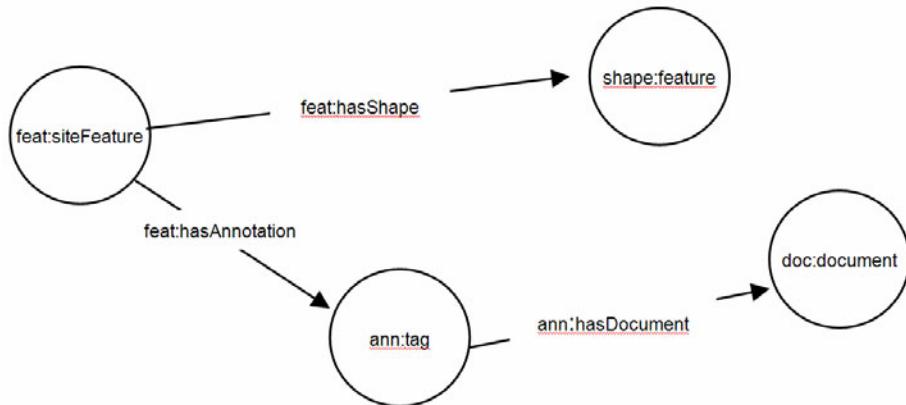
This has presented the integration of the spatial functions into a domain ontology via its adjustment. The ideas presented here could contribute to the development of analysis solution for the GIS technology. The combination of a rule language with spatial functions will add a new dimension in which users interpret their views. A layer in between the data layer and the visualization layer could be added in the existing GIS system which performs the ontological operations. This layer will act as the facilitating tool for the spatial knowledge base in the current system. The integration of such layer in the existing GIS system will support the system to carry out rule based analysis. Thus, the system can handle the analysis on the ground more dynamically as they could be used on fly through the rule engines. Thus, the system provides a firm base to GIS system by providing much needed dynamism to the system.

## References

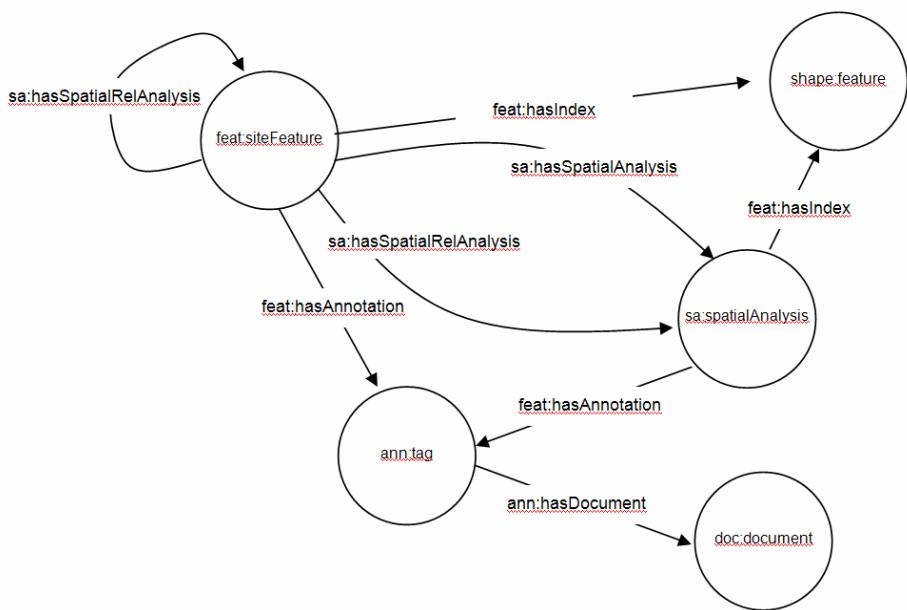
1. W3C, SWRL: A Semantic Web Rule Language Combining OWL and RuleML, <http://www.w3.org/Submission/SWRL/> (Last Visited: November 25, 2008)
2. Pan, J.Z., Horrocks, I.: OWL-Eu: Adding Customised Datatypes into OWL. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 153–166. Springer, Heidelberg (2005)
3. Smith, M.J., Goodchild, M.J., Longley, P.A.: Geospatial Analysis: A Comprehensive guide to Principles. Techniques and Software Tools. Metador (2007)
4. Bechhofer, S., Harmelen, F.V., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F., et al.: OWL Web Ontology Language. W3C Recommendation (2009), <http://www.w3.org/TR/owl-ref/> (Retrieved November 27, 2009)
5. Karmacharya, A., Cruz, C., Boochs, F., Marzani, F.: Support Of Spatial Analysis Through A Knowledgebase - A New Concept To Exploit Spatial Information Shown For Industrial Archaeology. In: The International 24th Cartographic Conference, Chili, November 16-21 (2009)
6. Karmacharya, A., Cruz, C., Boochs, F., Marzani, F.: ArchaeoKM: toward a better archaeological spatial datasets management. In: Computer Applications and Quantitative Methods in Archaeology (CAA), Williamsburg, Virginia, USA (2009)
7. Karmacharya, A., Cruz, C., Boochs, F., Marzani, F.: Managing Knowledge for Spatial Data – A Case Study with Industrial Archaeological Findings. In: Digital Heritage in the New Knowledge Environment: Shared Spaces & Open Paths to Cultural Content, Athenes, Grece (2008)
8. Cruz, C., Nicolle, C.: Ontology Enrichment and Automatic Population From XML Data. In: 4th ODBIS Workshop on Ontologies-based Techniques for DataBases in Information Systems and Knowledge Systems, Co-located with VLDB, August 23 (2008)
9. Boley, H.: The Rule Markup Initiative, <http://ruleml.org/> (Last Visited February 11, 2010)
10. Lieberman, J., Singh, R., Goad, C.: W3C Geospatial Ontologies – W3C Incubator Group Report. W3C (2009), <http://www.w3.org/2005/Incubator/geo/XGR-geo-ont-20071023/>
11. PostgreSQL, PostGIS Manual, PostgreSQL documentation (2008)

12. Wüst, T., Nebiker, S., Landolt, R.: Applying the 3D GIS DILAS to Archaeology and Cultural Heritage Projects-Requirements and First Results. Basel University of Applied Sciences, Muttenz, Switzerland
13. Cosmas, J., Itagaki, T., Green, D., Grabczewski, E., Waelkens, M., Degeest, R., et al.: 3D MURALE:A Multimedia System for Archaeology. In: Proc. ACM Virtual Reality, Archaeology and Cultural Heritage (VAST 2001) (November 2001)
14. Semantic Interoperability Community of Practice (SICoP), Introducing Semantic Technologies and the Vision of the Semantic Web (2005)
15. Roman, D., Klien, E., Skogan, D.: SWING – A Semantic Web Services Framework for the Geospatial Domain. In: Position Paper at the Terra Cognita 2006 - Directions to the Geospatial Semantic Web Workshop, Athens, USA (2006)
16. Cruz, I.F.: Geospatial Data Integration, ADVIS Lab, Department of Computer Science, University of Illinois, Chicago (2004)
17. Chaudhary, A., Sunna, W., Cruz, I.F.: Semi-automatic Ontology Alignment for Geospatial Data Integration. In: Egenhofer, M.J., Freksa, C., Miller, H.J. (eds.) GIScience 2004. LNCS, vol. 3234, pp. 51–66. Springer, Heidelberg (2004)
18. Tanasescu, V., Gugliotta, A., Domingue, J., Gutiérrez Villarías, L., Davies, R., Rowlatt, M., Richardson, M., Stinčić, S.: A Semantic Web Services GIS based Emergency Management Application. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 959–966. Springer, Heidelberg (2006)
19. Lieberman, J., Singh, R., Goad, C.: W3C Geospatial Ontologies – W3C Incubator Group Report. W3C,  
<http://www.w3.org/2005/Incubator/geo/XGR-geo-ont-20071023/>  
 (Last visited, June 23 2009)

## Appendix: Figures



**Fig. 1.** Basic ontological structure of ArchaeoKM



**Fig. 2.** Spatial adjustment within the ontology

# Computer-Based Dietary Menu Planning: How to Support It by Complex Knowledge?

Barbara Koroušić Seljak

Computer Systems Department, Jožef Stefan Institute,  
Jamova cesta 39, 1000 Ljubljana, Slovenia  
[barbara.korousic@ijs.si](mailto:barbara.korousic@ijs.si)

**Abstract.** Today too many patients are confronted with a problem of malnutrition. As finding and treating malnutrition early may help the patient gain or maintain weight, improve the patient's response to therapy, and reduce complications of treatment, different tools for nutrition screening, assessment, and treatment have been developed. In this paper, we introduce a computer-based method for dietary menu planning to be applied as a crucial part of the screening and assessment tools. As such a tool may be useful if and only if it is based upon a comprehensive domain knowledge, methodology for capturing and reasoning such knowledge is longed for. In the paper, we formalize the menu-planning problem, and describe the evolutionary algorithm for its solving.

## 1 Introduction

The history of nutritional science dates back to the end of the eighteenth century, when the French chemist Lavoisier discovered a relationship between the human metabolism of food and the process of breathing [3]. Since then, scientists, using scientific methods, have acquired a vast amount of information and knowledge regarding the role of food in the maintenance of good health, which argue in favor of the importance of adequate and balanced nutrition in all life periods.

Nutrition is particularly important for critically ill patients. Thanks to findings of a huge number of studies conducted around the world, many practitioners and their patients are becoming aware of the fact that for patient treatment food is as fundamental as drugs and other therapeutical methods [21]. Let us mention, for example, a recent observational study involving 16,290 patients aged 18 years and over from 748 wards, 256 hospitals and 25 countries, which reported that decreased food intake is a major independent risk factor for in-hospital mortality [7]. Another study by the European Society for Clinical Nutrition and Metabolism reported that the average prevalence of malnutrition<sup>1</sup> in European hospitals is 35%, with a range from 10-85% [9].

<sup>1</sup> Malnutrition is a state of nutrition in which a deficiency or imbalance of energy, protein and other nutrients causes measurable adverse effects on tissue or body form, function or clinical outcome [22].

### 1.1 How to Treat Malnutrition?

Finding and treating malnutrition early may help the patient gain or maintain weight, improve the patient's response to therapy, and reduce complications of treatment [1], [16]. *Nutrition screening* is the first step in assessing and treating nutrition problems. Its purpose is to identify an individual at high risk requiring nutritional support. There exist screening tools that address questions on the patient's recent weight loss, current body mass index, disease severity, and recent food intake in a simple and rapid way. The next step is *nutritional assessment* that is important for detailed diagnosis of acute and chronic malnutrition. It is more complex than screening and should include the following principles: history and examination, disease status, functional assessment, laboratory tests, and fluid balance. There are many methods and indexes based on the assessment methods. However, their interpretation and correlation can still be problematic.

Modern information and communication technologies may be applied to effectively implement screening and assessment tools. In this paper, an example of such a tool as part of an eHealth system for clinical nutrition support is described. Its main feature is personalized dietary-menu planning based on data and information acquired by the nutrition screening and assessment.

The paper is organized as follows: Section 2 presents a brief survey of computer-based methods for menu planning, a formulation of the menu-planning problem, and an introduction to evolutionary computation and multi-objective optimization; Section 3 describes an evolutionary algorithm for menu planning; Section 4 gives an evaluation of the approach; and Section 5 gives concluding remarks.

## 2 Computer-Based Menu Planning

Menu planning is an art that one often learns by a costly heuristic or trial-and-error process. A well-trained professional can cope with the complexity of regular menu planning, but as soon as one attempts to meet the needs of diverse groups, control costs and quality, and schedule production tightly for maximum utilization of labour and equipment time, the probability of success diminishes.

The need for computer-based methods, which facilitate the routine decisions in menu planning, has already been recognized four decades ago<sup>2</sup>. As a long-term result, the quality of information and data necessary for computerization of the menu planning process has become available to many institutions, and its apparent feasibility has thus increased. The EuroFIR platform [10], for example, is such an example of a comprehensive, coherent and validated databank for the distribution of food composition data (FCD). It has been implemented as a decentralized computer system through a network of local FCD storages under the control of a local authority. This system supports data interchange through the EuroFIR Web Services interfaces; the interchange format is based on the

---

<sup>2</sup> A more comprehensive review of the historical development of computer-based menu planning methods can be found in [17].

Extensible Markup Language (XML) [18] that is an open standard for structuring information commonly used for sharing structured data, especially via the Internet.

## 2.1 Knowledge Base

Besides *food composition data* personalized menu planning require other data and information, such as

- *personal data* (like weight, height, sex, age) and many others (like clinical records) acquired by the nutrition screening and assessment,
- *dietary reference values* for the intake of energy and nutrients,
- *meal formats* and other *cuisine rules*.

Dietary reference values are standards of the amounts of each nutrient needed to maintain good health [5]. These are established specifically for population groups and not for individuals, who differ in the daily amounts of nutrients they need. For most nutrients the measured average requirement plus 20% takes care of the needs of nearly everyone. For patients and individuals with special nutrition needs, dietary reference values are defined individually by dietitians and nutritionists, respectively, considering the personal data and medical knowledge [20]. Last but not least cuisine rules are important part of the knowledge base because every nation, ethnic group and even individual has its own food, which must be considered by menu planning otherwise its purpose is completely useless.

## 2.2 Formulation

Mathematically, menu planning can be reduced to a multidimensional knapsack problem (MDKP), which is a widely studied combinatorial optimization problem [15]: *Given foods of different values and volumes, the MDKP is to find the most valuable combination of foods that fits in a knapsack of fixed volumes. Values are defined subjectively with respect to food quality, cost and aesthetic parameters (comprising taste, consistency, color, temperature, shape and method of preparation). Volumes are defined by dietary recommendations and guidelines.*

We are given a knapsack of  $m$  volumes  $C_k, k = 1, 2, \dots, m$ , and  $n$  food items. Each item  $i$  has nine values  $v_{ik} \in \mathbb{N}^+, v_{ik} > 0, k = 1, 2, \dots, 9$ , and  $m$  volumes  $\omega_{ik} \in \mathbb{R}^+, \omega_{ik} > 0, k = 1, 2, \dots, m$ , one for each capacity. We are looking for a composition of  $t$  items,  $t < n$ , such that  $\sum_{i=1}^t \omega_{ik} x_i \leq C_k$  ( $\Phi$  can be  $\leq$  or  $\geq$ ,  $k = 1, 2, \dots, m, t \leq n$ ), and for which the total values

$$\sum_{i=1}^t v_{ik} x_i, k = 1, 2$$

are maximized, while

$$\sum_{i=1}^t v_{ik} x_i, k = 3$$

and

$$\sum_{j=1}^{n_{al}} \left| \sum_{i=1}^n h_{lj}(x_i) - \frac{\sum_{i=1}^n h(x_i)}{n_{al}} \right|, l = 4, 5, \dots, 9$$

are minimized, where  $n_{al}$  is the number of possible states of an aesthetic standard  $l$ . The functions used in the above objective function are defined as follows:

$$h_{lj}(x_i) = \begin{cases} 0, & \text{if } x_i = 0 \\ 1, & \text{if } x_i > 0 \wedge v_{il} = j \end{cases}, \quad i = 1, 2, \dots, n, \quad l = 4, 5, \dots, 9,$$

and

$$h(x_i) = \begin{cases} 0, & \text{if } x_i = 0 \\ 1, & \text{otherwise} \end{cases}, \quad i = 1, 2, \dots, n.$$

The parameter  $x_i \in [0.25P_i, 2P_i]$  denotes the quantity of the selected item  $i$  expressed in a unit (gram, milligram, microgram, milliliter, etc.). Its value is limited by the fractions of the item's portion size  $P_i$ .

The MDKP is easy to formulate, yet its decision problem is *NP-complete*<sup>3</sup>. Knapsack values and volumes of the MDKP instance for the menu-planning problem are linear, but highly-complex because they are weakly correlated. As there are at least two optimal solutions that are not indifferent to each other, the problem is multimodal. Another difficulty is that foods are selected from a FCD, which consists of several thousand items having tens of composition parameters. As a consequence, the decision space contains a large set of potential solutions to the menu-planning problem. Moreover, the problem landscape defined by the decision and the objective space contains several peaks.

There exist different approaches for computing an optimal solution of a MDKP including exact algorithms, approximation algorithms, and heuristic algorithms that try to tackle the problem's complexity. A considerable number of heuristic algorithms, which work "reasonably well" on many MDKP instances, but for which there is no proof that they are both always fast and always produce a good result, have already been developed and tested. These algorithms rely on different heuristics that may be greedy-type, relaxation-based, approximate dynamic programming, or metaheuristics.

A full-scale presentation of methods and techniques available for the solution of knapsack problems was provided in the book by Kellerer, Pferschy and Pisinger [15]. In addition, reviews of multi-constrained 0-1 knapsack problems, presenting a subset of MDKPs, and associated heuristic algorithms can be found in [4] and [8].

---

<sup>3</sup> In the complexity theory, NP-complete problems are the most difficult problems, which cannot be solved by exact software techniques in a deterministic polynomial time but require time that is superpolynomial in the input size.

### 2.3 Evolutionary Computation

We decided to apply a metaheuristic algorithm using an *evolutionary computation* technique to solve the menu-planning MDKP instance. Evolutionary computation is a subfield of artificial intelligence that involves numerical and combinatorial optimization problems. It uses iterative progress, such as growth or development in a population of potential problem solutions. The field comprises many techniques, mostly involving metaheuristic optimization algorithms, such as evolutionary algorithms (EAs), swarm intelligence and other often bio-inspired algorithms. These techniques rely on analogies to natural processes; some of them have been inspired by biological mechanisms of evolution. The first ideas developed in the sixties by Holland [14] and Fogel [12] have already reached a stage of some maturity. In the last decade, numerous algorithms taking inspiration from nature have been proposed to handle continuous optimization problems: real-coded genetic algorithms using some specific operators, evolution strategies using Gaussian mutations with adaptive or self-adaptive update strategies, and differential evolution, to name a few.

As real-world optimization problems may involve objectives, constraints and parameters, which constantly change with time, dynamic consideration using evolutionary computation methods have also raised a lot of interest within the last few years. For these dynamic and uncertain optimization problems the objective is no longer to simply locate the global optimum solution, but to continuously track the optimum in dynamic environments, or to find a robust solution that operates optimally in the presence of uncertainties. In our case, optimal solutions are healthy menus termed as

- *good* or *bad* in terms of multiple conflicting objectives, such as: cost, quality of ingredients, aesthetic standards, or other factors; and
- *feasible*, satisfying all the problem constraints that are defined by dietary recommendations and guidelines.

While classical deterministic optimization methods can at best find one solution in one simulation run, evolutionary techniques are more efficient in finding multiple trade-off optimal solutions in a single simulation run. These solutions have a wide range of values for each objective representing the multi-dimensional Pareto-optimal front, requiring an additional decision-making activity for choosing a single solution from the front.

## 3 Evolutionary Method

We applied the NSGA-II (*Elitist Non-Dominated Sorting Genetic Algorithm*) evolutionary algoritem by Deb [6] in a multi-level way to solve the MDKP of menu planning. In general, the NSGA-II is a multi-objective EA that can be characterized by the use of three ideas: Pareto dominance-based 'fitness' evaluation, diversity maintenance, and elitism. The main idea behind the method is to develop healthy meals and daily menus independently, guiding the optimization to overall Pareto optimal  $n$ -day menus. All objectives are treated as equally

important. The decision on the best compromise to be chosen among adequate Pareto-optimal solutions is made after the search by a human expert.

**Fitness evaluation.** In each generation, the fitness of the (global or local) population is evaluated using the following objective functions:

$$f_k(\mathbf{x}) = \frac{1}{\sum_{i=1}^n v_{ik} x_i}, k = 1, 2,$$

$$f_3(\mathbf{x}) = \sum_{i=1}^n v_{i3} x_i,$$

$$f_l(\mathbf{x}) = \sum_{j=1}^{n_{al}} \left| \sum_{i=1}^n h_{lj}(x_i) - \frac{\sum_{i=1}^n h(x_i)}{n_{al}} \right|, 1 \leq i \leq n, 4 \leq l \leq 9, \quad (1)$$

$$h_{lj}(x_i) = \begin{cases} 0, & \text{if } x_i = 0 \\ 1, & \text{if } x_i > 0 \wedge v_{il} = j \end{cases}, \quad 1 \leq i \leq n, 4 \leq l \leq 9,$$

$$h(x_i) = \begin{cases} 0, & \text{if } x_i = 0 \\ 1, & \text{otherwise} \end{cases}, \quad 1 \leq i \leq n.$$

where  $v_{i1}$  denotes the functionality of the food item  $i$ ,  $v_{i2}$  its quality in the season,  $v_{i3}$  the cost,  $v_{i4}$  the taste,  $v_{i5}$  the consistency,  $v_{i6}$  the color,  $v_{i7}$  the temperature,  $v_{i8}$  the shape,  $v_{i9}$  the method of preparation, and  $n_{al}$  the number of possibilities for the  $l$ -th aesthetic standard. The aim of the 'global' and the 'local' evolutionary algorithms is to *minimize* the objective functions of (1).

**Methods for repairing infeasible individuals.** We handle the menu-planning problem's constraints by using the following repair methods:

1. At the *meal level*, we first replace in each infeasible meal those courses that mostly contribute to the violation of constraints with similar but more appropriate ones (e.g. we replace beef broth with vegetable soup if there is a lack of fiber in a meal), and then convert infeasible solutions into feasible solutions using a deterministic local optimization procedure of linear programming. This procedure, based on the simplex method [2], refines the quantities of foods to satisfy the meal subproblem constraints;
2. At the *daily-menu* and the *n-days menu level*, we repair infeasible individuals by replacing critical meals that do not satisfy the constraints on the major food groups with more appropriate ones. Here, we use the problem-specific knowledge, considering a recommendation that i) a daily menu has to be composed of a certain number of food units from each major food group and ii) an *n*-days menu has to include a diverse set of foods from the major food groups. There may be limitations on frequency of red meat, fish, potatoe etc.

For this aim, we apply the Lamarckian repair scheme, in which replacements of critical meals are used to generate new 'offspring' [8].

The replacement of meal or menu elements that mostly contribute to the violation of constraints requires a prior sorting of elements. We apply the inverse non-dominated sorting, meaning that the replacing procedure starts replacing the elements from the last front and ends replacing the elements from the Pareto optimal front. As a matter of fact, because the procedure is greedy, the replacement is stopped as soon as the first improvement is achieved. In this way, the cost of repair is minimized.

## 4 Evaluation of the Method

As a demonstration, we applied the multi-level NSGA-II to a problem of planning optimal weekly menus for people without specific dietary requirements in a local hospital. We started the 'global' NSGA-II from an existing non-optimal weekly menu.

In Table 1, we list the parameters used to generate meals, daily menus and weekly menus by the multi-level NSGA-II. We ran the algorithm for 25 times to obtain the experimental results presented in Table 2. In Figure 1, a part of the feasible search space, whose shape is depicted for three objectives, but actually modified by nine objectives, is presented. A subset of the analysis results for a weekly menu generated by the multi-level NSGA-II is presented in Table 3. This

**Table 1.** Parameters

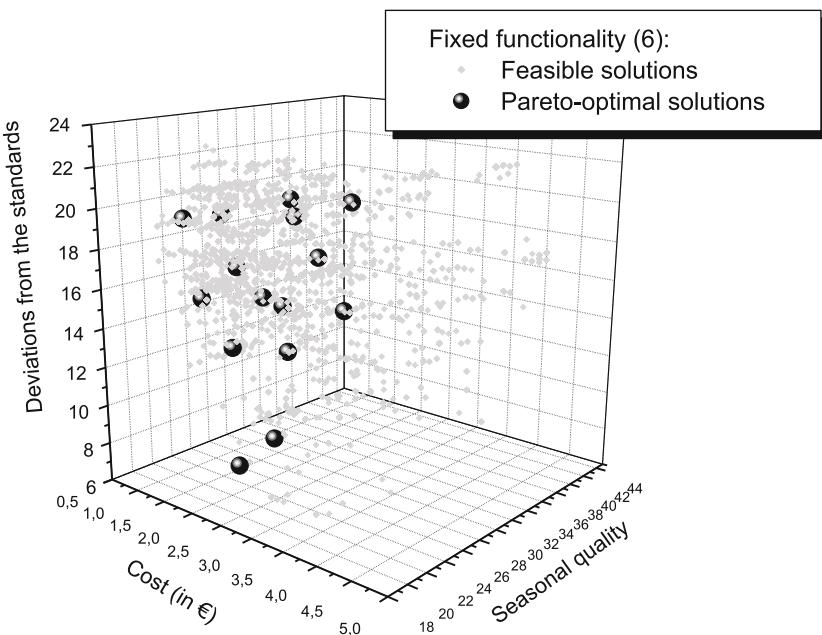
PARAMETER	THE WEEKLY-MENU LEVEL	THE DAILY-MENU LEVEL	THE MEAL LEVEL
Chromosomes length	7	5	10
Population size	100	100	100
Pool size	700	500	—
Crossover probability	0.7	0.7	0.7
Mutation probability	0.14–0.01	0.2–0.01	0.1–0.017
Selection type		Two-point crossover	
Crossover type		Linear descending mutation	
Number of iterations	24	18	35

**Table 2.** Experimental results

Percentage of infeasible solutions in each new generation	89
Percentage of successfully repaired infeasible solutions	65
Cost (EUR)	QUALITY IN SEASON
Best result	48
Median	28
Worst result	18
Mean value	28.3
Standard deviation	4.7
Functionality	
Best result	12
Median	6
Worst result	0
Mean value	5.8
Standard deviation	3.4

**Table 3.** Analysis results of a computer-generated weekly menu

	MEAN DAILY VALUES	DACH RECOMMENDED DIETARY ALLOWANCES	GOAL ACHIEVED (%)
Energy (kcal)	2,036	2,000	102
Proteins (% of energy)	16	10–15	✓
Lipids (% of energy)	28	15–30	✓
Carbohydrates (% of energy)	56	55–75	✓
Simple sugars (% of energy)	4.5	< 10	✓
Saturated fats (% of energy)	6.6	< 10	✓
Ratio of omega-6 to omega-3 fatty acids	3.9	5	✓
Dietary fibre (g)	33.6	30–40	✓
Cholesterol (mg)	160	300	✓
Sodium (mg)	2,500	550–2,400	104
Breads, cereal, rice, and pasta (no. of units)	11.2	11	102
Vegetables (no. of units)	4.7	5	94
Fruits (no. of units)	3	3	100
Milk, yogurt, and cheese (no. of units)	2	2	100

**Fig. 1.** Part of the problem's search space

weekly menu was generated with respect to the following requirements for the major food group of meat and its substitutes: white meat, legumes, fish and eggs once per week, and red meat three times per week.

The method's application in real-world cases is also given in the recently published paper [17].

## 5 Conclusions

This paper has provided a brief introduction to an extension of the computer-based dietary menu planning that is an important part of nutrition screening and assessment tools. Computer-based menu planning may greatly facilitate the process, however, it must be supported by a comprehensive domain knowledge. For this reason, we have created ontologies to model food composition data, dietary reference values, and cuisine rules. We have modeled the problem-specific knowledge by mapping data information from different sources, including food composition data provided by the EuroFIR XML Schemata, and dietary reference values from the D-A-CH and the ESPEN recommendations and guidelines. Cuisine rules has been automatically captured from the data-mining model built from the existing data on national recipes and daily and  $n$ -day institutional meals.

Our future work involves incorporation of this knowledge into the system and upgrade of the evolutionary computation heuristics for menu-planning with ontology reasoning. Gaál, Vassányi and Kozmann proposed a method for weekly dietary menu planning that considers expert knowledge on menu harmony by a domain ontology and combines evolutionary programming with ontology reasoning [11]. Such an approach has proved as efficient to be applied in combination with evolutionary computation.

**Acknowledgments.** The work presented in this paper is co-funded by the Ministry of Higher Education, Science and Technology of the Republic of Slovenia through the European Regional Development Fund.

## References

1. Barendregt, K., et al.: Diagnosis of malnutrition - Screening and Assessment. In: Sobotka, L. (ed.) Basics in Clinical Nutrition, 3rd edn., pp. 11–18. Galén, Prag (2004)
2. Bhatti, M.A.: Practical Optimization Methods. Springer, Heidelberg (2000)
3. Carpenter, K.J.: A Short History of Nutritional Science: Part 1-4. American Society for Nutritional Sciences. J. Nutr. 133, 638–645, 975–984; 3023–3032; 3331–3342 (2003)
4. Chu, P.C., Beasley, J.E.: A Genetic Algorithm for the Multidimensional Knapsack Problem. Journal of Heuristics 4, 63–86 (1998)
5. Referenzwerte für die Nährstoffzufuhr. D-A-CH Referenzwerte der DGE, GE, SGE/SVE, <http://www.dge.de/modules.php?name=Content&pa=showpage&pid=3>
6. Deb, K.: Multi-Objective Optimization using Evolutionary Algorithms. John Wiley & Sons, Ltd, Chichester (2001)
7. Hiesmayr, M., et al.: Decreased food intake is a risk factor for mortality in hospitalised patients: the NutritionDay survey 2006. Clin. Nutr. 28, 484–491 (2009)

8. Ishibuchi, I., Kaige, S.: Comparison of Multiobjective Memetic Algorithms on 0/1 Knapsack Problems. In: Cantú-Paz, E., Foster, J.A., Deb, K., Davis, L., Roy, R., O'Reilly, U.-M., Beyer, H.-G., Kendall, G., Wilson, S.W., Harman, M., Wegener, J., Dasgupta, D., Potter, M.A., Schultz, A., Dowsland, K.A., Jonoska, N., Miller, J., Standish, R.K. (eds.) GECCO 2003. LNCS, vol. 2724. Springer, Heidelberg (2003)
9. The European Society for Clinical Nutrition and Metabolism (ESPEN),  
<http://www.espen.org>
10. The European Food Information Resource Network (EuroFIR),  
<http://www.eurofir.net>
11. Gaál, B., Vassányi, I., Kozmann, G.: Application of Artificial Intelligence for Weekly Dietary Menu Planning. In: Studies in Computational Intelligence (SCI), vol. 65, pp. 27–48. Springer, Heidelberg (2007)
12. Fogel, L.J., Owens, A.J.: Artificial Intelligence through Simulated Evolution, NY John Wile (1966)
13. Golob, T., et al.: Slovenian Food Composition Tables - Meat and Meat Products. Department of Food Science and Technology, Biotechnical Faculty, University of Ljubljana (2006)
14. Holland, J.H.: A Logical Theory of Adaptive Systems - Informally Described, pp. 1–5. The University of Michigan, Ann Arbor (1961)
15. Kellerer, H., Pferschy, U., Pisinger, D.: Knapsack Problems. Springer, Heidelberg (2004)
16. Kondrup, J., et al.: ESPEN guidelines for nutrition screening 2002. Clin. Nutr. 22, 415–421 (2003)
17. Koroušić Seljak, B.: Computer-Based Menu Planning. Journal of Food Composition and Analysis 22(5), 1650–1655 (2009)
18. Pakkala, H., Korhonen, T.: XML Schemata for EuroFIR Metadata Transport Package 1.0. Version 1.1. The EuroFIR Technical Series (2009)
19. Pelto, G.H., Pelto, P.J., Messer, E.: Research Methods in Nutritional Anthropology. United Nations University Press, Tokyo (1989)
20. Rotovnik Kozjek, N., Milošević, M.: The Slovenian Guidelines for Clinical Nutrition. Ministry of Health of the Republic Slovenia (2007)
21. Sobotka, L.: Basics in Clinical Nutrition, 3rd edn. Galén, Prag (2004)
22. Stratton, R.J.: Malnutrition: another health inequality? Proceedings of the Nutrition Society 66, 522–529 (2007)

# Flexible Semantic Querying of Clinical Archetypes

Catalina Martínez-Costa, José Antonio Miñarro-Giménez,  
Marcos Menárguez-Tortosa, Rafael Valencia-García,  
and Jesualdo Tomás Fernández-Breis

Departamento de Informática y Sistemas, Universidad de Murcia, Spain  
{cmartinezcosta,jose.minyarro,marcos,valencia,jfernand}@um.es

**Abstract.** In the last years, a number of semantic biomedical systems have been developed. However, their query interfaces are not easy to use for biomedical researchers since they require expertise in semantic languages. Consequently, its practical usage is limited. In this paper, we address this issue by moving the complexity in the design of the semantic query from knowing such query languages to exploring the domain ontology. We also report how this system has been applied to query a semantic repository of clinical archetypes.

**Keywords:** Ontology, Clinical Archetypes, Semantic Querying.

## 1 Introduction

In the last years, there has been an increasing interest in the development and application of ontologies in biomedical domains. A clear example is the existence of communities such as the OBO Foundry [1], which attempt to develop a set of orthogonal biomedical ontologies that could support biomedical research. The Semantic Web is a next generation web in which automated processing of information will deliver more concise results to the user, and allow machines to perform time consuming tasks [2]. For the Semantic Web to work, information must be expressed and published with precise semantics, via ontologies. An ontology is a formalization of a knowledge domain, a set of concepts and their relationships, which can be used by machines to perform automated reasoning. The Web Ontology Language (OWL) [3], which is a W3C [4] official recommendation, is one of the most widely used ontology languages. OWL is designed to implement web interoperable ontologies, and it presents an optimal balance between decidability and expressiveness. Moreover, as stated in [5], Semantic Web technologies have been increasingly used for data integration in life sciences and provide a useful framework for translational medicine. In this way, different Semantic Web technologies such as RDF [6], OWL [3] and SPARQL [7] have been used for developing semantic biological solutions (see for instance Biogateway [8]). Another remarkable effort is Bio2RDF [9], which pursues to build a semantic coordinate system for bioinformatics, the so-called Semantic web atlas of post-genomic knowledge.

Most biomedical semantic systems provide query interfaces based on powerful semantic languages such as DL or SPARQL, because they allow for exploiting the semantics of the domain. Such languages are very useful because they permit to incorporate all the restrictions modeled in the ontology in the queries. These languages are not very difficult for users with a computing background or having some experience in relational query languages such as SQL, but they are currently far from being usable by the average biomedical researcher, who should be the final users of those semantic tools. Therefore, mechanisms for making query construction simpler for such users are required.

In this work, we address this problem and present our query system, which can be used without any knowledge of semantic query languages. The current version does not provide the ideal query interface for biomedical researchers, although it constitutes our first step, by moving the complexity from knowing SPARQL to knowing a particular domain ontology. Once the query is designed by the user, then the SPARQL query is automatically generated and executed. In the last years, our research group has developed methods for representing and managing Electronic Healthcare Records using ontologies. One of the results of such research effort has been a repository of OWL clinical archetypes. In this paper, we report how this system can be applied to querying such OWL repository and retrieving the desired information about existing clinical archetypes by making semantic queries.

## 2 Background

### 2.1 Ontology Guided Queries

Researchers have noticed that “*the casual user is typically overwhelmed by the formal logic of the Semantic Web*” [10]. This is due to the fact that users, in order to use ontologies, have to be familiar with [11]: (1) the ontology syntax (e.g. RDF, OWL), (2) some formal query language (e.g. SPARQL), and (3) the structure and vocabulary of the target ontology.

The approach taken by most researchers is the use of natural language interfaces (NLIs). NLIs aim to provide end-users with a means to access knowledge in ontologies hiding the formality of ontologies and query languages [12]. Thus, NLIs help users avoid the burden of learning any logic-based language offering end-users a familiar and intuitive way of query formulation.

However, the realization of NLIs involves several difficulties, one of such problems being that of linguistic variability and ambiguities. In recent years, Controlled Natural Language (CNL) [13] has received much attention due to its ability to reduce ambiguity in natural language. CNLs are mainly characterized by two essential properties [14]: (1) their grammar is more restrictive than that of the general language, and (2) their vocabulary only contains a fraction of the words that are permissible in the general language. These restrictions aim at reducing or even eliminating both ambiguity and complexity. In recent years, the utilization of NLIs and CNLs in the context of the Semantic Web has received much attention. Several

platforms have been developed to function as either natural language ontology editors or natural language query systems. Two good examples in the first category are CNL Editor [15] (formerly OntoPath [16]) and GINO [10]. Moreover, our research group has recently developed OWLPath [17], a CNL-based NLI that assists users in designing their queries. Such systems rely on the existence of a grammar that guides the process and also constrains the expressiveness of the queries.

## 2.2 Querying Clinical Archetypes

Dual model approaches for developing Electronic Healthcare Records (EHR) standards differentiate between information and knowledge. In such standards, archetypes are the knowledge layer and represent healthcare and application specific concepts such as the measurement of cholesterol, blood pressure and so on, and they are usually defined using the Archetype Definition Language (ADL). This language provides a concrete syntax for expressing them as text documents. Nowadays there are two major EHR standards based on this modeling approach, namely, ISO EN 13606 [18] and OpenEHR [19].

There is very limited work on the implementation of queries on archetypes. The openEHR Foundation is developing the Archetype Query Language (AQL), a declarative query language based on a path syntax. It has been specifically designed for querying any kind of archetype, not only the clinical ones. It is based on the idea that each concept and all its properties have to be identified in an unique way in the archetype. Thus, finding a concept or property requires the construction of a path by using identifiers and concept or property names. The query syntax is similar to SQL, therefore the clauses SELECT, FROM, WHERE, and ORDER BY provide the basic structure of AQL.

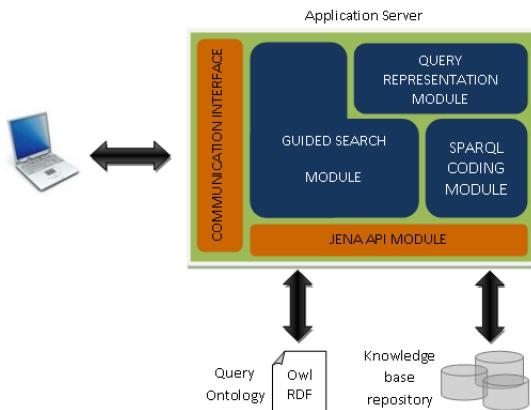
The ADL representation of archetypes has some limitations [20]. The use of ontologies for representing clinical archetypes offers some benefits against the use of ADL. Ontologies allow performing activities such as comparison, selection, classification and consistency checking in a more generic, easier and efficient way. Moreover, languages such as AQL do not offer the possibility of exploiting the semantics of archetypes, since the queries have also a syntactic nature.

Our research group has developed OWL ontologies for the referred standards. Their design required the semantic interpretation of both reference and archetype models of the standards. The resulting ontologies allow defining archetypes in a more legible and accessible way. They combine features from both archetype and reference models, and allow for publishing the archetype content as interrelated information, that is, a concept has all its properties accessible by itself and is directly connected with other concepts. In [20] we also reported how to transform ADL archetypes into OWL, and this allowed us to create our repository of OWL archetypes, which is included into our archetype management system, ArchMS [21]. However, the early versions of this system included query facilities that did not really exploit the full semantics of the domain. Recently, we applied OWLPath to guide the construction of such queries. This system demonstrated its usefulness for guiding the construction of queries that do not require the combination of many concepts through relations in the same query because that

would require a very complex grammar. Consequently, for biomedical domains, which are very complex and rich in relations, other solutions would be more practical.

### 3 The Ontology Guided Query System

In this section, we describe the system that we have developed to allow biomedical users to formulate complex semantic queries. Its architecture is depicted in Figure 1, and comprises (1) the communication interface, which includes the user interface, (2) the Jena module for accessing the semantic RDF/OWL repository, (3) the query ontology, and (4) the guided search subsystem.



**Fig. 1.** The architecture of the semantic flexible query system

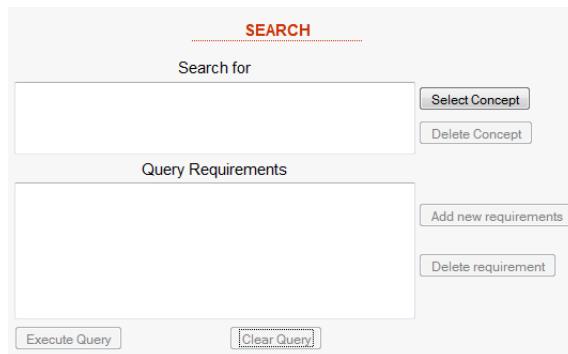
The web user interface has been implemented using Ajax, Javascript, Servlets and JSP. The communication interface holds much of the implemented logic/code to make it independent from the ontology guided search modules. The JENA API MODULE allows us to combine the knowledge-based repository and the ontology used for guiding the definition of the semantic query. Jena is an open source Java framework for building semantic web applications which provides us a programmatic environment for RDF, OWL and SPARQL. The ontology represents the conceptualisation of the knowledge domain of the repository. Next, the modules that implement the ontology guided search method are described:

- *Query representation module*: It stores the query defined by the user through the web interface. This is used by the Guided Search module for limiting the options offered to the user for extending the query, and by the SPARQL coding module for generating the SPARQL queries.
- *SPARQL coding module*: It transforms the query into an optimized SPARQL one. The generation of the query is not difficult given the representation provided at the query representation module, but the hardest part is sorting

the different type of condition clauses to make the SPARQL queries run fast enough. The most restrictive condition clauses are planned to be executed first.

- *Guided search module:* It assists users in the design of the query by managing the knowledge of the query ontology that is available for the query, and it is also responsible for gathering and returning the query results to the users.

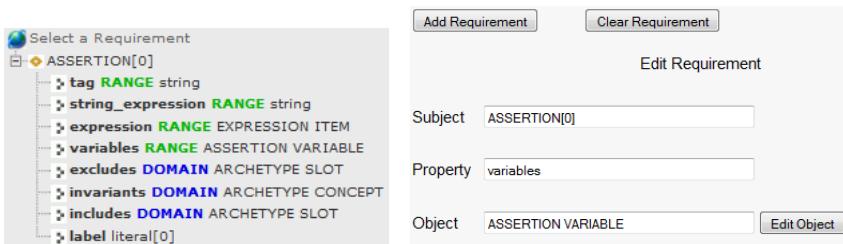
The Ontology Guided Search user interface is shown in Figure 2. The “Search for” area contains the ontology entities which users want to retrieve information about. By clicking on “Select Concept”, the ontological tree is shown to the user, who makes a choice and that concept is added into the query.



**Fig. 2.** The web interface for the design of the queries

The conditions and constraints on the properties for the current query are displayed in “Query Requirements” and can be added by clicking on “Add new requirement”. When this occurs, a list of allowed condition clauses is shown. The left part of Figure 3 shows the allowed condition clauses for a query about the concept “ASSERTION” of an ontology of archetypes. The concept “ASSERTION[0]” has some properties associated, such as “tag”, whose value is a String; “variables”, which associates instances of “ASSERTION” with instances “ASSERTION VARIABLE”; and “invariants”, which associates instances of the concepts “ARCHETYPE CONCEPT” and “ASSERTION”. Once a property is selected, new options are offered to the user to specify the value condition (see the right part of Figure 3). In case the property has associated a simple data type such as a string, the user can input the value. In case the property represents a relation with another ontology concept and this one can be specialized, then the user is given the possibility of choosing one of its sub-concepts by clicking on “Edit Object”. If the condition clauses refer to new instances of ontology concepts, new allowed condition clauses will be shown when adding new constraints to the query.

Once the query is executed, the results are displayed in a table whose columns are the ontology concepts selected into “Search for” and the rows are the clickable



**Fig. 3.** Defining a requirement

and browseable results. Figure 4 and 5 (see next section) are examples of such output.

## 4 Application of the Approach to Clinical Archetypes

In this section, we describe how this system has been used for designing queries about clinical archetypes. For this purpose, the ontologies previously defined in [22] were used. Archetypes are considered in such ontologies as individuals, so queries for searching particular archetypes can be defined. Next, two examples of queries are described.

### 4.1 Finding All Archetypes Written in Spanish

Table 1 shows the representation of the query after the selection of the concepts and properties in the user interface. This query means that we want instances of *ARCHETYPE CONCEPT* whose property *original\_language* is of *CODE PHRASE TYPE*, that defines the *code\_string* property whose value has to be “es”. The “Query requirement” area will constrain the archetype clinical properties, here it restricts the original language property of the archetype concept to be Spanish. The [i] suffixes indicate that they refer to different instances. For instance, the two requirements that contain *CODE PHRASE TYPE[1]* refer to the same instance.

Thus, the user has defined the query without using SPARQL. The corresponding SPARQL query is shown in Table 2. This query is automatically generated by the system. For this purpose, The “Search for” and “Query requirements” areas are changed into SELECT and WHERE SPARQL clauses. Each concept

**Table 1.** Query One:Find all archetypes that have been written in Spanish

<b>Search for:</b> ARCHETYPE CONCEPT[0]
<b>Query requirements:</b>
ARCHETYPE CONCEPT[0] → original_language → CODE PHRASE TYPE[1] CODE PHRASE TYPE[1] → code_string → String Type[2] String Type[2] → list → es

**Table 2.** Query One in SPARQL

```

@prefix ar: <http://klt.inf.um.es/~cati/ontologies/test/ISO13606-AR-v2.1.owl>.
SELECT ?ARCHETYPE_CONCEPT_0
WHERE {
?ARCHETYPE_CONCEPT_0 ar:original_language ?CODE_PHRASE_TYPE_1 .
?CODE_PHRASE_TYPE_1 ar:code_string ?String_Type_2 .
?String_Type_2 ar:list ?literal_3 .
FILTER (regex(?literal_3, "es")) .
}

```

ARCHETYPE CONCEPT[0]
CEN-EN13606-CLUSTER.Descripcion_de_la_medicacion.v1
CEN-EN13606-COMPOSITION.Historia_clinica_resumida.v1
CEN-EN13606-COMPOSITION.informe_de_alta-cirugia_general.v1
CEN-EN13606-COMPOSITION.informe_de_alta-medicina_digestiva.v1
CEN-EN13606-COMPOSITION.informe_de_alta.v1
CEN-EN13606-ENTRY.Colesterol.v1
CEN-EN13606-ENTRY.Creatinina.v1
CEN-EN13606-ENTRY.NortonMapeo.v1

**Fig. 4.** Query One: Results

and property selected graphically is identified in the ontology by means of its URI and are constrained by means of FILTER clauses. Figure 4 shows the list of results for this query. This list contains the identifiers of the archetype concepts written in Spanish that are stored in our semantic archetype repository. The interface allows for navigating the results and view their properties.

#### 4.2 Finding the Unit in Which the Systolic Blood Pressure Is Measured and Its Allowed Values

Table 3 shows the representation of the query after the selection of the concepts and properties in the user interface. Now, “Search for” contains three concepts: *ELEMENT*, to which the systolic blood pressure belongs, its *units* and its *value*, which are, respectively, a string and a real number. “Query requirements” define the properties that must meet the results. The resulting *ELEMENT* must contain “Systolic blood pressure” as *TERM DEFINITION* value. Then, we will retrieve its *value*, which is of *PHYSICAL QUANTITY* type and has the properties *units* and *value\_real*. The transformed SPARQL query is shown in Table 4.

Finally, the query results are shown in Figure 5. It includes the properties of the retrieved instances. The result depicted on the left is the units in which

**Table 3.** Query Two: Find the unit in which the systolic blood pressure is measured and its allowed values

Search for:
ELEMENT[0]
String Type[1]
Real Type[2]
Query requirements:
ELEMENT[0]→term_definitions→TERM DEFINITION[3]
TERM DEFINITION[3]→text→Systolic blood pressure
ELEMENT[0]→element_value→PHYSICAL QUANTITY[4]
PHYSICAL QUANTITY[4]→units→CODED SIMPLE VALUE[5]
CODED SIMPLE VALUE[5]→code_value→String Type[1]
PHYSICAL QUANTITY[4]→value_real→Real Type[2]

**Table 4.** Query 2 in SPARQL

```
@prefix ar: <http://klt.inf.um.es/~cati/ontologies/test/ISO13606-AR-v2.1.owl>.
SELECT
?ELEMENT_0
?String_Type_1
?Real_Type_2
WHERE {
?ELEMENT_0 ar:term_definitions ?TERM_DEFINITION_3 .
?TERM_DEFINITION_3 ar:text ?literal_4 .
FILTER (regex(?literal_4, "Systolic blood pressure")) .
?ELEMENT_0 ar:element_value ?PHYSICAL_QUANTITY_5 .
?PHYSICAL_QUANTITY_5 ar:units ?CODED_SIMPLE_VALUE_7 .
?CODED_SIMPLE_VALUE_7 ar:code_value ?String_Type_1 .
?PHYSICAL_QUANTITY_5 ar:value_real ?Real_Type_2 .
}
```

String Type	IntervalOrReal		
any_allowed	false	upper_bound	1000.0
list_open	false	any_allowed	false
list	mm[Hg]	lower_unbounded	false
		upper_included	false
		upper_unbounded	false
		lower_included	true
		lower_bound	0.0

**Fig. 5.** Query Two: Results

the systolic blood pressure is measured: mm[Hg]. The right one is the range of allowed values: 0 (*lower\_bound*) and 1000 (*upper\_bound*).

## 5 Conclusions

Providing easy and flexible mechanisms for querying biomedical semantic repositories is fundamental for its success. To date, many efforts have been put on the

semantic representation and integration of biomedical data, so very interesting semantic repositories have been generated. However, most of the currently available systems have problems for facilitating the definition of expressive, semantic queries. This is due to the fact that they require knowing complex languages such as SPARQL or limit the expressiveness of the queries, as it happens with most CNLs.

One of our research goals is to facilitate biomedical researchers the exploitation of semantic repositories, so flexible and easier to use query mechanisms are required. In this work, we present our initial results, that move the complexity from knowing SPARQL to working with a domain ontology. In fact, using a SPARQL-based query interface does not only require expertise in SPARQL but also expertise in the underlying ontology. Our system removes one of those two requirements for making semantic queries. However, the users still need to know about the ontology and the structure of the individuals they are searching for. Our impression is that the current interface is better for users than the ones based on SPARQL, although we need to conduct some experiments to obtain user satisfaction results that would support this claim.

We are currently working on making the query interface even simpler. Ideally, we would like to allow users to define semantic queries without noticing that they are using an ontology, by reducing to number of clicks and selections for adding a particular ontological entity or requirement into the query.

## Acknowledgments

This work has been possible thanks to the Spanish Ministry for Science and Education through the grant TSI2007-66575-C02-02. JA Miñarro is supported by the Fundación Séneca and the Servicio de Empleo y Formación through the grant 07836/BPS/07.

## References

1. <http://www.obofoundry.org/>
2. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. *Scientific American* 284(5), 34–43 (2001)
3. features: Owl web ontology language, <http://www.w3.org/TR/owl>
4. World wide web consortium, <http://www.w3c.org>
5. Bodenreider, O., Sahoo, S.S.: Semantic web for translational biomedicine: Two pilot experiments. *Proceedings of the AMIA Summit on Translational Bioinformatics*, 148 (2008)
6. Resource description framework (rdf), <http://www.w3.org/RDF>
7. Sparql query language for rdf, <http://www.w3.org/TR/rdf-sparqlquery>
8. Antezana, E., Blondé, W., Egaña, M., Rutherford, A., Stevens, R., De Baets, B., Mironov, V., Kuiper, M.: Biogateway: a semantic systems biology tool for the life sciences. *BMC Bioinformatics* 10(Suppl. 10), S11 (2009)
9. Bio2rdf, <http://bio2rdf.org/>

10. Bernstein, A., Kaufmann, E.: Gino - a guided input natural language ontology editor. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 144–157. Springer, Heidelberg (2006)
11. Wang, C., Xiong, M., Zhou, Q., Yu, Y.: Panto: A portable natural language interface to ontologies. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 473–487. Springer, Heidelberg (2007)
12. Kaufmann, E., Bernstein, A.: How useful are natural language interfaces to the semantic web for casual end-users? In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 281–294. Springer, Heidelberg (2007)
13. Schwitter, R.: A controlled natural language layer for the semantic web. In: Zhang, S., Jarvis, R.A. (eds.) AI 2005. LNCS (LNAI), vol. 3809, pp. 425–434. Springer, Heidelberg (2005)
14. Smart, P.R.: Controlled natural languages and the semantic web. Technical report, School of Electronics and Computer Science, University of Southampton., Technical Report ITA/P12/SemWebCNL (2008)
15. Namgoong, H., Kim, H.G.: Ontology-based controlled natural language editor using cfg with lexical dependency. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 351–364. Springer, Heidelberg (2007)
16. Jiménez-Ruiz, E., Llavori, R.B., Nebot, V., Sanz, I.: Ontopath: A language for retrieving ontology fragments. In: Meersman, R., Tari, Z. (eds.) OTM 2007, Part I. LNCS, vol. 4803, pp. 897–914. Springer, Heidelberg (2007)
17. Chirlaque, J.L., Valencia-García, R., García-Sánchez, F., Castellanos-Nieves, D., Fernández-Breis, J.T.: Owl-path: an ontology-guided natural language query editor. In: Eurocast (2009)
18. <http://www.iso.org/iso/home.htm>
19. <http://www.openehr.org>
20. Martínez-Costa, C., Menárguez-Tortosa, M., Fernández-Breis, J.T., Maldonado, J.A.: A model-driven approach for representing clinical archetypes for semantic web environments. Journal of Biomedical Informatics 42(1), 150–164 (2009)
21. Fernández Breis, J., Menárguez Tortosa, M., Martínez Costa, C., Fernández Breis, E., Herrero Sempere, J., Moner Cano, D., Sánchez Cuadrado, J., Valencia García, R., Robles Viejo, M.: A semantic web-based system for managing clinical archetypes. In: 30th Annual International IEEE EMBS Conference, pp. 1482–1485 (2008)
22. Fernández Breis, J., Menárguez Tortosa, M., Vivancos Vicente, P., Valencia García, R., Moner Cano, D., Maldonado, J.: Poseacle: An ontological infrastructure for managing clinical archetypes in semantic web environments. Medinfo. (2007)

# A Formal Domain Model for Dietary and Physical Activity Counseling

Erzsébet Mák<sup>2</sup>, Balázs Pintér<sup>1</sup>, Balázs Gaál<sup>1,\*</sup>, István Vassányi<sup>1</sup>,  
György Kozmann<sup>1</sup>, and Istvánné Németh<sup>2</sup>

<sup>1</sup> University of Pannonia, Department of Electrical Engineering and Information Systems, Veszprém, Hungary  
[bgaal@almos.vein.hu](mailto:bgaal@almos.vein.hu)

<sup>2</sup> Semmelweis University, Department of Dietetics and Nutrition Sciences, Budapest, Hungary

**Abstract.** Diet and physical activity planning is a complex process that usually comprises repetitive expert-patient discussions and multi-hour construction phases. Recent advances in artificial intelligence and improvements in CPU speeds make it now possible to enhance or even substitute the work of the dietary expert. Although research in this field began as early as the 1940s, no comprehensive domain model has been developed to date. Previous works reduced the problem to then solvable mathematical models, thus lessening the quality of the solution. Here, we present a novel domain model which can handle the multi-objective nature of the problem as well as the proper use of expert knowledge on dietary harmony. The model provides a base for the computerized planning of human-competitive solutions. An implementation of this model is employed in the nutrition and lifestyle counseling expert system *Menu-gene*.

**Keywords:** dietary menu planning, physical activity planning, domain model, artificial intelligence.

## 1 Introduction

Any computerized method which deals with the solution of a real-world problem has an understanding, a representation of the problem, also called a domain model. The detail and expressivity of the model highly determines the quality of the whole method, regardless of the algorithm that it runs. Diet and physical activity plans are made up of instructions which are to be carried out in a given time or time-frame (e.g.: ingest 2dl tomato soup at noon, perform one hundred push ups during the afternoon etc.). They have a lot of structural freedom, which makes them hard to formalize without imposing constraints on the expressivity of the model. Diet and physical activity plans share many similarities and can be handled with the same mathematical models, with the construction of

---

\* Corresponding author.

personalized diet plans being the more complex problem. Probably this is why computerized diet counseling is the more researched field of the two.

The problem of mathematically optimizing dietary plans appeared as early as the 1940s. In fact, one of the earliest applications of the famous simplex method [1] was the determination of an adequate diet that was of least cost, formed as a linear-programming problem [1]. Other random search and mathematical programming based methods appeared from the 1960s [2][3][4]. Later, more sophisticated artificial intelligence methods were developed mostly using Case-Based Reasoning (CBR) or Rule-Based Reasoning (RBR) or combining these two with other techniques [5][6]. A hybrid CBR-RBR system, CAMPER [7] integrates the advantages of the two independent implementations: the case-based menu planner, CAMP [8] and PRISM [9]. Whether or not it was stated explicitly, these methods used traditional mathematical programming models for inner representation, mostly the general Mixed Integer Linear Programming (MILP) model. Only the more recent multi-level Genetic Algorithm (GA) based approaches employed advanced problem representations. In [10], we used a hierarchically interconnected parallel evolution of MILPs, while in [11], Seljak used a sequential solution of connected MILPs. These works focused on practical solution methods, rather than on the exact definition of the problem.

In this paper, we describe a formal, hierarchical model for the menu/exercise planning problem. The novelty of the model is the multi-level structure and the uniform treatment of numerical and harmony assessment. The objectives are defined in Sect. 2. Then, in Sect. 3 we describe the domain model by iteratively refining a simple, single-objective model.

## 2 Objectives of Diet and Physical Activity Planning

Everyone has a common sense understanding of how a weekly diet or physical activity plan may look like, what it may contain, etc. We define the plans as hierarchical structures. The concepts Nutritional Structure (NS) and Activity Structure (AS) will denote a set of interdependent objects making up the plans. Examples of NSs and ASs objects are 'a daily training schedule', a 'vegetarian dinner', 'a glass of orange juice' etc. These objects belong to a certain domain level. Although the number of levels is just a parameter of the model, in the dietary implementation of the model we normally use the Nutritional Levels (NLs) of nutrients, foodstuffs, dishes, meals, daily menu plans, and weekly menu plans. Each Nutrition Hierarchy Object (NHO) of the NS belongs to one of these levels. The real-world meaning of NHOs are presented in Table 1. Those NHOs which represent plans are also termed Dietary Menu Plans (DMPs). In practice, a Dietary Menu Plan (DMP) is basically a list of dishes and their serving times and portions for a given time period, filling the slots of a menu pattern. A slot in a DMP can be either left empty, or filled with a NHO. Note that the duration of a DMP is in correspondance with its level i.e. a single meal, a daily plan or a weekly plan.

The structure of Physical Activity Plans (AS) is quite similar to that of NS. Activity Level (AL) is analogous to NL. We normally use the levels of *effect*,

**Table 1.** Real-world meaning of Nutrition Hierarchy Objects (NHOs) from corresponding NLs

NL	real-world meaning of the NHO
Week	menu plan for a week
Day	menu plan for a day
Meal	menu plan for a single meal
Dish	record of a recipe database
Foodstuff	record of a food composition database
Nutrient	column of a food composition table

*routine, exercise, day* and *week*. Note that the term *level* refers to some position in the hierarchy and not to the rate of exercise intensity. Each Activity Hierarchy Object (AHO) belongs to one of these levels. The real world meaning of AHOs is shown in Table 2.

**Table 2.** Real-world meaning of Activity Hierarchy Objects for the ALs

AL	real-world meaning of the AHO
Week	collection of exercises for a week
Day	collection of exercises for a day
Exercise	collection of exercise routines
Routine	a single exercise routine
Effect	amount of calorie burn, chemical increase, time taken

The Activity Timetable Plan (ATP) is analogous to the DMP. It is a list of routines performed at some point during the ATP. ATPs have slots to represent lower-level AHOs, just as this was the case in the nutrition domain with NHOs. An exercise consists of one or more routines, which have effects on the performer (energy burn, increasing dopamine, serotonin, etc.). The sum of the effects of the exercises should be close to the personal optima given by physical activity guidelines [12].

A DMP and an ATP, and any NHO and AHO as well, may have a virtually unlimited number of *attributes*, which influence their assessment. These attributes can be *quantifiable* (numerical) or *non-quantifiable* (nominal). For example, the amount of protein in an NHO is quantifiable and can be expressed with a numerical value, but the seasonality, with possible values of 'spring time dish', 'summer time dish', etc., is non-quantifiable. Non-quantifiable attributes can be expressed as numerical codes, but, contrary to quantifiable ones, these numbers do not cumulate when NHOs are combined. The NHOs and AHOs are assessed by rating their attributes and the attributes of the objects in their slots. Thus each NHO that a DMP contains influences the overall goodness of the DMP, because it contributes to the worth of upper-level NHOs it is built into. The same holds true for AHOs and ATPs in ASs.

Basically, the aim of the menu planning process is to have those NHOs selected for the solution, which collectively satisfy the expectations of the patient. Due to the analogy of DMPs and ATPs, the computer-based generation of these plans

can be carried out with the same approach. In the following, we will elaborate the formal definition of the DMP by iteratively developing an initial model.

### 3 The Proposed Model

We define four DMP types as follows. We start with the simplest, single-objective, single-level version. Then an extension which handles multiple objectives is given. After that, the problem is extended to multiple levels, which is an original contribution of this paper, as no multi-level model has been presented for the diet problem previously. In the final extension, the handling of expert knowledge and harmony is introduced, which has not been formalized before either.

**Simple Dietary Menu Planning Problem (S-DMPP).** An assignment  $\mathbf{A} := (a_{i,j})_{m \times n}$  of  $m$  objects to  $n$  slots is sought, where  $a_{i,j}$  is the number of instances of the  $i^{th}$  object assigned to the  $j^{th}$  slot. The *satisfiability* part of the S-DMPP problem implies:

$$c_{min} \leq \sum_{i=1}^m \sum_{j=1}^n (p_i \cdot a_{i,j}) \leq c_{max} \quad (1)$$

where  $p_i$  is the payload of the  $i^{th}$  object and  $c_{min}$  and  $c_{max}$  are the lower and upper constraints on the single objective of the S-DMPP. For example, the payload can be the energy or fiber content of an object. In many cases, lower and upper constraints are quite wide, with an optimal value for the payload somewhere in between. We use  $c_{opt}$  to represent this optimal value and state the *optimization* part of S-DMPP as a minimization problem (Equation (2)). Note that for  $c_{opt} : c_{min} \leq c_{opt} \leq c_{max}$ .

$$\underset{\mathbf{c}_{opt}}{\text{minimize}} \quad \left| c_{opt} - \sum_{i=1}^m \sum_{j=1}^n (p_i \cdot a_{i,j}) \right| \quad (2)$$

We define four types of S-DMPPs depending on how the assignments of objects to slots are restricted. Objects may belong to only one, or more than one slot, and slots may hold only one, or more than one instance of an object. The four subversions of the S-DMPP only differ in the permitted values of matrix  $\mathbf{A}$ .

**Multi-Objective Dietary Menu Planning Problem (MO-DMPP).** In the dietary practice, there are normally more than one aspects or objectives for a DMP to satisfy simultaneously. The number of theoretically possible objectives can even surpass one thousand, considering that the number of nutrition components is more than two hundred in a comprehensive Food Composition Database (FCDB), and yet there are other goals.

In this case, an assignment  $\mathbf{A} := (a_{i,j})_{m \times n}$  of objects to slots is sought, with

$$\mathbf{c}_{\min}^T \leq \sum_{i=1}^m \sum_{j=1}^n ([\mathbf{P}^T]_i \cdot a_{i,j}) \leq \mathbf{c}_{\max}^T \quad (3)$$

where  $\mathbf{P} := (p_{i,k})_{m \times o}$ , and  $p_{i,k}$  is the  $k^{th}$  payload of the  $i^{th}$  object, and  $\mathbf{c}_{\min}^T$  and  $\mathbf{c}_{\max}^T$  are column vectors, with each pair of their values in the same row representing the minimal and maximal constraint for a given payload.

Note that Equation (3) may be hard or impossible to satisfy considering the variety of recommended dietary intakes [13]. For this reason, we define the column vector  $\mathbf{c}_{\text{opt}}^T$  and the counterpart of Equation (2), omitted here for brevity. Thus, our goal is to minimize the deviation of the payloads from the optimum, even if the lower and upper constraints cannot be satisfied i.e. the solution is not feasible. Basically, Equation (3) presents the MILP model which has been the de facto inner representation of the diet problem in all of the early solvers like [3,9]. While this model is sufficient for planning DMPs with slots for a meal, it cannot represent the hierarchical structure of longer duration NSs.

**Hierarchical Multi-Objective Dietary Menu Planning Problem (HMO-DMPP).** We consider a DMP which describes a real-world menu plan (for example for a week), containing sublevel DMPs (plans for a day, plans for meals, etc.) in a hierarchical structure. We extend the MO-DMPP model to handle assignment matrices arranged in a multi-level and hierarchical structure.  $[\mathbf{A}_{l+1}]$  is an assignment of objects to slots considering an  $(l + 1)$ -level problem.  $[\mathbf{P}_{l+1}]$  is the payload matrix of this  $(l + 1)$ -level problem.  $\mathbf{A}_{l,u} := (a_{l(i,j)})_{m_{(l,u)} \times n_{(l,u)}}$  assignment matrix of level  $l$  objects and slots for the  $u^{th}$  subproblem of  $\mathbf{A}_{l+1}$  with the payload matrix  $[\mathbf{P}(\mathbf{A}_{l,u})]$ .

The assignments of the subproblems needn't be of the same dimension. Neither the number of slots, nor the number of objects need to be equal for the subproblems. This also holds for the payload vectors.

As the level  $l$  matrices build up the  $[\mathbf{A}_{l+1}]$  assignment matrix,  $f_{l,u}$  functions are needed which transform the number of  $r$  subproblems' assignment matrices height to match the height of  $[\mathbf{A}_{l+1}]$ . Basically the  $f_{l,u}$  functions, which are unique for each subproblem, are for adding objects which were not present in the assignment matrix but were at least in one of the other subproblems' matrices. For example, if a DMP for a day is considered, the objects for soups only appear in the meal level matrices of lunch and dinner, because by common sense, they are not for serving elsewhere. However, when the meal level matrices are used to form the daily level matrix, the objects for soups are added to the other meal level matrices too (breakfast, morning snack, evening snack) through the  $f$  functions. Note that these newly added objects would not be assigned to any slots in the modified  $f(\mathbf{A})$  matrices. Table 3 below summarizes the notations and variables.

$$[\mathbf{A}_{l+1}] = [f_{l,1}([\mathbf{A}_{l,1}]) \dots f_{l,r}([\mathbf{A}_{l,r}])] , 1 \leq l \leq q \quad (4)$$

We formulate the HMO-DMPP as follows, where  $q$  is the number of levels of the problem, and  $\mathbf{A}_q$  is the uppermost assignment matrix, i.e. it is not a subproblem of any other problem.  $[\mathbf{A}_q]$  must be evaluated recursively. The meaning of the variables and expressions are presented in Tables 3 and 4.

$$[\mathbf{A}_q] = [f_{q-1,1}([\mathbf{A}_{q-1,1}]) \dots f_{q-1,r}([\mathbf{A}_{q-1,r}(\mathbf{A}_q)])] \quad (5a)$$

**Table 3.** Meaning of the variables used to formalize the HMO-DMPP

Variable	Index	Variable(s)	Meaning
m	i		number of objects
n	j		number of slots
o	k		number of payloads
q	h,l		number of levels
r	u,v		number of subproblems

$$\mathbf{c}_{\min}^T(\mathbf{A}_q) \leq \sum_{i=1}^{m(\mathbf{A}_q)} \sum_{j=1}^{n(\mathbf{A}_q)} ([\mathbf{P}(\mathbf{A}_q)^T]_i \cdot a_{i,j}) \leq \mathbf{c}_{\max}^T(\mathbf{A}_q) \quad (5b)$$

**Algorithm 1. recursively-evaluate-problem([A<sub>h,v</sub>])**


---

```

for all  $l, u$  such that  $\exists \mathbf{A}_{l,u}^{h,v}$  do
    evaluate  $\mathbf{c}_{\min}^T(\mathbf{A}_{l,u}^{h,v}) \leq \sum_{i=1}^{m(\mathbf{A}_{l,u}^{h,v})} \sum_{j=1}^{n(\mathbf{A}_{l,u}^{h,v})} ([\mathbf{P}(\mathbf{A}_{l,u}^{h,v})^T]_i \cdot a_{i,j}) \leq \mathbf{c}_{\max}^T(\mathbf{A}_{l,u}^{h,v})$ 
    recursively-evaluate-problem([Al,uh,v])
end for

```

---

**Introducing Expert Knowledge and Harmony to the model.** From the sole point of view of ingredient quantification, the HMO-DMPP model formalizes the problem. If an object is assigned to a slot, its payloads take account in the summation and in the assessment of the assignment, which in turn is a DMP.

Taking the example of a meal plan and an assignment of a food object to a meal plan slot, the USDA SR21 FCDB object **Apple, raw, with skin** with medium size is assigned to the slot dessert. It is easily calculable from the data of the FCDB that, for example, how much energy (397kJ) and protein (0.47g) this object adds to the DMP.

This assignment holds implicit information for the human expert, therefore this knowledge has to be expressed by some means in the model. The **Apple, raw, with skin** with medium size is an apple, it is also a fruit, and also a dessert made of fruit. This is the taxonomic information of the component. This assignment also makes the DMP one which contains fruit, one which contains apple, and one which has fruit for dessert. Moreover, if this meal plan does not contain any meat, it belongs to the class of vegetarian meals with fruit, which information is derived from the combination of the meal's components. We use the function denoted with  $\rho$  (standing for harmony) to take an assignment matrix as parameter and transform it to a column vector with real values, with each row representing the value of a harmony payload. Each of the harmony payloads represents one of the following concepts:

**Taxonomy information** is inferred from the assignment matrix of each DMP and represents implicit and unquantifiable knowledge about that DMP.

**Table 4.** Meaning of the expressions used to formalize the HMO-DMPP

Expression	Meaning
$[\mathbf{A}_{l,u}^{h,v}]$	The assignment matrix $A$ represents a solution for a level $l$ problem, which is the $u^{th}$ subproblem of its level $h$ parent, which is the $v^{th}$ subproblem of its parent.
$[\mathbf{A}_{l,u}^h]$	The assignment matrix $A$ represents a solution for a level $l$ problem, which is the $u^{th}$ subproblem of its level $h$ parent. The same as the above one, but without denoting whether the level $h$ parent problem has a parent or not.
$[\mathbf{A}_{l,u}]$	The assignment matrix $A$ represents a solution for a level $l$ problem, which is the $u^{th}$ subproblem of its parent. The same as the above one, but without denoting the level of the parent.
$[\mathbf{A}_l]$	The assignment matrix $A$ represents a solution for a level $l$ problem. The same as the above one, but without denoting whether the problem is a subproblem
$m(\mathbf{A}_{l,u}^{h,v})$	the height (the number of rows) of the $[\mathbf{A}_{l,u}^{h,v}]$ matrix, namely the number of objects
$n(\mathbf{A}_{l,u}^{h,v})$	the width (the number of columns) of the $[\mathbf{A}_{l,u}^{h,v}]$ matrix, namely the number of slots

In the harmony vector, taxonomy information is represented with the values 0 or 1, 1 meaning that the DMP is a member of the taxonomy class.

**Combination information** for a DMP is inferred from the assignment matrices of the sublevel DMPs and represents the unquantifiable information the combination of the sublevel DMPs bear. In the harmony vector, taxonomy information is represented with the values 0 or 1.

**Combination harmony value** is calculated similarly as the combination information but results in a numerical value rating the harmony of the combination of the components making up the DMP (represented in Multi-Level Genetic Algorithm (muleGA) through objective values). In the harmony vector, taxonomy information is represented with real values.

For example, if the assignment matrix of an arbitrary DMP has 2 pieces of **Apple, raw, with skin** assigned to one slot, and 3 pieces of **Pear, raw, with skin** to another slot, and no other fruits and no meat at all, then the harmony vector will have the value 5 for the harmony payload entitled fruits (taxonomy information). If the DMP is for a lunch, then the harmony vector will have the value 1 for the harmony payload called “vegetarian lunch with fruit”, and 0 for the harmony payload called “dinner with fruit” (combination information).

The example presented the simplest case, when the NHO **Apple, raw, with skin** was selected to fill a specific slot in the DMP. In this case, the apple was a simple dish, which had only one ingredient, one apple. However, DMP slots usually assigned with NHOs representing dishes with more ingredients. Take the *soup* slot of a DMP for meal as an example, and assign *chicken soup* to it. The payload of this object can be calculated by summing the data of each component of the recipe provided by the Food Composition Database (FCDB).

Besides taxonomy and combination information, the harmony vector has to express the harmony of the objects assigned to slots of the same DMP through the *combination harmony values*. The harmony payloads for representing this kind of information could be called “harmony of the assignment”, “harmony of fruits in the assignment” and so on. A harmony payload will get zero value if there is no information, for example, on the harmony of fruits. Positive value will mean harmony, with a bigger value meaning more harmonizing components, while negative values are for representing disharmony. For example, it is not appetizing to serve tomato juice as a drink for a tomato based main dish. For each assignment matrix, the calculation of the harmony constraints (taxonomy information, combination information, combination harmony values) is done according to Equation (6).

$$\mathbf{h}_{\min}^T(A_{l,u}^{h,v}) \leq \rho(A_{l,u}^{h,v}) \leq \mathbf{h}_{\max}^T(A_{l,u}^{h,v}) \quad (6)$$

To express harmony in the model, the calculation of the harmony vector according to Equation (6) should be added to the HMO-DMPP model, namely to Equation (5) and to Algorithm 1.

Note that the complex calculation of the harmony vector is hidden in the function  $\rho$ . The computational complexity of the calculation of  $\rho$  is exponential in time in the function of the number of the slots ( $n$ ), because each subset of the objects associated with slots should get evaluated according to harmony. Not including the empty set, the number of subsets is  $2^n - 1$ , and this many checks are needed to calculate the value of each harmony payload. This makes the proof of the decision problem ‘whether the assignment’s payloads are within the constraints’ verifiable in exponential time. For NP complexity the proof would have to be verifiable in polynomial time by a deterministic Turing machine. Therefore, the introduction of harmony makes the decision problem harder than NP, a good candidate for evolutionary solving methods. Taxonomic and combinatorial data effectively extend the MILP model to express harmony. A similar extension of MILPs with first-order logic is presented in [14].

**Hierarchical Multi-Objective Dietary Menu Planning Problem with Harmony (HMO-DMPP-H).** The HMO-DMPP-H model extends the previous HMO-DMPP model with the concept of the harmony function, which is represented by  $\rho$ . The actual calculation of the harmony payloads, which can involve description logic inferencing and strenuous numerical calculations, is hidden behind the function  $\rho$ . The HMO-DMPP-H problem is formulated as follows, where  $q$  is the number of levels the problem has, and  $\mathbf{A}_q$  is the uppermost assignment matrix.  $[\mathbf{A}_q]$  must be evaluated recursively.

$$[\mathbf{A}_q] = [f_{q-1,1}([\mathbf{A}_{q-1,1}]) \dots f_{q-1,r(\mathbf{A}_q)}([\mathbf{A}_{q-1,r(\mathbf{A}_q)}])] \quad (7a)$$

$$\mathbf{c}_{\min}^T(A_q) \leq \sum_{i=1}^{m(\mathbf{A}_q)} \sum_{j=1}^{n(\mathbf{A}_q)} ([\mathbf{P}(\mathbf{A}_q)^T]_i \cdot a_{i,j}) \leq \mathbf{c}_{\max}^T(\mathbf{A}_q) \quad (7b)$$

**Algorithm 2.** recursively-evaluate-problem-with-harmony([ $\mathbf{A}_{h,v}$ ])

---

```

for all  $l, u$  such that  $\exists \mathbf{A}_{l,u}^{h,v}$  do
    evaluate  $\mathbf{c}_{\min}^T(\mathbf{A}_{l,u}^{h,v}) \leq \sum_{i=1}^{m(\mathbf{A}_{l,u}^{h,v})} \sum_{j=1}^{n(\mathbf{A}_{l,u}^{h,v})} ([\mathbf{P}(\mathbf{A}_{l,u}^{h,v})^T]_i \cdot a_{i,j}) \leq \mathbf{c}_{\max}^T(\mathbf{A}_{l,u}^{h,v})$ 
    evaluate  $\mathbf{h}_{\min}^T(\mathbf{A}_{l,u}^{h,v}) \leq \rho(\mathbf{A}_{l,u}^{h,v}) \leq \mathbf{h}_{\max}^T(\mathbf{A}_{l,u}^{h,v})$ 
    recursively-evaluate-problem-with-harmony([ $\mathbf{A}_{l,u}^{h,v}$ ])
end for

```

---

$$\mathbf{h}_{\min}^T(\mathbf{A}_q) \leq \rho(\mathbf{A}_q) \leq \mathbf{h}_{\max}^T(\mathbf{A}_q) \quad (7c)$$

**Implementation in MenuGene.** Using this model, we developed a Multi-Level Genetic Algorithm (muleGA) based solution method [15]. The muleGA handles and evolves all the assignment matrices in the hierarchy in parallel, and continuously tries to improve the object-to-slot assignments in the NS (or AS) to optimize the plans according to the quantitative and qualitative requirements. The calculation of the harmony vector is based on expert assigned weights on the co-occurrence of certain food and dish sets. Our results show that strict numerical constraints on any nutritional level can be satisfied by this GA-based method, and that harmony scores can be efficiently used to control the search process.

## 4 Conclusion

The HMO-DMPP-H model presented in the paper extends and improves previous representations. The two new feature are the recursive definition of the hierarchical structure of the problem, and the harmony vector encoding taxonomy and combination related information. Neither the multi-level hierarchy, nor the attributes of the harmony vector have been explicitly formalized previously. These extensions allow the formalization of any diet and physical activity planning problem and provide a base for comparing solution methods according to their completeness and specificity. The work presented was partially funded by the Hungarian National Research project No. OM-00191/2008 (AALAMSRK).

## References

1. Dantzig, G.B.: Linear programming and extensions. Princeton University Press, Princeton (1963)
2. Balintfy, J.L.: Menu planning by computer. Commun. ACM 7(4), 255–259 (1964)
3. Eckstein, E.F.: Menu planning by computer: the random approach. Journal of American Dietetic Association 51(6), 529–533 (1967)
4. Sklan, D., Dariel, I.: Diet planning for humans using mixed-integer linear programming. British Journal of Nutrition 70(1), 27–35 (1993)

5. Yang, N.: An expert system on menu planning. master's thesis, Department of Computer Engineering and Science Case Western Reserve University, Cleveland, OH (1989)
6. Hinrichs, T.R.: Problem-solving in open worlds: a case study in design. PhD thesis, Georgia Institute of Technology, Atlanta, GA, USA (1992)
7. Marling, C.R., petot, G.J., Sterling, L.S.: Integrating case-based and rule-based reasoning to meet multiple design constraints. Computational Intelligence 15(3), 308–332 (1999)
8. Marling, C.R., Petot, G., Sterling, L.: Planning nutritional menus using case-based reasoning. In: Working Notes of the AAAI Spring Symposium on Artificial Intelligence in Medicine, pp. 109–113 (1996)
9. Kovacic, K.J.: Using common-sense knowledge for computer menu planning. PhD thesis, Cleveland, Ohio: Case Western Reserve University (1995)
10. Gaál, B., Vassányi, I., Kozmann, G.: A novel artificial intelligence method for weekly dietary menu planning. Methods Inf. Med. 44(5), 655–664 (2005)
11. Seljak, B.K.: Dietary Menu Planning Using an Evolutionary Method. In: Proceedings of Intelligent Engineering Systems, INES 2006 (2006)
12. William, L., Haskell, I.-M., Lee, R.R., Pate, K.E., Powell, S.N., Blair, B.A., Franklin, C.A., Macera, G.W., Heath, P.D., Thompson, Bauman, A.: Physical Activity and Public Health: Updated Recommendation for Adults From the American College of Sports Medicine and the American Heart Association. Circulation 116(9), 1081–1093 (2007)
13. Dollahite, J., Franklin, D., McNew, R.: Problems encountered in meeting the recommended dietary allowances for menus designed according to the dietary guidelines for americans. J. Am. Diet. Assoc. 95, 341–347 (1995)
14. Gordon, G.J., Hong, S.A., Dudík, M.: First-order mixed integer linear programming. In: Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, UAI (2009)
15. Gaál, B.: Multi-Level Genetic Algorithms and Expert System for Health Promotion. PhD thesis, Department of Electrical Engineering and Information Systems, Faculty of Information Technology, Information Science and Technology PhD School of University of Pannonia, Veszprém, Hungary (December 2009)

# An Ontology Based Approach to Measuring the Semantic Similarity between Information Objects in Personal Information Collections

Lei Shi and Rossitza Setchi

School of Engineering, Cardiff University, Cardiff CF24 3AA, UK  
[{Leishi,Setchi}@cf.ac.uk](mailto:{Leishi,Setchi}@cf.ac.uk)

**Abstract.** This paper introduces a semantic approach to personal information management, which employs natural language processing, ontologies and a vector space model to measure the semantic similarity between information objects in personal information collections. The approach involves natural language processing, named entity recognition, and information object integration. In particular, natural language processing is used to detect meaningful and semantically distinguishable information objects within collections of personal information. Then, the named entities are extracted from these information objects and their features (such as weight and category) are used to measure the semantic similarity between them. Further research includes using the semantic similarity measure developed to index and retrieve information objects in a semantic based system for personal information management.

**Keywords:** ontology, named entity recognition, semantic similarity, personal information management, information object.

## 1 Introduction

People nowadays are surrounded by digital data such as texts, photographs, audio and video clips related to their work, social life and past. The amount of personal data is ever increasing and its organisation, management and maintenance has become a challenge. Personal Information Management (PIM) as defined in [1] “refers to both the practice and the study of the activities a person performs in order to acquire or create, organize, maintain, retrieve, use and distribute the information needed to complete tasks (work-related and not) and to fulfil various roles and responsibilities (as parent, employee, friend, member of community, etc.)”. The challenge is how to deal with fragmented data, heterogeneous documents, and interconnected themes, people, events, and activities.

This paper contributes to the research in the area of personal information management by proposing a semantic-based approach aimed at measuring the semantic similarity between information objects in personal information collections.

The paper is organised as follows. Section 2 reviews the use of semantic-based technologies in personal information management. Section 3 describes the approach developed which is further illustrated in Section 4 with experimental studies. Section 5 concludes the paper.

## 2 Related Work

The idea of supporting human memory with a personal digital library is often attributed to Vannevar Bush [2] who first described a system called Memex (from memory extender), which would store all his books, records, communications, and experiences.

Until recently PIM research was primarily focused on the organisation of information in databases. However, as highlighted by Kersten et al. [3], personal information cannot be efficiently handled by current database management systems. A new generation of databases is required to handle heterogeneous and distributed personal information. As stated by Abiteboul et al. [4], they will have to include advanced features such as integration of multi-modal data, information fusion, multimedia query, reasoning under uncertainty, personalization, etc. An example in this direction is iMeMex, a personal data space management system used to manage complex personal data spaces without much low-level data management effort [5].

More recent research focuses on the use of semantic technologies in information management and reuse. One of the most prominent developments, the semantic desktop, is an approach that facilitates the reuse and sharing of information in different applications [6]. Information items are treated similarly to semantic web resources: they are identified by URI and queried as RDF graphs. Ontologies are used to link the information items [7]. The NEPOMUK project [8] extends these ideas by developing a collaborative environment, which supports personal information management and sharing and exchange of information across social and organisation barriers. A promising development is the research in semantic information management and retrieval based on semantic associations. For example, SeMex organizes data in a semantically meaningful way by providing a domain model consisting of classes and associations between them [9]. Based on the logical view of personal information, SeMex allows the user to browse information by associations and access information composed on the fly by integrating personal and public data. None of these applications however use semantic similarity measures as a means of grouping similar information objects together.

There are two type of similarity measure related to this paper. The first one introduced by Rada [10] measures the distance between concepts in taxonomies with is-a type of relations. This method assumes that all links between the nodes have the same weight and then computes the distance between any two concepts as the shortest path between them. The main drawback of this method is the assumption that all links are of the same weight. Resnik's approach [11] addresses this deficiency, by proposing a similarity measure based on the use of WordNet and the information content concepts share, and their probability computed using the Brown corpus. Similar semantic similarity measures are also used in applications utilising domain specific ontologies such as Gene Ontology [12, 13].

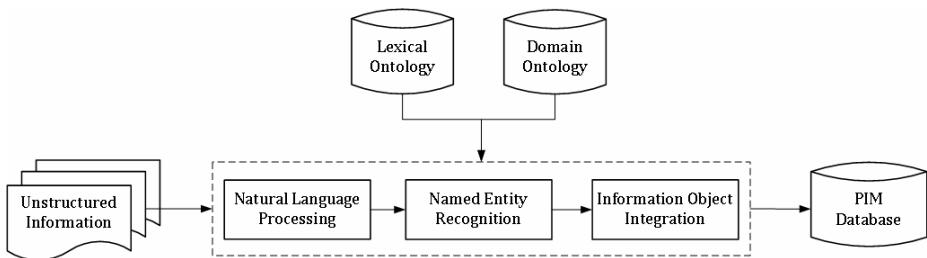
This research focuses on the relations between information objects in personal information collections which mimic complex relationships between people in real life. Such relationships normally have different strength which means that the relations between the nodes in the ontology of named entities of people should have different weight as well.

This paper contributes to the research in this area by proposing an ontology based approach to identifying the semantic relations and measuring the semantic similarity between the information objects in personal collections.

### 3 Semantic Similarity between Information Objects

#### 3.1 Conceptual Model

This research deals with personal information collected throughout the lifetime of a person. Each information item in the personal collection has its textual form. Even audio clips, videos, and photographs are represented as metadata, annotations, or descriptions regardless of their physical or virtual form and despite all possible data formats which might have been used. Furthermore, this paper treats all personal information as a collection of information objects. An information object [14, 15] is a data structure which represents an identifiable, semantically distinguishable, classifiable and meaningful instance of information. In this work, an *information object* is the atom element of personal information that must include at least one named entity, and can be distinguished from or related to other information objects by the named entity included in it (or other semantic features).



**Fig. 1.** Conceptual Model

The conceptual model (Fig.1) includes:

- Natural language processing involves text segmentation and word sense disambiguation; the former splits the information into information objects and the latter is applied to reduce the noise introduced by synonyms and polysemy. A lexical ontology is employed to facilitate the word sense disambiguation.
- Named entity recognition is an automatic way to extract named entities from the information objects using predefined categories such as ‘person’ and ‘location’.
- Information object integration groups similar information objects together based on their semantic similarity. Domain ontologies can be applied to facilitate information integration.

#### 3.2 Named Entity Recognition

Information extraction is based on using natural language processing to automatically extract entities, detect entity relations, and capture concepts of specific domains. Named

entity recognition is an essential subtask of information extraction, which deals with finding entities and classifying them in categories (e.g. organizations, persons, locations, dates, times, monetary values and percentages) [16]. Named entity recognition is considered a robust subtask of information extraction, with high accuracy.

In this research, named entity recognition uses ontologies to extract named entities and classify them using predefined entity categories. Each category is a container of entities with similar semantic attributes. For example ‘location’ includes names of locations and ‘person’ lists persons’ names. Domain ontologies may be applied to improve the accuracy. For example, an ontology of a person’s family and friends can provide essential cues and reduce ambiguity by indicating the type of relations between the persons in the family circle.

In this research, a dataset  $d$  contains several sets of information objects,  $d = \{io_1, io_2, \dots, io_n\}$ . Each set  $\{io_n\}$  is seen as a feature set of  $d$ . Each information object contains one or more named entities, i.e.  $io_n = \{ne_1, ne_2, \dots, ne_i\}$ . Each named entity belongs to a specific category. Assume  $\{c_k\}$  is the set of named entity categories, then  $c_k = \{ne_1^{c_k}, ne_2^{c_k}, \dots, ne_i^{c_k}\}$ . Therefore, an information object  $io_n$  is represented as

$$io_n = \bigcup_{k=1}^n c_k \{ne_1^{c_k}, ne_2^{c_k}, \dots, ne_i^{c_k}\}, \quad (1)$$

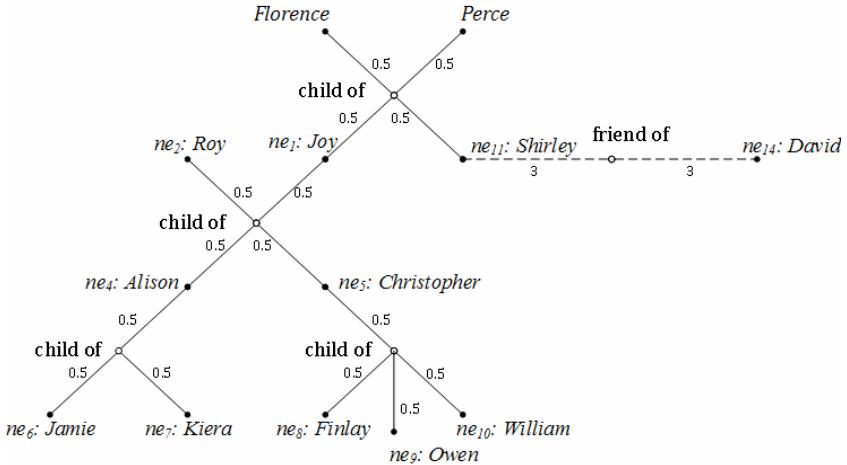
where  $k = \{1, 2, \dots, n\}$  is the number of categories used to group the information objects.

### 3.3 Ontology-Based Similarity Measures

Domain ontologies are used in this research to reduce the semantic ambiguity, structure the named entity categories and detect the semantic links between the information objects. Moreover, ontologies are used to measure the semantic similarity between the named entities using the distance between them.

Fig. 2 shows a family circle represented as an ontology. Let  $dist(ne_i, ne_j)$  is the distance between entities  $ne_i$  and  $ne_j$ ; According to [10], the  $dist(ne_i, ne_j)$  equals to the shortest path between them.

In the figure, the symbol ‘•’ represents a named entity, and it is treated as a node in the ontology. The symbol ‘◦’ denotes a ‘relationship’ that is applied to connect the entities together. Different relationship will be linked by paths with different length (weight). In this example, Joy and Shirley are sisters, Shirley and David are friends. As seen in Fig 2, the distance between each of these pairs is respectively 1 and 6, as “family” (indicated by a solid line) is a stronger link than “friends” (shown as a dashed line). The entity similarity between entities  $ne_i$  and  $ne_j$  in an ontology is defined as follows:



**Fig. 2.** An ontology of a family circle (friends and family)

$$\text{simi}(\text{ne}_i, \text{ne}_j) = [\ln(\text{dist}(\text{ne}_i, \text{ne}_j) + e)]^{-1} \quad (2)$$

onto

This implies that *entity similarity* is monotonically increasing as *distance* decreases. In this ontology, “ne<sub>4</sub>: Alison” is a child of “ne<sub>1</sub>: Joy” and *dist* (“Joy”, “Alison”) equals to 1, thus the *entity similarity* between those two entities according to (2) is 0.76; “ne<sub>6</sub>: Jamie” is a child of “Alison” and *dist* (“Joy”, “Jamie”) equals to 2, then the *entity similarity* between “Joy” and “Jamie” is 0.64. This result shows that “Joy” is closer to “Alison” (her daughter) than “Jamie” (her grandson).

### 3.4 Information Object Integration

The output of the information collection process is a set of information objects with no associations with each other. Information object integration aims at grouping them based on the semantic similarity measures explained in this section.

As mentioned before, the entities set  $\{\text{ne}_i^{c_k}\}$  is a feature set of  $io_j$ . Let  $w_{io_i}(\text{ne}_i^{c_k})$  denotes the *entity weight* of  $\text{ne}_i^{c_k}$ ,  $\text{count}(\text{ne}_i^{c_k})$  is the frequency of  $\text{ne}_i^{c_k}$ , and  $ne$  is the total number of named entities contained in  $io_i$ . A high value of the *entity weight* indicates a greater importance of the entity to the corresponding information object. Formally, the *entity weight* is:

$$w_{io_i}(\text{ne}_i^{c_k}) = \text{count}(\text{ne}_i^{c_k}) \times ne^{-1}. \quad (3)$$

For example, if an information object contains two named entities, “Alison” and “Christopher”, both from the same category, then the weight of these named entity in that information object is  $1 \times 2^{-1} = 0.50$ .

An information object can be represented as follows:

$$io_n = \bigcup_{k=1}^n c_k \{ w_{io_i}(ne_1^{c_k}), w_{io_i}(ne_2^{c_k}), \dots, w_{io_i}(ne_i^{c_k}) \}, \quad (4)$$

where  $k = \{1, 2, \dots, n\}$  is the number of categories used.

Then, the vector space model is applied to measure the similarity between two information objects  $io_i$  and  $io_j$ , based on their common entities all belonging to category  $c_k$ ,

$$\cos simi(io_i^{c_k}, io_j^{c_k}) = \frac{\sum_{n=1}^n [w_{io_i}(ne_n^{c_k}) \times w_{io_j}(ne_n^{c_k})]}{\sqrt{\sum_{n=1}^n w_{io_i}(ne_n^{c_k})^2} \times \sqrt{\sum_{n=1}^n w_{io_j}(ne_n^{c_k})^2}} \quad (5)$$

where  $n = \{1, 2, \dots, n\}$  is the number of entities contained in the information objects  $io_i$  and  $io_j$ .

In this research, entities from the same ontology are called *related entities*. As mentioned in section 3.3, any *related entities* have a *distance* and *entity similarity* based on that *distance*. Assume there are two information objects  $io_i^{c_k}$  and  $io_j^{c_k}$  that contain entities  $ne_m$  and  $ne_n$  respectively, where  $ne_m$  and  $ne_n$  is a pair of *related entities*.

Let  $simi(io_i^{c_k}, io_j^{c_k})$  denotes *related entities based similarity* of two information objects containing *related entities*, then

$$\underset{\text{related } (ne_m, ne_n)}{simi(io_i^{c_k}, io_j^{c_k})} = simi(ne_m, ne_n) \times w_{io_i}(ne_m^{c_k}) \times w_{io_j}(ne_n^{c_k}). \quad (6)$$

If two information objects contain more than one pair of *related entities*, only the closest pair is considered to avoid bias towards objects containing many *related entities*.

Next, the *ontology based similarity measure* for the general case of having named entities belonging to several categories is defined as:

$$\underset{\text{sem}}{simi}(io_i^{c_k}, io_j^{c_k}) = \underset{\text{cos}}{simi}(io_i^{c_k}, io_j^{c_k}) + \underset{\text{related } (ne_m, ne_n)}{simi}(io_i^{c_k}, io_j^{c_k}). \quad (7)$$

Computing the similarity between any two information objects facilitates their grouping. The next section illustrates the approach through an experimental study.

## 4 Experimental Study

This experiment uses a person's life story collection from which eight information objects are selected randomly (Table 1). A family tree has been manually created as an ontology to facilitate the similarity measure and clustering process (Fig.2).

**Table 1.** Information objects used in the experiment

$io_n$	Content
$io_1$	<i>Jackie and me in the playground at the College where Roy first saw me.</i>
$io_2$	<i>A visit to London Zoo with Alison and Christopher.</i>
$io_3$	<i>Christopher's Christening. Photo taken on the front lawn. Notting Hill</i>
$io_4$	<i>The Grandchildren together at the Anniversary: Jamie, Kiera, Finlay, Owen and William.</i>
$io_5$	<i>Roy took this photo of me when he was on leave. He was not home when I wore this dress for Shirley's wedding.</i>
$io_6$	<i>I made Alison's wedding dress without a pattern.</i>
$io_7$	<i>In Mrs Trim's garden when the snow stayed for weeks and weeks. Jim fell in the snow and got frost bite. Cardiff</i>
$io_8$	<i>David and me with Jim by our back door. Penarth</i>

Two named entity categories are utilised in this experiment,  $c_1$ : *Person* and  $c_2$ : *Location*. The result of the named entity recognition is shown in Table 2.

**Table 2.** Named entities and categories

$io_n$	$c_1: Person$	$c_2: Location$
$io_1$	$ne_1: Joy; ne_2: Roy; ne_3: Jackie$	$n/a$
$io_2$	$ne_4: Alison; ne_5: Christopher;$	$ne_1: London;$
$io_3$	$ne_5: Christopher;$	$ne_2: Notting Hill;$
$io_4$	$ne_6: Jamie; ne_7: Kiera; ne_8: Finlay; ne_9: Owen; ne_{10}: William;$	$n/a$
$io_5$	$ne_2: Roy; ne_1: Joy; ne_{11}: Shirley;$	$n/a$
$io_6$	$ne_1: Joy; ne_4: Alison;$	$n/a$
$io_7$	$ne_{12}: Trim; ne_{13}: Jim;$	$ne_3: Cardiff;$
$io_8$	$ne_{14}: David; ne_1: Joy; ne_{13}: Jim;$	$ne_4: Penarth;$

The *entity weight* matrix of the information objects is shown in Table 3. For example, the entity weight of  $ne_6: Jamie$  contained in  $io_4$  is 0.20 because this information object contains four more named entities, thus  $w_{io_4}(ne_6) = 1 \times 5^{-1} = 0.20$ .

Table 4 shows the similarity computed using formula (5) based on the weight of the common entities (i.e. without the ontology). For example, both  $io_1$  and  $io_5$  contain entities “ $ne_1: Joy$ ” and “ $ne_2: Roy$ ”, and they both have *entity weight* 0.33 (see Table 3). Applying formula (5),  $simi(io_1^{c_1}, io_5^{c_1})$  is 0.67, thus  $io_1$  and  $io_5$

are considered relevant in terms of  $ne_1$  and  $ne_2$ . On the other hand, some of the information objects such as  $io_1$  and  $io_4$  have no common entities, and their similarity is 0.

**Table 3.** Entity weight matrix

	$c_1$														$c_2$			
$io_n$	$ne_1$	$ne_2$	$ne_3$	$ne_4$	$ne_5$	$ne_6$	$ne_7$	$ne_8$	$ne_9$	$ne_{10}$	$ne_{11}$	$ne_{12}$	$ne_{13}$	$ne_{14}$	$ne_1$	$ne_2$	$ne_3$	$ne_4$
$io_1$	0.33	0.33	0.33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$io_2$	0	0	0	0.33	0.33	0	0	0	0	0	0	0	0	0	0.33	0	0	0
$io_3$	0	0	0	0	0.50	0	0	0	0	0	0	0	0	0	0	0.50	0	0
$io_4$	0	0	0	0	0	0.20	0.20	0.20	0.20	0.20	0	0	0	0	0	0	0	0
$io_5$	0.33	0.33	0	0	0	0	0	0	0	0	0.33	0	0	0	0	0	0	0
$io_6$	0.50	0	0	0.50	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$io_7$	0	0	0	0	0	0	0	0	0	0	0.33	0.33	0	0	0	0.33	0	0
$io_8$	0.25	0	0	0	0	0	0	0	0	0	0	0	0.25	0.25	0	0	0	0.25

**Table 4.** Similarity measure  $simi(io_i^{c_k}, io_j^{c_k})$

However, based on the family ontology,  $io_1$  and  $io_4$  contain two *related entities*: “ $ne_1$ : Joy” and “ $ne_6$ : Jamie”. Table 5 shows the semantic similarity computed using the weight of both the common and related entities, and formulae (5), (6) and (7). As mentioned above, the similarity of  $io_1$  and  $io_4$  is 0, if the calculation is based on formula (5). Using formula (6),  $simi(io_1^{c_1}, io_4^{c_1})$  is 0.07, which is the semantic similarity  $_{related(ne_1, ne_6)}$

of  $io_1$  and  $io_4$  when their *related entities* are considered.

**Table 5.** Semantic similarity measure  $simi(io_i^{c_k}, io_j^{c_k})$

Furthermore, in an ontology of locations, the *distance* between “ $ne_1$ : London” and “ $ne_2$ : Notting Hill” is 1. Within entity category  $c_2$ : *Location*, the similarity of  $io_2^{c_2}$  and  $io_3^{c_2}$ , i.e.  $simi(io_2^{c_2}, io_3^{c_2})_{related(ne_1, ne_2)}$  is 0.13. The semantic similarity is computed using (7) as follows:

$$simi(io_2^{c_k}, io_3^{c_k})_{sem} = [simi(io_2^{c_1}, io_3^{c_1})_{cos} + simi(io_2^{c_1}, io_3^{c_1})_{related(ne_m, ne_n)}] + [simi(io_2^{c_2}, io_3^{c_2})_{cos} + simi(io_2^{c_2}, io_3^{c_2})_{related(ne_1, ne_2)}] = (0.71 + 0) + (0 + 0.13) = 0.84.$$

The comparison of the data in Table 4 and Table 5 shows that latent semantic associations between information objects are detected based on the relation of the *related entities* in the ontologies. By calculating the semantic similarity of any two information objects, the information objects in the data collection can be clustered or distinguished based on their similarity or dissimilarity. Furthermore, the semantic similarity measure facilitates the understanding of the semantic characteristics of the information objects and their clusters. It provides a way for an improved analysis of their meaning, relations and organisation in personal information collections.

## 5 Conclusion

This paper introduces an approach to measuring the semantic similarity of information objects. This approach considers an information object as a set of classified named entities. The categories of entities are not limited to those mentioned in the paper, but are extendable. More ontologies can be applied to cover different domains. As shown by the experiment results, this approach can detect the latent semantic relations between information objects which do not have common named entities or words. Further work includes integrating this approach into a semantic personal information management system and using it in information object indexing and retrieval.

## References

1. Teevan, J., Jones, W.P., Bederson, B.B.: Personal information management. Communications of the ACM 49(1), 40–43 (2006)
2. Bush, V.: As We Think. The Atlantic Monthly 176(1), 101–108 (1945)
3. Kersten, M., Weikum, G., Franklin, M., Keim, D., Buchmann, A., Chaudhuri, S.: A Database Striptease or How to Manage Your Personal Databases. In: 29th Int. Conf. on Very Large Data Bases, Berlin, Germany, September 9-12 (2003)
4. Abiteboul, S., Agrawal, R., Bernstein, P., Carey, M., Ceri, S., Croft, B., et al.: The Lowell Database Research Self-assessment. Commun. ACM 48(5), 111–118 (2005)
5. Dittrich, J., Salles, M., Karakashian, S.: iMeMex: A Platform for Personal Dataspace Management. In: SIGIR PIM Workshop, Seattle, USA, August 10-11 (2006)

6. The Semantic Desktop as a foundation for PIM research. In: Sauermann, L., Grimnes, G., Roth-Berghofer, T. (eds.) *The Personal Information Management Workshop at ACM Conference on Human Factors in Computing Systems, CHI 2008*, Florence, Italy, April 5-10 (2008)
7. Sauermann, L., Bernardi, A., Dengel, A.: Overview and Outlook on the Semantic Desktop. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) *ISWC 2005*. LNCS, vol. 3729. Springer, Heidelberg (2005)
8. Groza, T., Handschuh, S., Moeller, K., Grimnes, G., Sauermann, L., Minack, E., Mesnage, C., Jazayeri, M., Reif, G., Gudjonsdottir, R.: The NEPOMUK Project - On the way to the Social Semantic Desktop. In: *I-Semantics 2007*, JUCS. pp. 201–211 (2007)
9. Cai, Y., Dong, X.L., Halevy, A., Liu, J.M., Madhavan, J.: Personal Information Management with SEMEX. In: *2005 ACM SIGMOD Int. Conf. on Management of Data*, Baltimore, Maryland, June 13-16 (2005)
10. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and Application of a Metric on Semantic Nets. *IEEE Trans. Systems, Man and Cybernetics* 19(1), 17–30 (1989)
11. Resnik, P.: Semantic Similarity in a taxonomy: An\_Information-Based Measure and its Application to Problems of Ambiguity in NL. *Journal of AI* 11(11), 95–130 (1999)
12. Lord, P., Stevens, R., Brass, A., Goble, C.: Investigating Semantic Similarity Measures across the Gene Ontology: The Relationship between Sequence and Annotation. *Bioinformatics* 19(10), 1275 (2003)
13. Pesquita, C., Faria, D., Falcão, A.O., Lord, P., Couto, F.M.: Semantic Similarity in Biomedical Ontologies. *PLoS Comput Biol.* 5(7), e1000443 (2009)
14. Setchi, R.: Enhanced Product Support through Intelligent Product Manuals, PhD Thesis, Cardiff University (2000)
15. Pham, D.T., Setchi, R.: Authoring Environment for Documentation Development. *IMechE Proceedings* 215(B), 877–882 (2001)
16. Chinchor, N.: Overview of Proceedings of the Seventh Message Understanding Conference (MUC-7)/MET-2, 7th MUC, Fairfax, VA, April 7-9 (1998)

# Ontology Based Graphical Query Language Supporting Recursion

Arun Anand Sadanandan, Kow Weng Onn, and Dickson Lukose

Artificial Intelligence Center, MIMOS Berhad,

Technology Park Malaysia

57000 Kuala Lumpur, Malaysia

{arun.anand, kwonn, dickson.lukose}@mimos.my

**Abstract.** Text based queries often tend to be complex, and may result in non user friendly query structures. However, querying information systems using visual means, even for complex queries has proven to be more efficient and effective as compared to text based queries. This is owing to the fact that visual systems make way for better human-computer communication. This paper introduces an improved query system using a Visual Query Language. The system allows the users to construct query graphs by interacting with the ontology in a user friendly manner. The main purpose of the system is to enable efficient querying on ontologies even by novice users who do not have an in-depth knowledge of internal query structures. The system also supports graphical recursive queries and methods to interpret recursive programs from these visual query graphs. Additionally, we have performed some preliminary usability experiments to test the efficiency and effectiveness of the system.

**Keywords:** Visual Query Languages, Visual Query Systems, Visual Semantic Query, Graphical Recursion, Semantic Web, Ontologies.

## 1 Introduction

Retrieving information by means of querying any RDF<sup>1</sup> or OWL<sup>2</sup> ontology requires a user to have in-depth knowledge in specialized query languages such as SPARQL<sup>3</sup> or PROLOG<sup>4</sup>. However, even in the case of a user being proficient in these languages, it is a challenging task to make complex and large queries due to the complexity of information stored in the knowledge bases. Furthermore, there are times when queries are hard to express in natural language form, especially when querying complex and large knowledge bases.

Visual Query Systems are techniques supporting query formulation using visual means. According to [1] Visual Query Systems are defined as systems for information retrieval using a visual representation to depict the domain of interest and express

---

<sup>1</sup> <http://www.w3.org/TR/REC-rdf-syntax/>

<sup>2</sup> <http://www.w3.org/TR/owl-features/>

<sup>3</sup> <http://www.w3.org/2001/sw/DataAccess/rq23/>

<sup>4</sup> <http://www.franz.com/agraph/support/documentation/current/prolog-tutorial.html>

related requests. These systems allow users to express queries visually and provide tools to allow users to interact with the system. The target users for these systems are typically novices, who are not concerned about the structure of ontology. Having said that, the Visual Query Systems could also benefit expert users, especially when it comes to dealing with complex queries and data structures. The potential advantages of these systems are syntax structure validity, improved efficiency and understanding as well as reduced training requirements [1].

## 1.1 Related Approaches

Several approaches have been proposed in the area of graphical query methods. In this section we will discuss some of the related work that is most relevant to the proposed method. As defined in [1], VQSS can be classified into Form-based, Diagram-based, Icon-based or a hybrid of the other three. The proposed system falls under the category of Diagram-based, as it uses geometric elements to construct queries.

Query By Diagram [2] and [3] uses a Diagram-based approach which builds the query on the E-R schema tables. It supports recursion through the operator closure-of using transitive property. Another approach is the Classification Query Language (CQL) [4], which uses a Query By Example [5] technique which uses a method of filling constants and sample values into the skeletons [4] using a table based interface. Another example of this is VISUAL [6], a graphical icon-based query language for scientific databases. VISUAL's query processing techniques are based on representing the relationships of the application domain. Some methods such as CQL[4] do not specify recursion explicitly in the queries, but is achieved by embedding recursion into basic concepts of classifications and set theory [4]. Another approach which does not explicitly represent recursion is GraphLog [7] which implies recursion through closure literals.

Several other visual query systems such as NITELIGHT [8], OntoVQL [9], iSPARQL [10] have been developed by the researchers. OntoVQL maps the visual query language into the nRQL[11] Language, whereas NITELIGHT and iSPARQL translates the query graphs in to SPARQL syntax. But these approaches do not extend support for recursion. Recursion is a very useful method to retrieve information from graphs of unknown depth without having to write a complex SPARQL query. Furthermore, the proposed system is not limited only to a specific semantic language such as SPARQL.

## 1.2 The Visual Semantic Query System

In this section, we will describe how the Visual Query engine constructs and processes the query graphs, followed by the system architecture and the corresponding modules.

The system enables users to query information stored in knowledge bases in RDF format using visual techniques. This is achieved by allowing the user to construct a graph representing the query and retrieve the answers in text form as well as graphical form. Once the user has constructed the query graph the system automatically converts the query graph into an internal representation that is required for retrieving the answers from the knowledge base. Subsequently, the answers in a specific format are

converted into text form as well as graphical forms. This method helps the user to search for complex information in the knowledge base intuitively without requiring the user to have in-depth understanding of the underlying knowledge base. The query construction process is enhanced by automatically filtering potential query elements. The system helps in minimising the mistakes that the user may make in query construction this way. For example, when the user selects a node in the query graph, the property list shall be filtered to show the range of properties that the particular concept is associated with. So this way the users are guided towards making a more accurate query even if the detailed structure of the ontology is not known.

The Visual Query system starts by loading an ontology. Subsequently the list of concepts and properties in the ontology are loaded. Then the user can start creating the query by selecting a known concept or an unknown concept. After this, more properties and concepts can be selected and added to the query graph. Once the graph is constructed, the user can execute the query. Following this the system converts the query graph structure into a SPARQL or PROLOG statement and sends the query statement to the query processing engine, which retrieves the answers. Then the answers are converted into readable textual representation as well as answer graphs and displayed on the screen. The system also supports utility functions to load/save graphs and to perform regular graph operations such as add nodes, delete node etc.

Another feature of the system is the support for recursive queries. This is driven by the need of retrieval of sub graphs with unknown depth from ontologies. This is achieved through recursive query functionality. We propose a method to graphically represent the idea of recursion using graph notation. Following this, the recursive graphs are converted into logical program structures handling recursion. Finally, the results of executing the query are manipulated to be displayed in a hierarchical representation of all iterations of recursion.

The system architecture and methodology of the proposed system are discussed in section 2, followed by the experiments and discussion of results in section 3.

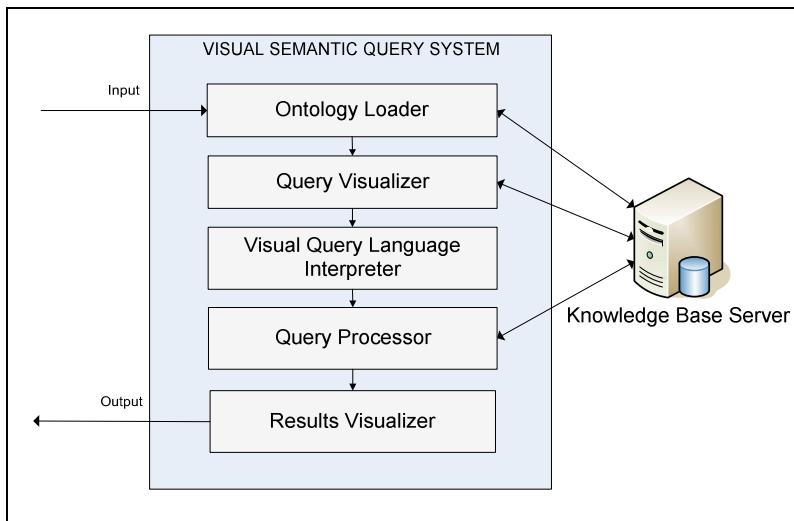
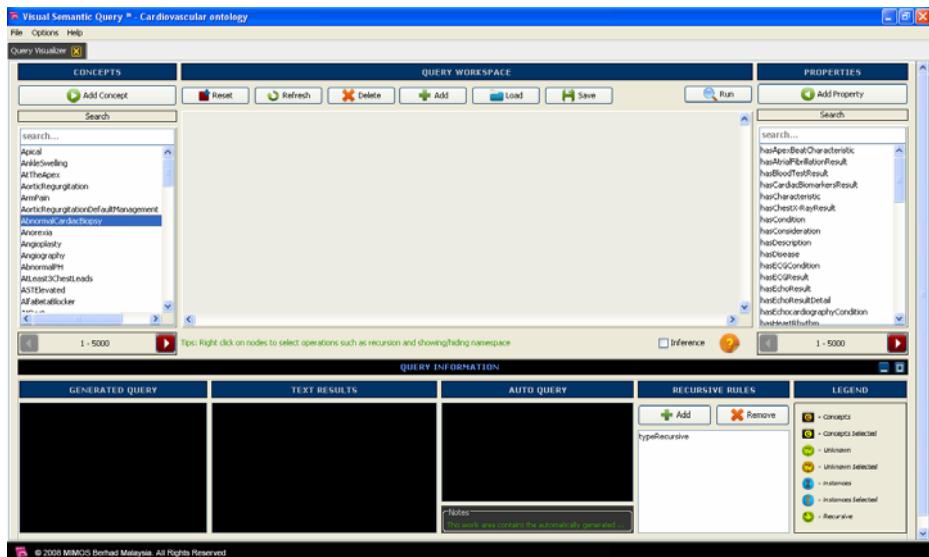
## 2 System Architecture

The system has a Graphical User Interface (GUI) where the input is given through user interaction with the graphical elements. The system output is in the form of result graphs, as well as readable text.

The Visual Semantic Query consists of several modules. These modules are described in more detail in the following subsections. The architecture diagram is shown in Fig. 1 where the several modules of the system are shown.

### 2.1 Ontology Populating Module

This module covers the first step of the user, where an ontology is selected for loading. The selected ontology is sent to the knowledge base processor to retrieve all the concepts, instances and properties in the ontology. Once the system receives the data it is loaded into the interface for user interaction. The concepts/instances in the ontology are loaded into a list of concepts on the left hand side and the properties are loaded into a list of properties on the right hand side. Note that initially only a few

**Fig. 1.** System architecture of the Visual Semantic Query system**Fig. 2.** The Visual Semantic Query Interface

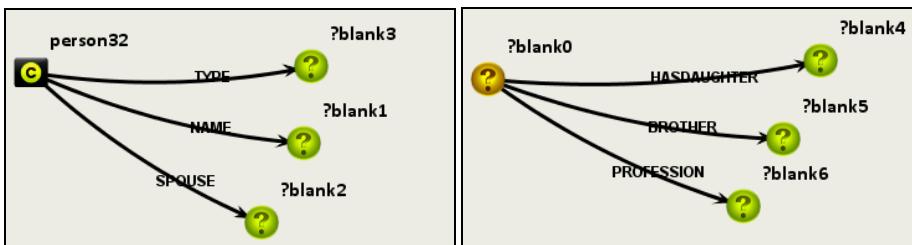
thousand concepts/instances and properties are loaded. How much of data needs to be displayed is controlled by a configurable parameter. The data is shown in alphabetical order, so the user may browse through and decide where the query should begin from. Alternatively user may also do a keyword search if he/she has an idea of what to query about. Users can access more information in the ontology by navigation buttons. These lists of concepts/instances and properties are dynamic in nature as the

information displayed will change depending on the user's actions. Since all the information is served on demand, the system is able to scale to handle very large ontologies. Indexing and caching techniques are used to make the retrieval faster. A screen capture of the system is shown in Fig. 2.

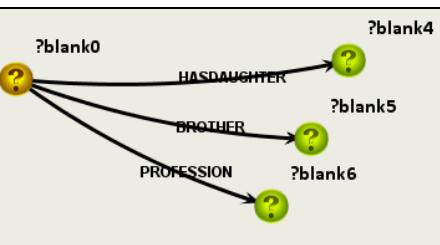
## 2.2 Query Construction Module

This module deals with the query graph construction phase. Since the system supports data in RDF format, the query graphs are based on the subject-predicate-object notation of RDF. Here subjects and objects are represented by nodes and predicates are represented as edges connecting the subject nodes to object nodes. The user has the option of starting the query graph construction process using the Visual Query language through two methods as follows:

Query creation from a known concept. The user starts by selecting a concept from the list of concepts by browsing/searching the concepts list and begin query construction using that concept. This will create a node in the query graph. Upon selecting the created node, the properties panel is populated with all the properties which are attached to that particular concept in the ontology. This is achieved by converting the user's request into a SPARQL query to fetch the corresponding properties of the selected concept. Now that a query is formed, the user may extend the query graph to make the query richer by adding more query elements in the form of properties and concepts. An example is shown in Fig. 3 where the graphical query is to find the properties *type*, *name* and *spouse* of a concept *person32*.



**Fig. 3.** Query creation from a known concept



**Fig. 4.** Query creation from an unbound node

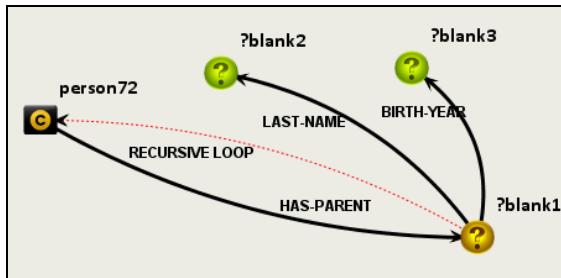
Query creation from an unbound node. The user starts by creating a query graph with an unknown concept indicated by an unbound node. This method can be used if user has information about only the properties and how they connect to each other in a query. Once the query is constructed user can keep on extending the graph using the steps described for the previous method. An example is shown in Fig. 4 where the query is to find the properties *brother* and *hasdaughter* of any concept with a constraint that the *profession* of that concept should be *senator*.

The system is also equipped with additional options such as saving query graphs and loading pre-stored queries.

### 2.3 Recursive Query Construction Module

This module deals with the recursive query construction and visualization. The query graphs are based on the subject-predicate-object triple notation, where subjects and objects are represented by nodes and predicates are represented as edges between the subjects and objects.

To construct recursive query graphs, there need to be an existing regular query graph. Construction of the regular query graph is as mentioned in section 2.2. Recursion is represented in terms of properties. For example, querying *child* property on query graph would translate to finding all descendants when applied recursively. Similarly, to find ancestors, the *parent* property that can be applied recursively. Using this idea, we represent recursion by using a graphical notation, a dotted arrow between the concepts connected by the property that the user has identified for recursion. This is shown in Fig. 5 below. The query translates to finding the *has-parent* property of concept *person77* recursively along the *last-name* and *birth-year* of the parent property. In other words, find all last name and birth year of all the ancestors of *person77*.

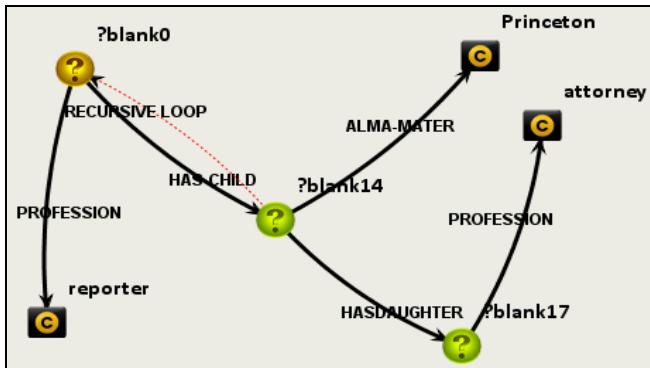


**Fig. 5.** Recursion visualization in a query graph

Following this, more concepts and properties may be added to the query graph to create more complex query elements and constraints. For example a complicated scenario like "*Find all the descendants of a person who is a reporter, with a constraint that all the descendants went to Princeton University, with an even further constraint to fetch each of those descendants' daughters who are law professionals*" can be expressed with a simple query graph as shown in Fig. 6.

### 2.4 Visual Query Language Interpretation Module

After constructing the query, the system interprets the query graph into a form that can be executable by the machine. This mapping from the query graph into a syntax specific textual representation language such as SPARQL/PROLOG is done by this module. Depending on the type of graphical language used, the system translates the graphical elements are construct a corresponding internal query language. In our system, the regular query graphs are converted into SPARQL queries and the recursive query graphs are converted into PROLOG statements. This is done by automatically creating the syntactical elements by utilizing the triple notation within the graph. The result of such an interpretation is shown in Fig. 7.



**Fig. 6.** Recursion visualization in a query graph for a complex query

**GENERATED QUERY**

```
Select DISTINCT ?blank1 ?blank3 ?blank2 WHERE
{
  <http://www.mimos.my/cardio.owl#ArmPain>
  <http://www.w3.org/2000/01/rdf-schema#label> ?blank1 .
  <http://www.mimos.my/cardio.owl#ArmPain>
  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> ?blank3 .
  <http://www.mimos.my/cardio.owl#ArmPain>
  <http://www.mimos.my/cardio.owl#hasDescription> ?blank2 .
}
```

**Fig. 7.** SPARQL query generated from a visual graph

## 2.5 Recursive Visual Query Language Interpretation Module

In this module, the steps involved in translating the graph into internal query structures are defined. Once the system identifies a graph to be a recursive query graph, the system automatically creates PROLOG rules using the manner in which the properties and concepts are connected to each other. An example of this is shown in Fig. 8. Like before, this process is also based on extracting the triple form from the query graph. After creating the rules, system creates recursive PROLOG queries by combining the rules that were identified earlier. An example of this is shown in Fig. 9. Additionally, once the recursive rules are created, they can be saved and reused in other queries as well. Recursive functionality such as this may be used to discover information, which may not be very obvious. For example, when it comes to discovering money laundering patterns in financial transactions, one may create a query representing a potential money laundering pattern and execute the query recursively to find information that is not directly linked.

## 2.6 Query Processing Module

This module is concerned with establishing connections with the knowledge base and executing the query statements created in the previous steps. This is a generic module which can be replaced with other knowledge base processor engines. The current implementation uses AllegroGraph [12] server as the knowledge base processor.

```

GENERATED QUERY

ancestor: (<-- (ancestor ?x ?y ?y ?arg3 ?arg4) (q ?x
!<http://www.franz.com/simple.owl#has-parent> ?y)(q ?y
!<http://www.franz.com/simple.owl#first-name> ?arg3)(q ?y
!<http://www.franz.com/simple.owl#profession> ?arg4))
ancestor: (<- (ancestor ?x ?y ?z ?arg3 ?arg4) (q ?x
!<http://www.franz.com/simple.owl#has-parent> ?z) (ancestor ?z ?y ?y ?arg3 ?arg4))

```

**Fig. 8.** SPARQL query generated from a visual graph

```

GENERATED QUERY

Prolog Query: (?x ?y ?arg0 ?arg1)(ancestor !<http://www.franz.com/simple.owl#person9> ?x
?y ?arg0 ?arg1)

```

**Fig. 9.** SPARQL query generated from a visual graph

## 2.7 Results Visualizing Module

Once the query processor receives the answers from the knowledge base, the next step is to convert the syntax specific results into readable answer forms and display the answers to the user in graphical form. An example of this is shown in Fig. 10.

```

TEXT RESULTS

ArmPain LABEL "Arm pain"
ArmPain TYPE Pain
ArmPain HASDESCRIPTION "Pain in the arms
usually on the inside of the left arm radiating
to the chest"

```

**Fig. 10.** Results in text form

## 2.8 Recursive Results Visualizer

This module converts the results of the query received from the knowledge base processor into visual result graphs. The result graphs are represented using a hierarchical view of all the iterations of recursion. The hierarchical graph is embedded in rings of concentric circles with nodes laid out on the circles and arcs connecting the nodes. The results graph can be manipulated in two ways as shown below.

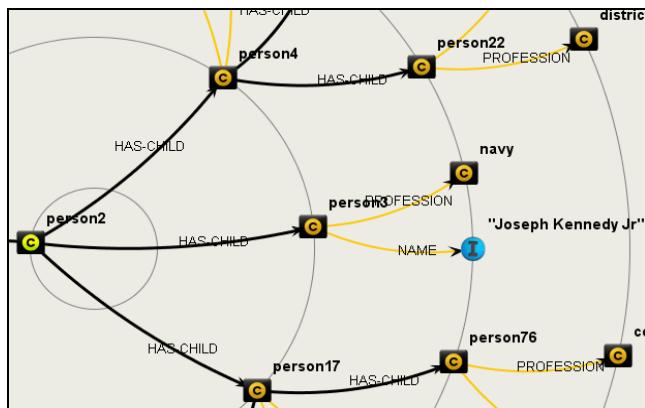
*Method 1.* All the iterations of recursion are shown.

- Iterate through the results of recursion.
- Convert the concepts and properties into hierarchical graph structure starting from the initial query node.
- Identify the depth of recursion.

- d) Assign concepts and properties of all the iterations of recursion, onto as many concentric circles as the number of depths of recursion. (Fig. 11)
- e) Display the graph.

*Method 2.* The results are displayed in a single level of recursion (Collapsed view).

- a) Iterate through the results of recursion.
- b) Convert the concepts and properties into hierarchical graph structure starting from the initial query node.
- c) Manipulate the results and storing them in a single level of depth, starting from the initial query node.
- d) Display the graph.



**Fig. 11.** Visualization of recursion results

### 3 Experiments and Results

The experiments were conducted on several knowledge bases. Since the system is a GUI application, usability tests were carried out with several representative users. We tested 6 people from novice users with no previous experience on the application but with basic ontology domain & semantic query knowledge. We also tested 2 people from expert user with knowledge on ontology domain and semantic query and application. The tests were conducted in a usability lab and involved observing test participants while they performed several tasks given.

We created a set of scenarios designed to test the ability of users in using the application. Each task scenario was designed to be as real-world as possible. There are total of seven test scenarios which will be given to all participants. They are loading an ontology, creating a simple query and run it, saving and loading query graphs, searching for a specific concept and create a graphical query using that concept by applying recursion to a property in the query.

Effectiveness and efficiency were measured by the successful completion of criteria breakdowns from scenario tasks. If they are matched, it will be marked as a ‘Yes’. A success mark is given the full credit of 100%. Criteria that are not matched will be given a ‘No’ mark. Unsuccessful task criteria can include events such as the user giving up, user requiring assistance from the facilitator or completing tasks incorrectly, etc. Partial credit will also be made available in the form of a ‘Partial’ mark which will allow for 50% credit. The usability score was measured using equation 1 below.

$$(Yes + (Partial \times 0.5)) / Total \times 100\% \quad (1)$$

The users were tested on three different aspects of usability namely effectiveness, efficiency and satisfaction. The evaluation was analyzed separately for novice and expert users. These tests were based upon the scenarios that were previously mentioned.

### 3.1 Effectiveness

The analysis of the effectiveness using success rate evaluation is done for the criteria as defined in the table below. The four criteria used for testing effectiveness were to measure if the user understood how to do the task, user did not take more than 3 steps to do the task, user did not need assistance to complete the task and if user succeeded to complete the task. Equation 2 shows the score for novice users, based on the formula in equation 1.

$$\begin{aligned} \text{Effectiveness (\%)} &= (146 + (7 \times 0.5)) / 168 \times 100\% \\ &= 88.99\% \end{aligned} \quad (2)$$

Equation 3 shows the score for expert users, based on the formula in equation 1.

$$\begin{aligned} \text{Effectiveness (\%)} &= (53 + (3 \times 0.5)) / 56 \times 100\% \\ &= 97.32\% \end{aligned} \quad (3)$$

### 3.2 Efficiency

The analysis of the efficiency using success rate evaluation is done for the criteria as defined in the table below. The four criteria used for testing efficiency were to measure if the user preformed the right action on the first try, if the user easily recovered from errors and mistakes that happened, if error and mistakes were minimal and if the user did not take much time to complete task. Using the same method as earlier, these were the findings for efficiency for novice users.

$$\begin{aligned} \text{Efficiency (\%)} &= (145 + (13 \times 0.5)) / 168 \times 100\% \\ &= 90.17\% \end{aligned} \quad (4)$$

The findings for expert users are summarized below.

$$\begin{aligned} \text{Efficiency (\%)} &= (53 + (3 \times 0.5)) / 56 \times 100\% \\ &= 97.32\% \end{aligned} \quad (5)$$

### 3.3 Satisfaction

Measures of satisfaction were taken using post questionnaires with participants. Using a 4 point Likert scale [13] with a negative weighting to 1 and a positive weighting to 4, each question answered by 6 novice users offers a possible positive response factor of 48 points and for 12 questions there are total of 288 points or 100% satisfaction. To establish the satisfaction rating for the VSQ application, we use the following equation:

$$\begin{aligned}
 \text{Satisfaction (\%)} &= \frac{\text{Answer Point} / \text{Total Point} \times 100\%}{(6)} \\
 &= \frac{221 / 288 \times 100\%}{(6)} \\
 &= \mathbf{76.74\%}
 \end{aligned} \tag{6}$$

Using Equation 4, the usability testing with six novice users resulted in a satisfaction rating of approximately 76%, while for two expert users the satisfaction value was 84% ( $80/96 \times 100\%$ ).

### 3.4 Usability Score

We have three metrics of usability (effectiveness, efficiency, and satisfaction), each expressed as a percentage. By averaging these three scores, we can define the usability of VSQ application as a number between 1 and 100. Based on this, the usability testing with six novice users resulted in a usability level of about 86 %, whereas expert users had about 93%.

In addition to user acceptance testing, performance tests were conducted on the system to measure the robustness and reliability. Load test scenarios were deployed for this and the system performed well for 300 concurrent virtual users.

## 4 Conclusion

In this paper we introduced a graphical query system for retrieving information from ontologies. Our system enables users with little prior knowledge on semantic languages to put together sophisticated and effective queries in a user friendly environment and obtain accurate results. One of the highlights of our system is the ability to handle recursive queries, thereby allowing retrieving complex information in an elegant and easy manner. Furthermore our system is designed in such a way that it can handle large datasets using fast indexing mechanisms. Finally usability tests as well as performance tests were conducted to evaluate the system. Although the user tests were done only on a small scale at the moment, we are planning to do more extensive testing in the future. Another limitation of the system is the ability to support other query constructs like ask, describe, etc. We would also like to incorporate query filters in the visual interface too. As part of future work, we would like to incorporate spatial and temporal queries using the graphical interface.

**Acknowledgments.** We would like to thank MIMOS Berhad for providing the funding and facilities to conduct this research. We would also like to thank the Usability Laboratory in MIMOS Berhad for helping us conduct the experiments with state of the art equipment and corresponding analysis.

## References

1. Catarci, T., Costabile, M.F., Levialdi, S., Batini, C.: Visual Query Systems for Databases: A Survey. *Journal of Visual Languages and Computing* 8(2), 215–260 (1997)
2. Angelaccio, M., Catarci, T., Santucci, G.: QBD\*: A Fully Visual Query System. *Journal of Visual Languages and Computing* 1(2), 255–273 (1990)
3. Angelaccio, M., Catarci, T., Santucci, G.: QBD\*: A Graphical Query Language with Recursion. *IEEE Transactions on Software Engineering* 16(10), 1150–1163
4. Järvelin, K., Niemi, T., Salminen, A.: The visual query language CQL for transitive and relational computation. *Data Knowl.* 35(1), 39–51 (2000)
5. Zloof, M.M.: Query-By-Example: operations on transitive closure, IBM, RC 5526 Yorktown Heights, NY (1975).
6. Balkir, N.H., Ozsoyoglu, G., Ozsoyoglu, Z.M.: A Graphical Query Language: VISUAL and Its Query Processing. *IEEE Trans. on Knowl. and Data Eng.* 14, 5 (2002)
7. Consens, M.P., Mendelzon, A.O.: GraphLog: A Visual Formalism for Real Life Recursion. In: Proceedings of 9th ACM SIGA CT-SIGMOID Symposium on Principles of Database Systems, pp. 404–416 (1990)
8. Russell, A., Smart, P.R., Braines, D., Shadbolt, N.R.: NITELIGHT: A Graphical Tool for Semantic Query Construction. In: Semantic Web User Interaction Workshop (SWUI 2008), Hosted by the 26th CHI Conference (CHI 2008), Florence, Italy (2008)
9. Fadhil, A., Haarslev, V.: OntoVQL: A Graphical Query Language for OWL Ontologies. In: International Workshop on Description Logics (DL 2007). Brixen-Bressanone, Italy (2007)
10. OpenLink iSPARQL, <http://demo.openlinksw.com/isparql/>
11. Haarslev, V., Möller, R., Wessel, M.: Querying the Semantic Web with Racer + nRQL. In: Proceedings of the Workshop on Description Logics 2004, ADL 2004 (2004)
12. AllegroGraph RDFStore, <http://www.franz.com/agraph/allegrograph/>
13. Likert, R.: A Technique for the Measurement of Attitudes. *Archives of Psychology* 140, 1–55 (1932)

# Using Concept Maps to Improve Proactive Information Delivery in TaskNavigator

Oleg Rostanin<sup>1</sup>, Heiko Maus<sup>1</sup>, Takeshi Suzuki<sup>2</sup>, and Kaoru Maeda<sup>2</sup>

<sup>1</sup> German Research Center for Artificial Intelligence, DFKI GmbH,  
Kaiserslautern, Germany

{Oleg.Rostanin,Heiko.Maus}@dfki.de

<sup>2</sup> Ricoh Co. Ltd,  
Yokohama, Japan  
{takeshi.suzuki,kaoru.maeda}@nts.ricoh.co.jp

**Abstract.** During the last decade, a plenty of approaches for intelligent user assistance in knowledge intensive working environments were developed. These solutions vary from a lightweight proactive information delivery (PID) based on a non-intrusive user observation to workflow-based assistance that requires formal modeling of processes, organizations, knowledge domains and task specific information needs. Whereas lightweight solutions have low precision and sometimes yet increase the users information overflow, approaches based on sophisticated modeling have severe problems with bootstrapping and maintenance. The work presented in the current paper aims to find an optimal integrated solution for user assistance in agile knowledge working environments that exploits a lightweight incremental modeling of task relevant knowledge and process know-how using concept maps and concept-based task tagging to improve the quality of PID results. The feasibility of the described approach was proved during the joint research project TaskNavigator conducted by Ricoh Co. Ltd and DFKI GmbH.

**Keywords:** concept map, concept-based task tagging, agile task management, lightweight modeling, proactive information delivery.

## 1 Introduction

During the last decade, a plenty of approaches appeared claiming to find proper solutions for intelligent user assistance in knowledge intensive working environments. According to [Holz et al 05], such approaches can be classified using two following dimensions: i) type of supported processes: from weakly structured to strictly structured; ii) type of information delivery: from lightweight to heavy-weight. [Holz et al 06] states that the knowledge work consists of both strictly structured processes that can be formally modeled and enacted using workflow management systems (WFMS) as well as agile processes (agile knowledge work, AKW) that are creative, innovative and very flexible that makes them difficult to be formalized, modeled in advance and re-used. The current work concentrates on the support of agile processes such as writing a newspaper article or

supervising a scientific thesis. These kinds of job require a lot of research work, creativity, talent and intuition. Although AKW is dynamic, it is required to be managed to be successfully completed in time. Task list management (TLM) tools are often used for flexible time management and planning in AKW environments.<sup>1</sup> Well integrated in the daily work, a TLM tool is getting an ideal place for intelligent information assistance required by a knowledge worker.

Proactive information delivery (PID) is a mechanism for user assistance that automatically identifies current user's information needs and makes proactive suggestions of information that can satisfy these needs. PID has to fulfill two main purposes: i) minimize users information overload by providing information precisely adapted to the current task's needs; ii) guarantee that an important document relevant to the task is not overlooked by the user. We distinguish light- and heavy-weight PID depending on the needed modeling effort [Holz et al 05]. In [Holz et al 06] we performed an experimental evaluation of the approach of lightweight PID based on the TLM system TaskNavigator. Unlike our lightweight approach, there are many heavyweight approaches to learning-on-the-job (see [Christl et al 08], [Rostanin et al 06]) aiming to educate users by providing information according to users' needs and competency level. These approaches ensure a high precision information delivery. However, their major bottleneck is a necessity of a relative large modeling effort that makes them difficult to introduce in an enterprise. The main goal of our latest research was to find and evaluate means that would allow to combine advantages of light- and heavy-weight PID (low modeling effort and high delivery precision).

The current paper analyses problems of the lightweight PID in TaskNavigator (section 2). Then, task tagging as easy way of task annotation is introduced (section 3). After that, using concept maps as a means for knowledge extraction and modeling and its usage in TaskNavigator to improve PID is discussed (see section 4). The results of a feasibility test of the developed PID approach are depicted in section 5. The paper is summarized in section 6.

## 2 Proactive Information Delivery in TaskNavigator

This section gives a short introduction into the TaskNavigator system and its PID feature. After that, problems in lightweight PID are identified.

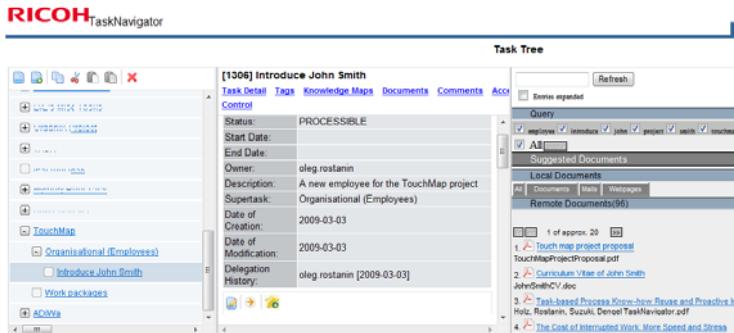
### 2.1 Lightweight PID in TaskNavigator

TaskNavigator is a web-based TLM system providing support for knowledge intensive business processes [Holz et al 06]. By the mechanism of task delegation, task comments and notification as well as flexible task structure management implied by work breakdown structure (WBS), TaskNavigator becomes a powerful tool for work coordination and collaboration in small distributed teams.

The core advantage of TaskNavigator is its PID feature. The main idea of PID is to deliver task-relevant information like documents with background

---

<sup>1</sup> <http://www.culturedcode.com/things/>, <http://www.workity.com/>



**Fig. 1.** Agile task management with TaskNavigator

information or e-mails related to the task proactively, i.e. without an explicit user request. The PID module of TaskNavigator generates a keyword-based query derived from the current task context (see 2.2) and sends a request to an external information retrieval (IR) system automatically. Results delivered from the IR system are sorted by their relevance score relative to the query and presented to the user in the TaskNavigator GUI (Fig. 1, rightmost frame). The user can navigate in the result list, open documents and attach them to the current task. The current implementation of the system uses the TRMeister database [Ikeda et al. 05] for document retrieval from the central TaskNavigator repository and Google Desktop Search API<sup>2</sup> to request users personal documents stored on a local computer. Currently, the central document repository contains the following information types: documents and web pages attached to tasks, tasks themselves (used for similar task retrieval) as well as shared network drives.

## 2.2 Task Context Model and PID Query Generation

The principle of work of lightweight PID is based on the suggestion that the context of a task can be described sufficiently by the task information (title, description) and the information attached to it (comments, documents and web pages), i.e. the task context or lightweight PID can be expressed as:

$$T_{ctx} = \{t, d, C, I\}, \quad (1)$$

where  $t$  is the task title,  $d$  its informal description,  $C = \{c_1, c_2, \dots, c_l\}$  is a set of task comments and  $I = \{i_1, i_2, \dots, i_k\}$  is a set of information objects attached to the task.

By using the TRMeister database, TaskNavigator can extract significant terms from the components of  $T_{ctx}$  and uses them as keywords for the automatically issued PID query.

<sup>2</sup> <http://code.google.com/intl/de-DE/apis/desktop/docs/queryapi.html>

*Example.* For the task “Introduce John Smith” with a description “John Smith is a new employee for the project TouchMap”, TRMeister will generate the following term set: introduce, john, smith, employee, project, touchmap. The query sent to the search engine will deliver documents satisfying the given keyword list, e.g. a project description or a John Smiths CV (Fig. 1, rightmost frame). These documents will be helpful to quickly complete the task.

### 2.3 Problems of Lightweight PID

The first real-world test of TaskNavigator showed great potential of lightweight PID as well as its bottlenecks: although, users were often positively surprised for getting useful e-mails or documents provided by PID that otherwise would have been surely overlooked, there were too many irrelevant documents delivered.

*Advantages.* The main advantage of lightweight PID is the low level of human effort to make it work: the user just needs to type a new task name in TaskNavigator to get PID results (compare to [Christl et al 08]).

*Problem (P1).* Statistics-based query generation used in lightweight PID can cause unsatisfactory quality of generated queries or search results:

(P1.1), *Weaknesses of statistical algorithms* The above example shows that the used statistical algorithm for keyword extraction removes the word “new” from the query as a stop word. However, the word “new [employee]” is essential here to express the task semantics.

(P1.2), *Compound search terms not supported* Still, even if the algorithm could identify the importance of the keyword “new” for the current task, this keyword does not have much sense alone without the keyword “employee”. In this case, we speak of compound search terms (phrases) that need to occur in the document in order the document to satisfy the query.

(P1.3), *Verbose descriptions* Verbose task descriptions cause that the generated query is spoiled. For the task “Apply a new MySQL-DB for TouchMap weblog” with a verbose description “To install a new wordpress software we need a separate database with the name wordpress on our mysql server” would generate the query “apply, mysql, db, touchmap, weblog, wordpress, install, software, separate, database ...” that would result in unpredictable PID results: depending on the IR engine, the result set could contain no or too many documents.

## 3 Task Tagging Improves PID

In this work we claim that:

(C1) introducing **implicit bottom-up modeling of the task context by task tagging** is feasible and can essentially contribute to the quality of PID.

Tagging is a wide-spread technology for lightweight classification and annotation of electronic resources by manually or automatically assigning keywords to them [Golder and Huberman 06]. Considering tasks in TaskNavigator as collaboratively annotated resources, we decompose *C1* into the following sub-claims:

(*C1.1*) Task tags can be used as keywords to refine a search query for task-related PID. Keywords defined by users do not cause problems *P1.1* and *P1.2* (if multi-word tags are allowed). Moreover, the implicit semantics behind task tags given by users will highlight the most important task aspects suppressing the problem of verbose task descriptions *P1.3*.

(*C1.2*) Provided the bag tagging model [Golder and Huberman 06] is utilized by TaskNavigator, where different users can tag tasks multiple times with the same tag, the popularity/relevance of task-related tags can be used to specify weights of single terms comprising a PID query. A weighted PID query will express the importance of each search term thus yet better defining the task semantics (contributes to solving *P1.3*).

(*C1.3*) Provided the list of tags of the parent task is easy available in the current task details, the parent task tags will ease the effort on current task tagging, increasing the system usability.

(*C1.4*) Task tags can be used to find similarly tagged resources (tasks or documents). This can be considered as a useful side-effect of the task tagging.

In order to implement this new vision on PID, the process of the task-specific information retrieval and delivery will be extended as follows: i) Propose possible tags to the user proactively (lightweight PID); ii) User accepts/rejects tag proposals or tags tasks manually (compound tags are allowed); iii) In the collaborative task management environment, the user can vote for or against the tag assigned by himself or by colleagues; iv) A new PID query is generated by TaskNavigator considering tags and tag votes as (compound) search terms and their weights in the query.

*Example.* For the task “Introduce John Smith” from the previous example, the user accepts the keyword “employee” as a tag and makes a compound tag “new employee” from it. Further, she composes a tag “john smith” from the two auto-generated keywords. These two phrases explicitly defined by the user will be given higher weights than for the auto-generated keywords that allows to better control the behavior of the IR engine and get more precise PID results.

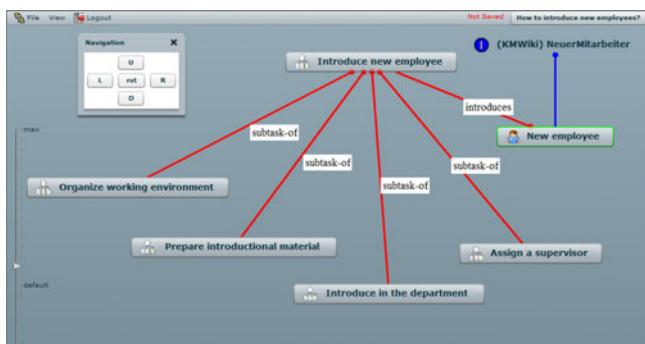
## 4 Lightweight Conceptual Modeling Using Concept Maps

Although task tagging can solve problems of lightweight PID, there are severe problems going along with tagging such as synonymity (**P2.1**), homonymity (**P2.2**), polysemy (**P2.3**) - see [Golder and Huberman 06]. In respect to the information retrieval, the problem of synonymy (includes synonyms, misspelling,

different writing styles, different languages) is the most critical. A standard way of solving the problem of synonymy and misspelling is to use mechanisms supporting controlled vocabularies during the tagging process. To solve homonymy and polysemy problems, more sophisticated ontological modeling of the task-relevant domains can be utilized. A sound ontological modeling of the task-related knowledge domains that can be used for precise PID [Christl et al 08] is effort consuming and in an open knowledge intensive working environment hardly possible. As an alternative, we offer a lightweight alternative for modeling tasks, task-relevant knowledge domains that partially includes the functionality of controlled vocabulary. The proposed solution is based on the idea of concept maps. Since the invention of concept maps by J. Novak in early 1970s [Novak 98], they found a lot of applications in education, research and industry. The intuitiveness of the approach allows the concept maps to be used by end users in non-IT industries. A simple data model behind standard concept maps compared to a formal ontology lacks a control of the vocabulary. Formally defined concept facets described by domains and ranges are also missing. These simplifications restrict using concept maps in the semantic web applications that require formal inferencing. We claim however that introducing light formalisms into the data model of concept maps will suffice to solve problems of tagged-based PID described above. Moreover, concept maps can help solving problems *P1-P3* complementary to simple task tagging.

#### 4.1 LeCoOnt Tool

LeCoOnt<sup>3</sup> is a web-based tool for collaborative concept mapping developed at DFKI. It is aimed to combine the graphical expressiveness and intuitiveness of concept maps, a simple but well-defined data model as well as vocabulary control to provide a universal platform for lightweight knowledge modeling using the concept map paradigm.



**Fig. 2.** LeCoOnt: A concept map “How to introduce a new employee?”

<sup>3</sup> <http://lecoont.opendfki.de>

Fig. 2 shows a screen of the LeCoOnt tool currently editing a concept map “How to introduce a new employee?”. The concept map defines a concept “New employee” as well as the task type “Introduce a new employee” consisting of several subtasks. Every concept in LeCoOnt consists of the following fields: i) a unique “uri” that allows to identify the concept; ii) a “stereotype” specifying the concept type; iii) a “label” (textual name, comparable with skos:prefLabel in SKOS notation); iv) “alternative label” consisting of a “;”-separated label list, e.g. abbreviations; v) “translations” consisting of a “;”-separated list of labels in different languages; vi) an informal textual description. The user can freely define relation names between the concepts. To avoid redundant relations, the user is supported by an auto-completion feature when creating a new relation between concepts that shows relation types already defined in the LeCoOnt database. Furthermore, LeCoOnt allows associating documents and URLs with any concept (e.g. a wiki page “NeuerMitarbeiter” for the concept “New employee”).

TaskNavigator integrates LeCoOnt as means to control the vocabulary used for task tagging. However users are allowed to create new tags for their tasks. These new tags are stored in the LeCoOnt database as unbound concepts that can be later reused for domain or task modeling.

## 4.2 Concept-Based PID and Task Modeling Using Concept Maps

Having introduced a controlled vocabulary maintained by LeCoOnt services, we are able to use it to identify which concepts from the LeCoOnt database match the current task context and not to rely on unstable results of statistics-based keyword extraction (see P1.1 – 3). Fig. 3 (right bottom) shows a proposal to use the concept “New employee” as a tag for the task “Introduce John Smith”. Tag proposals based on the LeCoOnt DB are generated by an information extraction (IE) engine iDocument [Adrian et al 08] that is integrated into LeCoOnt. The user can see the detailed concept information, like relations to other concepts or attached information objects in the advanced PID frame. The user can tag the current task with proposed concepts or attach concept information items to the task. Fig. 3 (left middle) illustrates tags accepted by the user and attached to the task “Introduce John Smith”. Attached concepts (concept-based tags) together with their alternative labels will be used by the PID engine to generate new PID queries. A simple PID query expansion enabled by the underlying concept model will ease the problem of synonymy (P2.1) when searching the documents. Relations of the tag to other concepts in the LeCoOnt database can be used to disambiguate meanings of keywords presented by tags and filter the resulting documents set delivered by PID (currently not implemented).

Whereas the task tagging represents a bottom-up approach to task modeling, the LeCoOnt tool can be used as means to lightweight top-down task modeling. On the Fig. 2 an informal model of the process “Introduce a new employee” created in LeCoOnt is shown. Having attached the concept “Introduce a new employee” as a task tag, a TaskNavigator user can decompose the task into sub-tasks according to the task model defined in the concept map. Created subtasks

The screenshot shows the LeCoOnt application interface. On the left, a task detail view for '[1306] Introduce John Smith' is displayed, containing fields for Description, Supertask, Date of Creation, Date of Modification, and Delegation History. Below this are sections for Tags (including 'Introduce new employee', 'Multitouch', and 'New employee') and Knowledge Maps (listing 'How to introduce new employees?'). On the right, a search interface is shown with a query builder and search results. The query builder shows terms like 'employee', 'introduce', 'john', 'project', 'smith', and 'touchmap'. The search results show a tree structure under '1. New employee' with nodes like '<any' and '<-introduces'. Below this, a list of documents is shown, including 'John Smith's CV' and '(KMWiki) NeuerMitarbeiter'.

Fig. 3. Concept-based PID

will be automatically tagged by corresponding concepts from the concept map and inherit information items attached to the concepts.

Thus, LeCoOnt is a tool supporting the capture and reuse of project-related knowledge. It makes task or domain modeling intuitive and simple. The modeling in LeCoOnt is supported by bottom-up task modeling implicitly done by task tagging. To yet simplify the modeling process, a semi-automatic generation of concept maps from TaskNavigator tasks trees can be realized.

#### 4.3 Extended Task Context Model

To summarize, we extended the definition ② as follows:

$$T'_{ctx} = \{t, d, C, I, T\}, \quad (2)$$

where  $T = \{t_1, t_2, \dots, t_l\}$  is a list of task tags referencing LeCoOnt concepts represented by the uri  $uri_i$ , labels  $L = \{l_{i,1}, \dots, l_{i,k}\}$  and an importance  $v_i$  relative to the task that is calculated as an average of user votes given for or against the task tag:  $t_i = \{uri_i, L_i, v_i\}$ .

Using the definition ②, the PID engine can generate queries of better quality that would deliver better results.

*Example.* The task “Provide introductory material for John Smith” tagged with concepts “Provide introductory material” (stereotype “task type”, voted as not important), “Digital paper” (voted as important, translation “digitales Papier”) will produce the following query: “digital paper” ( $w=100$ ), “digitales papier” ( $w=100$ ), “digital” ( $w=6$ ), “paper” ( $w=4$ ), “digitales” ( $w=6$ ), “papier” ( $w=4$ ), where  $w$  is a term weight in the query.

In implementing the tag-based PID we used the following heuristics: tags defined by users as important (just added to the task) get a weight of 100, very important (added and positively voted) get a weight of 200, less important (added but voted negatively) get the weight of 50, and not important (voted more than once negatively) get a weight of 0. Heuristics mentioned here were tested during the TaskNavigator feasibility study and were found by users as adequate. However, further experiments should be made to find the optimal configuration for PID query term weighting.

## 5 Evaluation

In order to show the feasibility of the approach, a case study was conducted at DFKI that lasted for 3 months. Totally, 11 subjects took part at the experiment: 4 students, 9 researchers and 2 consultants. During the case study, users created 376 tasks as well as attached 624 documents and 164 comments to their tasks. We classified users in two groups: i) users those who used TaskNavigator for part of their work and initiated 97% of the tasks; and ii) the rest with rather short usage period small number of own created tasks. The type of tasks conducted with TaskNavigator ranged from personal tasks such as workshop preparation or writing publications to project tasks such as project organization or customer relations. Over the case study period, 458 tags were added to tasks. During task tagging, 70 new concepts were created. Considering both numbers of tasks and given tags, each task got enriched description by 1.2 tags in average. Over 80% of tags were reused by some means, which means a number of tags being used in the system is fairly maintained to reduce risks introduced with tagging. Over half (54%) of the tags were automatically provided by the system. Finally, 24% of the tags were proposed by the concept-based PID and added to tasks by users.

For the controlled tasks, the subjects compared the query generated from the tasks textual context to the query generated from the concepts attached to the tasks. Once tags were available, usually the tag-based query terms were rated better. The overall impression of the subjects was, that both, lightweight and tag-based PID compliment each other, therefore, they should be used in combination.

## 6 Conclusion

The uniqueness of the TaskNavigator approach of concept-based PID is in using lightweight concept maps instead of formal ontologies to describe knowledge domains and support task tagging.

According to our case study, a bearable user effort spent for task tagging, either manual or supported by the system, allows to improve results of PID as well as to develop the corporative knowledge base. As a feasibility test with real users showed, both lightweight and extended PID approaches complement each other and should be used together. Whereas the concept based PID solves many problems of lightweight one, lightweight PID can help to solve the problem of a system cold start specific to tag-based PID: if there are few concepts available in the knowledge base, lightweight PID keyword proposals can be used to initialize it. Some conceptual aspects could not be tackled in the projects time frame: e.g., the PID engine used in this work considers neither different user skill and knowledge levels. Another critical issue is a seamless integration into the users workspace.

**Acknowledgments.** Parts of this work were supported by Ricoh Co. Ltd as well as German “Stiftung fuer Innovation Rheinland-Pfalz” (projects InnoWiss, TEAL, iDocument).

## References

- [Adrian et al 08] Adrian, B., Dengel, A.: Believing Finite-State cascades in Knowledge-based Information Extraction. In: Dengel, A.R., Berns, K., Breuel, T.M., Bömmarius, F., Roth-Berghofer, T.R. (eds.) KI 2008. LNCS (LNAI), vol. 5243, pp. 152–159. Springer, Heidelberg (2008)
- [Christl et al 08] Christl, C., Ghidini, C., Guss, J., Lindstaedt, S., Pammer, V., Scheir, P., Serafini, L.: Deploying semantic web technologies for work integrated learning in industry. A comparison: SME vs. large sized company. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 709–722. Springer, Heidelberg (2008)
- [Golder and Huberman 06] Golder, S.A., Huberman, B.A.: The structure of collaborative tagging. *Journal of Information Science* 32, 198–208 (2006)
- [Holz et al 05] Holz, H., Maus, H., Bernardi, A., Rostanin, O.: From Lightweight, Proactive Information Delivery to Business Process-Oriented Knowledge Management. *Journal of Universal Knowledge Management* (2), 101–127 (2005)
- [Holz et al 06] Holz, H., Rostanin, O., Dengel, A., Suzuki, T., Maeda, K., Kanasaki, K.: Task-Based Process Know-how Reuse and Proactive Information Delivery in TaskNavigator. In: Proc. CIKM 2006. ACM Conference on Information and Knowledge Management, Arlington, USA, November 6-11 (2006)
- [Ikeda et al 05] Ikeda, T., Mano, H., Itoh, H., Takegawa, H., Hiraoka, T., Horibe, S., Ogawa, Y.: TRMeister: a DBMS with High-Performance Full-text Search Functions. In: Proc. of the 21st Int. Conference on Data Engineering (ICDE 2005). IEEE, Los Alamitos (2005)
- [Novak 98] Novak, J.D.: Learning, creating, and using knowledge: Concept maps as facilitative tools in schools and corporations. Lawrence Erlbaum Associates, Mahwah (1998)
- [Rostanin et al 06] Rostanin, O., Ulrich, C., Holz, H., Song, S.: Project TEAL: Add Adaptive e-Learning to your Workflows. In: Proc. of I-KNOW 2006, Graz (2006)

# A Vocabulary Building Mechanism Based on Lexical Semantics for Querying the Semantic Web

Yu Asano<sup>1</sup> and Yuzuru Tanaka<sup>1,2</sup>

<sup>1</sup> Graduate School of Information Science and Technology, Hokkaido University

<sup>2</sup> Meme Media Laboratory, Hokkaido University

North 13, West 8, Sapporo, Japan

{asano, tanaka}@meme.hokudai.ac.jp

**Abstract.** This paper describes a new framework consisting of two mechanisms: one for building a rich vocabulary based on a lexical semantics, and the other for evaluating queries using such a vocabulary. A vocabulary built by the former mechanism has the following two features: (a) richness because of its expandability, and (b) the lexical-semantic definability of its words. Query expressions using such a rich vocabulary satisfy the following two properties: (c) no need to use nested queries, and (d) no need to use variables. In our framework, each vocabulary consists of basic words and derived words. A derived word can be defined as a character string label given to an expression that combines already defined words with operators. When someone defines a large vocabulary with all of its basic words being mapped to an ontology of the underlying Semantic Web, users can query this Semantic Web using this vocabulary.

**Keywords:** The Semantic Web, Vocabulary building mechanism, Lexical semantics, Query language.

## 1 Introduction

The World Wide Web Consortium proposed the Semantic Web standards such as RDF Schema (RDFS) [1], OWL Web Ontology Language (OWL) [2], Semantic Web Rule Language (SWRL) [3], and SPARQL Protocol and RDF Query Language (SPARQL) [4]. Some works publish resources extracted from a data set with semantic relations between them on the Web using vocabularies such as Dublin Core [5] and FOAF [6]. For instance, DBpedia [7] provides Wikipedia information, and DBLP provides information from academic computer science publications. They also allow us to query in SPARQL.

This paper describes two mechanisms: one for building a rich vocabulary based on a lexical semantics, and the other for evaluating queries using such a vocabulary. A vocabulary built by our proposed mechanism has the following two features: (a) richness because of its expandability, and (b) the lexical-semantic definition of its words. Query expressions using such a rich vocabulary satisfy

the following two properties: (c) no need to use nested query structures, and (d) no need to use variables.

Using the above vocabulary in queries for the Semantic Web simplifies the query sentence structure and removes the necessity of using variables from each query, which makes it much easier for users to query the Semantic Web. A vocabulary defined based on the lexical semantics is called a lexical vocabulary and consists of the following set of words.

- Basic words mapped to objects or relations among them in an object world
- Derived words defined as expressions that combine already defined words

In this paper, the Semantic Web including an ontology and a mechanism to evaluate a query using its ontology works as the object world. Each derived word is defined independently from the object world and can be rewritten by a combination of basic words. Therefore, a lexical vocabulary is reusable because once someone defines a lexicon of a large vocabulary with all of its basic words mapped to an ontology of the Semantic Web, users can query this Semantic Web using this vocabulary. The mapping is called the grounding. Our proposal focuses on using a rich vocabulary and uses simple syntax in query expressions.

Let us suppose that there are two ontologies: *Paper* with the properties *title* and *author*, and *Person* with the property *name*. Thus, a query “Find the titles of papers whose authors include a co-author of ‘Tim Berners-Lee’” is expressed by SPARQL as shown in Fig. ① (1). This expression uses property variables referring to a value of a property such as *?title*. SQL expressions shown in Fig. ②–(3) express the above same query when each above ontology is regarded as a database relation that has an attribute set consisting of the properties used in its ontology and *uri* referring to Uniform Resource Identifier (URI). According to SQL expression for a database, a complex query expression between more

<b>(1) Query in SPARQL</b>	<b>(2) Query in SQL (Nested structures)</b>
<pre>SELECT ?title FROM Paper FROM Person WHERE { ?paper1 title ?title.          ?paper1 author ?author1.          ?paper2 author ?author1.          ?paper2 author ?author2.          FILTER(?author1 != ?author2)          ?author2 name ?name          FILTER(?name = "Tim Berners-Lee")}</pre>	<pre>SELECT title FROM Paper WHERE author IN       (SELECT author        FROM Paper Y        WHERE uri IN              (SELECT uri               FROM Paper               WHERE author != Y.author               author IN                   (SELECT uri                    FROM Person                    WHERE name = "Tim Berners-Lee"))))</pre>
<b>(3) Query in SQL (Tuple variables)</b>	<b>(4) Query in proposed language QLLS</b>
<pre>SELECT X.title FROM Paper X, Paper Y, Paper Z, Person WHERE X.author = Y.author       AND Y.uri = Z.uri       AND Y.author != Z.author       AND Z.author = Person.url       AND Person.name = 'Tim Berners-Lee'</pre>	<pre>SELECT {Title} FROM Paper, Person WHERE Name of co_author of author       = "Tim Berners-Lee"</pre>

**Fig. 1.** Queries in different languages

than one relation is expressed by two methods: one is using a tuple variable such as  $?paper1$  in Fig. 1(1) and  $X$ ,  $Y$ , and  $Z$  in Fig. 1(3), and the other is using a nested structure like those shown in Fig. 1(2). On the other hand, the query in Fig. 1(4) uses a simple and short expression having no nested structure and no variables using words such as *co\\_author* and *author*. The derived word *co\\_author* is defined as the relation between people: “Person A who is an author of a paper written by person B and person A is not person B” using Eq. (1). This definition uses constructors such composition ( $:$ ), intersection ( $\&$ ), and negation ( $\neg$ ) as operators. The word *co\\_author* corresponds to the underlined part of each query of Fig. 1(1)–(3). A derived word *paper* is defined as the inverse ( $-$ ) of a word *author* using Eq. (2). Basic words *author*, *self*, *Title*, and *Name* are mapped to the above relations. In Fig. 1(4), *co\\_author* works like an adjective modifying *Name*, and therefore the former is called a modifier and the latter is called a noun. A noun starts with an upper case letter, and a modifier starts with a lower case letter.

$$\text{co\_author} \triangleq (\text{paper} : \text{author}) \& (\neg \text{self}) \quad (1)$$

$$\text{paper} \triangleq \text{author}^- \quad (2)$$

This paper is organised as follows. We begin with a lexical semantics in Sect. 2. Section 3 shows a Lexical Word Definition Language (LWDL). Section 4 describes a Query Language based on Lexical Semantics (QLLS) and a mechanism for evaluating the query expressions in QLLS. Then, Sect. 5 gives an example of a lexicon and a query evaluation. In Sect. 6, the proposed mechanism is compared to related works, before concluding in Sect. 7.

## 2 Semantics Based on a Universal Relation over the Semantic Web

This section shows a lexical semantics that works as a theoretical basis for evaluating queries using a vocabulary that is built based on the proposed mechanism.

Firstly, a universal relation over the Semantic Web is defined as conforming to a relational model of the database. In the Semantic Web, a triple set is used to describe the Web resources information as an ontology. The triple, which consists of a subject  $s$ , a predicate  $p$ , and an object  $o$ , is denoted by  $(s, p, o)$ . When  $G_i$  denotes a triple set of each objective domain and  $k$  is a number of domains, a set of all ontologies  $\mathbf{G}$  is defined as  $\{G_i \mid i \in \{1, \dots, k\}\}$ . Then  $S_{\mathbf{G}}$ ,  $P_{\mathbf{G}}$ , and  $O_{\mathbf{G}}$  respectively denote a set of all subjects, a set of all properties, and a set of all objects in  $\mathbf{G}$ . In the relational model of the database, any subset of  $V^{\Omega}$  is a relation over  $\Omega$  when  $\Omega$  is an attribute set and  $V$  is a value set. Any element of  $V^{\Omega}$  is a tuple and a function, the domain of which is  $\Omega$  and the range of which is  $V$ . Thus a function  $\Pi_{\Omega'} R$  is defined as  $\{\mu|_{\Omega'} \mid \mu \in R\}$  where  $\Omega' \subseteq \Omega$ . An expression  $\mu|_X : X \rightarrow V$  represents a restriction of a function  $\mu$  to  $X(\subseteq \Omega)$ . A universal relation  $U_{\mathbf{G}}$  is defined as a relation the attribute set  $\tilde{P}_{\mathbf{G}}$  of which consists of all elements of  $P_{\mathbf{G}}$  and *uri* referring to URI by Eqs. (3)–(5).

$$U_{\mathbf{G}} = \{ \mu \mid \mu \in (O_{\mathbf{G}} \cup \{\text{null}\})^{\tilde{P}_{\mathbf{G}}} \wedge \mu(uri) \in S_{\mathbf{G}} \wedge \mu(p_1) \in Value(\mu(uri), p_1, \mathbf{G}) \wedge \cdots \wedge \mu(p_n) \in Value(\mu(uri), p_n, \mathbf{G}) \} \quad (3)$$

$$Value(s, p, \mathbf{G}) = \begin{cases} Object(s, p, \mathbf{G}) & \text{if } Object(s, p, \mathbf{G}) \neq o \\ \{\text{null}\} & \text{if } Object(s, p, \mathbf{G}) = o \end{cases} \quad (4)$$

$$Object(s, p, \mathbf{G}) = \{ o \mid o \in O_{\mathbf{G}} \wedge \exists G \in \mathbf{G} ((s, p, o) \in G) \} \quad (5)$$

Over the universal relation  $U_{\mathbf{G}}$ , this paper describes a mechanism to define a vocabulary consisting of a noun set  $N$  and a modifier set  $M$  based on a basic noun set  $N_0$  and a basic modifier set  $M_0$  shown in Sect. 3 and a lexical semantics of words and constructors to evaluate a relation  $rel(X)$  conforming to the natural language semantics, where  $X \subseteq N$ .

A basic noun  $n$  is defined by grounding it to an element  $p$  of  $\tilde{P}_{\mathbf{G}}$  as Eq. (6). A function  $g_N$  is used for grounding a basic noun. This  $n$  is an element of a basic noun set  $N_0$ . For instance, a basic noun  $URI$  can be defined by grounding it to  $uri$ , which is an element of  $\tilde{P}_{\mathbf{G}}$  in Eq. (7). Thus,  $g_N(URI)$  is  $uri$ .

$$g_N(n) \triangleq p \quad (6)$$

$$g_N(URI) \triangleq uri \quad (7)$$

An information space  $\mathcal{R}$  over a noun set  $N$  is defined as  $\mathcal{R} \triangleq \lambda X / 2^N$ .  $rel(X)$ , when  $\lambda X / D$  denotes the lambda abstraction in the domain  $D$ . A set  $2^S$  denotes a power set of a set  $S$ . Suppose that  $\mathcal{R}_1$  and  $\mathcal{R}_2$  are copies of  $\mathcal{R}$ , we call  $\mathcal{R}_1$  a source information space, and  $\mathcal{R}_2$  a destination information space. A role of a modifier is a label added to each noun of  $\mathcal{R}_2$  for distinguishing nouns of  $\mathcal{R}_1$  and  $\mathcal{R}_2$ . A basic modifier  $a$  is defined as a constraint satisfying a relation  $p$  between a variable  $x$  referring to a value of noun  $n_s$  of  $\mathcal{R}_1$ , and a variable  $y$  referring to a value of noun  $n_d$  of  $\mathcal{R}_2$ . This relation  $p$  can be a property defined in the Semantic Web or a comparison operator. A set  $CO$  of available comparison operators is  $\{=, !=, <, >, <=, >=, substr, !substr\}$ . An operator  $substr$  expresses that its left-hand side is a substring of its right-hand side. Equation (8) is used for a constraint between two information spaces, and Eq. (9) for a constraint of only the destination information space where  $v$  is an element of a value set. Each symbol  $M_{0(A)}$  and  $M_{0(B)}$  denote a basic modifier set defined using Eqs. (8)–(9). A basic modifier set  $M_0$  is defined as  $M_{0(A)} \cup M_{0(B)}$ . After a basic modifier  $a$  is defined, the available noun set is expanded from  $N$  to  $N \cup a@N \cup a@(a@N) \cup a@(a@(a@N)) \dots$ , where  $a@N$  denotes  $\{a@n \mid n \in N\}$ . For instance, a basic modifier  $author$  can be defined as Eq. (10).

$$(A) g_M(a) \triangleq [n_s, n_d, x \ p \ y] \quad (8)$$

$$(B) g_M(a) \triangleq [-, n_d, y \ p \ "v"] \quad (9)$$

$$g_M(author) \triangleq [URI, URI, x \ author \ y] \quad (10)$$

An evaluation of a query using the above vocabulary needs to satisfy three semantic rule sets of a relation  $rel$  shown in Fig. 2. Description of these rules uses relational operations such as a selection ( $\sigma$ ), a cartesian production ( $\times$ ),

**(1) Grounding rule set**

- $\forall n_1 \dots \forall n_k \in N_0$   
 $(\text{rel}(\{n_1, \dots, n_k\}) = (n_1/g_N(n_1)) \dots (n_k/g_N(n_k)) \Pi_{\{g_N(n_1), \dots, g_N(n_k)\}} U_G)$

**(2) Semantic matching rule set**

(A constructor ‘◦’ corresponds to a relational operator ‘•’.)

Each constructor of +, &, – corresponds to each relational operator of  $\cup$ ,  $\cap$ ,  $-.$ )

- $\forall a \in M_0 \forall X \subseteq N \forall Y \subseteq N - a@N (\{g_M(a) \triangleq [n_s, n_d, x p y]\} \supseteq \{\text{rel}(a@X \cup Y) = \Pi_{a@X \cup Y} \sigma_{n_s p a@n_d} (\text{rel}(\{n_s\} \cup Y) \times (a@*/*) \text{rel}(\{n_d\} \cup X))\})$
- $\forall n \forall m \in N \forall X \subseteq N - \{(n \circ m)\}$   
 $(\text{rel}(\{(n \circ m)\} \cup X) = ((n \circ m)/n) \text{rel}(\{n\} \cup X) \bullet ((n \circ m)/m) \text{rel}(\{m\} \cup X))$
- $\forall a \forall b \in M \forall X \subseteq N \forall Y \subseteq N - (a \circ b)@N$   
 $(\text{rel}((a \circ b)@X \cup Y) = ((a \circ b)@X/a@X) \text{rel}(a@X \cup Y) \bullet ((a \circ b)@X/b@X) \text{rel}(b@X \cup Y))$

**(3) Substitution rule set**

- $\forall n \in N_D \forall X \subseteq N - \{n\} (\text{rel}(\{n\} \cup X) = (n/int_N(n)) \text{rel}(\{int_N(n)\} \cup X))$
- $\forall a \in M_D \forall X \subseteq N \forall Y \subseteq N - a^{(i)}@N$   
 $(\text{rel}(a@X \cup Y) = (a@X/int_M(a)@X) \text{rel}(int_M(a)@X \cup Y))$

**Fig. 2.** Three semantic rule sets of a relation  $\text{rel}(X)$  over a subset  $X$  of a noun set  $N$

and a projection ( $\Pi$ ). A symbol ‘/’ denotes a renaming operator. An expression ‘ $(n_1/n_2) \text{rel}(X)$ ’ means that  $n_1$  replaces  $n_2$  where  $X \subseteq N$  and  $n_2 \in X$ . If all nouns in  $X$  need to be replaced, a symbol ‘\*’ can be used. For instance, an expression ‘ $(a@*/*) \text{rel}(X)$ ’ means that an expression ‘ $a@$ ’ is added to each head of all nouns in  $X$ . A symbol ‘ $\supseteq$ ’ denotes a logical implication. Figure 2(1) is for grounding a basic noun set to a universal relation over the Semantic Web. Figure 2(2) shows a semantics of each operation between nouns and modifiers. A symbol ‘@’ is a constructor for modifying a noun by modifier. An expression ‘ $a@n$ ’ means that a modifier  $a$  modifies a noun  $n$ . Figure 2(3) is for substituting a definition  $def_w$  for its derived word  $w$  where  $w$  is defined as  $def_w$ . A function  $int_N$  is used for defining a derived noun, and  $int_N(n)$  denotes a definition of a noun  $n$ . A function  $int_M$  is used for a derived modifier.

The lexical semantics of each noun  $n$  and each modifier  $m$  are defined as Eqs. (11)–(12). We will introduce more constructors in Sect. 3.

$$\forall n \in N (\text{sem}(n) \triangleq \lambda X/2^N \cdot \text{rel}(\{n\} \cup X)) \quad (11)$$

$$\forall a \in M (\text{sem}(a) \triangleq \lambda X/2^N \cdot \lambda Y/2^{N-a@N} \cdot \text{rel}(a@X \cup Y)) \quad (12)$$

### 3 Vocabulary Building Mechanism in LWDL

This section describes a vocabulary building mechanism in LWDL based on the lexical semantics described in Sect. 2. This proposed mechanism is an extension of previous studies [8] [9].

Table 1 shows constructors for combining words. When a basic noun set  $N_0$  and a modifier noun set  $M_0$  are given by the definitions shown in Sect. 2, an element of each set is called a basic word, and a basic vocabulary  $B$  is defined as

**Table 1.** Constructors

	Noun	Modifier	Modifier-Noun
Unary		$\neg$ (Inverse), $\neg$ (Negation), $i$ (Exponentiation), $+$ (Transitive closure)	
Binary	$+$ (Union), $\&$ (Intersection), $-$ (Difference), $/$ (Group by)	$+$ , $\&$ , $-$ , $:$ (Composition)	@ (Modification)
Multiple	{ } (Set) Cf () (Function)		

**(1) Basic words**

- (1-1)  $N_0 \in N$   
 (1-2)  $M_0 \in M$

**(2) Words composed from more than one word**

- (2-1) For any  $a \in M$  and any  $n \in N$ , we assume that  $a@b \in N$ .  
 (2-2) For any  $n, m \in N$ , we assume that  $n+m, n\&m, n-m, n/m \in N$ .  
 (2-3) For any  $n_1, \dots, n_k \in N$ , we assume that  $\{n_1, \dots, n_k\} \in N$ .  
 (2-4) For any  $n_1, \dots, n_k \in N$  and any  $f \in F$ , we assume that  $f(n_1, \dots, n_k) \in N$ .  
 (2-5) For any  $a, b \in M$  and any  $i \in$  positive integer set, we assume that  $a+b, a\&b,$   
 $a : b, a-b, a^-, \neg a, a^i, a^+ \in M$ .

**(3) Derived words**

- (3-1) For any  $n \in N$  and any  $l \in \Sigma^+ - N$ , if  $l$  is defined as a derived noun  
 by giving a new label  $l$  to  $n: int_N(l) \triangleq n$ , we assume that  $l \in N$ .  
 (3-2) For any  $a \in M$  and any  $l \in \Sigma^+ - M$ , if  $l$  is defined as a derived modifier  
 by giving a new label  $l$  to  $a: int_M(l) \triangleq a$ , then we assume that  $l \in M$ .

- (4) Only those elements obtained by finite applications of the above rules (1)–(3)  
 are elements of the noun set  $N$  or the modifier set  $M$ .

**Fig. 3.** Definition of the noun set  $N$  and the modifier set  $M$ 

$N_0 \cup M_0$ . Let a symbol  $\Sigma^+$  denote a set of all strings. Figure 3 shows a definition of a noun set  $N$  and a modifier set  $M$ . A vocabulary  $\Delta$  is defined as  $N \cup M$ . A derived noun  $N_D$  is defined as  $(N - N_0) \cap \Sigma^+$ , a derived modifier  $M_D$  as  $(M - M_0) \cap \Sigma^+$ , and a derived vocabulary  $D$  as  $N_D \cup M_D$ . Derived words are defined using the definitions shown in Fig. 3 (3-1) or (3-2). A computable function set  $F$  includes functions such as  $sum(X)$  and  $count(X)$  where  $X \subseteq N$ . Using LWDL can build a rich vocabulary by defining a lot of words based on a finite number of basic words. An expression ‘ $(a : b)@n$ ’ is defined as ‘ $a@((b@n))$ ’ for any  $a, b \in M$  and any  $n \in N$ .

## 4 Query Language QLLS

This section describes a syntax of our proposed query language QLLS and a mechanism for evaluating QLLS queries.

### 4.1 Syntax of QLLS

A syntax of QLLS is shown in Eq. (13).

SELECT {any subset of a noun set  $N$ }

WHERE (any element of a conditional description set  $CD_{QL}$ ) (13)

The above set  $CD_{QL}$  is defined as follows where  $V$  is a value set.

- (i) For any  $x \in N$ , any  $y \in N \cup V$ , and any  $co \in CO$ , we assume that  $x \text{ co } y \in CD_{QL}$ .
- (ii) For any  $x, y \in CD_{QL}$ ,  $x \wedge y$ , we assume that  $x \vee y \in CD_{QL}$ .
- (iii) Only those elements obtained by finite applications of the above rules (i)–(ii) are elements of the conditional description set  $CD_{QL}$ .

## 4.2 Query Evaluation Mechanism

Here, this subsection describes rewriting rule sets for a query expression in QLLS satisfying three rules shown in Fig. 2. Also mentioned is a rewriting rule set for relational operations using a property of the Semantic Web in a condition.

Let a symbol  $q$  denote a QLLS expression,  $N_S(q)$  denote the noun set appearing in the SELECT phrase of  $q$ ,  $W(q)$  denote the conditional description in  $q$ , and  $N_W(q)$  denote the noun set appearing in the conditional description  $W(q)$ . Thus,  $q$  is evaluated by Eq. (14).

<b>(1) Rule set satisfying the grounding rule set</b>
• $\forall n_1 \dots \forall n_k \in N_0$
$(\text{rel}(\{n_1, \dots, n_k\}) \rightarrow (n_1/g_N(n_1)) \dots (n_k/g_N(n_k)) \Pi_{\{g_N(n_1), \dots, g_N(n_k)\}} U_G)$
<b>(2) Rule set satisfying the semantic matching rule set</b>
<b>(2-1) Basic modifier</b>
• $\forall a \in M_{0(A)} \forall X \subseteq N \forall Y \subseteq N-a@N$
$(\text{rel}(a@X \cup Y) \rightarrow \Pi_{a@X \cup Y} \sigma_{n_{sa} p_a a@n_{da}} (\text{rel}(\{n_{sa}\} \cup Y) \times (a@*/*) \text{rel}(\{n_{da}\} \cup X)))$
• $\forall a \in M_{0(A)} \forall X \subseteq N \forall Y \subseteq N-a^-@N$
$(\text{rel}(a^-@X \cup Y) \rightarrow \Pi_{a^-@X \cup Y} \sigma_{a^-@n_{sa} p_a n_{da}} (\text{rel}(\{n_{da}\} \cup Y) \times (a^-@*/*) \text{rel}(\{n_{sa}\} \cup X)))$
<b>(2-2) Operation between two nouns</b>
• $\forall n \forall m \in N \forall X \subseteq N-\{(n+m)\}$
$(\text{rel}(\{(n+m)\} \cup X) \rightarrow ((n+m)/n) \text{rel}(\{n\} \cup X) \cup ((n+m)/m) \text{rel}(\{m\} \cup X))$
<b>(3) Rule set satisfying the substitution rule set</b>
<b>(3-1) Derived noun</b>
• $\forall n \in N_D \forall X \subseteq N-\{n\} (\text{rel}(\{n\} \cup X) \rightarrow (n/int_N(n)) \text{rel}(\{int_N(n)\} \cup X))$
<b>(3-2) Derived modifier</b>
• $\forall a \in M_D \forall X \subseteq N \forall Y \subseteq N-a@N$
$(\text{rel}(a@X \cup Y) \rightarrow (a@X/int_M(a)@X) \text{rel}(int_M(a)@X \cup Y))$
<b>(5) Rule set for relational operations that have a property in a condition</b>
• $\forall X \subseteq N \forall n \in N \cap X \forall m \in (N \cap X) \cup V \forall p \in P_G$
$(\sigma_{n p m} \text{rel}(X) \rightarrow (n/uri)(m/p) \Pi_{\{uri,p\}} U_G \bowtie \text{rel}(X))$

Fig. 4. Rewriting rule sets for evaluating query expressions

$$\Pi_{N_S(q)} \sigma_{W(q)} \text{ rel}(N_S(q) \cup N_W(q)) \quad (14)$$

In Fig. 4, some of the rewriting rule sets are shown. Each constructor shown in Table 1 corresponds to a rewriting rule. Figure 4 (2) and (3) correspond to Fig. 3 (2) and (3). A symbol ‘ $\rightarrow$ ’ denotes that the left-hand side is rewritten to the right-hand side. In Fig. 4 (2-1),  $n_{sa}$  is the first argument of a definition of modifier  $a$ ,  $n_{da}$  is the second argument,  $p_a$  is a property or comparison operator in the third argument, and  $v_a$  is a value surrounded with double quotation marks in the conditional equation, which is the third argument. Figure 4 (5) shows a rewriting rule set for a selection operation that have a condition using a property defined in the Semantic Web. A symbol ‘ $\bowtie$ ’ denotes a natural join operator.

## 5 Examples of Generating and Evaluating Lexical-Vocabulary Queries

This section illustrates an example of building a lexical vocabulary in LWDL for searching for a paper over the Semantic Web, querying in QLLS using this vocabulary, and evaluating this query.

Let us assume that a property set  $P_G = \{dc : title, dc : date, dc : language, foaf : name, dc : creator, dc : references\}$  is defined in the Semantic Web.

<b>Lexical vocabulary</b>	
(1) $g_N(\text{Title}) \triangleq dc : title$	(9) $g_M(\text{reference}) \triangleq [URI, URI, x \ dc : references\ y]$
(2) $g_N(\text{Date}) \triangleq dc : date$	(10) $g_M(\text{english}) \triangleq [-, Language, y = "en"]$
(3) $g_N(\text{Language}) \triangleq dc : language$	(11) $int_N(\text{Number\_of\_paper}) \triangleq count(URI/author@URI)$
(4) $g_N(\text{Name}) \triangleq foaf : name$	(12) $int_M(\text{paper}) \triangleq author^-$
(5) $g_N(URI) \triangleq uri$	(13) $int_M(\text{citing\_paper}) \triangleq reference^-$
(6) $g_M(\text{since\_year}2000) \triangleq [-, Date, y \geq "2000"]$	(14) $int_M(\text{self\_citation}) \triangleq (author : paper) \& citing\_paper$
(7) $g_M(\text{self}) \triangleq [URI, URI, x=y]$	(15) $int_M(\text{co\_author}) \triangleq (paper : author) \& (\neg self)$
(8) $g_M(\text{author}) \triangleq [URI, URI, x \ dc : creator\ y]$	

### Queries

(Q.1) Find the titles of papers cited by ‘The Semantic Web’.

```
SELECT {Title} WHERE citing_paper@Title="The Semantic Web"
```

(Q.2) Find the full titles and author names of papers whose titles include ‘Semantic Web’ and that were published in 2000 or later.

```
SELECT {Title, author@Name}
WHERE Title substr "Semantic Web" ∧ Date >= "2000"
```

(Q.3) Find the titles of papers whose authors include a co-author of ‘Tim Berners-Lee’.

```
SELECT {Title} WHERE author : co_author@Name="Tim Berners-Lee"
```

(Q.4) Find the number of self-citations to ‘The Semantic Web’

(i.e. citations by one of the original paper’s authors).

```
SELECT {count(self_citation@URI/URI)} WHERE Title="The Semantic Web"
```

**Fig. 5.** Example of building a lexical vocabulary and querying using this vocabulary

Each prefix of *dc* : and *foaf* : is respectively the name space of Dublin Core and FOAF. Figure 5 shows lexical definitions of basic words using Fig. 5 (1)–(10) and derived words using Fig. 5 (11)–(15), and queries Fig. 5 (Q.1)–(Q.4) using this vocabulary.

According to the definition of composition (:) in Sect. 3, an expression ‘*author* : *co\_author*@*Name*’ in Fig. 5 (Q.3) equals ‘*author*@(*co\_author*@*Name*)’ and can be used as ‘*Name* of *co\_author* of *author*’ in Fig. 11 (4). On the other hand, Fig. 11 (1) shows the same query in SPARQL with many variables. Each query in QLLS can be evaluated by the proposed mechanism described in Sect. 4. For instance, Fig. 5 (Q.1) can be rewritten to the following expression by applying Eq. (4), Fig. 4 (3-2), (2-1), (1), and (5). By using this equation, a query in SPARQL for the objective Semantic Web can be generated automatically.

```
SELECT {Title} WHERE citing-paper@Title = "The Semantic Web"
→ Π_{Title} σciting-paper@Title="The Semantic Web"
(citing-paper@Title/reference^-@Title) Π_{Title, reference^-@Title}
((reference^-@URI/uri)(URI/dc : references) Π_{uri, dc:references} UG
× ((URI/uri)(Title/title) Π_{uri, title} UG
× (reference^-@ * / *) (URI/uri)(Title/title) Π_{uri, title} UG))
```

## 6 Comparison between Our Proposed LWDL and QLLS and Related Works

Table 2 compares our proposed LWDL and QLLS to related works such as OWL/SWRL, SPARQL, CLOnE [10], GINO [11], ACE [12], and ORAKEL [13] with respect to five items (a)–(f), where ‘n.a.’ denotes non applicability. According to Table 2, a combination of LWDL and QLLS is the only mechanism that achieves our goal of satisfying four features (a)–(d).

**Table 2.** Results of comparison between our proposal and related works

	OWL/SWRL	SPARQL	CLOnE	GINO	ACE	ORAKEL	LWDL	QLLS
(a)	Yes	n.a.	Yes	Yes	n.a.	n.a.	Yes	n.a.
(b)	No	n.a.	No	No	n.a.	n.a.	Yes	n.a.
(c)	n.a.	Yes	n.a.	Yes	Yes	Yes	n.a.	Yes
(d)	n.a.	No	n.a.	Yes	Yes	Yes	n.a.	Yes
(e)	No	n.a.	Yes	Yes	n.a.	n.a.	No	n.a.
(f)	n.a.	No	n.a.	Yes	Yes	Yes	n.a.	No

- (a) Definition of a rich vocabulary
- (b) Definition of a lexical vocabulary
- (c) No need to use nested structures in query
- (d) No need to use variables in query expressions
- (e) Use of a natural language like syntax in definition of a vocabulary
- (f) Use of a natural language like syntax in query expressions

Vocabularies defined in research proposals other than ours are not lexical vocabularies. Querying the Semantic Web based on a different ontology with a non-lexical vocabulary needs all elements of that vocabulary to be mapped to the base ontology. In contrast, a lexical vocabulary is reusable because querying using a lexical vocabulary needs only basic words to be mapped to the base ontology. OWL/SWRL can define the type, the domain, or the range of each property, while LWDL can not because our proposed mechanism focuses on building a rich vocabulary for querying. On the other hand, introducing a new property into an ontology using OWL/SWRL may need both of its definition in OWL and rule descriptions in SWRL. For this complex processes, other systems such as CLOnE and GINO have been developed. They allow us to use a natural language like syntax, but they also need the same processes to introduce a new property. GINO, ACE, and ORAKEL enable us to use a natural language like syntax in querying, but they do not provide a general-purpose vocabulary independent from the Semantic Web ontology.

## 7 Conclusion

This paper has proposed two mechanisms to build a large vocabulary based on a lexical semantics and to evaluate queries using this vocabulary for querying the Semantic Web. In our mechanism, a combination of words expresses a complex object or constraint. A new derived word can be defined by giving a new label to such an expression and by registering a pair of the given label and the corresponding expression in a lexicon. By repeating such a process, we can define a large number of derived words from a limited number of basic words. Our proposed vocabulary is reusable because, once we define a rich vocabulary, anyone can query a different Semantic Web using the same vocabulary by differently mapping only the basic words to the ontology of this new Semantic Web. The proposed query language need not use either nested query expressions or any variables. Each query can be evaluated based on our proposed lexical semantics.

## References

1. W3C: RDF Vocabulary Description Language 1.0: RDF Schema,  
<http://www.w3.org/TR/rdf-schema/>
2. W3C: OWL Web Ontology Language Overview,  
<http://www.w3.org/TR/owl-features/>
3. W3C: SWRL: A Semantic Web Rule Language Combining OWL and RuleML,  
<http://www.w3.org/Submission/SWRL/>
4. W3C: SPARQL Query Language for RDF,  
<http://www.w3.org/TR/rdf-sparql-query/>
5. Dublin Core Metadata Initiative: DCMI Metadata Terms,  
<http://dublincore.org/documents/dcmi-terms/>
6. W3C: FOAF Vocabulary Specification 0.97,  
<http://xmlns.com/foaf/spec/>

7. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A Nucleus for a Web of Open Data. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007)
8. Tanaka, Y.: Vocabulary Building for Database Queries. In: Goto, E., Nakajima, R., Yonezawa, A., Nakata, I., Furukawa, K. (eds.) RIMS 1982. LNCS, vol. 147, pp. 215–232. Springer, Heidelberg (1983)
9. Goto, F., Tanaka, Y.: Vocalog: An Extended Logic Programming Language Capable of Generic Concept-Description Using a Vocabulary. Transactions of Information Processing Society of Japan 33(4), 512–520 (1992)
10. Funk, A., Tablan, V., Bontcheva, K., Cunningham, H., Davis, B., Handschuh, S.: CLOnE: Controlled Language for Ontology Editing. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 142–155. Springer, Heidelberg (2007)
11. Bernstein, A., Kaufmann, E.: GINO - A Guided Input Natural Language Ontology Editor. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 144–157. Springer, Heidelberg (2006)
12. Bernstein, A., Kaufmann, E., Göhring, A., Kiefer, C.: Querying Ontologies: A Controlled English Interface for End-Users. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 112–126. Springer, Heidelberg (2005)
13. Cimiano, P., Haase, P., Heizmann, J.: Porting natural language interfaces between domains: an experimental user study with the ORAKEL system. In: Proc. of the 12th Int. Conf. on Intelligent User Interfaces, pp. 180–189 (2007)

# Designing a Knowledge Mapping Tool for Knowledge Workers

Heiko Haller and Andreas Abecker

FZI Forschungszentrum Informatik, D-76131 Karlsruhe, Germany

`firstname.lastname@fzi.de`

`http://www.fzi.de/ipe`

**Abstract.** We discuss design issues of iMapping, a novel approach for visually structuring information objects on the desktop. iMapping provides a user-interface layer on top of semantic desktop technologies which shall especially support personal knowledge management and personal information organization. iMapping allows for user-friendly articulation, semantic annotation and interlinking of personal knowledge, as well as its gradual semantic enrichment. In this paper, we elaborate on main GUI principles of the iMapping knowledge mapping and interaction design.

## 1 Introduction

Semantic technologies offer powerful new technological means for knowledge organisation and integration on the desktop, e.g., realized through the *Semantic Desktop* paradigm [10]. However, while such technologies, up to now, focus mainly on technological interoperability and integration, as well as consolidated *back-end* technology, by far not so much work has been devoted to the *front-end*, i.e. ergonomic user interfaces that allow to create, change and exploit semantic knowledge bases in an effective and user-friendly manner. On the other hand, cognitive science and instructional psychology have shown that visual mapping techniques (like *Concept Maps*, *MindMaps* [2], and *Spatial Hypertext* [11]) offer ways to intuitively structure fine-grained information objects. In this paper, we elaborate on some major design decisions for iMapping, a novel visual mapping approach which aims to combine the strengths of the established techniques, and is implemented using modern information technologies like deep zooming interfaces. iMapping is designed as a front-end to create and manipulate semantically annotated personal knowledge-bases; however, in this paper, for the sake of brevity, we focus on the GUI design decisions which can mostly be considered without much knowledge about the back-end behind.

## 2 Requirements

We analysed existing tools and approaches, as well as literature from user-interface design and cognitive science; this led to a set of functional requirements for visual knowledge mapping techniques and tools which shall be cognitively adequate for personal knowledge management (PKM) [6]. The main points are:

### Requirements for Knowledge Mapping Approaches

- Free placing: Items should be freely placeable anywhere and should maintain their positions—at least relative to their surrounding.
- Free relations:  
Allow to interlink items in different levels of formality:
  - Formalized / semantic links (like in ontologies)
  - Informally labeled links (like in concept maps)
  - Unlabeled links (just plain arrows or lines)
  - Free nodes that do not have any explicit relation to others (but may implicitly contain meaning through the relative positions [11])
- Annotations: Optionally hideable notes, marks or highlights in addition to the actual content itself.
- Overview / Abstraction: Following Shneiderman, “*Overview first, zoom and filter, then details on demand*” [14], which can be achieved through item clustering and hierarchical sub-maps.
- Scalability: The approach must conceptually be able to visually deal with large amounts of items.

### Requirements for Knowledge Mapping Software

- Simple editing: Adding or modifying items without much GUI interaction shall be possible.
- Connecting external content shall be possible (e.g., local files or Web pages).
- Focus / filter: deliberately narrowing down visibility to the essential.
- Integration of detail and context through smooth and steady zooming, overview functions, levels of detail.
- Interoperability with other related tools like PIM tools.

We do not see the use of formal semantics as a *requirement* for all visual knowledge mapping tools. But if formal modelling is supported, then the system should allow *incremental* formalization / structure evolution rather than *requiring* the user to make formal decisions always (cp. [13]).

## 3 Design of the iMapping Approach and Tool

We aimed at fulfilling to the greatest possible extent the requirements listed above, at the same time combining the core advantages of the classical mapping approaches (MindMaps, Concept Maps, and Spatial Hypertext), namely:

- visual knowledge representations with structural analogy to content (inherent in any of the mentioned approaches);
- easy hierarchical overall topology (like in mind-maps and as demanded by the requirement of abstraction);
- overview and scalability by integrating context and detail through zooming;
- facility for graph-based node-and-link representations as is common in concept maps and many other modeling approaches;

- allowing constructive ambiguity (as in Spatial Hypertext);
- possibility to express formal semantic statements; and
- querying semantic structures at the various levels of formality they have<sup>1</sup>.

This led to the basic features of iMapping:

### 3.1 Basic Interface Design: Nesting and Zooming

Starting with the *free placing* requirement, we allow items to be freely created at or moved to any position in the map. An iMap can be seen as a virtually infinite pin board. Usually (e.g., for personal note-taking or idea management), these items will be short text passages. The size typically varies from just a keyword to a short note or whole paragraphs. We also permit to use rich text marked up by a Wiki syntax. *Creating and adding content* to an iMap is done by clicking anywhere in the map and typing some text. It is always possible to add informal text items<sup>2</sup>.

A *basic hierarchical structure* is represented by visual inclusion: nested items are shown inside one another (see Fig. 4). Compared to a classical tree view like in Mind-Maps, this has the following benefits: (a) It leaves more freedom to place items according to Gestalt principles. (b) Node-and-link representations (like in classical concept maps) are still possible without visually interfering with connecting lines used for the hierarchical structure. (c) The layout principle stays the same on all levels of the hierarchy (in concentric trees like Mind-Maps, all sub-branches point in one direction). Like that, each item and each part of the map can be treated like a self-contained sub-map which largely helps clustering and modularization. (d) Nesting by inclusion (which is also widespread, e.g., in Venn diagrams, tree-maps), is closer to natural orientation where details are parts of their surrounding. In real life, when we want to see the big picture, we take a step back to see the surrounding context of something.

Of course, this kind of nesting is not new in itself. But traditional paper-and-pencil based mapping techniques can not cover many levels of hierarchy like this. Computer-based mapping approaches allow virtually infinite depth of nesting. In Zooming User Interfaces (ZUI) like iMapping, transitions between levels of hierarchy are made with a *smooth zooming function* that allows users to swiftly change perspective from overview to any detail or back. This opens up a virtually infinite amount of space for iMaps to grow over time, e.g., when used as personal knowledge repositories.

*Re-using existing named items* in other contexts can be done in a copy-and-paste manner or just by re-using their name in a new context. The user will be prompted to decide whether she wishes to create a logically separate item or re-use the existing one which would then have multiple visual instances. This is

---

<sup>1</sup> So far, this is only covered in the QuiKey extension of iMapping [5] which is not discussed in this paper.

<sup>2</sup> Although allowing for semantic knowledge management, it has been a fundamental design decision to never force the user to specify any semantic annotations or metadata, etc. Content can later be refined and formalized incrementally.

adequate if a logical entity item is relevant in different contexts like, e.g., one and the same person can occur in a sub-map of friends in a private context and also in a sub-map of co-workers in a work related context.

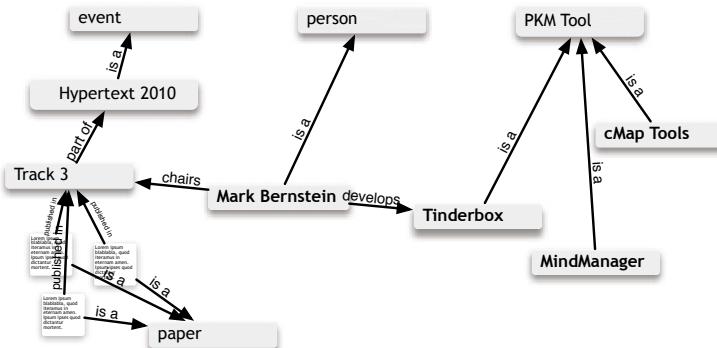
*Levels of detail:* Some information objects—especially text—are rather hard to recognize when they are scaled down to thumbnail size. So, composite items have two possible states: open and closed (or expanded and collapsed). Switching between these states is done manually per mouse click, or can take place automatically, depending on how large the object is displayed. This method is also sometimes referred to as “semantic zooming”. A longer text-item can be represented, e.g., by a keyword in collapsed state and with its full content in expanded mode. A more structured article shows from a distance only its title, when zoomed larger also some additional information like authors and date, then the table of contents and when zoomed to reasonable size, fades over to the full content. Vice versa, the two presentation modes also serve the purpose of reducing visual clutter and saving screen space: Items containing many children items can be *collapsed* such that the huge “belly” with the children is not visible. This also accounts for Shneiderman’s idea of showing *details only on demand*.

### 3.2 Links and Properties

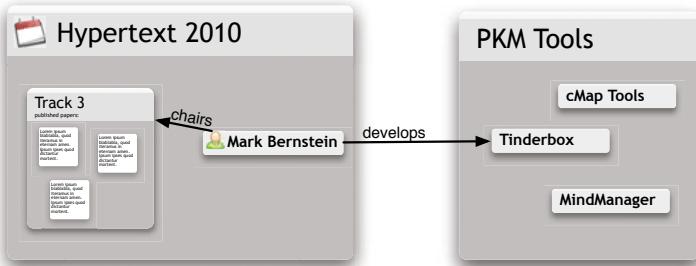
For *establishing link structures*, we distinguish:

1. *implicit linking:* Following the principle of Spatial Hypertext, items can be loosely placed in spatial relations to one another without explicitly linking them at all. As elaborated in [4] by Shipman, these spatial relations can be parsed to extract implicit relations, like sequence, grouping or hierarchy.
2. *explicit linking* on an item level (stating a relation between two objects); each of these can be mere navigational links or carry formal semantics if specified; in particular, we can distinguish four kinds of links and ways of establishing them:
  - (a) *labeled links:* If no pre-existing relation type is suitable, the user can always just enter the label of the link to be displayed, e.g., along the arrow. By that, a new relation type with that name is automatically created in the back-end.
  - (b) *unlabeled links:* Links do not have to be labeled at all.
  - (c) relating items by short *semantic statements* with the QuiKey tool [5]: When relating two items, the user gets a choice of existing relation types selectable by an incremental text search over their names, supporting reuse of existing relation types and avoiding misspellings.
  - (d) *hyperlinks:* Links do not have to be drawn as arrows; hyperlinks go from within the text content of one item to any other item.

*Avoiding Tangle.* In classic graph-based approaches, nodes usually have to be arranged in a layout that minimizes edge-crossings in order to reduce visual complexity. Of course, arranging a map with the goal to minimize line crossings leaves less freedom for arranging it by other criteria. But even when line crossings



**Fig. 1.** Example of a semantic net rendered as a flat graph



**Fig. 2.** The same structured information as in Fig. 1 but with less connecting lines. Most properties are visualized in more specific ways, namely by nesting and by showing known item types with an icon.

are reduced, maps with many nodes and links often suffer from tangled links. In iMaps, this problem is addressed on two levels: (a) by the general architecture that features specific visual properties and (b) on the user interface level by interactively showing links on demand only:

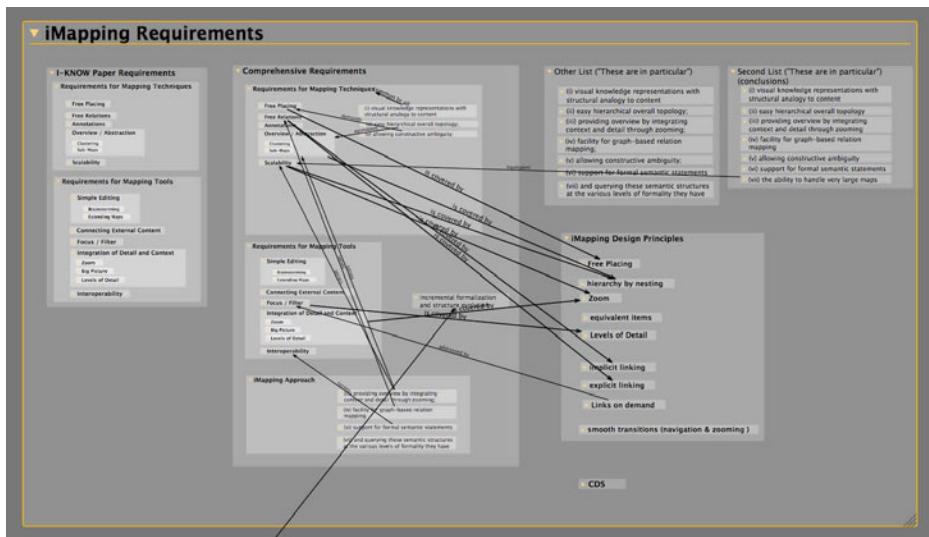
(a) *Specific Visual Properties Instead of Flat Graph.* Graph structures (e.g., E-R or UML diagrams) are capable to represent a very wide range of information structures. Even content which is not inherently graph structured can often be broken down into subject-predicate-object triples and thus represented as a graph. If such a graph is to be rendered visually, without further knowledge of the content's semantics, it appears natural to display it as a flat graph with each node one separate entity and each triple an arrow between two such entities. However, such graphs have a higher visual complexity and are thus harder than what can be done with some knowledge about the meanings of the properties (i.e. relation types) used. Figure 1 shows a small example of a concept-map-style semantic net, mapped out as a simple graph. Figure 2 shows the same structured

information but with less connecting lines needed. Most properties are visualized in more specific ways:

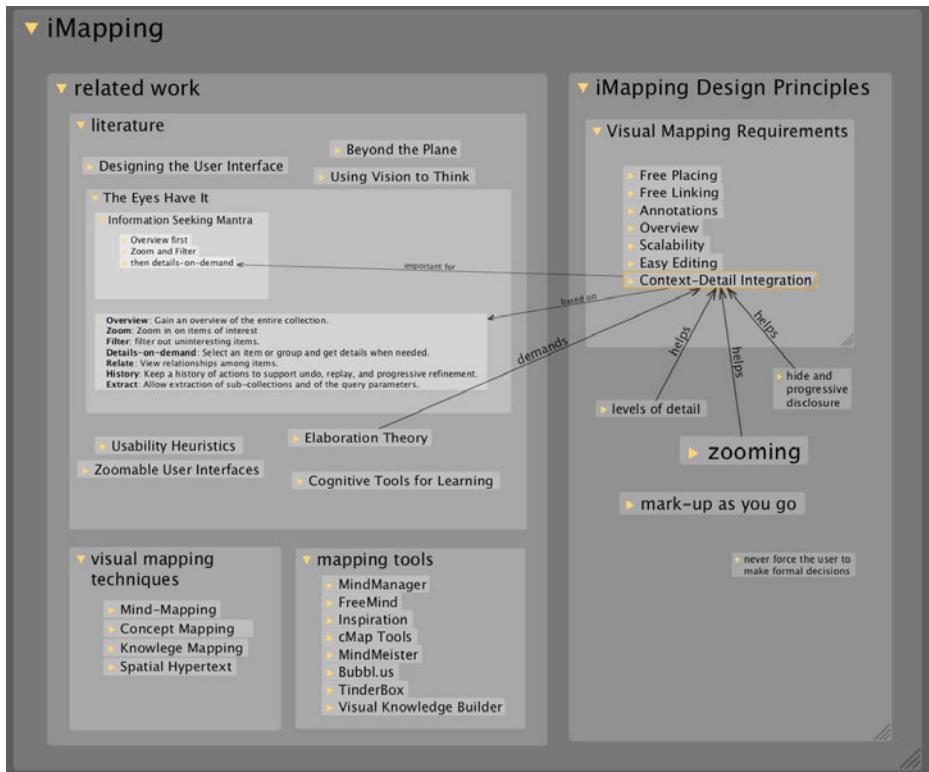
- As argued above, displaying hierarchy by nesting reduces visual clutter. Which of the properties are treated as hierarchical ones is in many cases subjective. But under a PKM perspective such decisions are up to the user.
- Often, groups of items share common properties—e.g., their type (like the “PKM-Tools”) or a relation to a common item (like the “Papers” in this example). Both cases can be depicted by nesting without the need of connecting lines between these items and their parents.
- Another way to show an item’s type, is an unobtrusive icon in the item’s head area. When an item is assigned to a user-defined type with no icon associated, the same principle can still be used with text—e.g., putting the type behind the item’s name in parentheses: “Mark Bernstein [Person]”.

This general principle of rendering many semantic properties in specific ways can be applied to many other semantic and visual properties.

(b) *Links on Demand*. As can be seen in Figure 3, while the above-mentioned measures reduce the number of lines needed, they still do not warrant untangled links. To further reduce visual clutter, it is recommended to only show items’ links on demand. Like that, the links of current interest are more salient and easier to visually integrate with the nodes they connect. In the current iMapping implementation, depending on user settings, links can be made permanent or only become visible when the item is hovered over or selected (see Fig. 4).



**Fig. 3.** Tangled Web Syndrome: When all links in a dense graph are visible, it is hard to distinguish them



**Fig. 4.** Part of an iMap (about iMapping). Only links from and to one item are visible.

*Guided Navigation.* While iMaps can be freely navigated by panning (scrolling) and zooming like it is known, e.g., from Google maps, our user tests have shown that users much prefer navigation guided by content structure like directly zooming to specific items by point-and-click or following links. In addition, a search function lets users jump to any item directly. However, any transition to another place in the iMap is carried out by smooth panning and zooming to maintain a sense of spatial orientation while moving through the map. Other aspects regarding the interaction design that we leave out here for the sake of brevity, comprise in particular the dragging of items and of links.

## 4 Related Work

There are many personal knowledge management tools—both research prototypes and industrial ones. Most closely related to iMapping are:

**Visual Knowledge Builder** [12] is a free research tool implementing the pure Spatial Hypertext approach. It differs from iMapping in several respects—most

notably in that iMapping also offers explicit visible links between items and a zooming facility to ease nested navigation.

**Tinderbox** [1] is a commercial tool and probably the most widely used spatial hypertext editor. It is a mature tool, rich in useful features and it is the closest to iMapping. While it seems to be mainly targeted at supporting authors' writing processes, it is also being used for personal knowledge management. Its structural approach is in many aspects comparable to iMapping—e.g. it is mainly based on freely placeable text-items that can be interlinked and nested into each other and that can have user-defined types of properties.

However, in Tinderbox, the focus is on one hierarchy level at a time. Nested items are regarded rather as separate sub-maps that need to be expanded one at a time, and links can only be made between sibling items. Also the intended way of use is to keep separate files for separate projects. This differs from the iMapping vision, which is targeted to support a continuously growing personal knowledge repository where everything can be semantically linked to everything else, that is represented on one infinite canvas and where several levels of hierarchy are seen simultaneously and transitioned fluently.

**Popcorn** [3] is an experimental personal knowledge base tool that combines the concept mapping approach with the principle of transclusion [9]. It has some strong similarities to iMapping, but it does not use a zooming approach to visually bind single sub-maps together.

Altogether, there are implementations of visual KM tools which come close to iMapping. But none offers the same set of design decisions and none has the explicit ambition of providing a front-end to semantic desktop technologies. Of course, empirical evaluation must show the practical usefulness of the iMapping design decisions.

## 5 Conclusions

**Implementation.** In this paper, we presented the central GUI-design decisions for iMapping which is a visual knowledge mapping tool that allows for (1) user-friendly articulation and semantical annotation, as well as (2) interlinking of personal knowledge, and (3) its gradual semantic enrichment. In this paper, we explicitly wanted to draw the attention to some basic *GUI-design* decisions of iMapping; of course, its *implementation* also offers technical (e.g., scalability) as well as conceptual challenges (e.g., which kind of semantic back-end). The design principles outlined in this paper have been implemented in a mature prototype presented in more detail in [8]. Design-wise, the current implementation (shown in Fig. 4) is characterised (i) by the shades design which underlines the nested structure of iMaps by grey tone shades; (ii) by an implementation of the details-on-demand principle where all controls for an item are only available when this item is selected / hovered.

**QuiKey Extension.** This iMapping implementation works together with the QuiKey extension [5] for fast keyboard-based input and search. Following the vision presented by Uffe Wiil [15], it is desirable to complement a visual tool by

techniques that provide map-independent access to the same structured information; hence, iMapping uses a semantic back-end that can be accessed during runtime by other tools. One such tool is QuiKey which provides a light-weight user interface for browsing and editing the semantic (meta) data of information items in a fine-granular way. At the same time, its incremental search functionality allows to jump to any item in the iMap very fast. Additionally, and in the same interaction paradigm, QuiKey allows to construct simple semantic queries as well as combining these simple queries to more complex ones in a step-by-step manner, thus allowing technically not too advanced users to build structured queries over a semantically annotated knowledge space.

**Evaluation.** The iMapping design has been based on extensive literature research combining ideas from cognitive ergonomics and useability research. During the implementation phase, specific design aspects have been refined with prototype users in a *formative evaluation* process within the NEPOMUK project. In [7] we explain an experimental, qualitative evaluation comparing iMapping with MindMaps for a typical mindmapping task (e.g., testing how well the GUI paradigm is suited to understand, keep in mind, and retrieve a large taxonomic information space); the results are encouraging and show the competitiveness of the current iMapping prototype with state-of-the-art mindmapping software; however, there is not yet statistically significant evidence that iMapping is *superior*. If this is the case, one has probably to investigate more typical iMapping use cases which involve extremely large, interlinked iMaps evolved over a longer period of time. Apparently, such situations can hardly be examined in lab experiments. Hence, at the time of finalizing this paper, the authors are conducting a longer-term usage study where about a dozen people uses iMaps for daily work over a longer period of time (several months).

**Future Work.** Besides further extensions and consolidation of the implementation and the non-trivial task of evaluation, there are still many challenges and opportunities; for instance, the deep integration of iMapping with selected other PKM and PIM tools, but also of other widespread mapping and visualisation approaches like Venn Diagrams, Outlines, or Treemaps; or the combination with new mobile and novel I/O devices (like iPhone, speech interfaces, or ePaper) increasingly being used by knowledge workers.

**Acknowledgement.** This work was partly supported by the European Commission in the project NEPOMUK – The Social Semantic Desktop (FP6-027705), by the German Federal Ministry of Economics and Technology (BMWi) in the THESEUS research programme, and by the German Federal State of Baden-Württemberg.

## References

- [1] Bernstein, M.: Collage, composites, construction. In: HYPERTEXT 2003: Proceedings 14th ACM Conference, pp. 122–123. ACM, New York (2003)
- [2] Buzan, T., Buzan, B.: The Mind Map Book: How to Use Radiant Thinking to Maximize Your Brain's Untapped Potential. Plume (1996)

- [3] Davies, S., Allen, S., Raphaelson, J., Meng, E., Engleman, J., King, R., Lewis, C.: Popcorn: the personal knowledge base. In: DIS 2006: 6th Conference on Designing Interactive systems, pp. 150–159. ACM, New York (2006)
- [4] Francisco-Revilla, L., Shipman, F.: Parsing and interpreting ambiguous structures in spatial hypermedia. In: HYPertext 2005: Proceedings 16th ACM Conference, pp. 107–116. ACM, New York (2005)
- [5] Haller, H.: Quikey—an efficient semantic command line. In: Pinto, S., Cimiano, P. (eds.) 17th Int. Conf. on Knowledge Engineering and Knowledge Management, EKAW 2010 (2010) (to appear)
- [6] Haller, H., Abecker, A.: Requirements for Diagrammatic Knowledge Mapping Techniques. In: Proceedings I-SEMANTICS 2009 (2009)
- [7] Haller, H., Abecker, A.: iMapping - A Zooming User Interface Approach for Personal and Semantic Knowledge Management. In: HYPertext 2010: Proceedings 21st ACM Conference (2010)
- [8] Haller, H., Abecker, A., Völkel, M.: A Graphical Workbench for Knowledge Workers. In: Cordeiro, J. (ed.) Enterprise Information Systems: 11th Int. Conf. ICEIS 2009. Springer, Heidelberg (2010)
- [9] Nelson, T.H.: The heart of connection: hypermedia unified by transclusion. Commun. ACM 38(8), 31–33 (1995)
- [10] Sauermann, L., Kiesel, M., Schumacher, K., Bernardi, A.: Semantic Desktop. In: Blumauer, A., Pellegrini, T. (eds.) Social Semantic Web, pp. 337–362. Springer, Heidelberg (2009)
- [11] Shipman, F., Marshall, C.: Spatial hypertext: An alternative to navigational and semantic links. ACM Computing Surveys 31(4) (1999a)
- [12] Shipman, F., Moore, J.M., Maloor, P., Hsieh, H., Akkapeddi, R.: Semantics happen: knowledge building in spatial hypertext. In: HYPertext 2002: Proceedings 13th ACM Conference, pp. 25–34. ACM, New York (2002)
- [13] Shipman, F.M., Marshall, C.C.: Formality considered harmful: Experiences, emerging themes, and directions on the use of formal representations in interactive systems. In: Computer Supported Cooperative Work, vol. 8, pp. 333–352. Kluwer Academic Publishers, Dordrecht (1999)
- [14] Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. In: VL 1996: Proceedings 1996 IEEE Symposium on Visual Languages. IEEE Computer Society, Los Alamitos (1996)
- [15] Wiil, U.K.: Hypermedia technology for knowledge workers: a vision of the future. In: HYPertext 2005: Proceedings 16th ACM Conference, pp. 4–6. ACM, New York (2005)

# Author Index

- Abe, Akinori III-307  
Abecker, Andreas I-660  
Abdul Maulud, Khairul Nizam IV-22  
Abe, Hidenao III-297  
Abe, Jair Minoro III-123, III-133, III-143, III-154, III-164, III-200  
Abu Bakar, Azuraliza IV-22  
Adachi, Yoshinori III-63, III-81  
Adam, Giorgos III-389  
Aguilera, Felipe II-591  
Ahmad Basri, Noor Ezlin IV-22  
Ain, Qurat-ul I-340  
Akama, Seiki III-133, III-143, III-164, III-200  
Albusac, J. IV-347  
Alechina, Natasha IV-41  
Alonso-Betanzos, Amparo I-168  
Alosefer, Yaser IV-556  
Álvarez, Héctor II-581  
Alvarez, Héctor II-591  
Alvez, Carlos E. II-44  
Alvi, Atif IV-576  
An, Dongchan II-302  
Aoki, Kumiko II-143  
Aoki, Masato IV-153  
Apostolakis, Ioannis III-23  
Appice, Annalisa III-339  
Artsugi, Masayoshi IV-220  
Asano, Yu I-649  
Ashida, Masaya II-611  
Asimakis, Konstantinos III-389  
Aspragathos, Nikos II-341  
Ayres, Gareth IV-566  
Azpeitia, Eneko II-495  
Azuma, Haruka III-273  
  
Baba, Norio III-555  
Baba, Takahiro III-207  
Babenyshev, S. II-224  
Babenyshev, Sergey I-230  
Baig, Abdul Rauf I-61  
Bajo, Javier IV-318  
Baralis, Elena III-418  
  
Bardone, Emanuele III-331  
Barry, Dana M. IV-200  
Bath, Peter II-163  
Baumgartner Jr., William A. IV-420  
Belohlavek, Radim I-471  
Belša, Igor II-21  
Ben Hassine, Mohamed Ali I-532  
Bermejo-Alonso, Julita I-522  
Bhatti, Asim I-5  
Biernacki, Paweł I-350, I-360  
Biscarri, Félix I-410  
Biscarri, Jesús I-410  
Bishop, Christopher I-3  
Blašković, Bruno II-292  
Bluemke, Ilona II-82  
Boochs, Frank I-576  
Borzemski, Leszek II-505  
Bouamama, Sadok II-312  
Bouché, P. IV-32  
Bouras, Christos III-379, III-389  
Bourgoin, Steve IV-410  
Bratosin, Carmen I-41  
Bravo-Marquez, Felipe II-93  
Bretonnel Cohen, K. IV-420  
Bruno, Giulia III-418  
Buckingham, Christopher D. IV-88  
Bukatović, Martin I-432  
Bumbaru, Severin I-188  
Byrne, Caroline IV-365  
  
Cambria, Erik IV-385  
Carpio Valadez, Juan Martín II-183  
Carrella, Stefano II-361  
Caspar, Joachim IV-402  
Cavallaro, Alexander I-290  
Chakkour, Fairouz IV-586  
Champesme, Marc II-351  
Chang, Jae-Woo I-511  
Charton, Eric IV-410  
Chiusano, Silvia III-418  
Chowdhury, Nihad Karim I-511  
Ciampi, Anna III-339  
Cîrlugea, Mihaela IV-613  
Ciruela, Sergio IV-70

- Clive, Barry IV-633  
 Cocea, Mihaela II-103, II-124  
 Condell, Joan IV-430  
 Cooper, Jerry IV-497  
 Corchado, Emilio S. IV-318  
 Corchado, Juan M. IV-318  
 Coyne, Bob IV-375  
 Cruz, Christophe I-576  
 Csipkes, Doris IV-603  
 Csipkes, Gabor IV-603  
 Cui, Hong IV-506  
 Culham, Alastair IV-517  
 Čupić, Marko I-100  
 Cuzzocrea, Alfredo III-426  
 Czarnecki, Adam II-533
- d'Amato, Claudia III-359  
 Dąbrowska-Kubik, Katarzyna III-369  
 Dalbelo Bašić, Bojana I-100, II-31  
 Da Silva Filho, João Inácio III-154  
 de Faria, Ricardo Coelho III-174  
 Debenham, John I-220  
 Deb Nath, Rudra Pratap I-511  
 Delgado, Miguel IV-70, IV-337  
 Dembitz, Šándor II-292  
 Dengel, Andreas I-290  
 Dentorras, Argyris II-331  
 De Paz, Juan F. IV-318  
 Detyniecki, Marcin I-544  
 Di Bitonto, Pierpaolo II-64  
 Diniz, Janaína A.S. III-182  
 Dino Matijaš, Vatroslav I-100  
 Dolog, Peter III-398  
 Domenici, Virna C. III-418  
 Doña, J.M. III-445  
 do Prado, Hércules Antonio III-174  
 Duarte, Abraham II-183
- Eckl, Chris IV-385  
 Ercan, Tuncay II-253, II-263  
 Ernst, Patrick I-290  
 Ezawa, Hiroyasu IV-280
- Fadzli, Syed Abdullah IV-240  
 Fahlman, Scott E. II-193  
 Fanizzi, Nicola III-359  
 Farago, Paul IV-623  
 Farquad, M.A.H. I-461  
 Feng, Haifeng I-544  
 Fernández-Breis, Jesualdo Tomás I-597
- Fernandez-Canque, Hernando IV-603  
 Fernández de Alba, José M. IV-328  
 Ferneda, Edilson III-174, III-182  
 Ferro, Alfredo III-438  
 Festila, Lelia IV-623  
 Festilă, Lelia IV-613  
 Figueiredo, Adelaide III-182  
 Flann, Christina IV-497  
 Fontenla-Romero, Óscar I-168  
 Forge, David II-351  
 Fortino, Giancarlo I-240  
 Fraire Huacuja, Héctor Joaquín II-183  
 Fujii, Satoru III-483, III-519  
 Fujiki, Takatoshi III-509  
 Fujiwara, Minoru IV-163  
 Fukuda, Taro III-473  
 Fukui, Shinji III-89  
 Fukumi, Minoru III-612  
 Fukumura, Yoshimi II-143, IV-190, IV-200  
 Fulcher, John II-454  
 Furutani, Michiko III-307  
 Furutani, Yoshiyuki III-307
- G.V.R., Kiran II-11  
 Gaál, Balázs I-607  
 Gabbar, Hossam A. II-427  
 Gagnon, Michel IV-410  
 Gartiser, N. IV-32  
 Gharahbagh, Abdorreza Alavi I-331  
 Ghofrani, Sedigheh I-331  
 Gibbins, Nicholas IV-594  
 Giddy, Jonathan IV-485  
 Giugno, Rosalba III-438  
 Glaser, Hugh IV-594  
 Gledec, Gordan II-292  
 Glez-Morcillo, C. IV-347  
 Goesele, Michael IV-402  
 Golemanova, Emilia II-253, II-263  
 Golemanov, Tzanko II-253, II-263  
 Gomez, Juan Carlos I-566  
 Gómez-Ruiz, J. III-445  
 Gómez Zuluaga, Giovanni II-601  
 Gonda, Yuuji IV-210  
 Gonzaga Martins, Helga III-154  
 Gonzalez B., Juan J. II-203  
 González, Yanira I-51  
 Görg, Carsten IV-420  
 Gotoda, Naka II-620, IV-145  
 Graczyk, Magdalena I-111

- Graña, M. IV-80  
 Grauer, Manfred II-399  
 Greaves, David IV-576  
 Grillo, Nicola III-426  
 Grimnes, Gunnar I-290  
 Grosvenor, Roger II-371  
 Grzech, Adam II-523  
 Guerrero, Juan I. I-410  
 Guerrero, Luis A. II-93, II-591  
 Gurevych, Iryna IV-402  
 Gutierrez-Santos, Sergio II-124
- Hacid, Mohand-Said III-426  
 Hagita, Norihiro III-307  
 Häkansson, Anne II-273, IV-60, IV-98, IV-124  
 Halabi, Ammar IV-527  
 Haller, Heiko I-660  
 Hamaguchi, Takashi II-381, II-417  
 Hamdan, Abdul Razak I-491  
 Hanabusa, Hisatomo IV-308  
 Handa, Hisashi III-555  
 Hangos, Katalin M. II-389  
 Hanser, Eva IV-430  
 Harada, Kouji III-637  
 Hardisty, Alex IV-485  
 Hartung, Ronald IV-124  
 Hartung, Ronald L. II-273  
 Hasegawa, Mikio IV-271  
 Hasegawa, Naoki IV-190  
 Hashimoto, Yoshihiro II-417  
 Hashizume, Aoi II-135  
 Haskkour, Nadia IV-586  
 Hattori, Akira IV-290  
 Havasi, Catherine IV-385  
 Hayami, Haruo IV-290  
 Hayashi, Hidehiko IV-475  
 Hayashi, Yuki IV-153  
 Heap, Marshall J. IV-517  
 Hernández, Carlos I-522  
 Hintea, Sorin IV-603, IV-613, IV-623  
 Hiratsuka, Yoshimune III-315  
 Hirokawa, Masakazu I-148  
 Hirokawa, Sachio III-207  
 Hirschberg, Julia IV-375  
 Hocenski, Željko I-300  
 Höppner, F. I-442  
 Horák, Aleš I-432  
 Horiguchi, Ryota IV-308  
 Hosseini, Mohammad Mehdi I-331
- Huang, Houkuan II-1  
 Hunger, A. II-114  
 Hunter, Lawrence E. IV-420  
 Hussain, Amir IV-385
- Iftikhar, Nadeem III-349  
 Igarashi, Masao III-622  
 Iijima, Chie III-264  
 Iijima, Morihisa IV-308  
 Ikeda, Mitsuru IV-163  
 Inoue, Akiya III-225  
 Inoue, Etsuko III-509  
 Inuzuka, Nobuhiro III-72  
 Iribe, Yurie II-143, IV-173  
 Ishida, Keisuke IV-190  
 Ishida, Yoshiteru III-628, III-637, III-645, III-652  
 Ishii, Naohiro III-97, III-104, III-113  
 Islim, Ahmed-Derar IV-527  
 Isokawa, Teijiro III-592  
 Iswandy, Kuncup II-361  
 Itoh, Toshiaki II-381  
 Itokawa, Tsuyoshi IV-220  
 Ito, Momoyo III-612  
 Itou, Junko III-473, III-527  
 Ivan, Lavallée I-452  
 Iwahori, Yuji III-63, III-81, III-89  
 Iwashita, Motoi III-225  
 Iwazaki, Tomonori IV-190
- Jabeen, Hajira I-61  
 Jaffar, M. Arfan I-340  
 Jain, Lakhmi C. II-454  
 Jakobović, Domagoj I-100  
 Jamil, Hasan III-408  
 Jantan, Hamidah I-491  
 Jascanu, Nicolae I-188  
 Jascanu, Veronica I-188  
 Jezic, Gordan I-261  
 Jia, Dawei I-5  
 Jiao, Roger I-131  
 Jimbo, Takashi III-97  
 Jimenez, L. IV-347  
 Jimenez-Molina, Angel II-54  
 Jing, Liping II-1  
 Jin, Zhe III-464  
 Ji, Xiaofei I-369  
 Jones, Andrew C. IV-485

- Kambayashi, Yasushi I-198  
 Kamide, Norihiro I-178, II-153  
 Kanda, Taki II-477  
 Kanematsu, Hideyuki IV-200  
 Karadgi, Sachin II-399  
 Karmacharya, Ashish I-576  
 Kasabov, Nikola I-1  
 Kastania, Anastasia N. III-43, III-53  
 Kasugai, Kunio III-81  
 Katarzyniak, Radosław I-271  
 Katoh, Takashi III-455  
 Kavaklı, Manolya II-214  
 Kawaguchi, Masashi III-97  
 Kawakatsu, Hidefumi III-281  
 Kawano, Hiromichi III-225  
 Kelly, Michael IV-11, IV-135  
 Khalid, Marzuki I-69, II-464  
 Kholod, Marina III-273  
 Kim, Daewoong IV-261  
 Kim, Hakin II-302  
 Kimura, Makito III-264  
 Kimura, Naoki II-381, II-409  
 Kipsang Choge, Hillary III-612  
 Kitasuka, Teruaki IV-220  
 Klawonn, Frank I-141, II-244  
 Ko, In-Young II-54  
 Kodama, Issei IV-475  
 Koffa, Antigoni III-53  
 Kogawa, Keisuke III-555  
 Kohlert, Christian I-321  
 Kohlert, Michael I-321  
 Kojima, Masanori III-572  
 Kojiri, Tomoko IV-153  
 Kolp, Manuel I-209  
 Komine, Noriyuki III-545, III-572  
 König, Andreas I-321, II-361  
 Koroušić Seljak, Barbara I-587  
 Kou, Tekishi II-409  
 Kouno, Shouji III-225  
 Kountchev, Roumen III-133, III-215  
 Kozierkiewicz-Hetmańska, Adrianna I-281  
 Kozmann, György I-607  
 Kratchanov, Kostadin II-253, II-263  
 Krišto, Ivan II-21  
 Kubo, Masao IV-298  
 Kumakawa, Toshiro III-315  
 Kunifugi, Susumu IV-457  
 Kunimune, Hisayoshi IV-210  
 Kurahashi, Wataru III-89  
 Kurdi, Mohamed-Zakaria IV-527  
 Kurosawa, Takeshi III-225  
 Kusaka, Mariko III-555  
 Kuwahara, Daiki IV-465  
 Lambert-Torres, Germano III-154  
 Lasota, Tadeusz I-111  
 Laterza, Maria II-64  
 Lawrenz, Wolfhard II-244  
 Lawrynowicz, Agnieszka III-359  
 Le, D.-L. II-114  
 Lee, Huey-Ming II-438  
 Lee, Hyun-Jo I-511  
 Lensch, Hendrik P.A. IV-402  
 León, Carlos I-410  
 León, Coromoto I-51  
 Lesot, Marie-Jeanne I-544  
 Lewandowski, Andrzej I-311  
 L'Huillier, Gaston II-93, II-581  
 Li, Chumping I-131  
 Li, Kai IV-173  
 Li, Yibo I-369  
 Li, You II-445  
 Lin, Lily II-438  
 Liu, Honghai I-369  
 Liu, Jin I-379  
 Liu, Jing II-214  
 Liu, Kecheng I-554  
 Liu, Lucing III-207  
 Liu, Xiaofan IV-41  
 Liu, Yang I-90  
 Liu, Ying I-131  
 Logan, Brian IV-41  
 Lopes, Helder F.S. III-164  
 López, Juan C. II-193  
 Loukakos, Panagiotis I-481  
 Lovrek, Ignac I-251  
 Ludwig, Simone A. IV-536  
 Lukose, Dickson I-627  
 Lunney, Tom IV-430  
 Luz, Saturnino IV-394  
 Ma, Minhua IV-430  
 Macía, I. IV-80  
 Mackin, Kenneth J. III-622  
 Maddouri, Mondher I-121  
 Maeda, Kaoru I-639  
 Maehara, Chihiro IV-261  
 Maeno, Hiroshi IV-163  
 Maezawa, Toshiki II-645

- Magnani, Lorenzo III-331  
 Magoulas, George D. II-103, II-124  
 Mahanti, Ambuj II-282  
 Maheswaran, Ravi II-163  
 Mahoto, Naeem A. III-418  
 Majumder, Sandipan II-282  
 Mák, Erzsébet I-607  
 Makino, Toshiyuki III-72  
 Małachowski, Bartłomiej IV-180  
 Malerba, Donato III-339  
 Mancilla-Amaya, Leonardo II-553  
 Marc, Bui I-452  
 Marín, Nicolás IV-70  
 Markos, Panagiotis II-331  
 Marteau, Pierre-François I-420  
 Martínez-Romero, Marcos II-74  
 Martínez-Costa, Catalina I-597  
 Martínez F., José A. II-173, II-203  
 Marzani, Franck I-576  
 Masoodian, Masood IV-394  
 Mat Ali, Nazmona I-554  
 Matić, Tomislav I-300  
 Matsubara, Takashi IV-298  
 Matsuda, Noriyuki II-637  
 Matsui, Nobuyuki III-592  
 Matsumoto, Hideyuki II-417  
 Matsuoka, Rumiko III-307  
 Matsushita, Kotaro III-622  
 Matsuura, Kenji II-620, IV-145  
 Mattila, Jorma K. IV-108  
 Maus, Heiko I-639  
 Mc Kevitt, Paul IV-430  
 McMeekin, Scott G. IV-633  
 Meddouri, Nida I-121  
 Mehmood, Irfan I-340  
 Mehmood, Rashid IV-566, IV-576  
 Mello, Bernardo A. III-182  
 Menárguez-Tortosa, Marcos I-597  
 Merlo, Eduardo II-581, II-591  
 Metz, Daniel II-399  
 Millán, Rocío I-410  
 Millard, Ian C. IV-594  
 Minaduki, Akinori IV-475  
 Miñarro-Giménez, José Antonio I-597  
 Mineno, Hiroshi II-135, III-535  
 Miranda, Gara I-51  
 Mishina, Takashi III-493  
 Misue, Kazuo IV-440  
 Mitsuishi, Takashi III-281  
 Miura, Hajime IV-190  
 Miura, Hirokazu II-637  
 Miura, Motoki IV-457, IV-465  
 Miyachi, Taizo II-645  
 Miyaji, Isao III-483  
 Miyoshi, Masato III-612  
 Mizuno, Tadanori III-572  
 Mizuno, Shinji II-143  
 Mizuno, Tadanori II-135, III-535  
 Mizutani, Masashi I-198  
 Moens, Marie-Francine I-566  
 Mohd Yatid, Moonyati Binti III-473  
 Molina, José Manuel IV-357  
 Möller, Manuel I-290  
 Molnar, Goran I-100  
 Monedero, Iñigo I-410  
 Moradian, Esmiralda IV-98, IV-124  
 Morihiro, Koichiro III-592  
 Morii, Fujiki I-390  
 Morita, Hiroki III-572  
 Moulianitis, Vassilis II-341  
 Moya, Francisco II-193  
 Muhammad Fuad, Muhammad Marwan I-420  
 Muhammad-Sukki, Firdaus IV-633  
 Mukai, Naoto IV-280  
 Müller, Ulf II-399  
 Munemori, Jun III-473, III-527  
 Munteanu, Cristian R. II-74  
 Murakami, Akira III-315  
 Murat, Ahat I-452  
 Musa, Zalili II-454  
 Nabi, Zubair IV-576  
 Nahavandi, Saeid I-5  
 Nakada, Toyohisa IV-449  
 Nakagawa, Masaru III-509  
 Nakahara, Takanobu III-244, III-273  
 Nakamatsu, Kazumi III-123, III-133,  
     III-143, III-164,  
     III-200, III-215  
 Namiki, Junji III-562  
 Naqi, Syed M. I-340  
 Naruse, Keitaro IV-298  
 Nasri, Chaker Abidi I-532  
 Nauck, Detlef I-141  
 Naveen, Nekuri I-80  
 Németh, Erzsébet II-389  
 Németh, Istvánné I-607  
 Nguyen, A.-T. II-114  
 Nguyen, D.-T. II-114

- Nguyen, Ngoc Thanh I-281  
 Niimura, Masaaki IV-210  
 Nishi, Daisuke II-637  
 Nishide, Tadashi III-473  
 Nishihara, Takanao II-645  
 Nishihara, Yoko III-315  
 Nishimura, Haruhiko III-592  
 Nishino, Kazunori II-143  
 Niskanen, Vesa A. IV-116  
 Niwa, Takahito III-113  
 Noda, Masaru II-381  
 Noguchi, Daijiro IV-163  
 Nonaka, Yuki IV-271  
 Nunohiro, Eiji III-622
- Obembe, Olufunmilayo IV-88  
 Obermöller, Nils II-244  
 Oehlmann, Ruediger III-290  
 O'Grady, Michael J. IV-365  
 O'Hare, Gregory M.P. IV-365  
 Ohsawa, Yukio III-315  
 Oikawa, Ryotaro I-198  
 Okada, Masashi III-104  
 Okada, Yoshihiro IV-251  
 Okada, Yousuke III-113  
 Okajima, Seiji IV-251  
 Okamoto, Takeshi III-628  
 Okumura, Noriyuki IV-51  
 Oliver, José L. I-31  
 Oltean, Gabriel IV-623  
 Omitola, Tope IV-594  
 Omori, Yuichi IV-271  
 Onn, Kow Weng I-627  
 Onogi, Manabu III-113  
 Ooshaksaraie, Leila IV-22  
 Orlewicz, Agnieszka II-82  
 Orłowski, Aleksander II-515  
 Orłowski, Cezary II-533, II-543, II-571  
 Othman, Zulaiha Ali I-491  
 Otsuka, Shinji II-620  
 Ounelli, Habib I-532  
 Oyama, Tadahiro III-612  
 Ozaki, Masahiro III-63  
 Ozell, Benoit IV-410
- Palenzuela, Carlos II-495  
 Paloc, C. IV-80  
 Park, Jong Geol III-622  
 Park, Seog II-302
- Pavón, Juan IV-328  
 Pazos, Alejandro II-74  
 Pazos R., Rodolfo II-183  
 Pazos R., Rodolfo A. II-173, II-203  
 Pedersen, Torben Bach III-349  
 Peláez, J.I. III-445  
 Pérez O., Joaquín II-173  
 Pereira, Javier II-74  
 Peter, S. I-442  
 Petre, Emil II-234  
 Petric, Ana I-261  
 Petrigni, Caterina III-418  
 Pham, Tuan D. I-379  
 Pintér, Balázs I-607  
 Podobnik, Vedran I-251  
 Porto-Díaz, Iago I-168  
 Poulopoulos, Vassilis III-389  
 Pratim Sanyal, Partha IV-506  
 Prickett, Paul II-371  
 Pudi, Vikram II-11  
 Pu, Fei IV-135  
 Puga Soberanes, Héctor José II-183  
 Puglisi, Piera Laura III-438  
 Pulvirenti, Alfredo III-438
- Raghavendra Rao, C. I-80  
 Raja, Hardik IV-485  
 Raju, S. Bapi I-461  
 Rambousek, Adam I-432  
 Rambow, Owen IV-375  
 Ramirez-Iniguez, Roberto IV-633  
 Rana, Omer F. IV-546, IV-556  
 Rango, Francesco I-240  
 Ravi, V. I-80, I-461  
 Ray, Sanjog II-282  
 Read, Simon II-163  
 Reicher, Tomislav II-21  
 Renaud, D. IV-32  
 Resta, Marina III-583  
 Richards, Kevin IV-497  
 Ríos, Sebastián A. II-93, II-581, II-591  
 Rodríguez, Manuel I-522  
 Rodríguez, Sara IV-318  
 Rogers, Bill IV-394  
 Rojas P., Juan C. II-203  
 Rojtberg, Pavel IV-402  
 Roos, Stefanie IV-536  
 Ros, María IV-337  
 Roselli, Teresa II-64

- Rossano, Veronica II-64  
 Rostanin, Oleg I-639  
 Rousselot, F. IV-32  
 Rouveyrol, Claire III-81  
 Różewski, Przemysław IV-180  
 Ruhlmann, Laurent IV-410  
 Russo, Wilma I-240  
 Rybakov, Vladimir I-230, II-224, III-323  
 Rygielski, Piotr II-523
- Sadanandan, Arun Anand I-627  
 Sadek, Jawad IV-586  
 Saha, Sourav II-282  
 Said, Fouchal I-452  
 Sakamoto, Ryuuki III-501  
 Salem, Ziad IV-586  
 San, Tay Cheng I-69  
 Sánchez-Pi, Nayat IV-357  
 Sanín, Cesar II-553, II-601  
 Sanin, Cesar II-563  
 Santaolaya S., René II-203  
 Santofimia, María J. II-193  
 Sanz, Ricardo I-522  
 Sasaki, Takuya III-455  
 Sato, Hiroshi IV-298  
 Sato, Hitomi IV-290  
 Sawamoto, Jun III-455  
 Sawaragi, Tetsuo I-2  
 Schäfer, Walter II-399  
 Schwarz, Katharina IV-402  
 Segawa, Norihisa III-455  
 Segura, Carlos I-51  
 Seifert, Sascha I-290  
 Selișteanu, Dan II-234  
 Șendrescu, Dorin II-234  
 Seta, Kazuhisa IV-163  
 Setchi, Rossitza I-481, I-617, IV-240  
 Shadbolt, Nigel IV-594  
 Shankar, Ravi II-11  
 Shi, Lei I-617  
 Shida, Haruki III-628  
 Shidama, Yasunari III-281  
 Shima, Takahiro III-519  
 Shimoda, Toshifumi II-143  
 Shimogawa, Shinsuke III-225  
 Shiraishi, Yoh III-493  
 Siddiqui, Raees II-371  
 Sidiropoulos, Sofia III-43  
 Sidorova, Natalia I-41
- Sierra, Carles I-220  
 Šilić, Artur II-21, II-31  
 Sintek, Michael I-290  
 Sitek, Tomasz II-571  
 Skorupa, Grzegorz I-271  
 Sofiane, Benamor I-452  
 Sohn, So Young IV-200  
 Soldano, Henry II-351  
 Sproat, Richard IV-375  
 Srivastava, Muni S. III-7  
 Stasko, John IV-420  
 Stewart, Brian G. IV-633  
 Suchacka, Grażyna II-505  
 Sugihara, Taro IV-457  
 Sugino, Eiji III-455  
 Sugiyama, Takeshi III-15  
 Sun, Fan I-90  
 Sunayama, Wataru III-235  
 Suzuki, Kenji I-148  
 Suzuki, Nobuo III-1  
 Suzuki, Takeshi II-645  
 Suzuki, Takeshi I-639  
 Suzuki, Yu IV-440  
 Szczerbicki, Edward II-515, II-553,  
     II-563, II-601  
 Szlachetko, Boguslaw I-311  
 Szolga, Lorant Andras IV-613
- Taguchi, Ryosuke IV-200  
 Takahashi, Hirotaka IV-190  
 Takahashi, Megumi III-519  
 Takahashi, Osamu III-493  
 Takai, Keiji III-254  
 Takano, Shigeru IV-251  
 Takeda, Kazuhiro II-381, II-417  
 Takeda, Kosuke III-501  
 Takeda, Masaki III-555  
 Takeshima, Ryo IV-230  
 Taki, Hirokazu II-611, II-637  
 Takimoto, Munehiro I-198  
 Tamura, Yukihiko III-235  
 Tanabe, Kei-ichi III-645  
 Tanaka, Jiro IV-440  
 Tanaka, Takushi III-190  
 Tanaka, Toshio II-620  
 Tanaka-Yamawaki, Mieko III-602  
 Tanaka, Yuzuru I-14, I-649  
 Telec, Zbigniew I-111  
 Tenorio, E. III-445

- Tipney, Hannah IV-420  
 Tokuda, Mitsuhiro IV-465  
 Tominaga, Yuuki IV-210  
 Tomiyama, Yuuki III-235  
 Torii, Ippei III-104, III-113  
 Toro, Carlos II-495  
 Torres, Claudio Rodrigo III-154  
 Tortosa, Leandro I-31  
 Tóth, Attila II-389  
 Tran, V.-H. II-114  
 Trawiński, Bogdan I-111  
 Tschumitschew, Katharina I-141, II-244  
 Tsogkas, Vassilis III-379  
 Tsuchiya, Seiji I-400, IV-1  
 Tsuda, Kazuhiko III-1  
 Tsuge, Satoru III-612  
 Tsuge, Yoshifumi II-409  
 Tsumoto, Shusaku III-297
- Uemura, Yuki IV-220  
 Ueta, Tetsushi IV-145  
 Uno, Takeaki III-244  
 Ushijima, Taketoshi IV-261
- Valencia-García, Rafael I-597  
 Vallejo, D. IV-347  
 Valsamos, Harry II-341  
 van der Aalst, Wil I-41  
 Vaquero, Javier II-495  
 Varlamis, Iraklis III-23, III-33  
 Vassányi, István I-607  
 Vázquez A., Graciela II-173  
 Vázquez-Naya, José M. II-74  
 Vecchietti, Aldo R. II-44  
 Velásquez, Juan D. II-93, II-581  
 Ventos, Véronique II-351  
 Verspoor, Karin IV-420  
 Vicent, José F. I-31  
 Vila, Amparo IV-337  
 Villanueva, David Terán II-183  
 Vychodil, Vilem I-471
- Wada, Yuji III-455  
 Wakayama, Yuki IV-251  
 Walters, Simon II-322  
 Wan, Chang I-501  
 Wan, Jie IV-365  
 Wang, Bo II-445  
 Watabe, Hirokazu I-400, IV-1
- Watada, Junzo II-445, II-454, II-485  
 Watanabe, Takashi III-123  
 Watanabe, Toyohide IV-153, IV-230  
 Watanabe, Yuji III-660  
 Watanabe, Yuta IV-475  
 Wautelet, Yves I-209  
 Whitaker, Roger I-4  
 White, Richard J. IV-485  
 Willett, Peter II-163  
 Wilton, Aaron IV-497  
 Woodham, Robert J. III-81, III-89  
 Wu, Dan IV-60, IV-124
- Xu, Guandong III-398
- Yaakob, Shamshul Bahar II-485  
 Yada, Katsutoshi III-244, III-254, III-273  
 Yamada, Kunihiro III-483, III-535, III-545, III-562, III-572  
 Yamada, Takayoshi I-158  
 Yamaguchi, Takahira III-264  
 Yamaguchi, Takashi III-622  
 Yamamoto, Hidehiko I-158  
 Yamamura, Mariko III-1, III-7  
 Yamazaki, Atsuko K. II-630  
 Yamazaki, Makoto IV-190  
 Yanagihara, Hirokazu III-7  
 Yanagisawa, Yukio III-622  
 Yano, Yoneo II-620, IV-145  
 Yaoi, Takumu III-483  
 Yasue, Kizuki II-409  
 Yatsugi, Kotaro IV-261  
 Yonekura, Naohiro III-97  
 Yoshida, Koji III-519  
 Yoshida, Kouji III-483, III-572  
 Yoshihara, Yuriko III-555  
 Yoshihiro, Takuya III-509  
 Yoshimura, Eriko I-400, IV-1  
 Yu, Chunshui IV-506  
 Yu, Jian II-1  
 Yuizono, Takaya III-464  
 Yukawa, Takashi IV-190  
 Yun, Jiali II-1  
 Yunfei, Zeng III-63  
 Yunus, Mohd. Ridzuan II-464  
 Yusa, Naoki III-572  
 Yusof, Rubiyah I-69, II-464

Yusof, Yuhanis IV-546  
Yuuki, Osamu III-535, III-562  
  
Zamora, Antonio I-31  
Zanni-Merk, C. IV-32  
Zhang, Haoxi II-563

Zhang, Hui I-131  
Zhang, Yan IV-11, IV-135  
Zhang, Yanchun III-398  
Zhou, Yi IV-135  
Ziółkowski, Artur II-543  
Zong, Yu III-398