Ngoc Thanh Nguyen
Chong-Gun Kim
Adam Janiak (Eds.)

LNAI 6591

Intelligent Information and Database Systems

Third International Conference, ACIIDS 2011
Daegu, Korea, April 2011
Proceedings, Part I

Springer

Lecture Notes in Artificial Intelligence 6591
Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

Ngoc Thanh Nguyen Chong-Gun Kim
Adam Janiak (Eds.)

Intelligent Information and Database Systems

Third International Conference, ACIIDS 2011
Daegu, Korea, April 20-22, 2011
Proceedings, Part I

Springer

Volume Editors

Ngoc Thanh Nguyen
Wrocław University of Technology
Institute of Informatics
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
E-mail: ngoc-thanh.nguyen@pwr.edu.pl

Chong-Gun Kim
Yeungnam University
Department of Computer Engineering
Dae-Dong, 712-749 Gyeungsan, Korea
E-mail: cgkim@yu.ac.kr

Adam Janiak
Wrocław University of Technology
Institute of Informatics, Automation and Robotics
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
E-mail: adam.janiak@pwr.wroc.pl

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-20038-0 e-ISBN 978-3-642-20039-7
DOI 10.1007/978-3-642-20039-7
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011923233

CR Subject Classification (1998): I.2, H.3, H.2.8, H.4-5, F.1, K.4

LNCS Sublibrary: SL 7 – Artificial Intelligence

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

ACIIDS 2011 was the third event of the series of international scientific conferences for research and applications in the field of intelligent information and database systems. The aim of ACIIDS 2011 was to provide an international forum for scientific research in the technologies and applications of intelligent information, database systems and their applications. ACIIDS 2011 was co-organized by Yeungnam University (Korea) and Wroclaw University of Technology (Poland) and took place in Deagu (Korea) during April 20–22, 2011. The first two events, ACIIDS 2009 and ACIIDS 2010, took place in Dong Hoi city and Hue city in Vietnam, respectively.

We received more than 310 papers from 27 countries over the world. Each paper was peer reviewed by at least two members of the International Program Committee and International Reviewer Board. Only 110 papers with the highest quality were selected for oral presentation and publication in the two volumes of ACIIDS 2011 proceedings.

The papers included in the proceedings cover the following topics: intelligent database systems, data warehouses and data mining, natural language processing and computational linguistics, Semantic Web, social networks and recommendation systems, collaborative systems and applications, e-bussiness and e-commerce systems, e-learning systems, information modeling and requirements engineering, information retrieval systems, intelligent agents and multi-agent systems, intelligent information systems, intelligent Internet systems, intelligent optimization techniques, object-relational DBMS, ontologies and knowledge sharing, semi-structured and XML database systems, unified modeling language and unified processes, Web services and Semantic Web, computer networks and communication systems.

Accepted and presented papers highlight the new trends and challenges of intelligent information and database systems. The presenters showed how new research could lead to new and innovative applications. We hope you will find these results useful and inspiring for your future research.

We would like to express our sincere thanks to the Honorary Chairs, Tadeusz Więckowski (Rector of Wroclaw University of Technology, Poland), Makoto Nagao (President of National Diet Library, Japan), and Key-Sun Choi (KAIST, Korea) for their support.

Our special thanks go to the Program Co-chairs, all Program and Reviewer Committee members and all the additional reviewers for their valuable efforts in the review process which helped us to guarantee the highest quality of the selected papers for the conference. We cordially thank the organizers and chairs of special sessions, who essentially contribute to the success of the conference.

We also would like to express our thanks to the keynote speakers (Hamido Fujita, Halina Kwaśnicka, Yong-Woo Lee and Eugene Santos Jr.) for their interesting and informative talks of world-class standard.

We cordially thank our main sponsors, Yeungnam University (Korea), Wroclaw University of Technology (Poland) and University of Information Technology (Vietnam). Our special thanks are due also to Springer for publishing the proceedings, and the other sponsors for their kind support.

We wish to thank the members of the Organizing Committee for their very substantial work, especially those who played essential roles: Jason J. Jung, Radosław Katarzyniak (Organizing Chairs) and the members of the Local Organizing Committee for their excellent work.

We cordially thank all the authors for their valuable contributions and other participants of this conference. The conference would not have been possible without them.

Thanks are also due to many experts who contributed to making the event a success.

April 2011

Ngoc Thanh Nguyen
Chong-Gun Kim
Adam Janiak

Conference Organization

Honorary Chairs

Tadeusz Więckowski	Rector of Wroclaw University of Technology, Poland
Key-Sun Choi	KAIST, Korea
Makoto Nagao	President of National Diet Library, Japan

General Co-chairs

Chong-Gun Kim	Yeungnam University, Korea
Adam Janiak	Wroclaw University of Technology, Poland

Program Chair

Ngoc Thanh Nguyen	Wroclaw University of Technology, Poland

Organizing Chairs

Jason J. Jung	Yeungnam University, Korea
Radosław Katarzyniak	Wroclaw University of Technology, Poland

Program Co-chairs

Dosam Hwang	Yeungnam University, Korea
Zbigniew Huzar	Wroclaw University of Technology, Poland
Pankoo Kim	Chosun University, Korea
Edward Szczerbicki	University of Newcastle, Australia
Tzung-Pei Hong	National University of Kaohsiung, Taiwan
Kiem Hoang	University of Information Technology, Vietnam

Special Session Chair

Bogdan Trawiński	Wroclaw University of Technology, Poland

Organizing Committee

Marcin Maleszka	Wroclaw University of Technology, Poland
Bernadetta Mianowska	Wroclaw University of Technology, Poland

Xuan Hau Pham	Yeungnam University, Korea
Adrianna Kozierkiewicz-Hetmańska	Wroclaw University of Technology, Poland
Anna Kozłowska	Wroclaw University of Technology, Poland
Wojciech Lorkiewicz	Wroclaw University of Technology, Poland
Hai Bang Truong	University of Information Technology, Vietnam
Mi-Nyer Jeon	Yeungnam University, Korea

Steering Committee

Ngoc Thanh Nguyen - Chair	Wroclaw University of Technology, Poland
Longbing Cao	University of Technology Sydney, Australia
Adam Grzech	Wroclaw University of Technology, Poland
Tu Bao Ho	Japan Advanced Institute of Science and Technology, Japan
Tzung-Pei Hong	National University of Kaohsiung, Taiwan
Lakhmi C. Jain	University of South Australia, Australia
Geun-Sik Jo	Inha University, Korea
Jason J. Jung	Yeungnam University, Korea
Hoai An Le-Thi	Paul Verlaine University - Metz, France
Antoni Ligęza	AGH University of Science and Technology, Poland
Toyoaki Nishida	Kyoto University, Japan
Leszek Rutkowski	Technical University of Czestochowa, Poland

Keynote Speakers

Hamido Fujita	Iwate Prefectural University, Japan
Halina Kwaśnicka	Wroclaw University of Technology, Poland
Yong-Woo Lee	University of Seoul, Korea
Eugene Santos Jr.	Thayer School of Engineering at Dartmouth College, USA

Special Sessions Organizers

1. *Multiple Model Approach to Machine Learning (MMAML 2011)*

Oscar Cordón	European Centre for Soft Computing, Spain
Przemysław Kazienko	Wroclaw University of Technology, Poland
Bogdan Trawiński	Wroclaw University of Technology, Poland

2. *International Workshop on Intelligent Management and e-Business (IMeB 2011)*

Chulmo Koo	Chosun University, Korea
Jason J. Jung	Yeungnam University, Korea

3. *Intelligent Cloud Computing and Security (ICCS 2011)*

Hyung Jong Kim	Seoul Women's University, Korea
Young Shin Han	SungKyunKwan University, Korea

4. *Modeling and Optimization Techniques for Intelligent Computing in Information Systems and Industrial Engineering (MOT-ISIE)*

Le Thi Hoai An	Paul Verlaine University – Metz, France
Pham Dinh Tao	National Institute for Applied Science-Rouen, France

5. *User Adaptive Systems for Mobile Wireless Systems (UAS 2011)*

Ondrej Krejcar	VŠB-Technical University of Ostrava, Czech Republic
Peter Brida	University of Žilina, Slovakia

6. *International Workshop on Intelligent Context Modeling and Ubiquitous Decision Support System (ICoM-UDSS)*

Kun Chang Lee	Sungkyunkwan University, Korea
Oh Byung Kwon	Kyunghee University, Korea
Jae Kyeong Kim	Kyunghee University, Korea

International Program Committee

El-Houssaine Aghezzaf	Ghent University, Belgium
Le Thi Hoai An	Paul Verlaine University - Metz, France
Costin Badica	University of Craiova, Romania
Maria Bielikova	Slovak University of Technology, Slovakia
Nguyen Thanh Binh	Hue University, Vietnam
Lydie Boudjeloud-Assala	Paul Verlaine University - Metz, France
Stephane Bressane	School of Computing, Singapore
Grażyna Brzykcy	Poznań University of Technology, Poland
The Duy Bui	Vietnam National University, Vietnam
Longbing Cao	University of Technology, Sydney, Australia
Frantisek Capkovic	Slovak Academy of Sciences, Slovakia
Oscar Castillo	Tijuana Institute of Technology, Mexico
Wooi Ping Cheah	Multimedia University, Malaysia
Jr-Shian Chen	Hungkuang University, Taiwan
Shyi-Ming Chen	National Taiwan University of Science and Technology, Taiwan
Suphamit Chittayasothorn	King Mongkut's Institute of Technology, Thailand
Tzu-Fu Chiu	Aletheia University, Taiwan
Kyung-Yong Chung	Sangji University, Korea
Alfredo Cuzzocrea	University of Calabria, Italy

Phan Cong-Vinh	London South Bank University, UK
Ireneusz Czrnowski	Gdynia Maritime University, Poland
Tran Khanh Dang	National University of Ho Chi Minh City, Vietnam
Paul Davidsson	Blekinge Institute of Technology, Sweden
Victor Felea	Alexandru Ioan Cuza University of Iasi, Romania
Jesus Alcala Fernandez	Universidad de Granada, Spain
Pawel Forczmanski	Zachodniopomorski Technical University, Poland
Dariusz Frejlichowski	Zachodniopomorski Technical University, Poland
Patrick Gallinar	UPMC, France
Irene Garrigós	University of Alicante, Spain
Hoang Huu Hanh	Vienna University of Technology, Austria
Jin-Kao Hao	University of Angers, France
Heikki Helin	University of Helsinki, Finland
Kiem Hoang	University of Information Technology, Vietnam
Tzung-Pei Hong	National University of Kaohsiung, Taiwan
Chenn-Jung Huang	National Dong Hwa University, Taiwan
Gordan Jezic	University of Zagreb, Croatia
Robert Kapłon	Wroclaw University of Technology, Poland
Shuaib Karim	Quaid-i-Azam University, Pakistan
Radosław Katarzyniak	Wroclaw University of Technology, Poland
Przemysław Kazienko	Wroclaw University of Technology, Poland
Cheonshik Kim	Sejong University, Korea
Joanna Kolodziej	University of Bielsko-Biala, Poland
Romuald Kotowski	Polish-Japanese Institute of Information Technology, Poland
Ondrej Krejcar	Technical University of Ostrava, Czech Republic
Dariusz Król	Wroclaw University of Technology, Poland
Tomasz Kubik	Wroclaw University of Technology, Poland
Kazuhiro Kuwabara	Ritsumeikan University, Japan
Halina Kwaśnicka	Wroclaw University of Technology, Poland
Raymond Lau	City University of Hong Kong, China
Eunser Lee	Andong National University, Korea
Zne-Jung Lee	National Taiwan University of Science and Technology, Taiwan
Chunshien Li	National Central University, Taiwan
Jose-Norberto Mazón López	University of Alicante, Spain
Marie Luong	Université Paris 13 Nord, France
Urszula Markowska-Kaczmar	Wroclaw University of Technology, Poland
Tadeusz Morzy	Poznań University of Technology, Poland
Kazumi Nakamatsu	University of Hyogo, Japan

Phi Khu Nguyen	University of Information Technology, Vietnam
Grzegorz Nalepa	AGH University of Science and Technology, Poland
Vincent Nguyen	The University of New South Wales, Australia
Toyoaki Nishida	Kyoto University, Japan
Cezary Orłowski	Gdansk University of Technology, Poland
Chung-Ming Ou	Kainan University, Taiwan
Jeng-Shyang Pan	National Kaohsiung University of Applied Sciences, Taiwan
Marcin Paprzycki	Systems Research Institute of the Polish Academy of Sciences, Poland
Do Phuc	University of Information Technology, Vietnam
Bhanu Prasad	Florida Agricultural and Mechanical University, USA
Witold Rekuć	Wroclaw University of Technology, Poland
Ibrahima Sakho	Paul Verlaine University - Metz, France
An-Zen Shih	Jinwen University of Science and Technology, Taiwan
Janusz Sobecki	Wroclaw University of Technology, Poland
Serge Stinckwich	Université de Caen Basse Normandie, France
Pham Dinh Tao	INSA of Rouen, France
Wojciech Thomas	Wroclaw University of Technology, Poland
Bogdan Trawiński	Wroclaw University of Technology, Poland
Hoang Hon Trinh	Ho Chi Minh City University of Technology, Vietnam
Hong-Linh Truong	Vienna University of Technology, Austria
K. Vidyasankar	Memorial University, Canada
Jia-Wen Wang	Nanhua University, Taiwan
Michal Wozniak	Wroclaw University of Technology, Poland
Xin-She Yang	National Physical Laboratory, UK
Wang Yongli	North China Electric Power University, China
Zhongwei Zhang	University of Southern Queensland, Australia

Program Committees of Special Sessions

Special Session on Multiple Model Approach to Machine Learning (MMAML 2011)

Hussein A. Abbass	University of New South Wales, Australia
Ajith Abraham	Norwegian University of Science and Technology, Norway
Jesús Alcalá-Fdez	University of Granada, Spain
Ethem Alpaydin	Bogaziçi University, Turkey
Oscar Castillo	Tijuana Institute of Technology, Mexico
Rung-Ching Chen	Chaoyang University of Technology, Taiwan

Suphamit Chittayasothorn	King Mongkut's Institute of Technology, Thailand
Emilio Corchado	University of Burgos, Spain
Oscar Cordón	European Centre for Soft Computing, Spain
José Alfredo F. Costa	Federal University (UFRN), Brazil
Mustafa Mat Deris	University Tun Hussein Onn Malaysia, Malaysia
Patrick Gallinari	Pierre et Marie Curie University, France
Lawrence O. Hall	University of South Florida, USA
Francisco Herrera	University of Granada, Spain
Frank Hoffmann	TU Dortmund, Germany
Tzung-Pei Hong	National University of Kaohsiung, Taiwan
Hisao Ishibuchi	Osaka Prefecture University, Japan
Yaochu Jin	University of Surrey, UK
Przemysław Kazienko	Wrocław University of Technology, Poland
Yong Seog Kim	Utah State University, USA
Frank Klawonn	Ostfalia University of Applied Sciences, Germany
Kin Keung Lai	City University of Hong Kong, Hong Kong
Mark Last	Ben-Gurion University of the Negev, Israel
Kun Chang Lee	Sungkyunkwan University, Korea
Chunshien Li	National Central University, Taiwan
Heitor S. Lopes	Federal University of Technology Paraná, Brazil
Edwin Lughofer	Johannes Kepler University Linz, Austria
Witold Pedrycz	University of Alberta, Canada
Ke Tang	University of Science and Technology of China, China
Bogdan Trawiński	Wrocław University of Technology, Poland
Olgierd Unold	Wrocław University of Technology, Poland
Michał Wozniak	Wrocław University of Technology, Poland
Zhongwei Zhang	University of Southern Queensland, Australia
Zhi-Hua Zhou	Nanjing University, China

The Third International Workshop on Intelligent Management and e-Business (IMeB 2011)

Chang E. Koh	University of North Texas, USA
Bomil Suh	Sookmyung Women's University, Korea
Hee-Woong Kim	Yonsei University, Korea
Vijay Sugumaran	Oakland University, USA
Angsana Techatassanasoontorn	Pennsylvania State University, USA
Yong Kim	Utah State University, USA
Shu-Chun Ho	National Kaohsiung Normal University, Taiwan
Jae-Nam Lee	Korea University, Korea
Namho Chung	Kyung Hee University, Korea
Joon Koh	Chonnam National University, Korea

Sang-Jun Lee	Chonnam National University, Korea
Kichan Nam	Sogang University, Korea
Bo Choi	Vanderbilt University, USA
GeeWoo Bock	SungKyungwan University, Korea
Dahui Li	University of Minnesota Duluth, USA
Jahyun Goo	Florida Atlantic University, USA
Shane Tombline	Marshall University, USA
Dale Shao	Marshall University, USA
Jae-Kwon Bae	Dongyang University, Korea
Hyunsoo Byun	Backsuk Art University, Korea
Jaeki Song	Texas Tech University, USA
Yongjin Kim	Sogang University, Korea

The Special Session on Intelligent Cloud Computing and Security (ICCS 2011)

Jemal. H. Abawajy	Deakin University, Australia
Ahmet Koltuksuz	Izmir Institute of Technology, Turkey
Brian King	Indiana University Purdue University Indianapolis, USA
Tatsuya Akutsu	Kyoto University, Japan
Minjeong Kim	University of North Carolina, USA
Hae Sang Song	Seowon University, Korea
Won Whoi Huh	Sungkyul University, Korea
Jong Sik Lee	Inha University, Korea
Chungman Seo	University of Arizona, USA
Il-Chul Moon	KAIST, Korea
Hyun Suk Cho	ETRI, Korea
JI Young Park	Ewha Women's University, Korea
Tae-Hoon Kim	Hannam University, Korea
Hee Suk Suh	Korea University of Technology, Korea
Lei SHU	Osaka University, Japan
Mukaddim Pathan	CSIRO, Australia
Al-Sakib Khan Pathan	International Islamic University Malaysia

Recent Advances in Modeling and Optimization Techniques for Intelligent Computing in Information Systems and Industrial Engineering (MOT-ISIE)

El-Houssaine Aghezzaf	Ghent University, Belgium
Le Thi Hoai An	Paul Verlaine University of Metz, France
Lydie Boudjeloud-Assala	Paul Verlaine University of Metz, France
Brieu Conan-Guez	Paul Verlaine University of Metz, France
Alain Gely	Paul Verlaine University of Metz, France

Jin-Kao Hao	University of Angers, France
Proth Jean-Marie	INRIA-Metz, France
Francois-Xavier Jollois	University of Paris V, France
Marie Luong	University of Paris 13, France
Do Thanh Nghi	Enst Brest, France
Vincent Nguyen	The University of New South Wales, Australia
Ibrahima Sakho	Paul Verlaine University of Metz, France
Daniel Singer	Paul Verlaine University of Metz, France
Pham Dinh Tao	Insa of Rouen, France
Bogdan Trawiński	Wrocław University of Technology, Poland

Special Session on User Adaptive Systems for Mobile Wireless Systems (UAS 2011)

Tipu Arvind Ramrekha	Kingston University, London, UK
Peter Brida	University of Zilina, Slovakia
Wei-Chen Cheng	Academia Sinica, Taiwan, Republic of China
Theofilos Chrysikos	University of Patras, Greece
Michael Feld	German Research Center for Artificial Intelligence (DFKI), Germany
Robert Frischer	VSB Technical University of Ostrava, Czech Republic
Jiri Horak	VSB Technical University of Ostrava, Czech Republic
Vladimir Kasik	VSB Technical University of Ostrava, Czech Republic
Aleksander Kostuch	Politechnika Gdanska / Sprint Sp. z o.o., Poland
Stavros Kotsopoulos	University of Patras, Greece
Jiri Kotzian	VSB Technical University of Ostrava, Czech Republic
Ondrej Krejcar	VSB Technical University of Ostrava, Czech Republic
Luca Longo	Trinity College Dublin, Republic of Ireland
Zdenek Machacek	VSB Technical University of Ostrava, Czech Republic
Juraj Machaj	University of Zilina, Slovakia
Norbert Majer	Research Institute of Posts and Telecommunications, Slovakia
Rainer Mautz	Swiss Federal Institute of Technology, Zurich, Switzerland
Marek Penhaker	VSB Technical University of Ostrava, Czech Republic
Stefan Pollak	University of Zilina, Slovakia
Bogdan Trawiński	Wrocław University of Technology, Poland
Jan Vyjidak	Cardiff University, UK
Vladimir Wieser	University of Zilina, Slovakia

International Workshop on Intelligent Context Modeling and Ubiquitous Decision Support System (ICoM-UDSS 2011)

Hyunchul Ahn	Kookmin University, Korea
Michael Beigl	Karlsruhe Institute of Technology, Germany
Seong Wook Chae	NIA, Korea
Namyong Cho	Samsung SDS Co., Korea
Do Young Choi	LG CNS, Korea
Hyng Seung Choo	Sungkyunkwan University, Korea
Namho Chung	Kyung Hee University, Korea
Avelino J. Gonzalez	University of Central Florida, USA
Yousub Hwang	University of Seoul, Korea
Hyea Kyeong Kim	Kyung Hee University, Korea
Jae Kyeong Kim	Kyung Hee University, Korea
Namgyu Kim	Kookmin University, Korea
Hye-Kyeong Ko	Korea Advanced Institute of Science and Technology, Korea
Ohbyung Kwon	Kyung Hee University, Korea
Dongwon Lee	Korea University, Korea
Hyun Jeong Lee	Korea University, Korea
Kun Chang Lee	Sungkyunkwan University, Korea
Sang Ho Lee	Sun Moon University, Korea
Sangjae Lee	Sejong University, Korea
Bong Won Park	Sungkyunkwan University, Korea
Hedda R. Schmidtke	Karlsruhe Institute of Technology (KIT), Germany
Young Wook Seo	NIPA, Republic of Korea
Stephan Sigg	Institute of Operating Systems and Computer Networks, Germany
Bogdan Trawiński	Wrocław University of Technology, Poland

Additional Reviewers

Trong Hai Duong	Inha University, Korea
Bernadetta Mianowska	Wroclaw University of Technology, Poland
Michał Sajkowski	Poznan University of Technology, Poland
Robert Susmaga	Poznan University of Technology, Poland

Table of Contents – Part I

Keynote Speeches

Intelligent Database Systems

Data Warehouses and Data Mining

Natural Language Processing and Computational Linguistics

Semantic Web, Social Networks and Recommendation Systems

Technologies for Intelligent Information Systems

Collaborative Systems and Applications

E-Business and e-Commerce Systems

Table of Contents – Part II

Intelligent Optimization Techniques

Rough Set Based and Fuzzy Set Based Systems

Intelligent Information Retrieval

Computer Vision Techniques

Multiple Model Approach to Machine Learning (MMAML 2011)

Intelligent Cloud Computing and Security (ICCS 2011)

Modelling and Optimization Techniques for Intelligent Computing in Information Systems and Industrial Engineering (MOT-ISIE)

User Adaptive Systems for Mobile Wireless Systems (UAS 2011)

International Workshop on Intelligent Context Modeling and Ubiquitous Decision Support System (ICoM-UDSS)

Virtual Doctor System (VDS): Reasoning Challenges for Simple Case Diagnosis Based on Ontologies Alignment

Hamido Fujita, Jun Hakura, and Masaki Kurematsu

Iwate Prefectuarl University,
Intelligent Software Systems Laboratory, 152-52 Sugo, Takizawa, Iwate-gun,
Iwate, 020-0173, Japan
{issam,hakura,kure}@iwate-pu.ac.jp

Abstract. Human computer Interaction based on emotional modelling and physical views, collectively; is investigated and reported in this paper. Two types of ontologies have been presented to formalize a patient state: Mental Ontology reflecting the patient mental behavior due to certain disorder and Physical Ontology reflecting the observed physical behavior exhibited through disorder. These two types of ontology have been mapped and aligned using a simple Bayesian Network for causal reasoning to define what we call as simple case diagnosis. We have constructed an integrated computerized model which reflects a human diagnostician as computer model and through it; an integrated interaction between that model and the real human user (patinet) us utlilized for 1st stage dignosis purposes.

1 Introduction

This paper is reporting on issues that we have disused in previous version on our project related to what we called virtual doctor system [2] [1], that is to design avatar that resemble a real human doctor and acts to interact with patient user to establish a diagnosis scenarios based on patient interactive procedural routine. The system outline is shown on Fig.1.

We have created a related technology, reflecting the state of art on creating a a system that resembles the user mental psychological behavior through a face, this concept is called as mental cloning [1]. The mental cloning is utilized to build the avatar reflecting a real person, the animated real-time images are created in real-time on this avatar for resembling the emotional behavior of that person reflected through this avatar in the same manner the real person interacts with certain world in similar invocation. This is represented by using that person ego state [3][6].

In this paper the system is expanded to resemble a medical doctor that interacts with human patient for medical diagnosis. The interoperability represented by utilizing the medical diagnosis of medical doctor in machine executable fashion based on patient interaction with virtual avatar resembling a real doctor. The Virtual Doctor

[1] A version of this paper is published in [2]; as Plenary talk CINTI_2010.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 1–13, 2011.

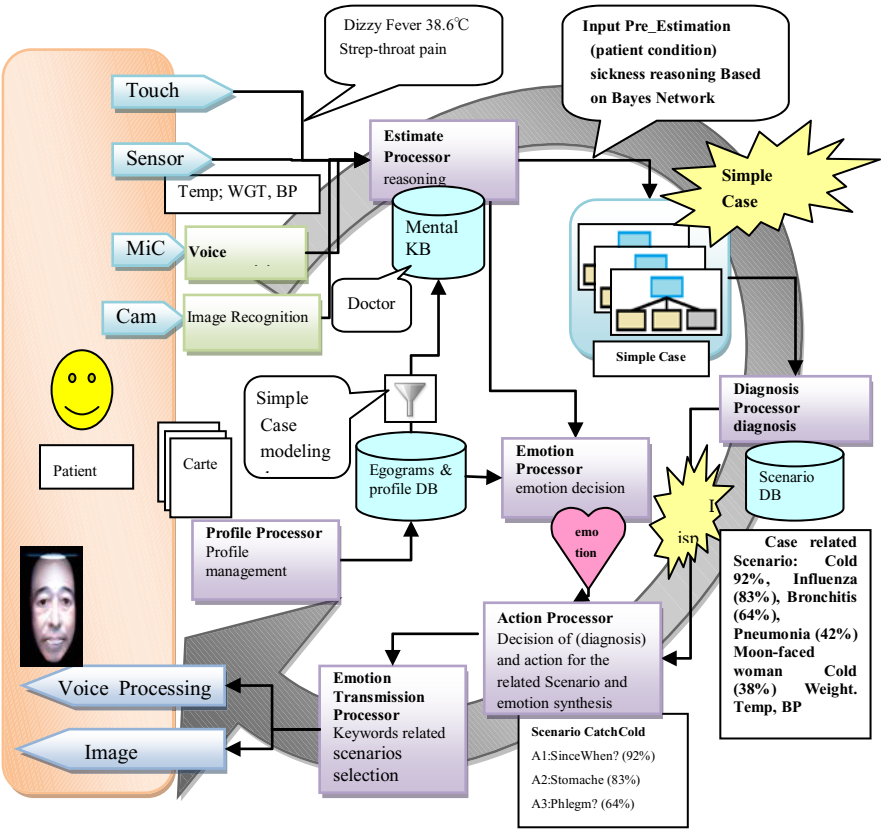


Fig. 1. The VDS system outline

System (VDS) is installed in a local hospital in Morioka (Iwate-Japan) where that doctor is regularly, practicing her medical diagnosis in real situation and environment. The avatar or VDS is working as a 1st glance diagnosis to classify patient based on the criticality and emergence based on examination parameters and diagnosis scenarios.

2 VDS Outline

Medical scenarios are defined on general guidelines formalization, and customization according to the subject doctor experience and related specialization on the course of medical practices. The system is to help the real doctor by filtering the outpatient (when they come to the hospital) waiting to see the real doctor. The virtual doctor sees patients by interacting with them and issues a decision making for simple cases and non-simple cases categorical analysis. The *simple medical case* in our context; is defined as the case that usually medical doctor reaches through medical diagnosis, such case is considered by medical Doctor namely A, as a state that the outpatient can be recovered by taking a rest, or simple medical supplement, a case resulted from

stress, heavy work or tiredness. The simple case treatment is in most cases is to ask the patient to rest and come back after few days if the recovery is not achieved or correlated physical (body) phenomenal sign is emerged, or stayed (not relieved). We have selected the *simple case* approach due to the following: We can test our system and its design reasoning framework. We use the system for helping the medical doctor to classify medical cases for outpatient based on criticality issues. Criticality issues are estimation of the outpatient sickness state. This is based on his/her mental and physical reasoning that is achieved (i.e., reached) collectively (inferred) by the VDS. The system outlined in this paper is structured and presented in Sec. 2; showing the system outline related to the reasoning aspect. We have two concepts that the system uses to do reasoning. One is called Physical Ontology, and the other is called Mental Ontology. We define the concept of "pain" as part of mental ontology; we implement the system using Owl (Ontology web language). The two represented ontologies are merged to yield a third ontology called Medical, oncology, representing the diagnosis processor in Fig. 1. The action processor is the decision making that uses Bayesian Network show on Fig. 2 to calculate the Conditional Probability Table (CPT) that assists the action processor to select the best scenarios that would be readout to the patient in the same manner as the real doctor does. On [6] we appended several video that demonstrate how the VDS interact with the patient and how the system's camera eyeing on patient face extracts values to reason through on patient state.

http://www.somet.soft.iwate-pu.ac.jp/system_news/All_News_600MB.rar:

TV news on our system, 600MB file you can download.

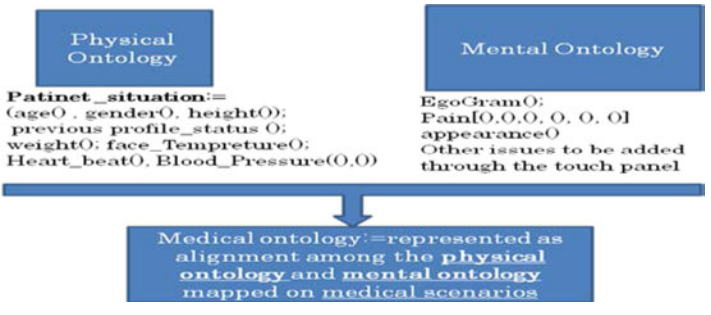


Fig. 2. Outline of VDS conceptual reasoning

3 Cognitive Based Intruction for VDS

There are several parts related to the VDS interaction modules that have already reported in published work related to voice face emotional recognition [4][5] and voice emotional recognition [6][8]. We have built a concept we call it as mental cloning [1] we could collect user emotion and mentality reflected through face and voice to understand the mental state of the user. These are used for creating output for the avatar and input data due to user emotional change (engagement) with the

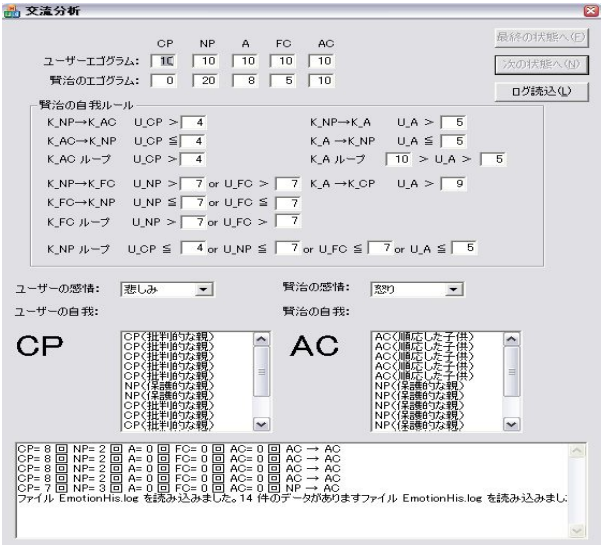


Fig. 3. Emotion Processor sample window

avatar [2]. The reasoning issues are not reported yet, and this paper is discussing the reasoning issues related to VDS application for diagnosis. We have restricted our diagnosis reported in this paper to *simple cases*. The context of the engagement is defined in advance (In this paper the context is medical diagnosis based on Doctor A). So collecting the user mental state is to have the system adapt to changes that would have the user be engaged with the system in a positive manner (forwarding interactive style of communication).

The avatar in our system is virtually constructed to resemble real Doctor or person who is the object the patient uses to interact. We built a mask face model for real doctor practicing his/her medical provision in hospital in our town [2][1]. Such face is used to interact with the user in emotional based manner [6]. The face would smile or else and act in emotional manner according to the context and engagement style of the user. The face mental background resembles an ego state reflected through the egogram resembled person (medical doctor) and represented in the system as a program [6]. We have studied this aspect and we created a program that can interact with the user using transactional analysis [3]. The face states are the primitive states that the system would select interactively according to the user engagement cognitive states (shown in Fig.3). The user ego state is also collected from the best match from the database based on what we called universal template. A set of egograms are stored in the system and indexed according to universal templates [3]. We have evaluated these universal templates based on experimenting them with Miyazawa Kenji avatar that experimented in A museum [6]. These stored classified ego grams are to work as a templates to test user ego states (emotion). User observed ego state is measured through a set of universal templates. The measurement of face parts movements are referenced (computed) to a indexed templates collected from many Japanese subject

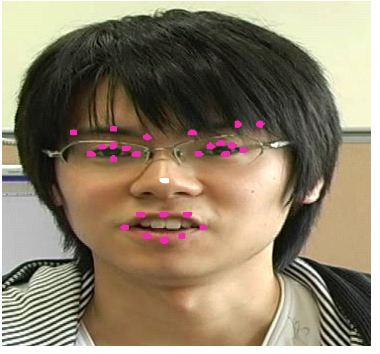


Fig. 4. User face seen by the system

(people) contributed in our experiment [3][5][6]. On [6] you can view (download to view) the movies showing how the system works and also for public news on the project, (in Japanese media). The system would test the mental states of the user based on these ego grams, and interact with the user based on instantiation of observed changes on the face parts collected due to emotional reasoning based engagement. Also, the same is done on the voice as well. The voice emotional features are examined to reflect the patient voice sound features like soaring throat related sound feature, or related expression to pain or else. The same also, is for expressing the dialogue with patient by doctors with emotional voice to patient synchronized with their mental situation.

However, in this paper; we have not presented patient voice emotional recognition or VDS generated emotional readout diagnosis.

Taken all these technologies into account we are reporting in this paper, using this established technology, we have examined and experimented reasoning related matter reflected to medical use-case provided by two medical doctors. We have built the system that ensemble a medical doctor interacting with the patient based on the established technology simply outlined above, to do diagnosis on patient at clinic or hospital in Morioka city, Japan. The system is working as a filter (or sorter) to do the 1st diagnosis based on provided medical guidelines specialized by these two nominated medical doctors working in that clinic. This is especially useful in Japanese local hospitals when patients usually wait for one to two hours to see the doctor (human physician). The system would assist the hospital to set patient into simple cases (with category, (without physical observation) and no-simple cases, (needs doctor observation).The issue that reported in here are related to new findings that we have collected in relation to VDS is using the mental cloning concept and the avatar technology together to construct a copy of medical doctor avatar [8][9]. We have experimented two types of medical doctors. Dr. A in hospital A, and Dr. B in Hospital B, both in Morioka city, Japan. These two hospitals' doctors' avatars and faces are constructed. The avatar face and voice with related diagnosis scenarios on "simple cases" have constructed. We have selected these two doctors based on the style of their diagnosis. As Dr. A uses patient appearance in reasoning and diagnosis (with certain physical touch), while Dr. B uses egogram based certification to analyze

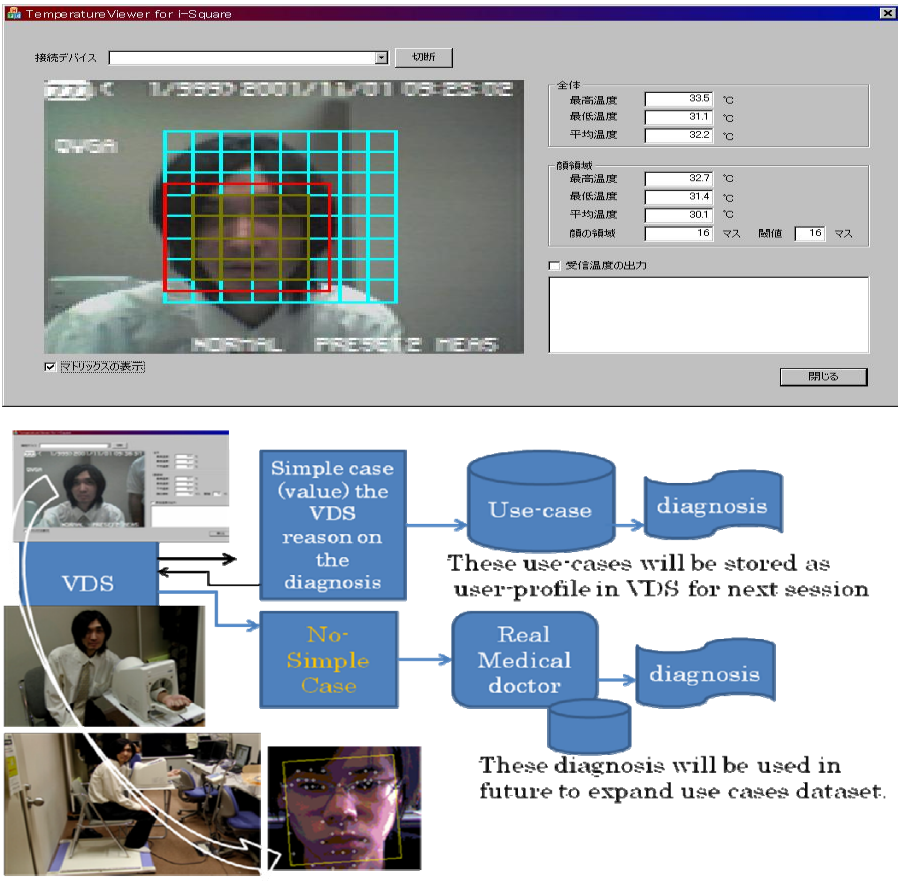


Fig. 5. The physical ontology and VDS related simple case

the patient mental states and do diagnosis on patients' condition by navigating in these states, through specific scenarios and networked style of decision making. She integrates all these decision based on her experience, represented as decision network style. These two instances of MD style of reasoning (diagnosis) are examined and represented in the VDS system based on provided instances of simple cases medical practices data.

3.1 Simple Case

We define what is the simple case, and what are the formal guidelines defining the simple case (Fig. 5). The relative customization of such simple cases is due to the doctor experiences in diagnosis. It is those cases that the MD concludes them as not critical symptoms, or symptoms that may need later on further observation, or situation that is not necessarily be recovered by medication or surgery or else. A relative medical advice, or supplementary medication supporting the medical case in

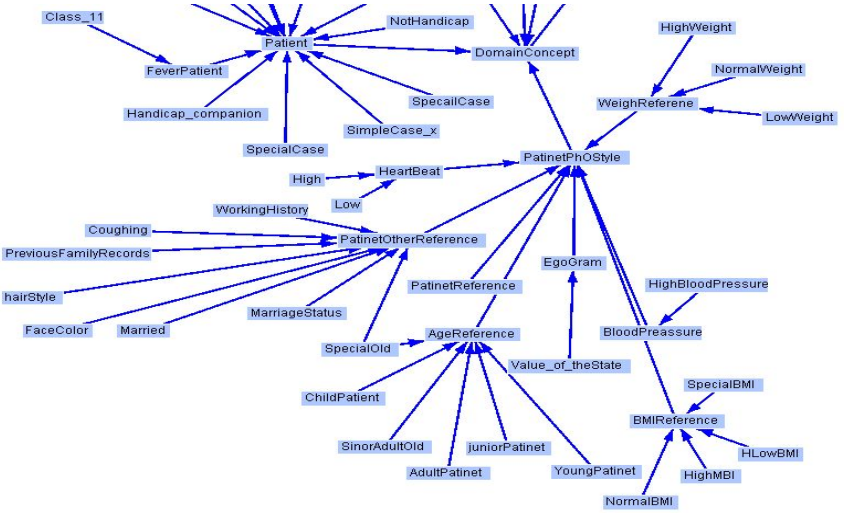


Fig. 6. Shown the created ontology visualized by growl

hand, or/and appointment to come again to confirm the sustainability of such case, are to be provided as MD for these called simple case.

Ontology can provide means to allow human to understand meanings of the elements and concepts or things in defining the problem, and also a means to reason on these classified items, through such semantically based representation. Semantic technologies contribute to provide machine-executable metadata for reasoning purposes.

3.2 Reasoning on Simple Cases Ontology

The conceptual reasoning framework is based on presenting two types of ontologies, reflecting patient (user) physical conceptual status as seen by the VDS, and as shown in Fig. 5. The simple cases are those medical cases that to be diagnosed by the MD as non-critical that can be resolved through a rest of else. To diagnosis these cases, MD needs to have collective views on patient situation integrated with physical coverage. We have automated these views using PhO and MeO enumerated on medical ontology. We have used a combination of probabilistic techniques, ontology representation and inference to determine the *simple case* we defined in our context. The target of the system is to identify the patient case as simple, with weight. (Simple_?), where? is: High, medium and low, or not-simple. The decision of taken as "not simple" means to go to the next room to see the real doctor. The related technical reasoning is also sent to the real doctor.

We have defined two types of ontologies (as shown in Fig.2).

Physical Ontology definition: PhO.
Mental Ontology Definition: MeO.

Each ontology represents causal relation articulated from physical view analysis, and mental view analysis. These are shown in Fig.6. and Fig.7.

Probabilistic model has been used to reference and infer to doctor diagnosis. These parts is taking use of alignment of the two defined ontologies, and do diagnosis based on probabilistic calculation to compute values that would be used to make the decision related to special cases, and this is modeled on Bayesian network. Inference uses the Bayesian network (as belief network) aligned and reflected on these two ontologies. On Fig. 5 we can see the big arrow at the left side. This arrow is reflecting to collected data through the user interface to articulate in collective manner the PhO issue, aligned with the MeO.

User situation classify the meaning of metadata based on (gender, egogram, age, history) [2]. The situations are represented by [gender, age, egogram] These are variables, acting as values that in a collective manner classify the medical scenarios of the mentioned two types of nominated Doctors examples. These values articulate to diagnosis in probabilistic manner to reflect the aligned mental view and physical view. The both are aligned to articulate on status of the outpatient in probabilistic manner to reason through on probabilistic combination of symptoms modeled as belief networks that is used to find the related remedy. The system shown in Fig. 1, there are example on probabilistic causal reasoning through Medical doctor approximation to treat the patient case.

The history is defined based on last state of the patient state. If the last state is simple, then we carried on. If not simple then continue with extracting data and send these to real doctor, for real-diagnosis.

The variables are those values defined by the PhO and MeO views. These variables are any values among 0 and 1. These values are representing the temperature in relation to threshold values (representing normal situation articulated on user situation). The total weight total of these values should be one.

Symptoms type Physical: Are those symptoms observed on the patient by devices or previous documented data (Fig. 5). In our system, we have the patient be seated on a chair with three types of devices that read: The body weight, temperature distribution on the face, and blood pressure. There are also other data that can be collected from previous history or document, referencing to previous physical state and articulate the new state.

Symptoms type Mental: These are the observed behavioral patterns on the patient face, articulated through templates to reflect the mental state of the patient, if she/he is in a pain or a sort of situation. These are classified according to the ontologies and as shown on Fig.6.

These above two situations each are reflected and represented on ontology reflecting the medical ontology specified by the two medical doctors and specialized by the difference in their ontology in patient diagnosis. The Symptoms reflected on Physical ontology are those reflected on mental ontology are mutually independent. The medical ontology represents the conceptual (abstract) view of medical diagnosis. The view is specialized by the doctor type, reflected as a model speciation, on the usage of the PhO and MeO in diagnosis. The simple case is defined in conceptual view and generalized form, the specialization due to the type definition of simple cases according to the doctor experiences is represented by on medical ontology. The style of reasoning diagnosis is also relative to the doctor diagnosis ontology (as a specialization to medical ontology).

The variables outcome would infer to the medical cases and invoke certain scenarios. These scenarios are explained to the outpatient as question or comments, expressed by the VDS in the same manner the real doctor does.

All variables values are computed based on mental and physical observation in the model. Variables are computed and collected by initiating scenarios with outpatient to collect these causes' values (probabilistic) for decision making.

With these medical doctors help we have established several cases resembling different types of configuration of variables. These configurations can resemble different types of simple cases. We could create more complicated configuration in the same manner, however this would make the Bayesian network more complicated. However, at the moment our interest is to build the system and test it in these two hospitals. There are several issues that need to examine from real experiment. Not only the inference complexity, but the practicality on having such system in medical practices for simple case scenarios needs to be investigated.

The probability weight of each symptom is calculated with weights representing the importance of that variable. All variables in the PhO, MeO have classified weight reflecting the structure of medical knowledge. The knowledge base of medical knowledge is categorized and classified base on these variables importance as part of the two defined ontologies. The evaluation is not numerical but subjectively qualitative and oracle.

We have created an evaluation as 30 outpatients, diagnoses were done by the system: 20 patients participated in the experiment: 5 simple diagnosed by the system among 20. The true Doctor diagnoses were 8 simple (the above 5 are included). Our doctors are looking to these outcomes, in aggressive manner as it has participated in assisting them to filter the patient cases. We could validate and tune the system based on several carried case studies to define to refine the simple case related issues through the two presented ontologies. This is to have MeO, be mapped with related aspect collected from the input to related aspect from the PhO, these two are aligned and mapped on medical ontology (MedO), in order to specify and map (alignment) between the above two ontologies through causal relation based on Bayesian Network which is proved to be useful in reasoning on imperfect knowledge like these related to medical diagnosis on patient 'cases.

The result would produce a keyword (or a set of data) or a statement as query that infer through the reasoner for the related scenarios that can reflect a nest of related constraints. The medical ontology alignment is a conceptual view of simple medical case. The simple case ontology is defined conceptually, in a generalized form; and the specialization on it is due to the type definition of different invariants related to simple cases extracted due to the doctor experiences. The style of reasoning diagnosis is also relative to doctor diagnosis's ontology.

A Bayesian network is a graphical model for probabilistic relationships among a set of variables represented using a directed acyclic graph (DAG), that its arrows represent causal influence among its nodes. The Noisy-OR network has the assumption that the model makes use of causal independence among the modeled causes and their common effect. The word noisy reflects that the interaction among the causes and the effect is not deterministic, so it is not possible to capture all the possible causes of an effect.

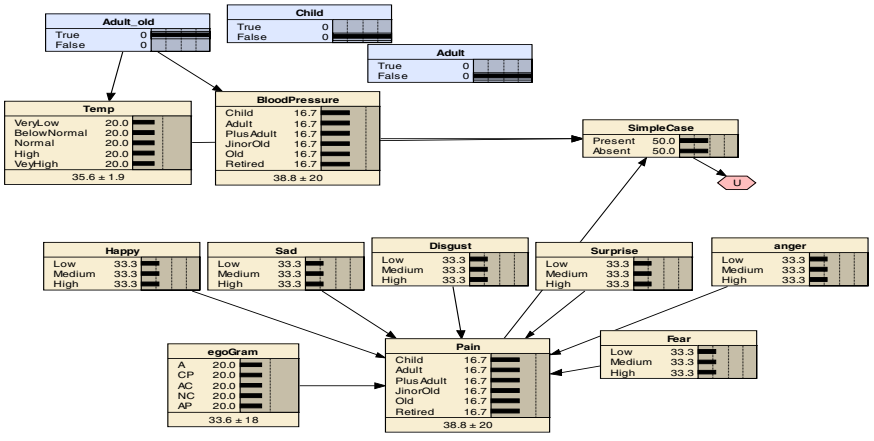


Fig. 7. Shows an example of belief network

We have used Netica[9] to construct the BN. Netica allows network construction and parameter learning from data. We derived the parameters for conditional probability values from medical studies related to simple cases observations. We have used causal independence model. As in the following section: This based on causal independence need to facilitate through the conditional probability theory $P(A_i \setminus f(A_i))$.

A diagnosis is instantiated from physical variables, (ph_1, ph_2,…ph_n) and mental variables (me_1, me_2,…me_n). These variables are mutual independent. There are also effect that articulated from each variables reflected from corresponding ontology, and all the summation of these effect would lead to medical scenario conclusion. The medical scenario is also can lead to other set of physical and mental variables and also lead to other medical scenario. It is an integrative causal reasoning based on Bayesian network.

$$P[(med(a) \setminus me_1,..me_n, ph_1,..ph_n)] =$$

$$\sum_i^n P(med(a) \setminus me_i) \sum_i^n P(med(a) \setminus ph_i) \prod_i^n P(ef_i \setminus me_i) \prod_i^n P(Ef_i \setminus ph_i)$$

where Ef or EF are Boolean value as either true or False mapped to the Med(a). Also these ef(i) and Ef(i) are conditionally independent between the both, as the signs (symptoms) resulted from the physical effect and mental effect are together due to certain disorder. These are not necessarily correlated by their relation on the medical diagnosis reflected in the above formula to establish decision making for causal reasoning on diagnosis.For example, *mild* blood pressure, with high_mild disgust, could lead to mild stress. It is a type of Simple case as reflected from MDr. A. We have noticed that; this is a sort of belief network for a noisy-OR type causal reasoning in context-specific independence [7].

The noisy-OR (ef(i)) is like a regular OR function, and all its parents are binary values in [0,1]. As we have the PhO and MeO symptoms related nodes are causal

independence related to ef(i), therefore, $P[(med(a) \setminus me_1,..me_n, ph_1,..ph_n)]$ is represented by terms of noisy hidden variables ef(i). Therefore we can calculate the conditional probability of all states reflected in Fig.8's example. The nodes in Fig.9, (believe networks concept based on the aligned ontology) are of two types: discrete nodes representing either symptoms observable or not. It is a binary value. There is continuous node representing values like temperature or weight (i.e., PhO), and these values are represented as conditional probability distribution. In other words, the temperature, for example is converted as high in relation to symptoms or as low. This is possible by storing conditional mean and variance in each decision node namely; the MeO (i) and phO(i).

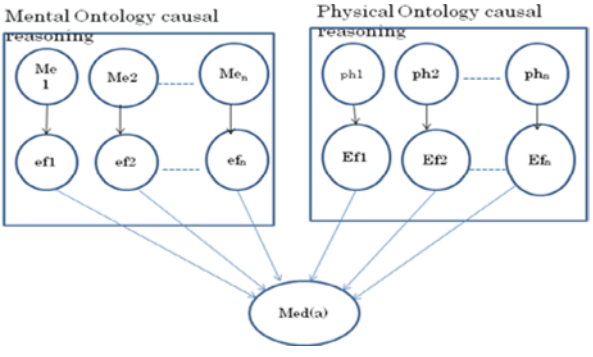


Fig. 8. The Bayesian network general concept

Fig.2. resembles the type of reasoning to the simple case causal based reasoning. The causal network calculates to reach the appropriate decision through the different dataset provided by these two doctors.

The analysis using the above formula is to calculate the probability of medical diagnosis based on collecting the effects: (ef(i)) related to MeO) and (Ef(i) related to PhO), respectively as shown on Fig.2. The collected computed probability is according to threshold values specifying the patient's condition as either simple case or not. For simple case, the system would advise the patient on his/her condition using a set of Simple cases categorized scenarios organized due to the physical and mental data, For example; temperature combination values (PhO) and disgust values (MeO). The conclusion will be presented to the patient as readout scenarios in the same manner as the real doctor does. For any other case, concluded as not simple the system may advise the patient to move to the next door to see the real doctor. The diagnosis scenarios and copy of it will be transferred as a file to the real doctor's computer. We have used Netica application [9] to build the Bayesian network reasoning engine. In addition, we use Netica Java application (API) to do the application that interact and reason on values for diagnosis pipilned with the OWL reasoning engine. The implementation has done using JAVA with Netica connected with Jess reasoner. The program resembling the two ontologies based inference (Fig.8) build the Bayesian Network in the memory for decision making and reflect

this in the reasoner such that to have the diagnosis processor (Fig.1) to select the appropriate scenarios according to Medical doctor's position participating in the analysis.

4 Conclusion

As Fig. 7 shows an example on how to use Netica to have the network learn the case data to do decision making based on samples collected from patient reading. In Fig. 7: The decision nodes specifying the user situation (age, gender), which is connected to continuous node that select the appropriate range of blood pressure (for example) according to age, due to provided tables that are reflected through these continuous node. The same is also valid for temperature. The belief network for pain (as Fig. 7 example), is a ranged values combination among 6 emotional primitives that are articulated (due to age and gender) and are to reflect on emotional status of pain in belief network using Netica. The reasoning is based on OWL-DL [2] by defining property, with restriction represented using OWL This would support ontology reasoning as Bayesian inference. The usage of Bayesian network or belief networks was to overcome the complexity in the reasoner when the concepts have some fuzzy or probabilistic types of restrictions.

The medical knowledge as case-based are embedded in OWL as property relation and restriction definitions, and are used to define the classification of mental axioms and physical axioms represented in TBOX. The individuals that have certain restriction and properties like male with age 60 has some values of regular cough. These types of restriction are modeled in DL as ABOX, as individual with certain properties. RCAER pro reasoner has been investigated and used on this aspect. We have used it as reasoner to classify the related individual and properties for certain individual based on certain restriction provided by the diagnosis. Racer was useful for limited set of axioms related to a collection of TBOX with certain property that could classify the related ABOX. The created graph that is used to bring the inference of TBX in relation to ABOX was rather time consuming and sometimes failure when the number of axioms are increased. So we need to limit the number of TBOX to use the RCAER style reasoning to create the network for deriving the CPT. Other efficient way reported in this paper was to employ the Bayesian network creating the CPT to select and compute the individual restriction based on the case-based reasoning using OWL-DL.

We have tested the system in one hospital where Dr. A is working. The system could conclude simple cases based on Dr. A provided case data. However, there are some problems that are majorly not technical and related to patient education to use such system. We found this is major obstacles however, this is not related to the technical aspect of the work. Combining the both otology in classifying medical data and related diagnosis as a representation, and use that representation to built belief network to learn on decision making using the Medical doctor experience, all these are unique characteristics of this work. To have a dataset on simple cases to set the decision making based on the two provided ontology utilized in medical application domain.

Acknowledgment

This research is supported by the Ministry of Internal Affairs and Communications of Japan under the Strategic Information and Communications R&D Promotion Programme (SCOPE). We would like to give our gratitude to the Medical Doctors Committee board of VDS of SCOPE project as the medical application board advisory of this research, who are providing experience and advises on their medical analysis on patients diagnosis, and simple cases outline senarios.

References

1. Fujita, H., Hakura, J., Kurematsu, M.: Intelligent human interface based on mental cloning-based software. International Journal on Knowledge-Based Systems 22(3), 216–234 (2009)
2. Fujita, H., Hakura, J., Kurematsu, M.: Invited Talk: Multiviews ontologies alignment for medical based reasoning ontology based reasoning for VDS. In: 2010 IEEE 11th International Symposium on Computational Intelligence and Informatics (CINTI), pp. 15–22 (November 18-20, 2010), doi:10.1109/CINTI.2010.5672279
3. Hakura, J., Kurematsu, M., Fujita, H.: An Exploration toward Emotion Estimation from Facial Expressions for Systems with Quasi-Personality. International Journal of Circuits, Systems And Signal Processing 1(2), 137–144 (2008)
4. Hakura, J., Kurematsu, M., Fujita, H.: Facial Expression Invariants for Estimating Mental States of Person. In: New Trends in Software Methodologies, tools and Techniques (SoMeT 2009). Frontiers in Artificial Intelligence and application series, vol. 199, IOS press, Amsterdam (2009), doi:10.3233/978-1-60750-049-0-518
5. Kurematsu, M., Ohashi, M., Kinoshita, O., Hakura, J., Fujita, H.: An Approach to implement Listeners Estimate Emotion in Speech. In: New Trends in Software Methodologies, tools and Techniques (SoMeT 2009). Frontiers in Artificial Intelligence and application series, vol. 199, IOS press, Amsterdam (2009), doi:10.3233/978-1-60750-049-0-531
6. `http://www.somet.soft.iwate-pu.ac.jp/system_news/All_News_600MB.rar` 600MB movie, `http://www.somet.soft.iwate-pu.ac.jp/system_news/System_flow_operation_mpeg.rar`, `http://www.somet.soft.iwate-pu.ac.jp/system_news/VDS_sample2.rar` 100MB
7. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Francisco (1988)
8. Kurematsu, K., Chiba, H., Hakura, J., Fujita, H.: A Framework of Emotional Speech Synthesize Using a Chord and a Scale. In: New Trends in Software Methodologies, tools and Techniques, SoMeT 2010, vol. 217, IOS press, Amsterdam (2010), doi:10.3233/978-1-60750-629-4-500
9. Netica software package: NORSYS Software Corporation, `http://www.norsys.com/`

Image Similarities on the Basis of Visual Content – An Attempt to Bridge the Semantic Gap

Halina Kwasnicka[1], Mariusz Paradowski[1], Michal Stanek[1], Michal Spytkowski[1], and Andrzej Sluzek[2]

[1] Wroclaw University of Technology, Wroclaw, Poland
{halina.kwasnicka,mariusz.paradowski,michal.stanek}@pwr.wroc.pl
[2] School of Computer Engineering, Nanyang Technological University, Singapore

Abstract. Image similarities is a useful concept regarding to the image retrieval on the basis of visual content of the images (CBIR - **C**ontent **B**ased **I**mage **R**etrieval). Because an image can have far more interpretations than text, visual similarity can be totally different from semantic similarity. We have developed similar images searching tools using global approaches as well as local approaches to find near similar images. In this paper we propose a method of bridging local and global levels, what should solve the problem of limited, non-adaptable dictionary when we use automatic annotations in a similar images retrieving task. Our far-away goal is to face the difficult problem with all current approaches to CBIR systems, connected with visual similarity: the semantic gap between low-level content and higher-level concepts.

1 Introduction

The concept of similarity plays a key role in image analysis and, more specific, in image *retrieval*. Viable formulation of *image similarity* allows effectively recognizing and retrieving images with related content. The general concept of image similarity is vague and may be defined in multiple different ways. Man is able to determine the mutual similarity of two images shown to him. Also he is able to find the similar image to a given one, but this image is *similar in his view*. So, the term *images similarity* is not precise, it is very subjective when is considered by people. Let us see the formal definition of word 'similarity', defined in the American Heritage Dictionary [29]: ***similarity*** – is quality or condition of being similar; resemblance. Following words are the synonyms: *likeness*, *similarity*, *similitude*, *resemblance*, *analogy*, *affinity*. These words denote agreement or conformity, the *likeness* implies close agreement, *similarity* and *similitude* suggest agreement **only in some respects** or **to some degree**, while *resemblance* refers to similarity in external or superficial details. *Analogy* means similarity "as of properties or functions, between things that are otherwise not comparable". The last word, *affinity* is a "likeness deriving from kinship or from the possession of shared properties or sympathies" [29]. What we want is **to design a computer system which will be able to find similar images to a given one**. Taking into account the described meanings, this task is very difficult. One

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 14–26, 2011.

can expect that such computer system will be imprecise and, possibly, it will require be tuned for particular users.

Summing the above, different people consider different images as similar and would like to get different retrieval results. Images may be considered similar if they: have the same interpretation, share the same object(s), evoke the same emotions, have identical spatial arrangement, share the same colors or textures, have identical fragments, etc. Some interpretations of the mentioned similarity concept may be modeled using the object *recognition paradigm*, i.e. intelligent techniques. To make the situation even more complex, we also need to take into account the human perspective and expectations [24]. Thus, machine learning paradigm seems to be a reasonable solution to the problem of image similarity measurement.

Application of object recognition paradigm to image retrieval may be successfully implemented using the similarity of recognized concepts. Instead of low-level, pixel based queries, the user is able to formulate meaningful, concept based queries [14]. This image retrieval scheme is sometimes referred to as ***Annotation Based Image Retrieval*** [11] in contrast to classic ***Content Based Image Retrieval***. Despite its multiple advantages, researchers point out the key disadvantage of such approach: the number of concepts is *predefined* and *finite* [28]. This property makes the object recognition based paradigm inapplicable when faced with infinite diversity of the surrounding world [7]. Effective image retrieval may require continuous creation of new concepts which describe the environment in a precise way.

The paper is a continuation of our research on the mentioned problem. We seek how to automatically create new concepts without any a'priori information, purely on a visual basis [23] and seamlessly integrate them into the notion of image similarity. The presented idea consists of multiple subcomponents, solving various subproblems, but it may not be yet considered as complete or functional. Thus, in this paper we do not give clear answers, but rather we present our most recent ideas. Some methods which are regarded as components of the proposed idea were developed and studied [1,17,27].

2 Global Image Analysis

The first component of the presented solution is a global image analysis method where we are interested in extracting general, holistic image features. Such features are easy to generalize and efficient for processing by intelligent approaches. We may simply build image recognition methods based on global features and accompanying labels. These recognition methods may be image distance based which turns to be a quite effective approach. Having a set of labels, the image similarity retrieval becomes a text based retrieval. However, as mentioned above, we face the problem of finite, limited and non-adaptable dictionary.

2.1 Global Image Distances

Automatic methods of images analysis define *image similarity* as a distance measure between images, which is a sum of distances between global visual features of considered images. To obtain the similarity between two images one can measure the distance between visual vectors in metric or probabilistic space. Minkowski, Cosine, Correlation, Mahalanobis or EMD are commonly used measures to calculate distances between visual features. The other approaches use divergence between image probabilistic models calculated for the set of visual features. In that category commonly used measure is Kullback-Leibler divergence or its symmetric version Jehnsen-Shannon divergence.

Visual features of an image define its certain visual property. Global features capture some overall characteristics of an image, as color, texture and shape. An image can be divided into a number of sub-images; in such approach, the whole image is described by a vector of features calculated for each sub-image. The global approach has one important advantage: the high speed, both features extraction and similarity measure calculation [6]. However, the global features are usually too rigid to represent an image.

The second approach is extraction of local features, computed for every pixel using its neighborhood. Additional step, features summarization must be performed. Often data set based on a distribution for each pixel is calculated in summarization step.

Some features from MPEG-7 standard, as histogram-based descriptors, spatial color descriptors and texture descriptors seem to be well suited for natural images retrieval [4,5,6].

2.2 Automatic Image Annotation

Automated **I**mage **A**nnotation (AIA) is a process which describes previously unseen image $\boldsymbol{Q}$ by a set of concepts $\{w_1, w_2, \cdots, w_N\}$ from the semantic dictionary $\boldsymbol{D}$. Word assignment can be made by finding the correlation between visual features which characterize query image $\boldsymbol{Q}$ and high-level semantics (concepts). AIA is an integral part of modern CBIR systems. Image annotations can be seen as a bridge between textual queries and visual image content.

Machine learning techniques used to solve the AIA problem can be split into *classification* based methods and *probabilistic modeling* methods. Classification methods lie on training classifiers to recognize if a given word is present within the proper description of the image. Different classifiers can be used in this approach, good result and speed one can obtain with decision trees [13,22].

Probabilistic modeling methods, such as Hierarchical Probabilistic Mixture Model (HPMM) [10], Translation Model (TM) [8], Supervised Multi-Class Labeling (SML) [3], Continuous Relevance Model (CRM) [14] and Multiple Bernoulli Relevance Models (MBRM) [9], try to find the probability density function of visual features associated to concepts. Parametric or non parametric estimation can be used in this approach.

Results obtained by AIA methods can be further improved by using filter methods which take into account word co-occurrence models [15], words relations in Word-Net [12] or our GRWCO [13] method which reduces the difference between expected and resulted word count vectors to reranking the output annotations. Recently, Makadia et al. have proposed a new method based on the hypothesis that similar images are likely to share the same annotations [19]. In this approach, an image annotation is a process of transferring most frequent labels from nearest neighbors. The method **does not solve the fundamental problem of determining the number of annotations** that should be assigned to the target image, it assumes that the optimal annotation length is given.

In our recent research, we have extended this approach. We have proposed PATSI (**P**hoto **A**nnotation **T**hrough Finding **S**imilar **I**mages) annotator which introduces transfer function [27] as well as an optimization algorithm which can be used to find both, the optimal number of neighbors and the best transfer threshold according to the specified quality measure [17]. PATSI consists of two main phases: *preparation* and *query*. In the first phase, for each image repeat: (1) split the image into a number of regions (sub-images); (2 calculate statistical visual features for every region (sub-image); (3) create the model of the image. In the query phase, do: (1) split the query image into regions (sub-images); (2) build a model of the query image; (3) calculate distances between the query image and all images in the dataset; (4) Select k most similar images; (5) Transfer all words (annotations) with a weight dependent on a position of a considered image in a similarity ranking list (how much the image is similar to the query image); (6) Select words with sum of weights greater than the assumed threshold t. These words are the annotations of the query image. The more detailed description of the PATSI algorithm one can find in [17] and [27].

2.3 Image Retrieval Using Annotations

In the PATSI approach, concepts from the most similar images are transferred to the query image using transfer function. Finding the k most similar images is performed by calculating the distance measure between visual features of a query image and images in the training set. The resulting annotation consists of all the words whose transfer values are greater than a specified threshold value t. The threshold value t influences the resulting annotation length. Optimal threshold value t^* and number of neighbors k must be found using an optimization process [17]. Images retrieval using PATSI is embedded into the method. A query image is an image for which the similar images should be found. The third task in the query phase is calculation of distances between the query image and all other images in a dataset. The images from the dataset are ranked with increasing distances and are presented to a user with this ordering. Jehnsen-Shannon divergence is calculated between models of images built onto image visual features. Visual features are treated as a realization of multivariate random variable described by multivariate Gaussian distribution. The parameters of that distribution are calculated using the Expectation Maximization

Algorithm (EM) [27]. All images were split by 20-by-20 grid splitter and for every cell a mean color value as well as a color deviation in RGB color space were calculated. Additionally, for all segments their center points, and mean Eigen values calculated on color Hessians were stored. PATSI annotation results using F-measure for MGV2006 [22] dataset with different visual features as well as different distance measures are presented in Table 1. For all visual features as well as distance measures we used exactly 19 most similar images in transfer process. All words with transfer value greater than $t = 1.2$ were then treated as the final annotation. PATSI annotator run with using distances in metric space

Table 1. F-measure of AIA on MGV2006 dataset using PATSI annotator with different feature sets and distance measures in the metric space

Visual Feature	Distance measure					
	Cannbe-ra	Chebys-hev	City-block	Correla-tion	Cosine	Eucli-dean
Auto Color Collerogram	0.20	0.16	0.18	0.17	0.17	0.17
CEDD	0.25	0.18	0.25	0.27	0.27	0.27
FCTH	0.24	0.17	0.25	0.23	0.23	0.24
Fuzzy Color Histogram	0.12	0.13	0.13	0.16	0.16	0.13
Gabor	0.06	0.06	0.06	0.09	0.09	0.06
General Color Layout	0.14	0.09	0.14	0.09	0.08	0.11
JPEG Coefficient Histogram	0.20	0.18	0.21	0.21	0.22	0.21
Tamura	0.15	0.14	0.15	0.15	0.15	0.15
CoOccurance matrix	0.17	0.07	0.17	0.17	0.18	0.16
RGB	0.20	0.10	0.20	0.20	0.18	0.23
HSV	0.21	0.09	0.21	0.19	0.17	0.19
RGB + DEV.	0.23	0.09	0.21	0.21	0.18	0.20
HSV + DEV	0.22	0.09	0.22	0.18	0.19	0.19
RGB + DEV + HES	0.23	0.10	0.22	0.18	0.18	0.20
HSV + DEV + HES	0.22	0.09	0.22	0.19	0.19	0.19
RGB + DEV + XY + HES	0.22	0.10	0.22	0.22	0.22	0.20
HSV + DEV + XY + HES	0.23	0.09	0.22	0.20	0.20	0.19

achieved highest results for CEDD visual feature, and Euclidean measure. The best mean F measure was also achieved with Euclidean distance. Very interesting results can be achieved using PATSI annotator with distance measure calculated in probabilistic space, see Table 2. Using Jehnsen-Shannon divergence allows us to significantly improve annotation results in comparison to results presented in Table 1 as well as for the other state-of-art methods [27].

Examples of annotations generated by PATSI for images from ICPR2004 database are presented in Table 3. This table contains also images identified as the most similar images to the considered one. We use Jehnsen-Shannon divergence to calculate distances between images. The PATSI annotator performance in comparison to other state-of-the art method was improved by 20 percentage points [17], achieving F-Measure equal to 78% for the best 27% words in the

Table 2. F-measure of AIA on MGV2006 dataset using PATSI annotator using Jehnsen-Shanon divergence in comparison to other state-of-art methods

Method	Precision	Recall	F-measure
PATSI(HSV + DEV)	0.33	0.38	0.36
PATSI(RGB + Dev)	0.40	0.44	0.42
PATSI(RGB + DEV + HES + XY)	**0.42**	**0.43**	**0.43**
FastDIM	0.24	0.16	0.19
FastDIM + GRWCO	0.34	0.34	0.34
MCML	0.32	0.24	0.27
MCML + GRWCO	0.38	0.37	0.37
CRM	0.39	0.34	0.36

dictionary of MGV2006 database [22]. The results suggest that for a small number of concepts AIA can be now treated as the effective image retrieval tool.

During experiments we have noticed that some of the features as well as distance measures are more suitable to detect some groups of words, while showing a weak performance for others. By combining them together we can increase overall annotation performance. Current research is focused on combining many similarity measures and visual features in one annotation transfer process. We have extended the PATSI algorithm to the multi-PATSI method which performs annotation transfer process based onto many similarity matrices calculated using different feature sets and different similarity measures. The results are combined into the final annotation based on the quality of particular annotators for specific words.

3 Local Image Analysis

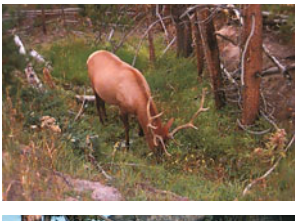
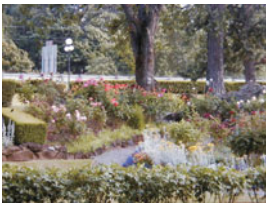
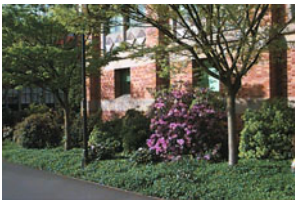
Local image analysis methods are built on the basis of local features, i.e., features calculated from very small image regions. Very popular and effective types of local features are *keypoints* [18,20]. Keypoints themselves are much harder to generalize (although such attempts exist, e.g., [21]) because they are much diversified along single objects. Yet, keypoints have a very nice property, they are able to capture the notion of sameness.

3.1 Image Matching

The goal of image matching is to detect whether two images share visually identical content. Image matching problem may be divided into many subproblems, such as: sub-image matching [16,30], image fragment matching [23], panorama recognition [2], etc. All these techniques provide high precision results.

Sub-image matching methods are able to determine if one image is a fragment of another image. Such approaches may be very useful for finding identical content in case where both images share only one common object. The key advantage of sub-image matching is the applicability of complex (even non-linear)

Table 3. PATSI annotation results for example images from ICPR2004 with their nearest neighbors

Original (query) image:	**Original (query) image:**	**Original (query) image:**
		
Original annotation: 'elk', 'greenery', 'ground', 'logs', 'tree', 'trunks' **Generated annotation:** 'elk', 'greenery', 'ground', 'logs', 'tree', 'trunks'	**Original annotation:** 'river', 'trees' **Generated annotation:** 'garden', 'grass', 'trees'	**Original annotation:** 'man', 'people', 'table', 'woman' **Generated annotation:** 'man', '*microphone*', 'woman', 'people',
Similar images:	**Similar images:**	**Similar images:**
		
		
		
		

geometrical models for the matching process. This allows finding objects seen from different viewpoints or even deformed ones. These methods may be effectively used to capture large objects, such as monuments, buildings. However, they are ineffective when faced with a problem of finding multiple fragments on both images.

Image fragment matching utilizes simpler geometrical models, but is able to find multiple identical objects on scenes with cluttered background. The disadvantage of this approach is the relative simplicity of applied geometry. Deformed or strongly non-planar objects are harder to capture. These methods may be effectively used to capture small object, such as bottles, books, boxes, etc.

Panorama recognition techniques assume that there is only one object of interest. This object is however captured only partially, i.e., different images contain different fragments of the object of interest. These methods may be used to capture huge objects, such as e.g. landscapes, cityscapes.

3.2 Automatic Visual Object Formation

The last, and the most important, fragment of our solution in low level vision refers to the concept of visual objects [7]. We have proposed a grouping method which is able to automatically form *visual object* [24,25]. It is based on the image matching methods discussed in Section 3.1. Having a high precision matching routine we may expect that the created groups are free of errors. The method is able (in a very limited way) to find meaningful visual objects purely on a visual basis, without any training data or supporting information. In fact it is an attempt to bridge *the semantic gap* [6,26].

The *automatic visual object formation* method has four major steps: (1) *preretrieval* to make the process more efficient; (2) *image matching* to find similarities within the set; (3) formation of *prototypes*, which are an intermediate structure [7], and finally; (4) formation of *visual objects*. In the first step we measure similarities between all images in the database. For further processing we select only the most similar ones. In the second step we perform image matching for all pairs of similar images within the set. As a result we get a set of (nearly all) similar image fragments found within the input collection. Because each image is matched with multiple other images, some image regions on a single image may have multiple different matches with other images. In the third step we group all these regions found within a single image. Created groups are called *prototypes*. In the last step we group all prototypes according to matching information between images. Resulting groups are called *visual objects* and they represent frequently repeating, matched fragments from the input collection. Exemplary *visual objects* found in a database containing both indoor and outdoor scenes are presented in Fig. 1. Although, each *visual object* consists of images containing the manifestation of the same underlying, physical object, this information is very useful. It allows formulating very specific queries, we may seek for such specific objects as a road sign, a flu-remedy pack, a model of a ship or a car, a monument, a mountain or landscape, etc.

Fig. 1. Exemplary visual objects are outlined on images from a processed image collection

4 Bridging Local and Global Level Vision

Having described all necessary components, let us now present the main idea of our current research. We envision that both, global and local image analysis routines cooperate together. We would like to utilize global approaches to provide an effective retrieval tool, and we would like to enforce it by the local approach to solve the problem of limited, non-adaptable dictionary. Let us assume that the dictionary used in the global image processing is hierarchical, e.g., it is a fragment of some larger *ontology*. Some concepts in the hierarchy may be *contradictory* and cannot exist together.

Given a small set of hierarchically arranged concepts (e.g. inside, outside, mountain, ship, building, sky) and a collection of images containing multiple instances of identical objects (however seen in different scenes and contexts) we would like to make the hierarchy of concepts more specific and precise. This idea is illustrated in Fig. 2. First we detect all visual objects using the object formation routine discussed in Section 3.2. Having all identical objects captured, we would like to link them into our existing hierarchy of objects. To do this, we employ intelligent, global image analysis techniques, e.g. classification, automatic image annotation. If needed, we may use a different intelligent technique on each level of hierarchy. Recognition process should take into account shapes of regions creating a visual object. We divide the image into three separate segments, each having a different *meaning* for the processed visual object. These three segments are: *interior*, *context* and *environment*, they are illustrated in Fig. 3. Having recognized objects on all images belonging to a single visual object, we may decide where to attach it within the concept hierarchy. Usually, various recognition or annotation methods have one of three possible outputs: precise concept probability values, roughly estimated concept scores or just a subset of concepts from the dictionary. All these output types have to be processed in a different way. We have designed three decision rules, one for each type of output. Each decision rule outputs a single support value s_x^w for each concept w and each image x containing the visual object.

Final decision regarding of linking the new concept within the hierarchy is made on the basis of decision rule outputs. An averaged concept support values s_w is calculated and possible contradictions in the hierarchy are solved. Contradictory concepts in each level of hierarchy are modeled as a set of sets Z (multiple different rules on each level of hierarchy). Each set Z_i contains all contradictory concepts. In case there are two or more contradictory concepts, the ones with

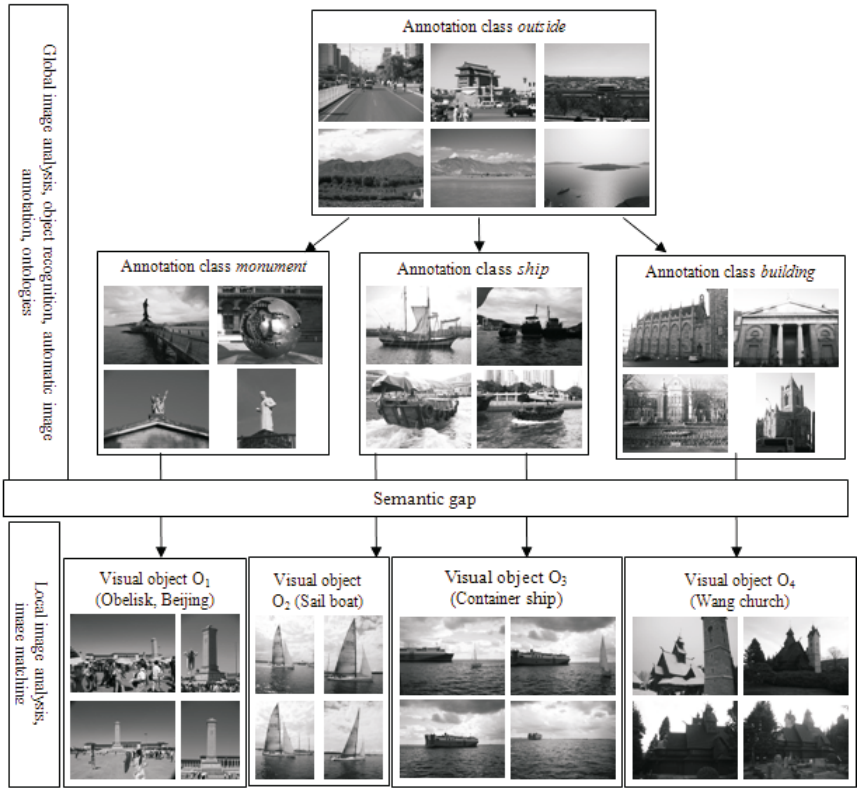


Fig. 2. The concept of bridging local and global level vision

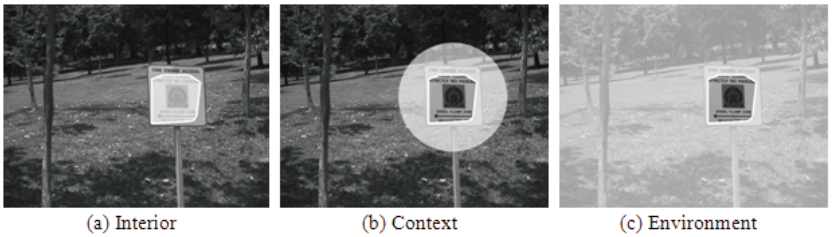


Fig. 3. Three different annotation regions for visual objects

the larger support are chosen by the decision rule. In case the decision rule d_w accepts the visual object it is processed deeper and deeper within the hierarchy.

5 Conclusion

The paper shows briefly the results of our methods concerning similar images retrieval using PATSI annotation algorithm (now we are testing multi-PATSI

method) and the method of images matching – it detects whether two images share visually identical content. The important part of our research in low level vision refers to the concept of visual objects. We have proposed a grouping method which is able to automatically form visual object, this approach is based on the image matching methods. Our method is able to find meaningful visual objects purely on a visual basis.

Currently we try to join global and local image analysis routines. Global approaches should provide efficient retrieval tool, but it can work only on limited dictionary, with all words well represented in a training set. Such a dictionary can contain words from a given ontology, i.e., the dictionary consists of hierarchically arranged concepts. Captured in low level analysis identical objects can be linked into a hierarchy of concepts (objects) by global image analysis techniques, e.g., automatic image annotation method.

Our future plans concern with the above mentioned problem. Initial set of decision rules are proposed, but we do not have experimental results. Of course, all sub-methods in the proposed approach should work very well. Having weak one part of the method we are not able to obtain good final results. So, we plan to improve our global method (e.g., multi-PATSI method) as well as the automatic visual object formation methods. These two research topics will be conducted in parallel with studies on the 'bridge' method that should allow for filling up the semantic gap, perhaps even to a limited extent.

All the presented researches are dedicated to searching similar images, although we still have a problem with understanding the concept *images similarity*. Meaning of *similarity of images* still causes problem, however more of us can easily indicate the similar images within a not large collection of images. It is important that those images are usually similar in *the view of particular user*, and therefore the *term images similarity is not precise*. In our group we have developed computer program, called SIMILARIS, and a set of images used with this program. The main aim of that research is defining a kind of baseline – measures of images similarity when these images are evaluated by people. That data can be than used to find the efficient measure of image similarity. After finishing the testing phase and our preliminary study, the program SIMILARIS together with used collection of images will be published on the server with free access to researchers. Researchers on CBIR systems focus on building systems with the very high precision, but the fundamental question still remains without answer: is it possible to obtain CBIR systems with high precision and recall measures? The studies with SIMILARIS should help to find answer to the above question.

Acknowledgments. This work is partially financed from the Ministry of Science and Higher Education Republic of Poland resources in 2008-2010 years as a Poland-Singapore joint research project 65/N-SINGAPORE/2007/0.

References

1. Broda, B., Kwasnicka, H., Paradowski, M., Stanek, M.: MAGMA: efficient method for image annotation in low dimensional feature space based on Multivariate Gaussian Models. In: Ganzha, M., Paprzycki, M. (eds.) Proc. of the IMCSIT. Polish Information Processing Society 2009, pp. 131–138 (2009)
2. Brown, M., Lowe, D.G.: Recognising panoramas. In: Ninth IEEE International Conference on Computer Vision, vol. 2, pp. 12–18 (2003)
3. Carneiro, G., Chan, A., Moreno, P., Vasconcelos, N.: Supervised learning of semantic classes for image annotation and retrieval. IEEE Trans. on Pattern Analysis and Machine Intelligence 29(3), 394–410 (2007)
4. Chatzichristofis, S., Boutalis, Y.: Cedd: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. Computer Vision Systems, 312–322 (2008)
5. Chatzichristofis, S.A., Boutalis, Y.S.: Fcth: Fuzzy color and texture histogram – a low level feature for accurate image retrieval. In: Interactive Services, Intern. Workshop on Image Analysis for Multimedia, pp. 191–196 (2008)
6. Datta, R., Joshi, D., Li, J., Wang, J.: Image retrieval: Ideas, influences, and trends of the new age. ACM Transactions on Computing 40(2) (2008)
7. Dickinson, S.J., Leonardis, A., Schiele, B., Tarr, M.J. (eds.): Object Categorization: Computer and Human Vision Perspectives. Cambridge University Press, Cambridge (2009)
8. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.A.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2353, pp. 97–112. Springer, Heidelberg (2002)
9. Feng, S.L., Manmatha, R., Lavrenko, V.: Multiple bernoulli relevance models for image and video annotation. In: Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1002–1009 (2004)
10. Hironobu, Y.M., Takahashi, H., Oka, R.: Image-to-word transformation based on dividing and vector quantizing images with words. Neural Networks in Boltzmann Machines 4 (1999)
11. Inoue, M.: On the need for annotation-based image retrieval. In: Proc. of the Information Retrieval in Context (IRiX), A Workshop at SIGIR 2004, pp. 44–46 (2004)
12. Jin, Y., Khan, L., Wang, L., Awad, M.: Image annotations by combining multiple evidence and wordnet. In: Proc. of the 13th Annual ACM Intern. Conf. on Multimedia (2005)
13. Kwasnicka, H., Paradowski, M.: Resulted word counts optimization-a new approach for better automatic image annotation. Pattern Recognition 41(12) (2008)
14. Lavrenko, V., Manmatha, R., Jeon, J.: A model for learning the semantics of pictures. In: Proc. of Neural Information Processing Systems (NIPS). MIT Press, Cambridge (2003)
15. Llorente, A., Motta, E., Roger, S.: Image annotation refinement using web-based keyword correlation. In: Proc. of the 4th International Conference on Semantic and Digital Media Technologies, Berlin, pp. 188–191 (2009)
16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)

17. Maier, O., Stanek, M., Kwasnicka, H.: PATSI - photo annotation through similar images with annotation length optimization. In: Klopotek, M.A., et al. (eds.) Intelligent information systems 2010, pp. 219–232. Pub. House of Univ. of Podlasie (2010)
18. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. In: Proc. of British Machine Vision Conference, pp. 384–393 (2002)
19. Makadia, A., Pavlovic, V., Kumar, S.: A new baseline for image annotation. In: Proc. of the 10th European Conference on Computer Vision, pp. 316–329 (2008)
20. Mikolajczyk, K., Schmid, C.: Scale and Affine Invariant Interest Point Detectors. International Journal of Computer Vision 60, 63–86 (2004)
21. Monay, F., Quelhas, P., Odobez, J.M., Gatica-Perez, D.: Contextual Classification of Image Patches with Latent Aspect Models. EURASIP Journal on Image and Video Processing, Article ID 602920 (2009), doi:10.1155/2009/602920
22. Paradowski, M.: Methods of automatic annotation as an efficient tool for images collections describing. PhD thesis, Wroclaw Univ. of Technology (2008) (in Polish)
23. Paradowski, M., Sluzek, A.: Automatic Visual Object Formation using Image Fragment Matching. In: Proc. of 2010 Intern. Multiconf. on Computer Science and Information Technology, vol. 5, pp. 45–51, IEEE Catalog CFP1064E-CDR (2010)
24. Russell, R., Sinha, P.: Perceptually-based comparison of image similarity metrics. Technical report, AIM-2001-014, CBCL-201 (2001)
25. Sluzek, A., Paradowski, M.: A Vision-based Technique for Assisting Visually Impaired People and Autonomous Agents. In: Proc. of 3th Intern. Conference on Human System Interaction, pp. 653–660 (2010)
26. Smeulders, A.W.M., Gupta, A.: Content-Based Image Retrieval at the End of the Early Years. Transactions on Pattern Analysis and Machine Intelligence 22(12), 1349–1380 (2000)
27. Stanek, M., Broda, B., Kwasnicka, H.: PATSI — Photo Annotation through Finding Similar Images with Multivariate Gaussian Models. In: Bolc, L., Tadeusiewicz, R., Chmielewski, L.J., Wojciechowski, K. (eds.) ICCVG 2010. LNCS, vol. 6375, pp. 284–291. Springer, Heidelberg (2010)
28. Tadeusiewicz, R., Ogiela, M.R.: The New Concept In Computer Vision: Automatic Understanding of the Images. Proc. of Artificial Intelligence and Soft Computing, 133–144 (2004)
29. The American Heritage® Dictionary of the English Language. 4th edn. copyright ©2000, by Houghton Mifflin Company. Updated in 2009, Houghton Mifflin Company Pub. (2000)
30. Yang, D., Sluzek, A.: A low-dimensional local descriptor incorporating TPS warping for image matching. Image and Vision Computing 28(8) (2010)

Architecture for a Parallel Focused Crawler for Clickstream Analysis

Ali Selamat and Fatemeh Ahmadi-Abkenari

Software Engineering Research Group, UTM Knowledge Economy Research Alliance & Software Engineering Department, Faculty of Computer Science & Information Systems, Universiti Teknologi Malaysia (UTM), 81310 UTM Johor Baharu Campus, Johor, Malaysia
aselamat@utm.my, pkhoshnoud@yahoo.com

Abstract. The tremendous growth of the Web poses many challenges for all-purpose single-process crawlers including the presence of some irrelevant answers among search results and the coverage and scaling issues regarding the enormous dimension of the World Wide Web. Meanwhile, more enhanced and convincing algorithms are on demand to yield more precise and relevant search results in an appropriate amount of time. Due to the fact that employing the link based Web page importance metrics in search engines is not an absolute solution to identify the best answer set by the overall search system and because employing such metrics within a multi-processes crawler bears a considerable communication overhead on the overall system, employing a link independent Web page importance metric is required to govern the priority rule within the queue of fetched URLs. The aim of this paper is to propose a modest weighted architecture for a focused structured parallel crawler in which the credit assignment to the discovered URLs is performed upon a combined metric based on clickstream analysis and Web page text similarity analysis to the specified mapped topic(s).

Keywords: Clickstream analysis, Focused crawlers, Parallel crawlers, Web data management, Web page importance metrics.

1 Introduction

The dimension of the World Wide Web is being expanded by an unpredictable speed. As a result, search engines encounter many challenges such as yielding accurate and up-to-date results to the users, and responding in an appropriate timely manner. A centralized single-process crawler is a part of a search engine that traverses the Web graph and fetches any URLs from the initial or seed URLs, keeps them in a queue and then in an iterated manner - according to an importance metric - selects the first most important K URLs for further processing. A parallel crawler on the other hand is a multi-processes crawler in which upon partitioning the Web into different segments, each parallel process is responsible for one of the Web fractions. Since due to the enormous size of the Web, a single-process crawler is not capable of reaching an acceptable download rate, employing a parallel crawler within a search engine architecture is scalable. Besides, different parallel processes could run at geographic

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 27–35, 2011.

distant location to download the pages on different zones, so, upon applying the parallel crawler, the load on the overall network could be reduced [9]. The bottleneck in the performance of any crawler is applying an appropriate Web page importance metric.

In this paper we employ a clickstream-based metric as a heuristic; the hypothesis is the existence of a standard upon which the authorized crawlers have legal rights to access the server log files. We first review on the literature of focused crawlers and the existing Web page importance metrics by defining the drawbacks of each of them. Then, we briefly discuss our clickstream-based metric since it has been thoroughly discussed in a companion paper. Next, the application of the clickstream-based metric within the architecture of a focused parallel crawler will be presented. So, the objective of this paper is to propose an architecture for a focused-based parallel crawler in which prioritizing the crawl frontier is based on a function of clickstream analysis combined with a text analysis approach. The text analysis part of the derived formula helps the identification of authoritative Web content in newly uploaded pages and the pages in the dark part of the Web. Besides, the suggested architecture answers this question that in the absence of a central coordinator, how and in which order, the overall crawler reads the ordered queue of the parallel processes to achieve the most important discovered pages by parallel processes at the earliest time of result organizing.

2 Related Works

The related works on focused crawlers and linked based importance metrics are discussed as follows;

2.1 Focused Crawlers

There are two different classes of crawlers known as focused and unfocused. The purpose of unfocused crawlers is to attempt to search over the entire Web to construct their index. As a result, they confront the laborious job of creating, refreshing and maintaining a database of great dimensions. While a focused crawler limits its function upon a semantic Web zone by selectively seeking out the relevant pages to predefined topic taxonomy and avoiding irrelevant Web regions as an effort to eliminate the irrelevant items among the search results and maintaining a reasonable dimensions of the index. A focused crawler's notion of limiting the crawl boundary is fascinating because "a recognition that covering a single galaxy can be more practical and useful than trying to cover the entire universe" [6].

The user information demands specification in a focused crawler via exemplary Web documents instead of by keyword-based document. Therefore, a mapping process is performed by the system to highlight (a) topic(s) in the pre-existing topic tree which can be constructed based on human judgment [6]. The core elements of a traditional focused crawler are a classifier and a distiller. While the classifier checks the relevancy of each Web document's content to the topic taxonomy based on the naïve Bayesian algorithm, the distiller finds hub pages inside the relevant Web regions by utilizing a modified version of HITS algorithm. These two components, together determine the priority rule for the existing URLs in a priority based queue called crawl frontier [6], [13], [15], [16].

2.2 Link-Dependent Web Page Importance Metrics

As stated above, a centralized crawler or each parallel process within a parallel crawler retrieves URLs and keeps the links in a queue of URLs. Then, in the next step, due to the time and storage constraints, a crawler or a parallel process must decide which most important *K* URLs to process first according to one Web page importance metric. There are diverse Web page importance metrics, each views the importance of a page from a different perspective such as outgoing or incoming link enumeration, text analysis or location angles including Backlink count, PageRank, HITS, forward link count, location metric and content-query similarity checking metrics [10]. But the most well-known category is the link-based metrics. PageRank metric as a modification to Backlink count that simply counts the links to a page, calculates the weighted incoming links in an iterated manner and considers a damping factor which presents the probability of visiting the next page randomly [3],[10]. The *TimedPageRank* algorithm adds the temporal dimension to the PageRank by considering a function of time $f(t)$ ($0 \leq f(t) \leq 1$) in lieu of the damping factor. The notion of TimedPageRank is that a Web surfer at a page i has two options: 1) Randomly choosing an outgoing link with the probability of $f(t_i)$ and 2) Jumping to a random page without a link with the probability of $1\text{-}f(t_i)$. For a completely new page within a Web site, an average of the TimedPageRank of other pages in the Web site is used [19]. The HITS metric views Web page importance in its hub and authority scores. A Web page with high hub score is a page that points to Web pages with high authority scores and a Web page with high authority score is a page that has been pointed to by Web pages with high hub scores.

The HITS metric has some drawbacks including the issue of topic drift, its failure in detecting mutually reinforcing relationship between hosts, and its shortcoming to differentiate between the automatically generated links from the citation-based links within the Web environment. Due to the fact that pages to which a hub page points to are not definitely around the original topic, the problem of topic drift is formed. The second problem occurs when a set of documents in one host points to one document on another host. As a result, the hub score of pages on the first host and the authority score of the page on second host will be increased. But this kind of citation cannot be regarded as coming from different sources. Finally, Web authoring tools generate some links automatically that these links cannot be regarded as citation based links. Although the literature includes some modifications to HITS algorithm such as the research on detecting micro hubs, neglecting links with the same root, putting weights to links based on some text analysis approach or using a combination of anchor text with this metric, there is no evidence of a complete success of these attempts [2], [4], [5], [7],[8].

3 Proposed Clickstream Based Importance Metric

Since the area of Web usage mining considers the utilization of server log files, the usage of clickstream analysis is roughly ignored and little attention to the research has been allocated to it as a Web page importance metric despite its advantages. The literature includes the research on clickstream data for e-commerce objectives [12], [14]. In a companion paper, we proposed the clickstream analysis for Web page importance determination [1]. Besides, that paper has included the application of this proposed metric in a focused crawler. Furthermore, the application of this metric in a crawler with only a

parallel structure has been discussed in the companion paper [17]. Clickstream-based importance metric is computed according to the total duration of all visits per page during the observation span of time. In other words, the log ranking for a page (LR_p) is the total duration of server sessions per page (D_{sp}) as shown in Equation (1) [1];

$$LR_p = D_{sp} \tag{1}$$

According to the link independent nature of clickstream-based metric, in a crawler based of this metric, the need for link enumeration will be removed. As a result, in a parallel crawler, the communication overhead, together with parallel processes that inform each other of the existence of links will be eliminated [1]. Also upon employing the clickstream importance metric, the calculated importance of each page is precise and independent from the downloaded segments of the Web. In our approach, by employing a clickstream-based metric within the crawler, we will go beyond noticing the page importance in its connection pattern. Instead, the page credit computation is performed according to a function of a simple textual log files in lieu of working with matrices of high dimensions. Because the clickstream analysis has no relation with the page content, we combine it with a text analysis approach to make a robust decision on priority determination for the items stored in a crawl frontier.

4 Architecture of a Crawler Based on Clickstream Importance Metric

Within a crawler with a focused structure let's consider G as a Web Graph with physically dispersed nodes, C as a tree-shaped topic directory, c as each topic node, $D(c)$ as example documents associated with topic c, c^* as the highlighted mapped node, p as a Web page and $R_{c^*}(p)$ is a measure of content relevancy page p to c^*. Since it is not very easy for users to issue an effective search request, the user enters the interface and imports the documents (Web pages) of his/her interest to the search system. We call this set as D_u. Upon analyzing the imported documents, the system highlights (a) node(s) in the existing topic taxonomy tree through the mapping process. Figure 1 depicts the mapping process from the corpus to (a) node(s) in topic taxonomy tree.

Let's call the mapped highlighted node, c^*. The mapping process will be done using an *Inverted List* with special heed to the terms in title, italic and bold faced and headings in the imported documents by considering the weighting scheme of *TF-IDF*. So given a set of user imported documents of $D=\{d_1, \ldots, d_n\}$ with a unique identifier for each document, there is a vocabulary V containing all the distinct terms in the semantic region in which the focused crawler is specialized, such as the field of computer science. The Inverted list version is $<id_i, w_j, [o_1,\ldots, o_k]>$ in which the id_i is the document unique identifier, the w_j is the weight of each term j and the rest is the offset of the term j in the document i [14]. In order to use less memory space for the Inverted List, a method of compression is used to represent the document unique identifier since it is the most space consuming section of an Inverted List. The method of compression could be *Elias Delta coding* in which the representation is shorter for large integers such as documents numbers in comparison to other compression methods like Unary coding or Elias Gamma coding [14]. Furthermore the system will add any distinct terms from the corpus which is not included in the vocabulary into a temporary vocabulary with the exact version of the above Inverted List. Then a

Semantic Checker will be used to determine the semantically related terms from this temporary vocabulary with the terms in the main vocabulary of *V*. Then the not semantically related words with high importance rate will be used by the system to edit the topic taxonomy. After the detection of (a) node(s) in the topic taxonomy tree, the pre-existing examples (Web pages) associated with the node(s) will be added to the user imported examples to form the crawling list or *CL* as shown in Equation (2) [1];

$$CL = D_u + D(c^*) \tag{2}$$

The *CL* list in our approach is considered as the second level seed URL and should be divided among parallel processes by a Web partitioning function. Due to the fact that all of the pages of a Web site are not the descriptive in nature and usually one or a few of them could illustrate the user's topics of interest, the probability that the Web pages inside the D_u and $D(c^*)$ belonging to different Web sites is high. Therefore with regard to the discussed advantages of the site-hash-based approach, we choose the site-hash-based partitioning function to divide the second level seed URLs among parallel processes. As a result, the selected partitioning function could distribute the crawling list among parallel processes in a balanced manner.

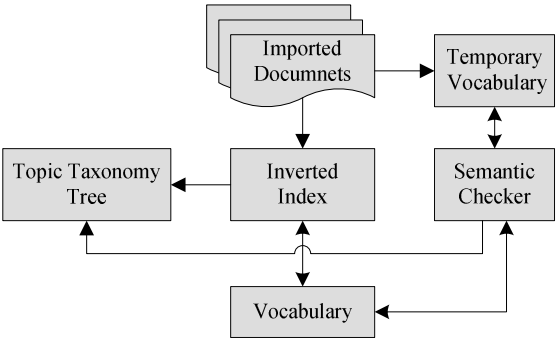


Fig. 1. Mapping process of user imported documents to (a) node(s) in topic taxonomy tree

As depicted in Fig. 2, in the proposed architecture for a focused parallel crawler, each parallel process has four elements namely *distiller*, *classifier*, *crawler* and *coordinator*. The crawler element is responsible for the process of fetching any unvisited URLs from the allocated second seed URLs in an iterated manner and populates them in the crawl frontier. The crawl frontier is a priority-based list corresponding to the Best-First crawling. Also a time-stamp list of visited URLs or *crawl history* is maintained by the overall crawler to keep those URLs pages that have been fetched, as a way to decrease the overlaps among different parallel processes. The history can be maintained in disk for post crawling evaluation or in memory for fast look-up [14]. Besides, a section of *duplication detector* is maintained by each parallel process to prevent the occurrence of duplicate URLs in the crawl frontier through maintaining a separate hash-table. Figure 2 only shows one of the parallel processes because of the space limitation. In the next step, there should be an algorithm which guides the priority rule inside the crawl frontier to choose the most important URLs for further processing.

For having a robust decision on the importance of each existing page in the crawl frontier, the distiller and classifier sections together govern the priority rule. The distiller element is responsible to calculate the page importance based on the clickstream analysis. We call the result as *LR(p)* short for ranking of page *p* based on log file contents. Intermittently the classifier measures the importance of page based on the content relevancy of page *p* to the mapped topic(s) of *c**. The result of this part will be reflected in the computed value of $R_{c*}(p)$ as $0 \le R_{c*}(p) \le 1$. The $R_{c*}(p)$ exactly like that in the traditional focused crawler is based on the naïve Bayesian algorithm as shown in Equation (3) [15];

$$R_{c*}(p) = \sum_{c*} \Pr(c* \mid p) \tag{3}$$

Now a combined importance metric is applied for credit assignments to the items in the frontier by the coordinator section of each parallel process. As discussed earlier, the presence of a central coordinator causes an inevitable load on the network due to the unavoidable communication between this component and parallel processes. Besides, the existence of a central section within a system causes maintenance complication with regard to the dependency of the whole system to this element on the upgrade occasions. Hence in our proposed architecture, we eliminate the presence of a central coordinator and instead, there is a smaller section inside each parallel process to harmonize the result of each parallel process with the results of other parallel processes.

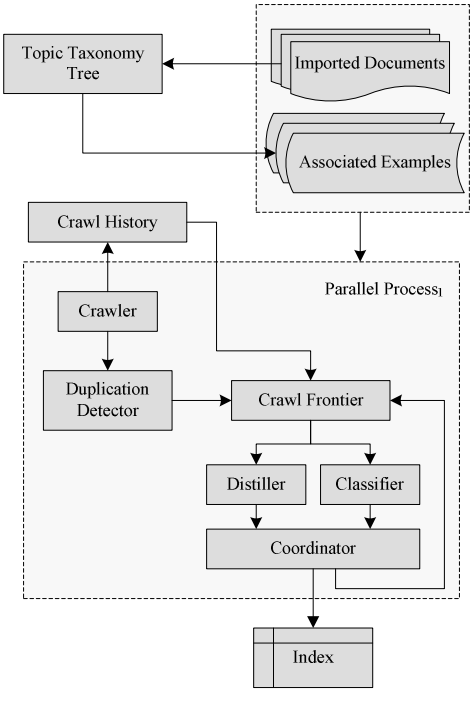


Fig. 2. Architecture of a focused parallel crawler

Therefore the coordinator of each parallel process is responsible for combining the two measures of *LR(p)* and $R_{c*}(p)$ to yield an importance measure of *I(p)*. As *LR(p)* is from the second type and it has a colossal value, to leverage it with the $R_{c*}(p)$, it is converted into hours. Moreover due to the fact that some pages with high *LR(p)* scores may contain news and be of low descriptive nature for user information demands, therefore in the computation of *I(p)*, an emphasis factor of *E* empowers the $R_{c*}(p)$ to make it more compelling than *LR(p)*. So the *I(p)* is computed as shown in Equation (4);

$$I(p) = LR(p)/3600 + E \times R_{c*}(p) \tag{4}$$

Upon the computation of *I(p)* by the coordinator section for each Web page in the crawl frontier, the queue of each parallel process is reordered based on the descending value of *I(p)*. Now each parallel process knows the importance of each page in its crawl frontier. The question raised in the absence of a central coordinator is that in which order the overall crawler fetches the ordered queue of each parallel process. To address this issue, a matter of communication transfer is vital among parallel processes to substitute the role of the central coordinator and to yield an organized and integrated result by the overall crawler. To achieve this objective, let's consider *K* as the size of the crawl frontiers or the number of Web pages in the queue of each parallel process in a way that it is the same for all parallel processes. As part of communication, parallel processes should inform each other of the importance of the pages in their queue. If during the communication, the average of *I(p)* for all the *K* URLs is transferred, it means that a queue with few number of very high important pages and many low important pages may have the same average as a queue with many medium important pages. Missing the very high important pages contributes to inaccuracy and misjudgment of the overall crawler. To prevent this problem, the communication among parallel processes in our approach, consists of sending the average of *I(p)* for a fraction of the pages or the *K/L* size of the frontier in lieu of transmitting the average of *I(p)* for *K* URLs. Therefore if *m* is the number of parallel processes in the parallel crawler and *n* is the size of each information nugget which is due to transfer and *L* is the number of frontier divisions, the notification overhead is calculated as shown in Equation (5);

$$Notification\ overhead = n \times m \times (m-1) \times L \tag{5}$$

As depicted in figure 3 the format of each nugget is in a way that the first position from the left is a flag bit. It is set to one if the average of *I(p)* belongs to the last *K/L* set of that parallel process' queue and zero if vice versa. The rest of the nuggets consist of the correspondent parallel process identification number, the *K/L* identification number or the start position of *L* within the queue and the last part is the average of *I(p)* for the *K/L* set respectively. Equation (6) shows the formulation for *n* since it is a dependent variable to *m* and *L* in which α is the size of I_{avg} value in bits.

$$n = 1 + \alpha + Log_2^{mL} \tag{6}$$

Upon receiving the notification nuggets from other parallel processes, the coordinator section of each parallel process compares the *I(p)* average in the received nuggets with that of itself. The coordinator of the parallel process with the best *I(p)* average sends the URLs of the corresponding *K/L* set to the index section and produces a new nugget

for the next *K/L* set. The production of the nuggets by parallel processes is done in a synchronous manner in each *t* seconds intervals. When the time of *t* arrives, only the parallel process which sends a set of URLs to the index is eligible to produce a new nugget and this nugget will be compared to the old received ones. A parallel process that sends its last set of URLs to the index - with the flag bit set to one - does not produce any new notification nugget and the comparison process will be performed among the other parallel processes. This is a normal run of the procedure and any other variation of this rule produces an erroneous condition. Figure 4 depicts an example of a parallel crawler with four parallel processes in which L=4. For the matter of simplicity it only depicts the nuggets received by parallel process$_1$ and the first transmitted nugget from parallel process$_1$ in the first iteration of notification transmission.

Fig. 3. The structure of the notification nuggets

According to Equation (5) to decrease the notification overhead, the number of parallel processes should not be high. Google employs a few numbers of parallel processes of power two [19]. Moreover the size of each nugget should be maintained as small as possible. Also considering the L as a big number causes more numbers of notification nugget transfer among parallel processes and as a result more overhead on the overall system is produced. Upon considering few numbers of parallel processes and an appropriate measure for L, the size of each notification nugget will be influenced according to Equation (6).

5 Conclusion

In this paper we propose a crawl frontier prioritizing metric based on clickstream analysis and text analysis within an introduced architecture for a focused structured parallel crawler. The reasons to choose this framework are the limited topic specific search boundary of a focused crawler and its ability to yield more precise answers to the users' information demand and the optimized download rate of a parallel crawler in comparison to a centralized crawler. Besides, in our approach, parallel processes collaborate with each other in the absence of a central coordinator section in order to minimize the inevitable communication overhead and to make parallel processes to operate more independently.

Acknowledgment

The authors wish to thank Ministry of Higher Education Malaysia (MOHE) and Universiti Teknologi Malaysia (UTM), for funding the related research.

References

1. Ahmadi-Abkenari, F., Selamat, A.: Application of Clickstream Analysis in a Tailored Focused Web Crawler. Journal of Communications of SIWN,The Systemic and Informatics World Network (2010)
2. Bharat, K., Henzinger, M.R.: Improved Algorithms for Topic Distillation in a Hyperlinked Environment. In: Proceeding of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 104–111 (1998)
3. Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks 30(1-7), 107–117 (1998)
4. Chakrabarti, S.: Mining the Web. In: Discovering Knowledge from Hypertext Data. Morgan Kaufmann, San Francisco (2003)
5. Chackrabarti, S.: Integrating Document Object Model with Hyperlinks for Enhanced Topic Distillation and Information Extraction. In: Proceeding of the 13th international World Wide Web Conference (WWW 2001), pp. 211–220 (2001)
6. Chackrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A.: Mining the Link Structure of the World Wide Web. IEEE Computer 32(8), 60–67 (1999)
7. Chackrabarti, S., Dom, B., Raghavan, P., Rajagopalan, S., Gibson, D., Kleinberg, J.: Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text. In: Proceeding of the 7th international World Wide Web Conference, WWW 2007 (1998)
8. Chakrabarti, S., Van den Berg, M., Dom, B.: Focused Crawling: A New Approach to Topic Specific Web Resource Discovery. Computer Networks 31(11-16), 1623–1640 (1999)
9. Cho, J., Garcia-Molina, H.: Parallel Crawlers. In: Procceding of 11th International Conference on World Wide Web. ACM Press, New York (2002)
10. Cho, J., Garcia-Molina, H., Page, L.: Efficient Crawling through URL Ordering. In: Procceeding of 7th international Conference on World Wide Web (1998)
11. Diligenti, M., Coetzee, F.M., Lawrence, S., Giles, C.L., Gori, M.: Focused Crawling using Context Graph. In: Procceeding of the 26th VLDB Conference, Cairo, Egypt, pp. 527–534 (2000)
12. Giudici, P.: Applied Data Mining, Web Clickstream Analysis. ch.8, pp. 229–253. Wiley Press, Chichester (2003) ISBN: 0-470-84678-X
13. Kleinberg, J.: Authoritative Sources in a Hyperlinked Environment. Journal of the ACM 46(5), 604–632 (1999)
14. Liu, B.: Web Data Mining, Information Retrieval and Web Search. ch.6, pp. 183–215. Springer Press, Heidelberg (2007) ISBN: 3-540-37881-2
15. McCallum, A., Nigam, K.: A Comparison of Event Models for Naïve Baes Text Classification. In: Procceeding of the AAAI-1998 Workshop on Learning for Text Categorization (1998)
16. Menczer, F., Pant, G., Srinivasan, P.: Topical Web Crawlers: Evaluating Adaptive Algorithms. ACM Transactions on Internet Technology 4(4), 378–419 (2004)
17. Selamat, A., Ahmadi-Abkenari, F.: Application of Clickstream Analysis as Web Page Importance Metric in Parallel Crawlers. In: Procceeding of the International Symposium on Information Technology (ITSIM 2010), Kuala Lumpur, Malaysia (2010)
18. Srivastava, A.N., Sahami, M.: Text Mining, Classification, Custering and Applications. CRC Press, Boca Raton (2009)

A Model for Complex Tree Integration Tasks

Marcin Maleszka and Ngoc Thanh Nguyen

Institute of Informatics, Wroclaw University of Technology, Wybrzeze Wyspianskiego 27, 50370 Wroclaw, Poland
{marcin.maleszka,ngoc-thanh.nguyen}@pwr.wroc.pl

Abstract. The common approach to integrating XML documents is based on existing formal structures, not originally designed to integration tasks. In this paper we propose a Complex Tree model designed from the beginning to integration tasks, capable of representing most tree structures. The Complex Tree model is defined on both Schema and Instance level, to better work in practical situations. The integration task for Complex Trees is also defined on both levels. A set of explicitly stated criteria for integration is given, to better design future integration algorithms, in respect of the desired aim of integration process. Finally a simple integration algorithm is presented, based on selected criteria.

Keywords: XML integration, complex tree, integration task, integration criteria.

1 Introduction

The XML file format and it derivatives are today the *de facto* standard of data and knowledge storage [14]. The XML itself has evolved, moving from its first schema – DTD – to more modern approaches, like XSD, DSD and more [5]. Each of those solutions was designed to make the XML (both on schema and instance level) more robust and more practical. Those structures are therefore the basis of many enterprises [3].

The same enterprises occasionally need to integrate those specifically designed structures, i.e. in e-business [3], thus making it necessary to develop tools to automate the process. Various tools for this task exist, each develop for a specific practical, or theoretical purpose [9][13]. Not many of those however are based on the integration theory, most arising from the practical need and only using excerpts of the theory.

This paper is based on an alternate approach. First, a structure allowing for defining various hierarchical structures is defined, with accordance to the integration theory and an integration task is defined. Then criteria are laid down for the integration task, based on the practical approaches (where the criteria are not explicitly stated) and on the theory (where the criteria are too general). The paper concludes with a simple integration algorithm, with more to follow in the future work of the authors.

This paper is organized as follows: Section 2 contains a short survey of existing integration approaches, both for practically used XML and other hierarchical structures; Section 3 provides a model of a hierarchical structure (tree) used in this

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 36–46, 2011.

paper, as well as a definition of an integration task for this model; Section 4 gives an overview of various simple integration criteria; Section 5 provides a simple integration algorithm based on the definitions of previous sections; the paper concludes with a discussion on future work to be done in this area.

2 Related Works

Some of the chronologically oldest approaches to hierarchical structures (trees) integration may be derived from the works of evolutionary biologists seeking methods to integrate differing results from multiple data sources (integrating different evolutionary trees). The structures they operated on in the 1980s and later are *n*-trees, trees with labeled only leaves. Different approaches created basing on general integration theory used some additional structures for integration, like clusters [7] and their variations [4][15], so-called Maximum Agreement SubTrees [10] or triads [1]. Each of those methods, while well developed, was useful only for the narrow field it was designed for.

For fully labeled trees, an area much closer to practical XML documents, some work was done on integrating multilevel classification trees [6]. There some problems known from integration theory, like knowledge inconsistency and the need to determine a consensus, became more visible.

The main area of work in integrating hierarchical structures was done in the most practical area, that is the integration of XML documents and their schemas (general hierarchical schemas, DTD schemas, XSD schemas and more). These approaches vary from using a specific tree grammar [2] to using graph representation [11]. The latter may be based on other sub-representations, like path-based approaches used in XML search engines [8] (there the aim of integration is for the final structure to be able to provide the same answer for a specific path query, thus all paths must remain unchanged during the process).

In their survey on automatic schema matching Rahm and Bernstein [13] detail solutions used also when integrating hierarchical schemas (XML and DTD). The classification they provide shows examples of using single elements or whole sub-structures for matching, matching according to cardinality or some auxiliary information about the element. These are in fact implicitly stated criteria for integration – as defined in section 4.

3 Hierarchical Structures Integration

In this section we will introduce the hierarchical structure model designed specifically for integration purposes, which is capable of representing different tree structures (XML, its schemas, n-trees, some ontologies, etc.) – the Complex Tree. First, the general description of the structure will be presented. In 3.2, the distinction between Complex Tree Schema and Instance will be drawn. In the rest of this section integration tasks will be defined for both structures.

3.1 Complex Tree

For purposes of this work, a model of hierarchical structure (called Complex Tree) has been defined. This model provides the ability to represent most practically used structures (XML, its schemas, n-trees, some ontologies), while retaining relatively simple for integration purposes. It also allows representation of both schema and instance level structures with virtually no changes in structure.

Definition 1. Complex Tree (CT) is a tuple

$$CT = (T, S, V, E)\ , \tag{1}$$

where:

T – a set of node types (i.e.: *root, leaf*)

$$T = \{t_1, t_2, t_3, \dots\}\ , \tag{2}$$

S – a function defining the number and type of attributes for each type of nodes (i.e. nodes of the type *root* always have only 1 attribute *date_created*)

$$S(t_i) = (a_{i1}, a_{i2}, \dots, a_{in_i})\ , \tag{3}$$

V – a set of nodes (vertices), where each node is a triple

$$V = (l, t, A)\ , \tag{4}$$

in which the consecutive elements represent the label of the node, the type of the node ($t \epsilon T$) and the set of attributes and their values for this node (as defined by the S function; note that in some cases the attribute values will be null, i.e. if the hierarchical structure represents a schema, not the actual document)

E – is the set of edges in the structure

$$E = \{e = (v, w), v \in V, w \in V\}\ . \tag{5}$$

3.2 Complex Tree Schema and Instance

The Complex Tree definition given above is a general one, when practical uses differentiate between Complex Tree Schema and Complex Tree Instance (i.e. as generalizations of XSD and XML). While the rest of this paper is based on this general definition, this chapter will explain the specific versions of the Complex Tree.

Complex Tree Schema (CTS), like database schema or XML Schema, is a structure that describes the possible outlook of multiple structure it is representing (Complex Tree Instances (CTI). As such, the CTS has less actual nodes and no attribute values – no actual data. On the other hand the CTS may have some types that will not occur in the instance of this schema.

Complex Tree Instance, like the actual database table or XML document, is a structure that stores the data or knowledge. As such, it must not have all the types from the schema, but all the attributes must have determined values (even if it represents unknown) and the structure of the nodes may become very complex.

We represent the *i*-th Complex Tree Schema as $CT^{(i)}$, where:

$$CT^{(i)} = (T^{(i)}, S^{(i)}, V^{(i)}, E^{(i)})\ . \tag{6}$$

Each CTS may have multiple Complex Tree Instances, we represent the *j*-th CTI of *i*-th CTS as as $CT^{(i)}_{(j)}$, where:

$$CT^{(i)}_{(j)} = (T^{(i)}_{(j)}, S^{(i)}_{(j)}, V^{(i)}_{(j)}, E^{(i)}_{(j)}) \ . \tag{7}$$

Two distinct definitions of the same structure require different, although similar, approaches to defining the Integration Task. These are presented in the following two sections.

3.3 Integration Task for CTS

For a structure defined as in section 3.1, with respect to the notation in 3.2, the integration Task of Complex Tree Schema is given as follows:

The input of the integration process is *n* Complex Trees $CT^{(1)}$, $CT^{(2)}$, …, $CT^{(n)}$:

$$\begin{gathered} CT^{(1)} = (T^{(1)}, S^{(1)}, V^{(1)}, E^{(1)}) \\ \vdots \\ CT^{(n)} = (T^{(n)}, S^{(n)}, V^{(n)}, E^{(n)}) \ . \end{gathered} \tag{8}$$

The output of the integration process is one Complex Tree Schema *CT**, connected with input structures by a group of criteria.

$$CT^* = (T^*, S^*, V^*, E^*) \ . \tag{9}$$

The parameters of the integration task, are the integration criteria $K=\{K_1,K_2,\ldots K_n\}$ tying *CT** with CT_1 and CT_2, each at least at a given level $\alpha_1,\ldots,\alpha_n$

$$K_i\left(CT^* \middle| CT^{(1)}, CT^{(2)}, \ldots, CT^{(n)}\right) \geq \alpha_i \ . \tag{10}$$

Alternatively the integration process may be defined as a function *I*:

$$I: CTS^n \rightarrow CTS \ , \tag{11}$$

where *CTS* is the space of all possible Complex Tree Schemas.

Introduction of explicit integration criteria to the definition of the integration task is crucial, as without those the result of the process would not hold any relation to the input. The criteria also help define the aim of the integration – whether keeping all the information or the precise structure is most important (for more details see section 4).

Note also that integration defined as above automatically has one criterion met, without explicitly stating so, that is keeping the same type of structure as the output, as the input structures. Thus this criterion is not necessary to be defined.

3.4 Integration Task for CTI

For the same Complex Tree Schema multiple Instances are allowed, with different actual data (i.e. it is possible to have multiple XML documents with the same XSD). In this section we simplify the notation (only one CTS is used), that is instead of notation:

$$CT^{(i)}{}_{(j)} = (T^{(i)}{}_{(j)}, S^{(i)}{}_{(j)}, V^{(i)}{}_{(j)}, E^{(i)}{}_{(j)}) \ , \tag{12}$$

we will use:

$$CT_j = (T_j, S_j, V_j, E_j) \ . \tag{13}$$

The Integration Task for CTI may be now defined as follows:

The input of the integration process is *n* Complex Trees CT_1, CT_2, …, CT_n from the same schema.

$$\begin{gathered} CT_1 = (T_1, S_1, V_1, E_1) \\ \vdots \\ CT_n = (T_n, S_n, V_n, E_n) \ . \end{gathered} \tag{14}$$

The output of the integration process is one Complex Tree CT^*, connected with input structures by a group of criteria.

$$CT^* = (T^*, S^*, V^*, E^*) \ . \tag{15}$$

The parameters of the integration task, are the integration criteria $K=\{K_1,K_2,\ldots K_m\}$ tying CT^* with CT_1, CT_2,…,CT_n, each at least at a given level $\alpha_1,\ldots,\alpha_n$

$$K_i(CT^*|CT_1, CT_2, \ldots, CT_n) \geq \alpha_i \ . \tag{16}$$

Alternatively the integration process may be defined as a function *I*:

$$I: CTI^n \rightarrow CTI \ , \tag{17}$$

where *CTI* is the space of all possible Complex Tree Instances.

The final note to the previous section holds true for Integration of CTI.

4 Integration Criteria

The integration task defined in section 3.3. requires explicitly given criteria, that have to be met at given levels. In this section some basic criteria will be defined, based both on criteria used for XML integration and theoretically defined criteria used for general integration.

4.1 Completeness

The definition of the Complex Tree structure presented in section 3.1 allows for multiple definitions of the completeness criterion. Those would be the *definitions completeness*, *structure completeness* and *data completeness*. It is also possible to merge all those criteria into a *general completeness*, based on some parameters (given by the expert for each problem).

Definition 2. Definitions Completeness is a criterion measuring if all the types and their descriptions (T and S) from the input structures remain after the integration.

$$C_d(CT^*|CT_1, CT_2) = \frac{1}{card\{t:t\epsilon T_1 \cup T_2\}} \sum_{t\epsilon T_1 \cup T_2} \left[\frac{1}{2} m_T(t, T^*) + \frac{1}{2 \cdot card\{a:a\epsilon S_1(t) \vee S_2(t)\}} \sum_{a\epsilon S_1(t) \vee S_2(t)} m_S(a, S^*(t))\right] \tag{18}$$

where $m_T(t,T)$ determines whether element t exists in T and $m_S(a,S(t))$ determines whether a is an attribute of t according to S.

Definition 3. Structure completeness is a criterion measuring if all the nodes (identified by types and labels) from the input structures remain after the integration.

$$C_s(CT^*|CT_1, CT_2) = \frac{1}{card\{v:v\epsilon V_1 \cup V_2\}} \sum_{v\epsilon V_1 \cup V_2} m_V(v, V^*) , \tag{19}$$

where $m_V(v,V)$ determines whether element v exists in V.

Definition 4. Data completeness is a criterion measuring if all the data (attribute values) from the input structures remain after the integration.

$$C_a(CT^*|CT_1, CT_2) = \frac{1}{card\{v:v\epsilon V_1 \cup V_2\}} \sum_{v\epsilon V_1 \cup V_2} \left[\frac{1}{card\{a:a\epsilon A(v)\}} \sum_{a\epsilon A(v)} m_A(a(v), a^*(v))\right] , \tag{20}$$

where $m_A(a(v),a^*(v))$ determines the percentage of attributes of v from input trees occurring in the output trees.

Definition 5. General completeness is a criterion measuring if all the other completes measures are met, each with a given weight to the final result.

When considering CTS the first element is the most important and the parameter γ is 0 (there are no attribute values), while when considering CTI the two last elements are the most important.

$$C(CT^*|CT_1, CT_2) = \alpha C_d(CT^*|CT_1, CT_2) + \beta C_s(CT^*|CT_1, CT_2) + \gamma C_a(CT^*|CT_1, CT_2) . \tag{21}$$

4.2 Precision

Definition 6. Precision is a criterion measuring if no new elements were introduced during the integration and if no duplicate information is in the output tree.

In case of CTS the first element is more important, while in case of the CTI the second element should have higher weight.

$$P(CT^*|CT_1, CT_2) = \alpha \frac{card\{t:t\epsilon T_1 \cup T_2\}}{card\{t:t\epsilon T^*\}} + \beta \frac{card\{v:v\epsilon V_1 \cup V_2\}}{card\{v:v\epsilon V^*\}}. \quad (22)$$

The literature presents some similar criteria called optimality and understandability, which represent that each element after integration represents corresponding elements from both input trees and that each element after integration represents at most one element from each input tree. Here, these criteria are redefined to better fit the definition of the Hierarchical Structure proposed in section 3.1.

4.3 Optimality

The definition of optimality criterion in this section bears little resemblance to the optimality known in literature. Instead it is based on a criterion known as 1-optimality, that has not been used for tree structures often.

Definition 7. Optimality is a criterion measuring how close the output tree of the integration process is to the input trees, in terms of a given tree distance.

$$M(CT^*|CT_1, CT_2) = \frac{\min_H \left(d(CT,CT_1)+d(CT,CT_2)\right)}{d(CT^*,CT_1)+d(CT^*,CT_2)}, \quad (23)$$

where $d(H_1,H_2)$ is a distance measure for Complex Trees. A proposition of such distance is given below.

Definition 8. Complex Tree Distance is a weighted average of distances between the types of both CTs, the type attribute functions of both CTs and tree structures of both CTs.

In case of CTS the first two elements of the sum are more important, while for the CTI the tree distance is the most important.

$$d(CT_1, CT_2) = \alpha \cdot d_T(T_1, T_2) + \beta \cdot \frac{1}{card\{T_1 \cup T_2\}} \sum_{i=1}^{card\{T_1 \cup T_2\}} d_S(S_1(t_{1i}), S_2(t_{2i})) + \gamma \cdot d_{VE}((V_1, E_1), (V_2, E_2)), \quad (24)$$

where
distance between sets of types is measured as the relation of number of corresponding types in both sets to the overall number of types:

$$d_T(T_1, T_2) = 1 - \frac{card\{t:t\epsilon T_1 \wedge t\epsilon T_2\}}{card\{t:t\epsilon T_1 \vee t\epsilon T_2\}}, \quad (25)$$

distance between type attribute functions S is measured as the number of attributes attributed to the same types by both:

$$d_S(S_1(t_1), S_2(t_2)) = \begin{cases} 1 - \frac{card\{a:a\in S_1(t_1) \wedge a\in S_2(t_2)\}}{card\{a:a\in S_1(t_1) \vee a\in S_2(t_2)\}} & \text{if } t_1 = t_2 \\ 0 \text{ otherwise} \end{cases}, \quad (26)$$

distance between trees (directed graphs) is measured using a fast tree distance measure; in this case the tree edit measure introduced in [12], due to its linear complexity:

$$d_{VE}\big((V_1, E_2), (V_1, E_2)\big) = \frac{I \cdot c_i + D \cdot c_d + \sum_{r \in R} c_r(v_{1r}, v_{2r})}{N \cdot c_i + M \cdot c_d}, \tag{27}$$

in which:

N and *M* are the numbers of nodes in first and second Complex Tree, respectively,

I and *D* are the numbers of insert and delete operations required to transform first tree to the second, respectively,

c_i and c_d are the costs of insertion and deletion operation, respectively (here assumed to be 1),

R is the set of all pairs of nodes to be replaced,

$c_r(v_1,v_2)$ is a function determining the cost of replace operation between two nodes, normally a complex operation due to character of this meta-relationship, here for simplicity we assume it is equal to 0 if nodes are identical, equal to 1/3 if the nodes have the same type, equal to 2/3 if the nodes have the same type and label, and if they share identical attribute values it is equal to:

$$c_r(v_{1r}, v_{2r}) = 1 - \frac{card\{a : a \in A_1 \wedge a \in A_2\}}{3card\{a : a \in A_1 \vee a \in A_2\}}. \tag{28}$$

5 Simple Integration Algorithm

The aim of defining explicit integration criteria for the Integration Task is to make the decision about selecting an actual algorithm for a specific practical task simpler for the user. Even the few criteria presented in section 4 allow for multiple approaches to integration, for example completeness equal to 1 will almost never occur at the same tame as optimality equal to 1. In some tasks the first criterion would be more important, in others – the second. In this section we will present a simple algorithm that matches a given set of criteria at desired levels, in order to show that it is possible to design algorithms based on given requirements.

The simple integration algorithm was designed to fit the following criteria:

- Integration is done on schema level;
- General completeness must be equal to 1, with parameters α=2/3, β=1/3 and γ=0, respectively (see note in section 4.1);
- Precision equal to 1, with parameters α=½ and β=½, respectively;
- Non-zero optimality (note: this is always assured).

The description of the integration task is already provided in section 3.3. Below, only the algorithm will be presented, in a version for two input CTS's (n = 2).

Input: $CTS^{(1)}$, $CTS^{(2)}$
Output: CTS^*
BEGIN
$T^*:=T^{(1)}$
For each type t in $T^{(2)}$ do

If there is no matching (similar) type t present in T^*, add t to T^*
otherwise go to next t

For each type t in CTS^* do

1. For each a in $S^{(1)}(t)$ do
   Add a to $S^*(t)$
2. For each a in $S^{(2)}$(t) do
   If a is not present in $S^*(t)$, add a to $S^*(t)$
   otherwise go to next a

$V^*=V^{(1)}$
$E^*=E^{(1)}$
For each node v in $V^{(2)}$ in post-order:

1. If v is the root element of $CTS^{(2)}$, and the root of CTS^* is different than v, and if v is not a child of the root of CTS^*, add v as a child of the root of CTS^*; go to next node
2. If v is the root element of $CTS^{(2)}$ matching the root of CTS^*, go to next node
3. If v is present in CTS^*
   If v is a child of node w in $CTS^{(2)}$ and is not a child of the same node in CTS^*, add v as a child of node w in CTS^*
   otherwise go to next node
4. If v is not present in CTS^*
   Locate node w in CTS^* that has a matching (similar) node w as the parent of node v in $CTS^{(2)}$
   Add node v as a child of node w in CTS^*

END

The algorithm was designed specifically for low computational complexity. The fist loop section takes no more than $O\left(card\{T^{(2)}\}\right)$, which is linear. The same holds for the second loop section (which may be restricted from top by $\max\{a: a \in S^{(1)}(t) \vee a \in S^{(2)}(t), t \in T^{(1)} \cup T^{(2)}\}$). While the last loop section may appear complex, only case selection takes place and the complexity is $O\left(card\{V^{(2)}\}\right)$. The overall computational complexity of the algorithm is then linear and restricted by the number of elements in the structures, specifically $O\left(card\{T^{(2)}\} \cdot \left(1 + \max\{a: a \in S^{(1)}(t) \vee a \in S^{(2)}(t), t \in T^{(1)} \cup T^{(2)}\}\right) + card\{V^{(2)}\}\right)$.

5.1 Example

For an example of this simple algorithm, we will use the following structures:

$CT_1 = (T_1, S_1, V_1, E_1)$, $CT_2 = (T_2, S_2, V_2, E_2)$
T_1={root, leaf}, T_2={root, branch, leaf}, $\forall_t: S_1(t) = S_2(t) = \emptyset$
V_1={ (root,root,-), (a,leaf,-), (b,leaf,-) }
V_2 ={ (root,root,-), (a,leaf,-), (b,leaf,-), (c,leaf,-), (x,branch,-) }
E_1 = { (root,a), (root,b) }, E_2 = { (root, a), (root, x), (x,b), (x,c) }

The integration algorithm first joins the sets of node types, this results in:
$T^* = \{root, branch, leaf\}$

The function S determining the attributes in both Complex Trees is null, as there are no attributes present. The S^* thus remains null.

The process of integrating both tree structures takes place for V and E simultaneously. The result of this process is in this example:

$V^* = \{ (root,root,-), (a,leaf,-), (b,leaf,-), (b,leaf,-), (c,leaf,-), (x,branch,-) \}$
$E^* = \{ (root,a), (root, b), (root, x), (x, b), (x,c) \}$

The node labeled as b appear twice in the output structure, otherwise it is similar to T_2. The general completeness is equal to 1, the precision is also equal to 1 and the optimality is greater than 0.

6 Future Work

This paper presented only a general idea and basic concepts of the integration task with clearly defined criteria and with a structure designed just for the integration purposes.

The criteria described in section 4 are the most basic ones, deriving both from the study of XML integration in practical enterprises and from the general theory of integration, independent of the data structure. It is possible to develop more criteria, both theoretically and based on the existing solutions and integration requirements.

The simple algorithm presented in section 5 is the obvious solution for the criteria requirements stated for its creation. The same criteria may be the base for better (i.e. faster) algorithms or algorithm compatible with other, more precise criteria. With different set of base criteria, different algorithms may also be proposed. Specifically, the existing algorithms, once the criteria for them are explicitly stated, may have the possibility of improvement.

Acknowledgements. This paper was partially supported by Polish Ministry of Science and Higher Education under grant no. N N519 407437.

References

1. Adams, E.N.: N-Trees as Nestrings: Complexity, Similarity, and Consenss. Journal of Classification 3, 299–317 (1986)
2. Arenas, M., Libkin, L.: A Normal Form for XML Documents. ACM Transactions on Database Systems 29(1), 195–232 (2004)
3. Bae, J.K., Kim, J.: Integration of heterogeneous models to predict consumer behavior. Expert Systems with Applications 37, 1821–1826 (2010)
4. Barthelemy, J.P., McMorris, F.R.: The Median Procedure for n-Trees. Journal of Classification 3, 329–334 (1986)
5. Bonifati, A., Ceri, S.: Comparative Analysis of Five XML Query Languages. ACM SIGMOD Record 29(1) (2000)

6. Danilowicz, Cz., Nguyen, N.T.: Methods for choice of representation of ordered partitions and coverings, Wroclaw (1992)
7. Day, W.H.E.: Optimal Algorithms for Comparing Trees with Labled Leaves. Journal of Classification 2, 7–28 (1985)
8. Delobel, C., Reynaud, C., Rousset, M.C., Sirot, J.P., Vodislav, D.: Semantic integration in Xyleme: a uniform tree-based approach. Data & Knowledge Engineering 44, 267–298 (2003)
9. Do, H.-H., Melnik, S., Rahm, E.: Comparison of Schema Matching Evaluations. In: Chaudhri, A.B., Jeckle, M., Rahm, E., Unland, R. (eds.) NODe-WS 2002. LNCS, vol. 2593, pp. 221–237. Springer, Heidelberg (2003)
10. Farach, M., Przytycka, T.M., Thorup, M.: On the agreement of many trees. Information Processing Letters 55, 297–301 (1995)
11. Lian, W., Cheung, D.W., Mamoulis, N., Yiu, S.M.: An Efficient and Scalable Algorithm for Clustering XML Documents by Structure. IEEE Transactions on Knowledge and Data Engineeing 16(1) (January 2004)
12. Maleszka, M., Mianowska, B., Prusiewicz, A.: On some approaches to reduce the computational costs of similarity measures between XML trees. In: Nguyen, N.T., Kolaczek, G., Gabrys, B. (eds.) Knowledge Processing and Reasoning for Information Society, pp. 165–180. Exit, Warszawa (2008)
13. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. The VLDB Journal 10, 334–350 (2001)
14. Routledge, N., Bird, L., Goodchild, A.: UML and XML Schema. In: Zhou, X. (ed.) Conferences in Research and Practice in Information Technology, vol. 5 (2002)
15. Stinebrickner, R.: s-Consensus Trees and Indices. Bulletin of Mathematical Biology 46, 923–935 (1984)

Prototype of Object-Oriented Declarative Workflows

Marcin Dąbrowski[1], Michał Drabik[1], Mariusz Trzaska[1], and Kazimierz Subieta[1,2]

[1] Polish-Japanese Institute of Information Technology
{mdabrowski,mdrabik,mtrzaska,subieta}@pjwstk.edu.pl
[2] Institute of Computer Science Polish Academy of Sciences

Abstract. While in the traditional workflow processes the control flow is determined statically within process definitions, in declarative workflow processes the control flow is dynamic and implicit, determined by conditions that occur in the workflow data and the service environment. The environment consists of *active objects*, which play a double role. On the one hand, they are persistent data structures that can be queried and managed according to the syntax and semantics of a query language. On the other hand, active objects possess executable parts and represent workflow processes or tasks. The approach is motivated by features that are desirable in complex and less regular business processes: (1) the possibility of dynamic changes of process instances during their run, (2) mass parallelism of process instances and their components and (3) shifting the availability of resources that workflows deal with on the primary plan as a mean for triggering instances of process tasks. The paper presents the prototype of an object-oriented declarative workflows on a comprehensive example with roots in a real business case.

Keywords: workflow, object-oriented, declarative, query language, active object, dynamic workflow change, ODRA (Object Database for Rapid Application development), SBQL (Stack-Based Query Language).

1 Introduction

The workflow technology is a well developed domain with many commercial successes, which include such standards as BPEL [2], BPMN [8] and XPDL [13]. Nevertheless, there are still problems that undermine applications of workflow management systems in important business domains; in particular:

- Dynamic changes in workflow instances during their run. Dynamic workflow changes are the subject of many research papers, e.g. [1, 4, 5, 9, 10]. Although valuable results are achieved the problem in general still remains unsolved.
- Parallel execution of tasks within workflow processes. Currently, the parallelism is achieved by explicitly programmed splits and joins (AND, OR, XOR). In many cases such a parallelism is insufficient, for instance, when a process is to be split into many subprocesses, but their number is unknown in advance.
- Aborting a process or some of its parts. Currently, such situations are handled manually, with possibilities of inconsistencies and non-optimal human action.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 47–56, 2011.

- Resource management. In currently developed systems the control flow (a la Petri net) is on the primary plan. Resources necessary for execution (people, money, etc.) are on the secondary plan. But just availability, unavailability, planning and anticipating required resources are the main factors that should determine the process control flow.
- Tracking and monitoring. These activities should concerns the entire workflow environment and all running process instances, including databases that support workflows, the state of resources, anticipation of availability of resources, etc. For this reason the core for tracking and monitoring should be a query language (such as BPQL [6]) with the full algorithmic power rather than predefined tools.
- Parallel execution of workflow instances and their parts on many (hundreds) servers.
- Transaction processing. Classical implementations based on ACID properties and 2PC/3PC protocols are insufficient for workflows. Interactive business processes cannot be reversed, hence the attitude to transaction aborting should be changed.

In our last project we have investigated a new workflow paradigm that has the potential to overcome the above difficulties. We assumed that workflow instances can be changed during their run, hence they should possess a double nature. On the one side they are to be executable processes. On the other side, they should be considered database structures that are described by some conceptual schema and can be queries and manipulated as usual (nested) database objects.

The second assumption was inherent parallelism of all workflow processes and their parts. We avoid explicit splits and joins. Instead, we assume synchronization of parallel processes by special constructs of a query language. In this way our workflow instances remind PERT (Program Evaluation and Review Technique) networks rather than Petri nets. PERT naturally describes dependencies between tasks within non-computerized human activities and can be formalized using the object-oriented approach. Such a workflow system we describe as "declarative", because the control flow is not determined explicitly, but through declarative queries. Sequences of tasks can be supported by tasks' states and conditions on the states.

The third assumption is shifting the resource management on the primary plan. Resources (available, planned, anticipated) are reflected in the database. The control flow of process instances can be determined by conditions addressing resources.

In this way we came to the idea of *active objects*, which have the mentioned above double nature. Active objects are persistent data structures that are described by a database schema and can be queried and manipulated according to the syntax and semantics of a query language (in this role SBQL [12]). On the other hand, active objects possess active (executable) parts. We distinguish four kinds of such active parts: *firecondition*, *execution code*, *endcondition* and *endcode* (in this role SBQL too). An active object waits for execution until the time when its firecondition becomes true. After that, the object's execution code is executed, and all its active sub-objects are put into the *waiting-for-execution* state (and perhaps executed if their fireconditions become true). Execution of the execution code of a given active object is terminated when either all the actions are completed (including active sub-objects) or its endcondition becomes true. After fulfillment of an endcondition some actions

might be necessary (e.g. aborting transactions), thus an optional endcode. Each active object is an independent unit that can be manipulated by SBQL functionalities (updated, deleted, etc.). Active objects can be nested. In this way they can represent workflow processes, their tasks, subtasks, etc. Active parts can also be updated; their parsing, type checking, optimization and compilation are performed on-the-fly. Bindings are mostly dynamic.

The paper [9] presents a framework for formalizing process graphs and updating operations addressing such a graph. There are valuable observations concerning the necessity of dynamic workflow changes for real business processes and the necessity of strong discipline within the changes to avoid violation the consistency of the processes. Numerous authors follow the ideas of this paper. The fundamental difference of our approach is that we do not determine explicitly the process control flow graph. It is on the secondary plan, determined dynamically and implicitly by fireconditions and endconditions. In majority of cases the control flow graph can be different depending on a runtime state of the workflow, database and computer environment. The problem of the necessity of various control flow graphs for the same business process is one of the motivations for the research presented in [9], but it is not easy to see how such a feature can be achieved within the proposed formal workflow model. In our case the feature is an inherent property of the idea.

A declarative workflow deals with a database schema that describes executable data structures representing the processes, thus by definition enabling all updating operations that are provided within the assumed programming language SBQL. To restrict undesirable changes that may violate the consistency of processes we can use the semi-strong typing system that is implemented for SBQL. This of course may not be enough for more sophisticated situations. For this goal we plan to implement facilities that are well-know from relational systems, such as user rights, integrity constraints, business rules and triggers.

In [3] we describe in detail the concept of active object and related issues. To check the concept we have implemented three different prototypes. This paper is the first description of the third prototype [11], most advanced and with no previous tradeoffs concerning the new idea and current workflow technologies. The prototype is still a proof-of-a-concept rather a usable tool. More research and financial support is necessary to turn it into a product.

The rest of the paper is organized as follows. Section 2 presents basic assumptions and the architecture of the prototype. Section 3 presents how dynamic instance modifications can be performed. The presentation is based on a real example of a workflow that was taken from the experience in developing a bank system supporting credit processes. Section 4 concludes.

2 Prototype of Object-Oriented Declarative Workflow

The prototype [11] is built upon the ODRA system [7] and a Web-based application for manipulating prototype functionalities. The Web part uses the Groovy and Grails technologies. A workflow server part is written in Java.The prototype can be tested

using a Web application called SBQL4Workflow. It allows for all administrative tasks like creating process definitions, manipulating them, instating processes, freezing parts of a running workflow and more. A GUI generation module is based on the core Grails framework technology called GSP (Groovy Server Pages). It is similar to JSP (Java Server Pages). A client side is equipped with advanced AJAX controls to allow dynamic loading of a process tree and manipulating workflow objects minimizing the need to reload web pages. The SBQL code editor with syntax highlighting that is included into GUI makes the work with workflows much easier.

The ProcessMonitor is a Java application that can be run as a separate thread on a separate machine. It periodically checks (basing on timeouts) each ProcessInstance. Then, according to the values retrieved from condition codes, the ProcessMonitor executes the execution code of the process.

The prototype is build using the standard three layer approach. A middle layer consists of the Application Logic and ODRA Wrapper. The corresponding API allows for work with workflow objects. It is used not only by GUI and the ProcessMonitor but can be used by any Java program, so writing a different client application is possible. The ODRA Wrapper is a wrapper between the JOBC (Java Object Base Connectivity) library that is used to access the ODRA DBMS through queries and Java business objects used by the Application Logic. All workflow data are stored on the ODRA DBMS.

The process objects represent structures created by the workflow programmer. Once a process is initiated, all data, including the data of sub-processes, are copied to the corresponding ProcessInstance objects. The parent-children bidirectional pointer, combined with other SBQL query operators, gives a great flexibility in expressing conditions and codes. For instance:

- Find all my children (the code is written with regard to one particular Process).
- Find my parent.
- Find a process with a given status.
- Find a process with a given name.

These constructs can be easily combined for more complex search, for instance:

- Find a child that has a certain name and status.
  children **where** *name* = 'foo' **and** *status* = *processStatus.FINISHED*
- Check if all my children have the status ‘Finished’.
  exists(*children* **where** *status* = *ProcessStatus.FINISHED*)
- Find my “brother” (using *parent.children*).
- Find all my “nephews” (using *parent.children.children*).

To allow processes to store ad-hoc additional data we have provided the Attribute class with a set of methods in the Process and ProcessInstance classes addressing attributes. The attributes can be easily used to control the flow and allow communication between the processes (as one Process can query another Process attributes and change their values).

getAttribute('contractSigned')='true'

The code example presents the access to the attribute named ‘contractSigned’

setAttribute('mailSent'; *sendMail*('foo@bar.com'; 'Mail content'))

The code sends an email and stores the result (success or failure) as a process attribute.

3 Dynamic Instance Modifications

After creating a process instance for any business reason it can be the subject of changes without changing the corresponding process definition. Changes can be performed after launching an instance. Changes can also concern process definitions. Our prototype has the following options concerning changes within workflows:

- Editing and modifying process definitions;
- Instantiating process instances according to the definitions;
- Editing and modifying a process instance by editing its core attributes such as name, fire condition, end condition, execution code, end code, etc.;
- Running any SBQL program (having updates, inserts, deletions, etc.) in order to manipulate the entire workflow environment, including nested active objects representing processes, a resource database and any other persistent or volatile objects that are available within the environment. The programs include SBQL queries as expressions.

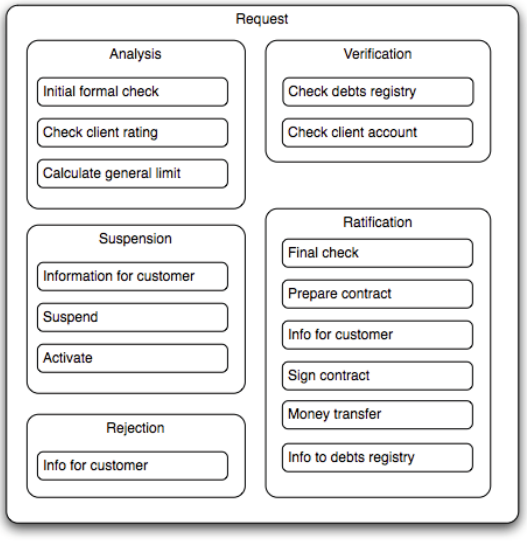


Fig. 1. ”Bank credit” workflow instance initial structure

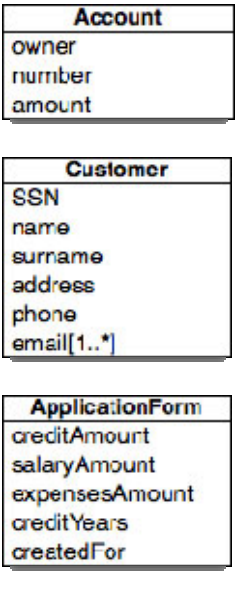


Fig. 2. Additional resources objects

Changing a process instance may require further changes of other instances to ensure consistency of the corresponding business process. Our prototype offers much flexibility in controlling process instances without altering other instances, mainly by preparing more generic fireconditions and endcondition that are insensitive to some changes of active objects. For instance, an endcondition can test completion of all corresponding sub-processes with the use of a universal quantifier. In many cases, however, altering a process instance may require some actions on other instances. These actions can be nested within a transaction.

To demonstrate the possibility of dynamic instance modification we have created a comprehensive example of real business processes concerning issuing and granting bank credits for customers. The structure (schema) of the process presented in Fig.1. All presented SBQL codes are tested on the prototype.

Example 1. *It demonstrates how to insert a new process into a workflow instance structure, without the need of changing the already working process instances details*. Letters in brackets correspond to status of a process instance: FINISHED, ACTIVE, WAITING.

Assume a bank credit process in progress. At the end of it the money that the customer has requested is transferred to his/her account. However after the transfer the customer has decided to change the target account. In this situation we can correct the working workflow instance by inserting an additional subprocesses that will do the requested operation. To achieve the goal we have to find a workflow instance that should be modified and create a new process called „*New account money transfer*" in it. It will have two attributes to store the value of an old and new account number, named respectively „*oldAccountNr*", "*newAccountNr*".

After inserting the new process its status is set to „*Waiting*". The next step is to set the proper firecondition. It should check whether exists the process "Money transfer" with the status equal to "FINISHED" at the same hierarchy level. It can be found within the children of "Ratification" process (the parent of the "*New account money transfer*"). The purpose of the newly inserted process is to withdraw the money from the old account and transfer it into the new one. The sequence of actions that needs to be performed is shown on the activity diagram below:

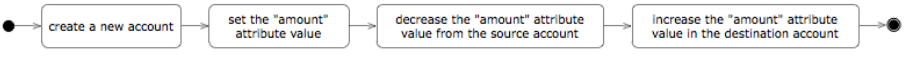


Fig. 3. Set of actions performed by "*New account money transfer*" execution code

The execution code creates a new account object with the number delivered from the „*newAccountNr*" attribute of this process instance. Now we should find out the information about the amount of money that should be transferred. This information is a part of the „*ApplicationForm*" object which is available, so the task will be to find the application form assigned to the current customer and obtain the „*creditAmount*" value. To make this value available for further processing it will be saved as a value of a newly created attribute called „*amount*". Now it is possible to withdraw the money from the old account. To do that the right Account object should be found (the account number is the value of the process instance „*oldAccountNr*" attribute), and

the value of its „*amount*" attribute should be decreased by a value of this process instance „*amount*" attribute. Then the new Account object should be found (the account number is the value of the process instance „*newAccountNr*" attribute) and its "amount" attribute should be increased by the value of the process instance "*amount*" attribute.

Before making any changes to a working workflow instance it should be suspended so that the state before and after applying the change is consistent. Knowing the current state of all process instances we can assume that a newly created process should start when the „*Money transfer"* is finished, and should end when the transfer operation between accounts is complete. In this case the insertion of a new process instance doesn't influence any other process, the construction of „*Ratification"* end condition ensures that it will not finish before every of its child finishes.

Example 2. *This example demonstrates the possibility of modification of a running workflow instance structure in order to meet new requirements. It shows how the proper written execution code can modify behavior of the workflow instance and how the workflow administrator can influence the behavior.*

The customer has decided to increase the credit amount just before signing the contract. In that case there is no need to restart the whole workflow instance, but only some of the processes.

The activities that the workflow administrator have to perform are the following:

1. Suspend a chosen workflow instance.
2. Add new process instance "*Increase credit*" (as a child of "*Ratification*").
3. Delete process instances that are no longer required ("*Verification*" - child of "*Request*", and "*Initial Formal Check*" - child of "*Analysis*").
4. Change the conditions of other involved process instances to conform to the new structure.
5. Resume workflow instance.

The purpose of the newly created "*Increase credit*" process instance is to change the "*creditAmount*" attribute of the "*ApplicationForm*" object associated with the customer. Activity diagram shown below gives an overview of the actions performed by execution code:

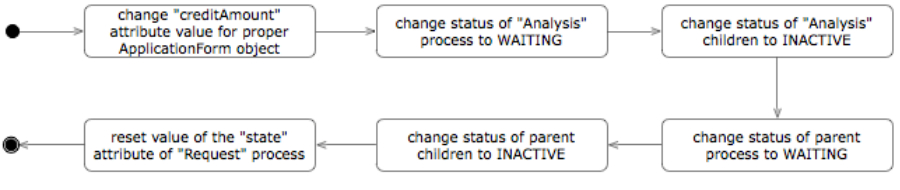


Fig. 4. Set of actions performed by "*New account money transfer*" execution code

It also resets the "Analysis" and "Ratification" process instances in order to perform their tasks once more. It is done by changing their status to "*Waiting*" so the process monitor will include them when checking the candidates to activate. The children of this processes should also be included with this difference that their status will be changed to "*Inactive"*.

Apart from process instances there are also attributes that values should be set to the previous state. This concerns the "*state*" attribute of the "*Request*" process instance, which holds the information about the current status of the application form and needs to be reset.

When the "*Increase credit*" will perform the given task it's no longer needed in the system and to ensure that it will act only once, we can create such an end code that will delete it.

The next step is to get rid of unnecessary process instances such as "*Initial formal check*" and "*Verification*", because there is no need to repeat them when only the amount of the credit is changed. After that the conditions of some process instances have to be adjusted. The "*Check client rating*" will start as soon as "*Analysis*" is active instead of start after the "*Initial formal check*" finishes.

All of the statements that concern "*Initial formal check*" should also be removed from the "*Analysis*" end condition.

The "*Ratification*" will no longer start after the "*Verification*" but as soon as the "*Analysis*" finishes.

Now the workflow instance is ready to properly handle the updated application form and perform suitable tasks in order to complete the request.

Example 3. *It demonstrates how to apply a modification that affects several process instances of the same kind.*

The modifications are to be applied to the „*bankLimit*" attribute of the „*Calculate general limit*" process to 700000. Changing the process definition is straightforward through the GUI tool. However, changing manually all of the working instances in this way is awkward and can be error prone. For this reason we create an SBQL statement which will access the workflow environment and will do the necessary modifications. The statement finds all of the instances of the "*Calculate general limit*" subprocess, which has an "*Inactive*" or "*Waiting*" status, and then updates the value of the "*bankLimit*" attribute to the new value.

An SBQL statement which updates the "Calculate general limit" "bankLimit" attribute in the proper workflow instances:

(*ProcessInstance* **where** *name* = 'Calculate general limit' **and** (*status=ProcessStatus.INACTIVE* **or** *status=ProcessStatus.WAITING*)).(*setAttribute*('bankLimit';'700000'))

Example 4. *It demonstrates how the working process instance can dynamically create new process instances.*

The commonly considered case in a process definition with parallel subprocesses is that a process instance is to be split into a fixed number of subprocesses. In many business situations the case leads to severe limitations, because the number of the subprocesses is known only during the execution of the instance. In such a case we should provide an option to create new subprocess instances dynamically, within the execution code of the process instance. To show this possibility we consider example where there is a need to send an e-mail with some information to the customer. During filling a request form a customer can provide some alternative e-mail addresses and we want to ensure that our e-mail will be delivered to all of them. The process

responsible for the contact with the customer is "*Information for customer*" so we will modify it to provide required functionality. During the execution of this process instance it will create as many children process instances sending e-mails as required.

The execution code of the "*Information for customer*" creates a process instance for each of an e-mail address of a current customer:

(*self* **as** *p*).(((*Customer* **where** *SSN=parent.parent.ProcessInstance.getAttribute*('customerSSN')) . *email* **as** *e*). (**create** *ProcessInstance*(...)))

Then we can populate the execution code for the newly created process instances in such way that it will send mail for one given address. If an e-mail was sent successfully it will create an attribute "*mailSent*" with value 1 to hold the information which will be used later to decide if the process should end or restart. Below there is a part of the execution code of a dynamically created process instance that sends e-mail to a given address and sets an attribute value depending on the result:

setAttribute('mailSent';*sendMail*('x@x.x';'Dear Customer, your application is to be corrected.'));

Part of the execution code of a dynamically created process instance that restarts it when sending mail has failed:

if(*getAttribute*('mailSent')=0) **then** (*status* := *ProcessStatus.WAITING*);

4 Conclusion

We have presented the idea of an object-oriented declarative workflow management system that is especially prepared to achieve an important goal: the possibility of changing process instances during their run. We have discussed consequences of such a requirement and have argued that such a revolutionary feature cannot be achieved on the ground of traditional approaches to workflows based on specification of control flow graphs. Our idea allows to achieve next important features, such as mass parallelism of processes and flexible resource management. The idea is supported by the working prototype that shows its feasibility. In the paper we present comprehensive examples showing how a declarative workflow can be defined and how it can be dynamically changed. The examples have shown the feasibility of the idea of declarative workflows for real business cases. The prototype is still under development.

References

1. van der Aalst, W.M.P.: Generic workflow models: How to handle dynamic change and capture management information? In: Proc. 4th Intl. Conf. on Cooperative Information Systems (CoopIS 1999), Los Alamitos, CA (1999)
2. Andrews, T., et al.: Business Process Execution Language for Web Services, Version 1.1. OASIS (2003)
3. Dąbrowski, M., Drabik, M., Trzaska, M., Subieta, K.: Dynamic Changes of Workflow Processes (September 2010) submitted to publication
4. C.A.Ellis, C.A., Keddara, K., Rozenberg, G.: Dynamic change within workflow systems. In: Proc. ACM Conf. on Organisational Computing Systems, COOCS 1995 (1995)

5. Ellis, C.A., Keddara, K., Wainer, J.: Modelling workflow dynamic changes using time hybrid flow. In: Workflow Management: Net based Concepts, Models, Techniques and Tools (WFM 1998), Computing Science Reports, vol. 98(7), Eindhoven University of Technology (1998)
6. Momotko, M., Subieta, K.: Process query language: A way to make workflow processes more flexible. In: Benczúr, A.A., Demetrovics, J., Gottlob, G. (eds.) ADBIS 2004. LNCS, vol. 3255, pp. 306–321. Springer, Heidelberg (2004)
7. ODRA (Object Database for Rapid Application development): Description and programmer manual (2008), http://www.sbql.pl/various/ODRA/ODRA_manual.html
8. OMG. Business Process Modeling Notation (BPMN) specification. Final Adopted Specification. Technical Report (2006)
9. Reichert, M., Dadam, P.: ADEPTflex: Supporting dynamic changes of workflow without loosing control. Journal of Intelligent Information Systems 10(2), 93–129 (1998)
10. Sadiq, S., Orlowska, M.E.: Architectural considerations in systems supporting dynamic workflow modification. In: Jarke, M., Oberweis, A. (eds.) CAiSE 1999. LNCS, vol. 1626, Springer, Heidelberg (1999)
11. SBQL4Workflow Prototype Implementation (May 2010), http://tomcat.pjwstk.edu.pl:8080/ProjectWorkflow/newsitem/list
12. Subieta, K.: Stack-Based Architecture (SBA) and Stack-Based Query Language, SBQL (2008), http://www.sbql.pl/
13. WfMC, WorkFlow process definition interface – XML Process Definition Language. WfMC TC 1025 (Draft 0.03a), May 22 (2001)

Extraction of TimeER Model from a Relational Database

Quang Hoang and Toan Van Nguyen

Hue University of Sciences
77 Nguyen Hue, Hue City, Vietnam
hquang@hueuni.edu.vn, toan.fiit@gmail.com

Abstract. Related to the problem of temporal database design, we can design the relational target model from the TimeER model. Extraction of the TimeER model from a relational model is called reverse engineering of the relational model. Solving this problem will facilitate an upgraded temporal information system. That means we will investigate a conceptual model which is used to design the temporal relational model. This approach of the extraction is based on the characteristics of the set of attributes, the primary key, and the set of foreign keys of the relational schema in the temporal relational model. Thereby, we propose conversion algorithms relating to the identification of the components existing within the TimeER model respectively.

Keywords: Database reverse engineering, Temporal database design, Temporal conceptual schema, Relational model.

1 Introduction

To solve the problem of temporal conceptual schema design, the research community has developed many different temporal ER models such as TERM, RAKE, MOTAR, TEER, STEER, ERT, TimeER [6], [7]. The TimeER model (Time-Extended-EER) is constructed that extends the EER model to provide built-in support for capturing temporal aspects more sufficiently compared to other models. On that basis, we can design temporal logical data models. Related to the TimeER model, there have been proposals for the conversion methods from the TimeER model to the relational target model [1], [2], [5].

Another issue arising from this is that if we need to upgrade an information system, then we need to modify the TimeER model (the conceptual model) to match the requirements of the real world. However, suppose that we could not define the TimeER for any reason. That means we need to investigate the TimeER model which is used to design the temporal relational model (the logical model). Extraction of the conceptual model from the logical model is called *reverse engineering* of the logical model [9].

In addition, solving this problem will facilitate the conversion of the relational model to other database model. Especially, it is the data model for temporal

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 57–66, 2011.

XML documents. One of the techniques to extract XML documents from a relational model is that we can use the TimeER model as an intermediate conversion result.

Thus, this paper will focus on the development of an algorithm to extract the TimeER model from a temporal relational model. This approach of the extraction is based on the characteristics of the set of attributes, primary key, and a/the set of foreign keys of the relational schema in the temporal relational model. Thereby, we propose conversion algorithms relating to the identification of the components existing within the TimeER model respectively.

This paper will be structured as follows: Section 2 will give an overview of the components of the TimeER model. Section 3 will provide a mapping algorithm to convert the TimeER model to a lexically-based relational target model. Section 3 presents the basis for the theory proposed in Section 4, which discusses the method to extract the TimeER model from a relational model. Finally, in Section 5, a conclusion and a discussion of future work will be then given.

2 An Overview of the TimeER Model

The TimeER model is developed from the EER model [5]. This model provides built-in support for capturing the following temporal aspects: the lifespan of an entity (denoted LS), the valid time of a fact (denoted VT), and the transaction time of an entity or a fact (denoted TT). As defined, temporal aspects of the entities in the database can be either the lifespan (LS), or the transaction time (TT), or both the lifespan and the transaction time (LT). The temporal aspects of the attributes of an entity can be either the valid time (VT), or the transaction time (TT), or both the valid time and the transaction time (BT). Moreover, because a relationship type can be seen as an entity type or an attribute, the designer then can define the temporal aspects supported with this relationship type if necessary.

Components of the TimeER model

- Entity types. An entity type is represented graphically by a rectangle, while a weak entity type is by a double rectangle. If the lifespan, or the transaction time, or both of them of the entity type is captured, it is indicated by placing a LS, or a TT, or a LT, respectively, behind the entity type name.

- Attributes. A single-valued attribute is represented by an oval, while a multi-valued attribute is by a double oval. Different from a simple attribute, a composite attribute is represented by an oval connected directly to the components of the composite attribute.

We can give an assumption that each single-valued composite attribute is replaced with a set of simple attributes. Therefore, any attribute of an entity type or a composite attribute is the one of the following attribute types: single-valued simple attribute, or multi-valued simple attribute, or multi-valued composite attribute.

If the valid time, or the transaction time, or both of them is captured, this is indicated by placing a VT, or a TT, or a BT (BiTemporal), respectively, behind the attribute name.

- Relationship types. A relationship type is represented by a diamond. For each relationship type it can be decided by the database designer whether or not to capture the temporal aspects of the relationships of the relationship type. If some temporal aspect is captured for a relationship type we term it temporal; otherwise, it is called non-temporal.

- Superclass/subclass relations. As in the EER model, the TimeER model offers support for specifying superclass/subclass relations. It is not possible to change the temporal support of the inherited attributes, but it is possible to add attributes and to further expand the inherited temporal support of the class itself.

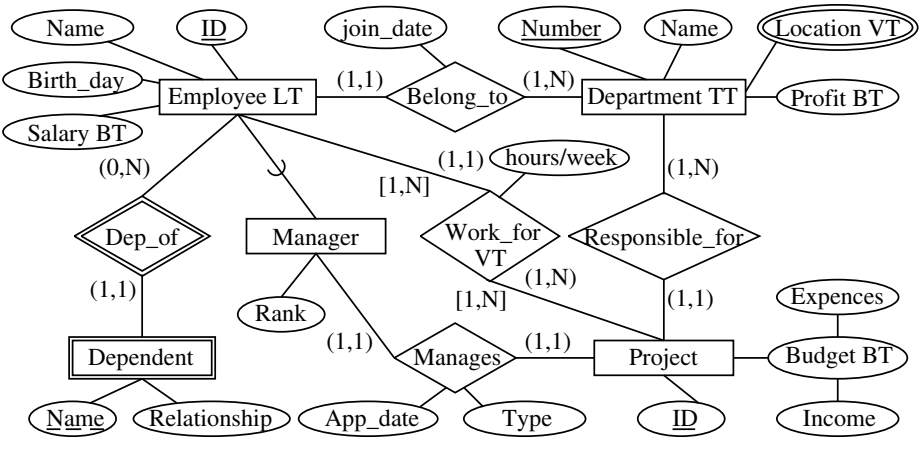


Fig. 1. TimeER diagram of a company database [5]

3 A Mapping Algorithm from the TimeER Model to the Relational Model

This section presents the mapping 7-step algorithm that transforms the components of a TimeER model to some relations, primary key constraints and foreign key constraints. The advantage of this mapping algorithm is that it can allow extension mapping for nested temporal multi-valued composite attributes of an entity type in the TimeER model [1].

Step 1. Mapping of entity types not participating in a superclass/ subclass relationship

For each entity type E, not participating in a superclass/subclass relationship and having non-temporal single-valued attributes $A_1, A_2, ..., A_n$, we consider the following cases.

a) Mapping of regular entity type: If E is the regular entity type whose key is $ID(E)$ (we assume that $|ID(E)| = 1$), then create a relation $R(E)$, called the *primary relation* representing the entity type E, which includes the attributes $ID(E) \subset \{A_1, A_2, ..., A_n\}$. The primary key of $R(E)$ is $ID(E)$.

b) Mapping of weak entity type: Let E be the weak entity type of the identifying relationship S with the owner entity type E'. It is supposed that E has the partial key $X \subset \{A_1, A_2, ..., A_n\}$. We then create the primary relation $R(E)$ that includes the attributes $FK \cup \{A_1, A_2, ..., A_n\}$, where FK is the foreign key referencing the relation $R(E')$. The primary key of $R(E)$ is $FK \cup X$.

It is assigned that the foreign keys of relations are indicated by the symbol "f.k." following the attribute names.

With the case where E is the temporal entity type (life span/transaction time), we create a new relation called the time relation representing the entity type E, denoted $TR(E)$, which includes the attributes $FK' \cup T$, where FK' is the foreign key referencing the relation $R(E)$. Note that T is the timestamp attributes depending on temporal support of E indicated by an asterix (*) in Table 1.

Table 1. Abbreviation used for temporal support of entity types and relationship *types*

(a) if * = LS then T = {LSs, LSe}
(b) if * = TT then T = {TTs, TTe}
(c) if * = LT then T = {TTs, TTe, LSs, LSe}

Consider $T' \subset T$ which is the set of underlined attributes in above table, the primary key of $TR(E)$ then is $FK \cup T'$.

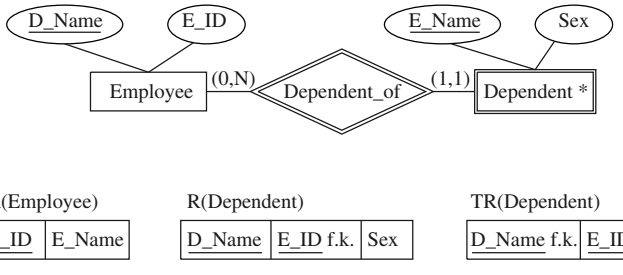


Fig. 2. Mapping of entity types not participating in a superclass/subclass relationship

Step 2. Mapping of entity types participating in a superclass/ subclass relationship

For each superclass/subclass relationship where superclass E has subclasses $S_1, S_2, ..., S_n$, we create the primary relation $R(E)$ referencing the entity type E to represent the superclass E. Suppose each subclass S_i has a set of added

non-temporal single-valued attributes X_i, we then create n new relations $SR(S_i)$, called *sub relations* representing the entity type S_i, which includes the attributes $FK \cup X_i$ (with $i = 1..n$) and the primary key is FK, where FK is the foreign key referencing the relation $R(E)$.

As in Step 1, if E or $S_1, S_2, ..., S_n$ are the temporal entity types, we then create the new time relations representing these entity types.

Step 3. Mapping of temporal single-valued attributes of an entity type

For each temporal single-valued attribute A of E, if the temporal support of A is indicated by an asterix (*), we create the time relation $TR_A(E)$ representing attribute A of E, which includes the attributes $FK \cup A \cup T$, where FK is the foreign key referencing the relation $R(E)$, and T is the timestamp attributes referencing character (*) in Table 2.

Table 2. Abbreviation used for temporal support of attributes and relationship types

(a) if * = VT then T = {VTs, VTe}
(b) if * = TT then T = {TTs, TTe}
(c) if * = BT then T = {TTs, TTe, VTs, VTe}

Consider $T' \subset T$ which is the set of underlined attributes in Table 2, the primary key of $TR_A(E)$ then is $FK \cup T'$.

Step 4. Mapping of multi-valued attributes

For each multi-valued attribute A of the entity type E in PNF (Partitioned Normal Form), or similarly, A is a multi-valued attribute of the composite attribute B, let R' be the relation representing the entity type E (or the composite attribute B). Mapping of the multi-valued attribute A to the corresponding relation then is recursively by considering the following cases.

a) A is the simple attribute. Consider the following possibilities:

- If A is the non-temporal attribute, we then create a new relation representing the attribute A, denoted $R_A(E)$ (or $R_A(B)$), which includes the attributes $FK \cup A'$, where FK is the foreign key referencing the relation R', and A' is the attribute used to store values of the multi-valued attribute A, referred to as the corresponding attribute to A. The primary key of $R_A(E)$ (or $R_A(B)$) then is $FK \cup A'$.

- If A is the temporal attribute, we then create a temporal relation representing the attribute A, denoted $TR_A(E)$ (or $TR_A(B)$), which includes the attributes $FK \cup A' \cup T$ and the primary key is $FK \cup A' \cup T'$, where FK is the foreign key referencing the relation R' , and A' is the corresponding attribute to A. Besides, T and T' are defined similarly as in Step 3.

b) A is the composite attribute. If A is the composite attribute which has the set of non-temporal single-value attributes X and the partial key K, we then create a new relation representing the attribute A, denoted $R_A(E)$ (or $R_A(B)$),

which includes the attributes $FK \cup X$ and the primary key is $FK \cup K$, where FK is the foreign key referencing the relation R'.

If A is the temporal attribute, we then add a temporal relation $TR_A(E)$ (or $TR_A(B)$), which includes the attributes $FK' \cup T$ and the primary key is $FK' \cup T'$, where FK' is the foreign key referencing the relation $R_A(E)$ (or $R_A(B)$). Besides, T and T' are defined similarly as in Step 3.

In the case where the composite attribute A has some temporal single-valued attributes, for each temporal single-value attribute C, we add a time relation $TR_C(A)$ representing the attribute C that includes the attributes $FK'' \cup C \cup T$ and the primary key is $FK'' \cup T'$, where FK'' is the foreign key referencing the relation $R_A(E)$ (or $R_A(B)$). Besides, T and T' are defined similarly as in Step 3.

Step 5. Mapping of non-temporal relationship types

Mapping of non-temporal relationship types between entity types is performed similarly to normal mapping (EER-to-Relational mapping [3]).

Step 6. Mapping of temporal binary relationship types without attribute

Consider the temporal binary relationship type S which does not have its own attribute and S is the relationship between E_1 and E_2. We then create the time relation representing the temporal binary relationship type S, noted as $TR(S)$, which includes attributes $FK_1 \cup FK_2 \cup T$, where FK_1 and FK_2 respectively are the foreign keys referencing the relations $R(E_1)$ and $R(E_2)$. Besides, T is defined in Table 1 or Table 2, depending on the temporal support specified for the relationship type S. The primary key of $TR(S)$ is $ID(E) \cup T'$, where $T' \subset T$ is as defined in Table 1 or Table 2. In addition, depending on the structural constraints (min, max) on participation of entity types in S, $ID(S)$ is defined as follows:

- If S is the relationship 1 - 1 then $ID(S) = FK_1$ or $ID(S) = FK_2$
- If S is the relationship 1 - *many* then $ID(S) = FK_2$
- If S is the relationship *many* - 1 then $ID(S) = FK_1$
- If S is the relationship *many* - *many* then $ID(S) = FK_1 \cup FK_2$

Step 7. Mapping of temporal binary relationship types with attribute

Consider the temporal binary relationship type S between two entity types E_1 and E_2 with their attributes X. We then create two relations as follows:

- *A* relation representing the binary relationship type S, denoted $R(S)$, which includes attributes $FK_1 \cup FK_2 \cup X$, where FK_1 and FK_2 respectively are the foreign keys referencing the relations $R(E_1)$ and $R(E_2)$. The primary key of $R(S)$ is $ID(S)$ defined similarly as in Step 6.

- *A* time relation representing the temporal aspect of the relationship type S, denoted $TR(S)$, which includes attributes $FK \cup T$ and the primary key is $FK \cup T'$, where FK is the foreign key referencing the relation $R(S)$. Besides, T is the timestamp attributes depending on the temporal support of the relationship S, where T and T' are also defined in Table 1 or Table 2.

In the case where the relationship S has some temporal attributes, for each temporal attribute A we create a time relation, denoted $TR_A(S)$, which includes the attributes $FK' \cup A \cup T_A$ and the primary key is $FK' \cup T'_A$, where FK' is the foreign key referencing the relation $R(S)$. Besides, T_A is the timestamp attributes referencing the temporal support specified for the attribute A. T_A and T'_A are described as in Table 2.

Note that mapping of other temporal relationship types which have or do not have their own attribute, such as recursive relationship type, n-ary relationship type, is performed similarly to the mapping in Step 6 or Step 7.

4 Extraction TimeER Model from Relational Model

Algorithm of extraction of the TimeER model from a relational model is defined as follows.

Input: The temporal relational model DB is the set of relations DB in which $R \in DB$ we can define the set of attributes U_R, the primary key PK_R, and the set of foreign keys FK_R.

Output: TimeER model. The TimeER model to extract is called satisfaction of reverse engineering, if this model is used together with the algorithm considered in Section 3, we can obtain a set of relations DB.

The TimeER model to extract is assumed to be used without a weak entity type. This assumption is acceptable because we can map a weak entity type with a multi-valued composite attribute of the respective owner entity type if the weak entity type is not involved in any other relationship in the model TimeER.

Extraction algorithms are implemented through the use of rules in turn to identify the components in the TimeER model based on the characteristics of the set of attributes, the primary key, and the set of foreign keys of the relation in the temporal relational model. The components include: entity type, temporal aspects of entity type, temporal attributes of an entity type, relationship between entity types, temporal aspects of a relationship, temporal attributes and non-temporal attributes of a relationship.

The proposed rule to identify one of the components is carried out under a general principle as follows. The condition needed to build a component is identified so that it satisfies the normal conversion algorithm for that component (referred to in Section 3), but does not satisfy for the rest. This allows us to prove the correctness of these rules with the method of exclusion.

To ease the construction of identification rules, we first perform the sub-group relations in the DB as follows. DB_1 is called the set of all relations $R \in DB$, which does not contain the timestamp attribute. In contrast, DB_2 is called the set of all relations $R \in DB$, which contains the timestamp attribute. We have: $DB_1 \cup DB_2 = DB$. Next, we convert each relationship type $R \in DB$ into a corresponding temporary entity with the same name and the same file attributes U_R, denoted $E(R)$.

Justification: According to the algorithm considered in Section 3, each type of entity in the TimeER model has a corresponding relation R containing all

the non-temporal single-valued attributes of that entity. Thus, in the reverse engineering, each relation R is mapped to a temporary entity type. However, we use the term "temporary entity type" because the relations in the $DB_1 \cup DB_2$ are those that would later be identified as the other components of the model TimeER (such as: relationships, multi-valued attributes, ...).

Method: Extraction of the TimeER model from DB_1 and DB_2 is then implemented through the use of algorithms in turn as follows.

Algorithm 4.1. (Extraction of the TimeER model from the relations without timestamp attribute)[1]

For each relation $R \in DB_1$ **do**

If: $PK_R \subseteq \bigcup_{FK \in FK_R} FK$ and $\exists R' \in DB_2, \exists FK' \in FK_{R'}$: FK' referencing R;

Then: R is identified to represent the temporal relationship type with attribute;

Else:

If: $|PK_R| = 1$;

Then:

R is identified to represent the entity type;

If: R has the foreign key $FK \in FK_R$ referencing R;

Then: R is identified to represent the non-temporal recursive 1-1/1-many relationship type;

Else

If: $PK_R \in FK_R$ referencing $R' \in DB_1$ and $R' \neq R$;

Then: $E(R)$ is the subclass of $E(R')$;

Else:

If: R has the foreign key $FK \neq PK_R$ referencing $R' \in DB_1$;

If: FK has unique constraint;

Then: R is identified to represent the non-temporal binary 1-1 relationship type;

Else: R is identified to represent the non-temporal binary 1-many relationship type;

Else $|PK_R| > 1$:

If: $PK_R \subseteq \bigcup_{FK \in FK_R} FK$

Then: R is identified to represent the non-temporal relationship type of degree n (n = 1: recursive many-many relationship type; n = 2: binary *many-many* relationship type; and n > 2: n-ary relationship type);

Else:

If: $PK_R \supset FK$, FK is the only foreign key of R;

Then: R is identified to represent the multi-valued attribute;

Endfor;

[1] Indentation style applies to **If-Then-Else** and **If-Then** statements.

Algorithm 4.2. (Extraction of the TimeER model from the relations with timestamp attribute)

For each relation $R \in DB_2$ do

If: R has the foreign key FK referencing $R' \in DB_1$ so that the relation R' has been identified as the relation to represent the temporal relationship type with attribute;

If: $U_R = FK \cup T$, where T is the set of timestamp attributes;

Then: R is identified to represent the temporal aspects of the relationship $S(R')$. The temporal aspects of the relationship depend on the attribute names in T;

Else $U_R \neq FK \cup T$: R is identified to represent the temporal attribute of the relationship $S(R')$. The temporal aspects of this attribute depend on the attribute names in T.

Else R has no foreign key referencing $R' \in DB_1$ so that R has been identified as the relation to represent the temporal relationship type with attribute:

If: $|FK_R| > 1$;

Then: R is identified to represent the temporal relationship $S(R')$ which does not have its own attribute. The temporal aspects of the relationship depend on the attribute names in T.

Else $|FK_R| \leq 1$:

If: R has the only foreign key FK;

If: $U_R = FK \cup T$, where T is the set of timestamp attributes;

Then: R is identified to represent the temporal aspects of the temporary entity type $E(R')$. The temporal aspects of the temporary entity type $E(R')$ depend on the attribute names in T;

Else $U_R \neq FK \cup T$:

If: $PK_R = FK \cup T'$, where $T' \subset T$;

Then: R is identified to represent the temporal single-value attribute of $E(R')$. The temporal aspects of this attribute depend on the attribute names in T;

Else $PK_R \neq FK \cup T'$: R is identified to represent the temporal multi-valued simple attribute of $E(R')$. The temporal aspects of this attribute depend on the attribute names in T.

Endfor;

5 Conclusion

In this paper we have proposed a method of extraction of the TimeER model from a relational model which is based on the characteristics of the set of attributes, the primary key, and the set of foreign keys of the relational schema in the temporal relational model. This approach is practical for temporal relational database existing since it is based only on the metadata defined by data definition language in the relational model (CREATE TABLE and ALTER TABLE statements in SQL).

We have done the design and made successful installation of this extraction on SQL 2005 Database Management System.

However, the completeness of the rules on the algorithms in Section 4 is only fitting if we use the conversion method presented in Section 3.

The reduction of the assumptions of reverse engineering is certain to affect this conversion method by property "not only" of the conversion problem in Section 3. For example, extracting a weak entity type in the model TimeER is not taken into account. The reason is that a weak entity type is also essentially seen as a multi-valued composite attribute of the respective owner entities. Such assumptions are unavoidable, while strengthening our ability to automate the extraction algorithm. However, logically this extraction method is acceptable because we can prove that for any input of a relational database, there is a corresponding database in the TimeER model.

Our research concerns the application of extraction of the TimeER model from a relational model to perform the conversion of the temporal relational model into other models, specially Temporal XML documents by using the TimeER model as an intermediate conversion result.

References

1. Quang, H., Thanh, H.T.: Extension of Method for Converting TimeER Model to Relational Model. Journal of Computer Science and Cybernetics 25(3), 246–257 (2009)
2. Quang, H., Thanh, H.T.: A Mapping Algorithm from TimeER Model to Relational Model. In: Proceedings The Second Hanoi Forum on Information - Communication Technology, Hanoi, December 11-13, pp. 37–45 (2008)
3. Elmasri, R., Navathe, S.B.: Fundamentals of Database Systems, 5th edn. Addison-Wesley, Reading (2007)
4. Jensen, C.S., Snodgrass, R.T.: Temporal Data Management. IEEE Transactions on Knowledge and Data Engineering 11(1), 36–44 (1999)
5. Jensen, C.S.: Temporal Database Management. Dr.techn. thesis, Aalborg University (2000), http://www.cs.auc.dk/~csj/Thesis/
6. Gregersen, H., Jensen, C.S.: Temporal Entity-Relationship Models - a Survey. IEEE Transactions on Knowledge and Data Engineering 11(3), 464–497 (1999)
7. Gregersen, H., Jensen, C.S.: Conceptual Modeling of Time-varying Information. TIMECENTER Technical Report TR-35 (1998)
8. Torp, K., Snodgrass, R.T., Jensen, C.S.: Effective Timestamping in Databases. VLDB Journal 8(4), 267–288 (2000)
9. Chiang, R.H.L., Barron, T.M., Storey, V.C.: Reverse Engineering of Relational Databases: Extraction of an EER Model from a Relational Database. Data & Knowledge Engineering 12, 107–142 (1994)

Certain Answers for Views and Queries Expressed as Non-recursive Datalog Programs with Negation

Victor Felea

"Al.I.Cuza" University of Iasi
Computer Science Department, 16 General Berthelot Street, Iasi, Romania
felea@infoiasi.ro
http://www.infoiasi.ro

Abstract. In this paper, we study the problem to compute certain answers in case when view definitions are expressed as non-recursive datalog programs with negation and queries are expressed as semi-positive non-recursive datalog programs with negation. Two situations are analyzed: the open world assumption (OWA) and the close world assumption (CWA). Associated to a view, and an extension of the view, a tree is constructed, which is useful to specify a method to compute certain answers.

Keywords: Datalog program, negation, query, view, certain answer.

1 Introduction

View-based query processing is a problem of computing the answer to a query based on a set of views. This problem is very important in application areas, such as query optimization, data warehousing, data integration. Two basic approaches to view-based query processing are known: query-rewriting and query-answering. These approaches are discussed in [4]. In the second approach, so-called certain tuples are computed. The problem of query-answering is the following: given a query on a database schema, and a set of views over the same schema, can we answer the query using only the answers to the views? This problem was intensively studied in the literature. Thus, in [1] the authors give the complexity of the problem of answering views using materialized views, where the languages for view definitions and queries are: conjunctive queries with inequalities, positive queries, datalog program and first order logic. In [12], some applications of the problem of answering queries using views, and algorithms are specified. The problem of view-based query processing in the context where databases are semistructured and both the query and the view are expressed as regular path queries, are studied in [6] and [7]. In [4], the authors analyze the complexity of query answering in the presence of key and inclusion dependencies. The answering query problem, in case when queries and views are in conjunctive form with arithmetic comparisons, is studied in [2]. In [13], the authors define so-called

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 67–77, 2011.

"relative containment", which formalizes the notion of query containment relative to the source that occur to in a data-integration systems. The problem of answering queries, using materialized views in the presence of negative atoms in view, is studied in [7]. In [8], the authors give a system that computes consistent answers to Datalog(disjunctive logic programming) queries with stable models semantics. A semantic model to compute consistent query answers is given in [9]. In [3], the authors apply logic programming based on answer sets to the retrieving consistent information from a possible inconsistent database. The problem of answering datalog queries using views is undecidable ([10]). In [14], the problem of whether a query Q can be answered using a set of views is studied.

Concerning the certain answers, the complexity of finding certain answers is discussed in [1], considering the case when views and queries are expressed in languages: conjunctive form, conjunctive form with inequalities, non-recursive datalog, datalog, first-order formula. In case when the query is expressed in datalog, and does not contain comparison predicate, and the views are in conjunctive form, the set of certain answers can be obtained using so-called query plan, which is a datalog program where extensional relations are the source relations [10]. The query plan that produces all certain answers is called the maximally-contained plan, and is defined in [10]. In [11], the authors show that the problem of computing the certain answers for union of conjunctive queries with inequalities is in $coNP$. In the paper [6], the authors study the problem of answering a query based on precomputed answers (that can be certain answers) for a set of views, in the context of Description Logic. By our best knowledge, the problem to compute the certain answer set in case when the negation occurs in views or query was not considered in literature up to now.

In this paper, we study the problem of computing certain answers in case when the views are expressed by non-recursive datalog programs with negation and the query is a non-recursive semi-positive datalog program.

2 Basic Definitions and Notations

Let Dom be a countable infinite domain for databases. The elements of Dom are called constants. Let $\mathcal{V}$ be a view expressed as a non-recursive datalog program having $V(\overline{x})$ as the head of a rule that is called the main rule. Each rule from $\mathcal{V}$ can contain negated literals. Let us denote by $EDB(\mathcal{V})$, $IDB(\mathcal{V})$, and $Rel(\mathcal{V})$ the set of extensional, intensional symbols, all symbols from $\mathcal{V}$, respectively. A rule that has $f(\overline{z})$ as its head is called a definition of f. Assume that for each $f \in IDB(\mathcal{V})$, there exists a single definition of f. The rules can contain free variables (these variables are those from the head of rule), existentially quantified variables (these variables occur in the rule body), constants that appear in the body of the rule. There are two restrictions about variables or constants that occur in the definition of an intensional symbol. The first one is: each variable that occurs in the head of rule it also must appear in the positive part of the body of rule. This is called the safe property of the view definition. The second one: each variable or constant that occurs in the negated part of the definition

of f, must occur in its positive part. This property is called "safe negation". A query Q is considered as a non-recursive semi-positive datalog program (the semi-positive datalog program is a datalog program where each definition symbol contains in its negated part only extensional symbols). The rules of Q satisfy the restrictions like as those for $\mathcal{V}$. Assume that $EDB(Q) \subseteq EDB(\mathcal{V})$. In the following definition we give the notion of certain answer.

Definition 1. *Let Q be a query expressed as a non-recursive semi-positive datalog program and $\mathcal{V}$ a view expressed as a non-recursive datalog program, having $V(\overline{x})$ as the head of the main rule. Let $I = \{\overline{w}_1, \ldots, \overline{w}_m\}$ be an extension of the view $\mathcal{V}$. The tuple $\overline{t}$ having arity$(\overline{t})$ = arity$(\overline{x})$, is a certain answer for I, $\mathcal{V}$, and Q under OWA, if $t \in Q(D)$ for all databases D defined on Dom such that $I \subseteq \mathcal{V}(D)$, and $Rel(D) \subseteq Rel(\mathcal{V})$. The tuple $\overline{t}$ is a certain answer for I, $\mathcal{V}$, and Q under CWA, if $\overline{t} \in Q(D)$ for all databases D defined on Dom such that $I = \mathcal{V}(D)$, and $Rel(D) \subseteq Rel(\mathcal{V})$.*

Intuitively, a tuple is a certain answer of the query Q, if it is an answer for any of the possible database instances which are consistent with the given extensions of the view. In the case OWA, the relation $I \subseteq \mathcal{V}(D)$ is equivalent to $\overline{w}_i \in V(D)$, for each $i, 1 \leq i \leq m$. Assume that all values from I belong to Dom. Let us denote by C the values from I, and the constants from $\mathcal{V}$. Let Y be the set of all variables that are existentially quantified in the rules of $\mathcal{V}$. Let π be a partition on the set $C \cup Y$, and $Class_\pi$ the set of all classes defined by π. A partition π is called a $C - partition$, if for two distinct constants c_1 and c_2, we have $[c_1]_\pi \neq [c_2]_\pi$, hence two distinct constants occur in two distinct classes with respect to π. Through the paper we use only $C - partitions$, so in the paper when we write *partition*, we mean $C - partition$. For a partition π defined on $C \cup Y$, we consider a mapping from $C \cup Y$ into $Class_\pi$, denoted η_π and defined as follows: $\eta_\pi(t) = [t]_\pi$, where $[t]_\pi$ denote the class that contains t. The mapping η_π is naturally extended for a vector $\overline{w} = (t_1, \ldots, t_r)$ having the components from $C \cup Y$, by $\eta_\pi(\overline{w}) = (\eta_\pi(t_1), \ldots, \eta_\pi(t_r))$. For an atom $R(\overline{w})$, we consider $\eta_\pi(R(\overline{w})) = R(\eta_\pi(\overline{w}))$. For a set S of atoms having the form $R(\overline{w})$, we define $\eta_\pi(S) = \{\eta_\pi(R(\overline{w}))|R(\overline{w}) \in S\}$. For a database D defined on Dom, we consider $val(D)$ the set of all values that occur in the atoms of D. Formally, $val(D) = \{v|\exists R(\overline{w}) \in D, v$ is a component of $\overline{w}\}$. The view $\mathcal{V}$ is said consistent with I under OWA if there exists a database D over Dom such that $I \subseteq \mathcal{V}(D)$. Let us denote by $f_1 o f_2$ the composition of the two mappings f_1, f_2, where $(f_1 o f_2)(t) = f_2(f_1(t))$. Now, we need to define a formula corresponding to a view definition. Let f be from $IDB(\mathcal{V})$ and its definition having the form:

$$f(\overline{z}) : -S_1(\overline{z}, \overline{t}_1), \ldots, S_n(\overline{z}, \overline{t}_n), \neg S_{n+1}(\overline{z}, \overline{t}_{n+1}), \ldots, \neg S_{n+p}(\overline{z}, \overline{t}_{n+p}) \qquad (1)$$

The vector $\overline{z}$ contains all free variables from this definition, the vector $\overline{t}_j$ contains all existentially quantified variables from $S_j(\overline{z}, \overline{t}_j)$. Let us denote by $\overline{y}_j$ the vector of all variables that occur in $S_j(\overline{z}, \overline{t}_j)$. Let $S_{n+j}(\overline{y}_{n+j})$ be an atom that occurs in the negated part of $f(\overline{z})$. If the symbol S_{n+j} occurs in the positive part of f with indexes $\alpha_1, \ldots, \alpha_q$, then we have: S_{n+j}=S_{α_i}, for each $i, 1 \leq i \leq q$,

and $S_{n+j} \neq S_\beta$, for each $\beta \in \{1,\ldots,n\} - \{\alpha_1,\ldots,\alpha_q\}$. We associate to the atom $S_{n+j}(\overline{y}_{n+j})$ a formula denoted ϕ^j, and defined as follows: $\phi^j = (\overline{y}_{n+j} \neq \overline{y}_{\alpha_1}) \wedge \ldots \wedge (\overline{y}_{n+j} \neq \overline{y}_{\alpha_q})$, where the expression $(\overline{y}_l \neq \overline{y}_s)$ denotes the following disjunction: $(y_l^1 \neq y_s^1) \vee \ldots \vee (y_l^r \neq y_s^r)$, the tuples $\overline{y}_l$ and $\overline{y}_s$ having the form: $\overline{y}_l = (y_l^1,\ldots,y_l^r)$, $\overline{y}_s = (y_s^1,\ldots,y_s^r)$. In case when S_{n+j} does not occur in the positive part of f, then we consider $\phi^j = TRUE$.

Definition 2. *The formula ϕ corresponding to f, denoted $\phi(f)$ is the conjunction of all formulas ϕ^j, $1 \leq j \leq p$, that means: $\phi(f) = \phi^1 \wedge \ldots \wedge \phi^p$.*

Now, we formally define the logic value of a formula for a given partition.

Definition 3. *Let π be a partition defined on $C \cup Y$ and $\phi(f)$ the formula constructed for f as we mentioned above. The logic value of $\phi(f)$ for π, denoted $\pi(\phi(f))$ is recursively defined as follows:*
(i) If $\phi \equiv (t \neq t')$, where t and $t' \in C \cup Y$, then $\pi(\phi) = TRUE$ if there is no class E from $Class_\pi$ such that $t,t' \in E$, i.e. $[t]_\pi \neq [t']_\pi$.
(ii) $\pi(\phi_1 \wedge \phi_2) = \pi(\phi_1) \wedge \pi(\phi_2)$, $\pi(\phi_1 \vee \phi_2) = \pi(\phi_1) \vee \pi(\phi_2)$.

3 The Construction of $TREE(I, \mathcal{V})$

Firstly, we need to give a definition that precises some notions which will be used in the construction of a tree corresponding to I and $\mathcal{V}$.

Definition 4. *Let π be a partition, T a database defined on $Class_\pi$, and f an intensional symbol from $\mathcal{V}$. Let I' be the projection of T on f, i. e. $I' = T[f]$. Let $I' = \{f(\overline{z}_1),\ldots,f(\overline{z}_h)\}$ and $f(\overline{z}) :- f_1(\overline{z},\overline{t})$ the definition of f in $\mathcal{V}$, where $\overline{t}$ contains all existentially quantified variables from the right-hand part of the definition of f. We denote by $pos(f(\overline{z}))$, $(neg(f(\overline{z})))$, the set of all positive (negated) atoms from the definition of $f(\overline{z})$. We consider the substitutions of the vector $\overline{z}$ with $\overline{z}_j$, $1 \leq j \leq h$ in the view definition of f and for two distinct indexes j and l we take the existentially quantified variable sets disjoint.*
(a) We denote these rules by $Ext(f, I')$, i.e.

$$f(\overline{z}_j) :- f_1(\overline{z}_j, \overline{t}_j), 1 \leq j \leq h.$$

(b) The set of all existentially quantified variables from $Ext(f, I')$ will be denoted by $ExtVar(f, I')$, i.e. $ExtVar(f, I') = \cup_{1 \leq j \leq h} \overline{t}_j$.
(c) Two databases defined on $Class_\pi$, denoted $DPos(f, I')$ and $DNeg(f, I')$, and specified as follows:
$DPos(f, I') = \cup_{1 \leq j \leq h} pos(f(\overline{z}_j))$, $DNeg(f, I') = \cup_{1 \leq j \leq h} neg(f(\overline{z}_j))$.

Associated to I and $\mathcal{V}$, we construct a tree denoted $TREE(I, \mathcal{V})$, where its root is denoted by α_0. For each node α of the tree, we associate so-called "basic elements" and so-called "calculated elements". The "basic elements" are:

(a) A set of intensional symbols from $\mathcal{V}$, denoted $RInt(\alpha)$,
(b) A set of constants and variables, denoted $OldCV(\alpha)$,

(*c*) A partition defined on $OldCV(\alpha)$, denoted $\pi(\alpha)$, and
(*d*) A database defined on $Class_{\pi(\alpha)}$, denoted $OldD(\alpha)$. These elements are also called inherited, because they are calculated for the precedent node.

The "calculated elements" associated to α, using Definition 4 are the following:
(*e*) For each symbol f from $RInt(\alpha)$ we compute $I' = OldD(\alpha)[f]$, $E1 = Ext(f, I')$, $D1 = DPos(f, I')$, $D2 = DNeg(f, I')$, $E2 = ExtVar(f, I')$.
(*f*) A new set of constants and variables $NewCV(\alpha) = OldCV(\alpha) \cup E2$,
(*g*) A new database defined on $Class_{\pi(\alpha)}$, $NewD(\alpha) = (OldD(\alpha) - I') \cup D1$.
(*h*) A database defined on $Class_{\pi(\alpha)}$ containing all atoms from all nodes γ situated on the path from α_0 to α, $DTNeg(\alpha) = DTNeg(\alpha') \cup D2$, where α' is the immediate predecessor of α.
(*i*) The set of all partitions defined on $NewCV(\alpha)$, that are extensions of $\pi(\alpha)$. This set is denoted $Partition(\alpha)$.
(*j*) For each partition π_1 from $Partition(\alpha)$, we compute the set $\mathcal{M}_{\pi_1}$, specified in Definition 5.
(*k*) A variable $TERM(\alpha)$ having the values: 0 when α is not terminal, 1 in case when α is terminal, but $RInt(\alpha) \neq \emptyset$, 2 when the node is terminal, but the database $OldD(\alpha)$ is inconsistent with $\mathcal{V}$, 3 otherwise.

Definition 5. *Let π be a partition, T a database defined on $Class_\pi$, $f \in IDB(\mathcal{V})$, $I' = T[f]$, $NewCV$ a set of constants and variables, π_1 a partition defined on $NewCV$, that is an extension of π, and $Ext(f, I')$ the datalog program specified in Definition 4. We define two databases on $Class_{\pi_1}$ denoted $T^{min}_{\pi_1}$ and $T^{max}_{\pi_1}$ and defined as follows:*

$$T^{min}_{\pi_1} = \eta_{\pi_1} DPos(f, I'), \tag{2}$$

$$T^{max}_{\pi_1} = \{\eta_{\pi_1} R(\overline{w}), R \in Rel(DPos(f, I')) \text{ and } \overline{w} \text{ has components from } NewCV\} - \eta_{\pi_1} DTNeg(\alpha) \tag{3}$$

A set of databases defined on $Class_{\pi_1}$, denoted $\mathcal{M}_{\pi_1}$ will be defined as follows:

$$\mathcal{M}_{\pi_1} = \{T_1 | T^{min}_{\pi_1} \subseteq T_1 \subseteq T^{max}_{\pi_1}\} \tag{4}$$

Corresponding to the root α_0 we take the following: $RInt(\alpha_0) = \{V\}$, $OldCV(\alpha_0) = C$, $\pi(\alpha_0)$ is the discrete partition on C, $OldD(\alpha_0) = \{V(\overline{w}_1), \ldots, V(\overline{w}_m)\}$, where $I = \{\overline{w}_1, \ldots, \overline{w}_m\}$, and $DTNeg(\alpha_0) = \emptyset$. Now, we specify some details about the construction of $TREE(I, \mathcal{V})$. Suppose we have constructed the node α having the elements described above. For each tuple having the form (f, π_1, T_1), where $f \in RInt(\alpha)$, $\pi_1 \in Partition(\alpha)$, and $T_1 \in \mathcal{M}_{\pi_1}$, we construct an immediate successor of α, denoted β, such that: $RInt(\beta) = RInt(\alpha) - \{f\} \cup IDB(Ext(f, I'))$, $OldCV(\beta) = NewCV(\alpha)$, $\pi(\beta) = \pi_1$, $OldD(\beta) = T_1$, $DTNeg(\beta) = DTNeg(\alpha) \cup DNeg(f, I')$.

In case when the view definition of f does not contain existentially quantified variables ($ExtVar(f, I') = \emptyset$), then the immediate successors of α are constructed for each vector (f, π_1, T_1) , where $f \in RInt(\alpha)$, $\pi_1 = \pi(\alpha)$ and $T_1 \in \mathcal{M}_{\pi_1}$. In case when the node α is terminal, but $RInt(\alpha) \neq \emptyset$, then we

called this node "failed node". In case when α is terminal and $RInt(\alpha) = \emptyset$, then it is need to test this node for consistency. Let $D = OldD(\alpha)$. For each $f \in IDB(\mathcal{V})$, we compute the answer of f for D, denoted $f(D)$. The test for this node α is the following:

$$(\cup_{f \in IDB(\mathcal{V})} f(D)) \cap DTNeg(\alpha) \neq \emptyset \vee DTNeg(\alpha)[EDB(\mathcal{V})] \cap OldD(\alpha) \neq \emptyset \quad (5)$$

In case when the node α satisfies this condition, then $TERM(\alpha) = 2$ otherwise $TERM(\alpha) = 3$. We remark that the answer of f for D is computable because the datalog program $\mathcal{V}$ is non-recursive. One can specify a recursive procedure denoted $CONSTR$, which for a given node α constructs all successors of α in $TREE(I, \mathcal{V})$. The parameters of the procedure $CONSTR$ are α, and those specified in $(a) - (d)$, and $DTNeg$. The main program to construct $TREE(I, \mathcal{V})$ could be the following:

BEGIN
$C1 = C$; $T = \{V(\overline{w}_1), \ldots, V(\overline{w}_m)\}$; π_0 is the discrete partition on $C1$;
$RInt = \{V\}$; $DTNeg = \emptyset$; $TERM(\alpha_0) = 0$;
CALL $CONSTR(\alpha_0, RInt, C1, \pi_0, T, DTNeg)$;
END

Example 1. Let us define a view $\mathcal{V}$ by the following two rules, and $I = (0)$ an extension of $\mathcal{V}$:
$V(x) : -f(x, z_1), \neg f(z_1, x)$, $f(t_1, t_2) : -R(t_1, z_2), R(z_2, t_2), \neg R(t_1, t_1)$.
The $TREE(I, \mathcal{V})$ will have three levels (the root is considered on level 1). For its root α_0 we have: $RInt(\alpha_0) = \{V\}$, $OldCV(\alpha_0) = \{0\}$, $\pi(\alpha_0) = \{\overline{0}\}$, $OldD(\alpha_0) = V(0)$. In this example we denote by $\overline{t_1 \ldots t_h}$ the class that contains the elements t_1, ...,t_h. We have: $f = V$, $I' = V(0)$, $Ext(f, I')$ consists of the rule: $V(0) : -f(0, z_1), \neg f(z_1, 0)$, $DPos(f, I') = f(0, z_1)$, $DNeg(f, I') = f(z_1, 0)$,$ExtVar(f, I') = \{z_1\}$, $NewD(\alpha_0) = \{f(0, z_1)\}$, $NewCV(\alpha_0) = \{0, z_1\}$, $DTNeg(\alpha_0) = \{f(z_1, 0)\}$. There are two partitions defined on $NewCV(\alpha_0)$, namely: $\pi_1 = \{\overline{0}, \overline{z_1}\}$ and $\pi_2 = \{\overline{0z_1}\}$. The formula ϕ associated to the extension $Ext(f, I')$ is $\phi = (z_1 \neq 0)$. This formula is satisfied only by π_1. For π_1, we compute the set $\mathcal{M}_{\pi_1}$ and if we denote by $M_1 = \{f(\overline{0}, \overline{0}), f(\overline{z_1}, \overline{z_1})\}$, and by $\mathcal{P}(M_1)$ the set of all subsets from M_1, then $\mathcal{M}_{\pi_1} = \{\{f(0, z_1)\} \cup S | S \in \mathcal{P}(M_1)\}$. Thus the node α_0 has four immediate successors. For $S = \emptyset$, let us denote this node by β_1. This node will be extended with two successors, denoted β_{11} and β_{12}, for $RInt = f$, and the partition $\pi_3 = \{\overline{0}, \overline{z_1 z_2}\}$. For $RInt = \{f\}$, and the partition $\pi_4 = \{\overline{0}, \overline{z_1}, \overline{z_2}\}$, we obtain 64 successors. For the nodes β_{11} and β_{12}, we have $TERM(\beta_{11}) = 3$, but $TERM(\beta_{12}) = 2$. In a similar manner, we obtain the elements associated to other nodes from $TREE(I, \mathcal{V})$.

4 Some Properties of Nodes from $TREE(I, \mathcal{V})$

In the following, we point out some properties of databases associated to the nodes from $TREE(I, \mathcal{V})$.

Proposition 1. *Let α be a node from $TREE(I, \mathcal{V})$ having $TERM(\alpha) \neq 3$. Let τ be an injective mapping from $Class_{\pi(\alpha)}$ into Dom, and $D' = \tau(OldD(\alpha))$. We have: $\mathcal{V}(D') = \emptyset$.*

In the following, we consider as mappings τ only those that preserve constants, i. e. $\tau([c]_\pi) = c$. We remark that the view $\mathcal{V}$ is consistent with I iff there exists a terminal node in $TREE(I, \mathcal{V})$ having $TERM(\alpha) = 3$. Now, we give a result about the relation between the answers of a query Q, expressed by a semi-positive non-recursive datalog program, for two databases defined on Dom.

Proposition 2. *Let D_1 and D be two databases defined on the schema $S = EDB(\mathcal{V})$ such that $D_1 \subseteq D$, and for each atom $R(\overline{w}) \in D - D_1$, there is a component v from $\overline{w}$ such that $v \notin val(D_1)$. Then, for each query Q expressed as a semi-positive non-recursive datalog program having $EDB(Q) \subseteq S$, we have $Q(D_1) \subseteq Q(D)$.*

Proof. Since the query Q is non-recursive, we can define for each relational symbol S from Q a level, denoted $level(Q)$, and define as follows:

(i) For each $S \in EDB(Q)$, we take $level(S) = 0$.
(ii) If f_j, $1 \leq j \leq s + t$ occur in the right part of the definition of f from (1), and $max\{level(f_j) | 1 \leq j \leq s + t\} = h$, then $level(f) = h + 1$.

We show by induction on $level(f)$ the statement: $f(D_1) \subseteq f(D)$. Firstly, let f be a symbol such that $level(f) = 1$, then from (ii) we have $level(f_j) = 0$, hence f_j are EDB-symbols. Let $\overline{w}$ be from $f(D_1)$. This implies there exists a substitution θ from the variables that occur in the definition of f into Dom such that $\theta f_j(\overline{z}_j) \in D_1$, $1 \leq j \leq n$ and $\theta f_{n+i}(\overline{z}_{n+i}) \notin D_1$, $1 \leq i \leq p$ and $\theta\overline{z} = \overline{w}$. We must show that: $\theta f_{n+i}(\overline{z}_{n+i}) \notin D$. Assume the contrary, then there exists i such that $\theta f_{n+i}(\overline{z}_{n+i}) \in D$. Since by the induction hypothesis $\theta f_{n+i}(\overline{z}_{n+i}) \notin D_1$, we obtain there exists a component v from $\theta\overline{z}_{n+i}$ such that $v \notin val(D_1)$. On the other hand, using the safeness property regarding the negation, we get: all components of $\theta\overline{z}_{n+i}$ belong to $\theta(\cup_{1 \leq j \leq n}\overline{z}_j)$, therefore all components of $\theta\overline{z}_{n+i}$ belong to $val(D_1)$, which is a contradiction. Thus, $w \in f(D)$.

For the inductive step, assume that $f_i(D_1) \subseteq f_i(D)$, for each symbol f_i, having $level(f_i) \leq h$. Let f be a symbol having $level(f) = h+1$, and f has the definition specified in (1). By the hypothesis of induction, we have $f_i(D_1) \subseteq f_i(D)$. Since the query Q is semi-positive, we have: $f_{n+i} \in EDB(Q)$, $1 \leq i \leq p$. As in case when $level(f) = 1$, we obtain $f(D_1) \subseteq f(D)$. □

Lemma 1. *Let $OldCV$ be a set of constants and variables, and D a database defined on Dom having $Rel(D) \subseteq Rel(\mathcal{V})$. Let θ be a substitution from $OldCV$ into Dom. Then, there exist a partition π defined on $OldCV$, an injective mapping τ_θ from $Class_\pi$ into Dom such that $\theta = \eta_\pi o \tau_\theta$.*

Proof. We define the partition π as follows: $t_1 \pi t_2$ iff $\theta(t_1) = \theta(t_2)$. The mapping τ_θ is as follows: $\tau_\theta([t]_\pi) = \theta(t)$, for each $t \in OldCV$. The statement $\theta = \eta_\pi o \tau_\theta$ results immediately. □

Proposition 3. *Let OldCV be a set of constants and variables, θ a substitution from OldCV into Dom. Let π be the partition of OldCV corresponding to θ (by Lemma 1). Let OldD be a database defined on* Class_π *and $f \in RInt(OldD)$. Let $I' = OldD[f] = \{f(\overline{z}_1), \ldots, f(\overline{z}_s)\}$. Let D be a database defined on Dom, having $Rel(D) \subseteq Rel(\mathcal{V})$. Let E be the extension of f and I', i.e. $E = Ext(f, I')$. Let $\phi(E)$ be the associated formula to E, and $V_1 = ExtVar(f, I')$, the set of all existentially quantified variables from E. We consider the definition of f expressed as in (1). Assume that there exists a substitution θ from OldCV into Dom such that $\theta f(\overline{z}_j) \in f(D)$, for each j, $1 \leq j \leq s$. Moreover, assume that $I \subseteq V(\tau_\theta(OldD))$, where τ_θ is the injective mapping corresponding to θ (Lemma 1). The the following statements hold:*

(i) There exists a substitution θ_1 from $NewCV = OldCV \cup V_1$ into Dom*, such that $\theta_1 S_k(\overline{z}_j, \overline{t}_{jk}) \in S_k(D)$, $1 \leq k \leq n$, $1 \leq j \leq s$. For two different indexes j and m the sets of existentially quantified variables are disjoint, i.e. $(\cup_{1 \leq l \leq n} \overline{t}_{jl}) \cap (\cup_{1 \leq l \leq n} \overline{t}_{ml}) = \emptyset$. Let $DPos(f, I') = \cup_{1 \leq k \leq n} \cup_{1 \leq j \leq s} S_k(\overline{z}_j, \overline{t}_{jk})$. Moreover, the substitution θ_1 is an extension of θ.*
(ii) $\theta_1 S_{n+e}(\overline{z}_j, \overline{t}_{jn+e}) \notin S_{n+e}(D)$, $1 \leq e \leq p$, $1 \leq j \leq s$.
(iii) The partition π_1 corresponding to the substitution θ_1 (Lemma 1) defined on $NewCV$ is an extension of π, and satisfies $\pi_1(\phi(E)) = TRUE$.
(iv) There exist a database $D' \subseteq D$, and a database T on Class_{π_1} *from $\mathcal{M}_{\pi_1}$ such that $D' = \tau_{\theta_1}(T)$.*
(v) The database D' specified above satisfies: $I \subseteq \mathcal{V}(D')$, and for each atom $S(\overline{w})$ from $D - D'$ the vector $\overline{w}$ contains at least a component that does not belong to V', where $V' = \theta_1(NewCV)$.

Proof. The statements (i) and (ii) follow from the relations $\theta f(\overline{z}_j) \in f(D)$, $1 \leq j \leq s$ and the definition of f. Since the substitution θ_1 is an extension of θ, it results that π_1 is an extension of π. Since the substitution θ satisfies the relations $\theta f(\overline{z}_j) \in f(D)$, we obtain that $\pi_1(\phi(E)) = TRUE$ (the statement (iii)). For the statements (iv) and (v) we consider a database defined on *Dom* as follows: $D' = \{S(\overline{v}) | S(\overline{v}) \in D,\ S \in Rel(NewD)$ and $\overline{v}$ contains only values from $V'\}$. The mapping τ_{θ_1} is bijective from $Class_{\pi_1}$ into V'. Let $T = \tau_{\theta_1}^{-1}(D')$. Let us show that $T \in \mathcal{M}_{\pi_1}$. From the statement (i), we obtain: $\theta_1 DPos(f, I') \subseteq D'$, which implies $T_{\pi_1}^{min} \subseteq T$, applying the substitution $\tau_{\theta_1}^{-1}$, and using the relation $\tau_{\theta_1}^{-1} o \theta_1 = \eta_{\pi_1}$. By Definition 5, we have: $T_{\pi_1}^{min} = \eta_{\pi_1} DPos(f, I')$, and $T_{\pi_1}^{max} = \{S(\overline{w}) | S \in Rel(DPos(f, I'))$, $\overline{w}$ has components from $Class_{\pi_1}\} - \eta_{\pi_1} DT$, where DT is $DNeg(f, I')$. We obtain that $T \subseteq T_{\pi_1}^{max}$, hence $T \in \mathcal{M}_{\pi_1}$.

To show the statement (v): By the relations specified in (i) and (ii), $V' = \theta_1(NewCV)$ and by the definition of the database D', we obtain: $\theta_1 S_k(\overline{z}_j, \overline{t}_{jk}) \in S_k(D')$, $1 \leq k \leq n$, $1 \leq j \leq s$, $\theta_1 S_{n+e}(\overline{z}_j, \overline{t}_{jn+e}) \notin S_{n+e}(D')$, $1 \leq e \leq p$, $1 \leq j \leq s$. These statements imply $\theta_1 f(\overline{z}_j) \in f(D')$, $1 \leq j \leq s$. Thus, using the hypothesis $I \subseteq V(\tau_\theta(OldD))$, we get $I \subseteq V(\tau_{\theta_1}(T))$, i.e. $I \subseteq V(D')$. The second part of the statement (v) results from the definition of D'. □

Theorem 1. *Let D be a database defined on Dom such that $I \subseteq V(D)$, and $Rel(D) \subseteq EDB(\mathcal{V})$. Then, there exist a terminal node α in $TREE(I, \mathcal{V})$*

having $TERM(\alpha) = 3$, *a substitution* θ *from* $OldCV(\alpha)$ *into* Dom, *an injective mapping* τ_θ *defined on* $Class_{\pi(\alpha)}$ *and having its values in* Dom *such that the following assertions are true:*

(a) $I \subseteq V(D')$, *where* $D' = \tau_\theta(OldD(\alpha))$.
(b) *For each atom* $R(\overline{w})$ *from* $D - D'$, *the vector* $\overline{w}$ *has at least a component that does not belong to* $V' = \theta(NewCV(\alpha))$.

Proof. Firstly, we assign the elements to the root α_0 of $TREE(I, \mathcal{V})$. We take $OldCV = C$, the set of all constants from I, θ the unit substitution: $\theta(c) = c$, for each $c \in C$. It results that the partition π corresponding to θ (Lemma 1) is the discrete partition, $OldD = \{\eta_\pi V(\overline{w}_1), ..., \eta_\pi V(\overline{w}_m)\}$. Let $f = V$ and $I' = OldD$. The substitution θ satisfies: $\theta V(\overline{w}_i) \in V(D)$, because $I \subseteq V(D)$. Since $\tau_\theta(OldD) = I$, we have $I \subseteq V(\tau_\theta(OldD))$. Thus, we can use Proposition 3, because the necessary conditions are satisfied. Applying Proposition 3 for the node α_0, we pass from the node α_0 to a successor β of α_0 from TREE$(I, \mathcal{V})$, and for the node β the hypothesis of Proposition 3 are again satisfied. Let $\alpha_0, \alpha_1, ..., \alpha_q = \alpha$ be the path from the root α_0 to α. Let $\theta(\alpha_i)$ be the substitution computed for the node α_i, using Proposition 3. Since $\theta(\alpha_{j+1})$ is an extension of $\theta(\alpha_j)$, $0 \leq j < q$, and the statements (a) and (b) from Proposition 3 are true for each node α_j, $0 \leq j \leq q$, we obtain that $TERM(\alpha) = 3$. □

Theorem 2. *Let* α *be a node terminal from* $TREE(I, \mathcal{V})$ *having* $TERM(\alpha) = 3$. *Let* τ *be an injective mapping from* $Class_{\pi(\alpha)}$ *into* Dom. *Let* D' *be the database defined on* Dom *such that* $D' = \tau(OldD(\alpha))$. *Then we have:* $I \subseteq \mathcal{V}(D')$.

Proposition 4. *Let* $NewCV$ *be a set of constants and variables and* π *a partition defined on* $NewCV$. *Let* T *be a database defined on* $Class_\pi$, *and* τ *an injective mapping from* $Class_\pi$ *into* Dom. *Then we have:* $\tau(Q(T)) = Q(\tau(T))$, *for each query* Q *expressed as non-recursive datalog program and having* $EDB(Q) \subseteq EDB(\mathcal{V})$.

5 Computing of Certain Answers

In this section, we give some results necessary to compute all certain answers corresponding to I, $\mathcal{V}$ and Q. Firstly, we need to give a definition that points out a class of tuples, called $C - tuples$.

Definition 6. *Let* $OldCV$ *be a set of constants and variables,* π *a partition defined on* $OldCV$, *and* $\overline{w}$ *a tuple on* $Class_\pi$. *The tuple* $\overline{w}$ *is called a* $C - tuple$, *if each of its components contains a constant.*

Proposition 5. *Let* α *be a terminal node from* $TREE(I, \mathcal{V})$, *with* $TERM(\alpha) = 3$, $\pi(\alpha)$ *the partition corresponding to* α, $T = OldD(\alpha)$ *the database associated to* α, *with elements from* $Class_{\pi(\alpha)}$. *Let* $T_a = Q(T)$ *be the answer of* Q *for* T *and a tuple* $\overline{w}_a$ *from* T_a.

(a) If the tuple $\overline{w}_a$ is not a $C-tuple$, then for each injective mapping τ from $Class_{\pi(\alpha)}$ into Dom, that preserves the constants, the tuple $\tau(\overline{w}_a)$ is not a certain answer under OWA for I, $\mathcal{V}$ and Q.
(b) If all tuples from T_a are not $C-tuples$, then there are not certain answers under OWA.

We remark that for a $C-tuple$ $\overline{w}_a$ from T_a, there exists only one injective mapping that preserves constants, denoted τ_0, and defined by: $\tau_0(t) = c$ if $c \in [t]_{\pi(\alpha)}$. Moreover, the tuple $\tau_0(\overline{w}_a)$ is a candidate for the set of certain answers.

We can specify a procedure called $CertAnsO$ that computes the set of certain answers. To compute certain answers under CWA, we consider only the terminal nodes α from $TREE(I, \mathcal{V})$ having $TERM(\alpha) = 3$ and satisfying the restrictions: $V(T)$ only contains $C-tuples$ and $\tau_0 V(T) = I$, where $T = OldD(\alpha)$. Thus one can obtain easily a procedure, that computes the certain answers under CWA.

6 Conclusion

In this paper, we have presented a method to compute the certain answers in case when views are expressed as non-recursive datalog programs with negation and queries as non-recursive semi-positive datalog programs with negation. For the future work, it is interesting to analyze whether our method can be extended in cases when views and queries are expressed by other types of datalog programs.

References

1. Abiteboul, S., Duschka, O.M.: Complexity of answering queries using materialized views. In: PODS, pp. 254–263 (1998)
2. Afrati, F., Li, C., Mitra, P.: Rewriting queries using views in the presence of arithmetic comparisons. Theoretical Computer Science 368, 88–123 (2006)
3. Arenas, M., Bertossi, L., Chomicki, J.: Answer set for consistent query answering in inconsistent databases. Theory and Practice of Logic Programming 3, 394–424 (2003)
4. Call, A., Lembro, D., Rosati, R.: Query rewriting and answering under constraints in data integration systems. In: IJCAI 2003, pp. 16–21 (2003)
5. Call, A., Lembro, D., Rosati, R.: On the Decodability and Complexity of Query Answering over Inconsistent and Incomplete Databases. In: PODS, pp. 260–271 (2003)
6. Calvanese, D., Giacomo, G. De, Lenzerini, M., Vardi, M.Y.: View-based query processing. On the relationship between rewriting, answering and losslessness. Theoretical Computer Science 371, 169–182 (2007)
7. Calvanese, D., Giacomo, G. De, Lenzerini, M., Vardi, M.Y.: Answering Regular Path Queries Views. In: Proc. of the 16th IEEE Int. Conf. on Data Engineering, ICDE, pp. 389–398 (2000)
8. Caniupan, M., Bertossi, L.: The consistency extractor system: Answer set programs for consistent query answering in databases. Data and Knowledge Engineering 69, 545–572 (2010)

9. Chomicki, J.: Consistent Query Answering, Recent Developments and Future Directions., http://cse.buffalo.edu/~chomicki/papers-iicis0
10. Duschka, O.M., Genesereth, M.R., Levy, A.: Recursive Query Plans for Data Integration. Journal of Logic Programming 43(1), 49–73 (2000)
11. Fagin, R., Kolaitis, P.G., Popa, L., Miller, R.J.: Data Exchange: Semantics and Query Answering. In: Calvanese, D., Lenzerini, M., Motwani, R. (eds.) ICDT 2003. LNCS, vol. 2572, pp. 207–224. Springer, Heidelberg (2002)
12. Flesca, S., Greco, S.: Rewriting queries using views. IEEE Trans. Knowledge Data Engineering 13(6), 980–995 (2001)
13. Millstein, T., Halevy, A., Friedman, M.: Query containment for data integration systems. JCSS 66, 20–39 (2003)
14. Nash, A., Segoufin, L., Vianu, V.: Views and Queries. Determinacy and Rewriting. ACM TODS 35(3) (2010)

Data Deduplication System for Supporting Multi-mode

Ho Min Jung*, Won Vien Park, Wan Yeon Lee,
Jeong Gun Lee, and Young Woong Ko

Dept. of Computer Engineering, Hallym University,
Okcheon-dong, Chunchon-si, Gangwon-do, Korea
{chorogyi,wonvien,wanlee,jeonggun.lee,yuko}@hallym.ac.kr
http://os.hallym.ac.kr

Abstract. The implementation approaches of data deduplication system divide into several modes including SBA(source-based approach), ILA(in-line approach) and PPA(post-process approach). Currently, most commercial systems are implemented and operated in an ILA and PPA approach, and some researchers have focused on the SBA approach. As data deduplication systems are widely used, to choose an appropriate mode considering operation environment becomes more and more important than ever. Because the overhead of each mode and resource usage wasn't fully studied, in some operating environments, the deduplication mode can lead to inefficiency and poor performance. In this study, we propose a data deduplication system supporting multi-mode. The proposed system can be operated in a mode that a user specifies during system operation, therefore, this system can be dynamically adjusted under consideration of system characteristics. In this paper, we operate the proposed system with the SBA, ILA and PPA mode, respectively, and we present the measurement results with a comparative analysis of the mode-specific performance and overhead.

Keywords: Deduplication, backup server, multi-mode, storage system.

1 Introduction

As the computer technology is rapid and widespread development, most information becomes digitalize and the amount of data is rapidly increasing. The amount of digital data around the world will be about 2052 Exabyte by 2012. Therefore, efficient data management will be more important and the data deduplication technique is regarded as an enabling technology. The data deduplication is a specialized data compression technique for eliminating redundant data to improve storage utilization. Generally, in the deduplication process, duplicate data is eliminated, leaving only one copy of the data to be stored, along with references

* This research was financially supported by the MEST and NRF through the Human Resource Training Project for Regional Innovation. And also was supported by Basic Science Research Program through the NRF funded by the MEST(2010-0016143).

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 78–87, 2011.

to the unique copy of data. With a help of data deduplication mechanism, the required storage capacity can be reduced since only the unique data is stored. In practical way, input data is divided into 4KB or more large blocks and given a hash value for each block. If the hash values are same between blocks, we regard that the blocks are identical. Therefore, before the data blocks are saved to storage system, we can eliminate duplicate data blocks in a file or between files. So, the data deduplication technique can be used efficiently on multiple versions of data such as a backup system. Generally, the functionalities of deduplication system are composed of hashing, comparing, input/output data blocks, searching and network transmission. The architecture of data deduplication system varies based on composing mechanism of the functionalities [1]. The deduplication system is classified as a source-based approach (SBA) and a target-based approach (TBA). If deduplication function is performed on client, it is classified into the source-based approach. If the key functionality of deduplication is performed on a deduplication server, it is classified into the target-based approach. The target-based approach is divided into an ILA and a PPA. In an ILA mode, the system processes a data deduplication function immediately when it receives data blocks. On the other hand, in a PPA mode, the system stores data blocks on the temporary storage at first, and it processes the data deduplication work if there are enough time and resource to do deduplication work.

The SBA, ILA and PPA are currently implemented several commodity product and are widely used. However, there is no concrete result that shows pros and cons of each method on different environments. For example, if a mobile user using 3G network tries to transfer a file, the user prefers to minimize network traffic because the user have to pay money for the amount packet data. In this situation, a SBA mode is suitable for the user, because it reduces network bandwidth. In addition, each method can cause a degradation and inefficiency of system resources because the patterns of system resource usage are different. For example, if the client machine is busy for processing CPU intensive job, then it prefers deduplication mode that uses less CPU resource. In this paper, we propose a multi-mode data deduplication system for file server. The proposed system can dynamically change deduplication mode by user. We measured the performance of deduplication system by operating SBA, ILA and PPA, respectively. The rest of this paper is organized as follows. In Section 2, we describe related works about deduplication system. In Section 3, we describe the operation mechanism of each deduplication mode, SBA, ILA and PPA. In Section 4, we explain the design principle of proposed system and implementation details for the deduplication system. In Section 5, we show performance evaluation result for each mode and we conclude and discuss future research plan.

2 Related Work

Venti [2] is a block-level network storage system which is similar to the proposed system. Venti identifies data blocks by a hash of their contents, because of using a collision-resistant hash function (SHA1) with a sufficiently large output, the

data block can be used as the address for read and write operations. The Low-Bandwidth File System [3] makes use of Rabin fingerprinting [4] to identify common blocks that are stored by a file system client and server, to reduce the amount of data that must be transferred over a low bandwidth link between the two when the client fetches or updates a file.

Zhu et al.s work [5] is among the earliest research in the inline storage deduplication area. They present two techniques that aim to reduce lookups on the disk-based chunk index. First, a bloom filter [6] is used to track the chunks seem by the system so that disk lookups are not made for non-existing chunks. Second, upon a chunk lookup miss in RAM, portions of the disk-based chunk index are prefetched to RAM. Lillibridge et al. [7] use the technique of sparse indexing to reduce the in-memory index size for chunks in the system at the cost of sacrificing deduplication quality. The system chunks the data into multiple megabyte segments, which are then lightly sampled (at random based on the chunk SHA-1 hash matching a pattern), and the samples are used to find a few segments seen in the recent past that share many chunks. Obtaining good deduplication quality depends on the chunk locality property of the dataset whether duplicate chunks tend to appear again together with the same chunks.

DEDE [8] is a decentralized deduplication system designed for SAN clustered file systems that supports a virtualization environment via a shared storage substrate. Each host maintains a write-log that contains the hashes of the blocks it has written. Periodically, each host queries and updates a shared index for the hashes in its own write-log to identify and reclaim storage for duplicate blocks. Unlike inline deduplication systems, the deduplication process is done out-of-band so as to minimize its impact on file system performance. HYDRAstor [9] discusses architecture and implementation of a commercial secondary storage system, which is content addressable and implements a global data deduplication policy. Recently, a new file system, called HydraFS [10], has been designed for HYDRAstor. In order to reduce the disk accesses, HYDRAstor uses bloom filter in RAM.

3 Multi-mode Deduplication

In this section, we describe SBA, ILA and PPA that are used on deduplication system and present the feature of each mode.

3.1 SBA Mode

In SBA mode, data deduplication process is performed in the client side and the client sends only non-duplicated files or blocks to deduplication server. First, the client performs file deduplication process by sending file hash data to server. The server checks file hash data from file hash list on DBMS. Second, if there is no matching file hash data in the server, the client starts block hashing.

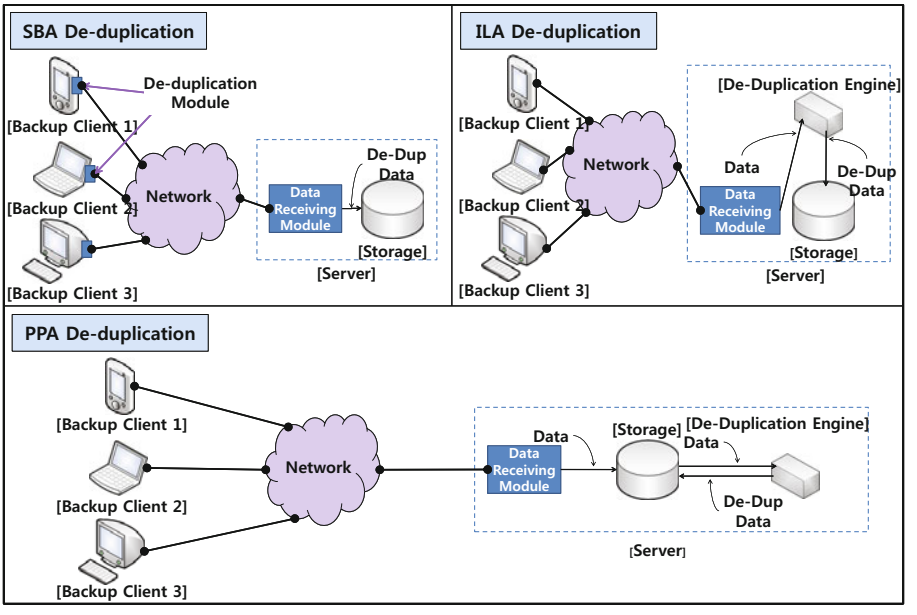


Fig. 1. The Conceptual Diagram of Deduplication Mode

The client divides a file into several blocks and hashes each block. The list of hash data is delivered to the server. Third, the server checks duplicated blocks by comparing the delivered hash data to hash data in the server. The server makes non-duplicated block list and send to the client. Finally, the client tries to send the non-duplicated data blocks to the server. In this mode, we can save network bandwidth by reducing the number of data blocks, therefore it is useful for a mobile device using 3G network. However, this approach requires more intensive CPU loads and the client may consume extensive system resource.

3.2 ILA Mode

ILA method performs elimination of duplicated data on server side when data is transferred from client to server. The client sends backup data to server and the server process deduplication work on the fly. Therefore, the client should send enough data blocks for handled in server side. The ILA mode can reduce disk storage of the system because the system does not need to preserve additional storage system for temporal data storing. In addition, this mode have decreased the entire working time by the immediate deduplication work and can save the DR(Disaster Recovery)-Ready time. However, it has drawback for managing the server not efficiently when large amounts of clients access to the server and the CPU resource of the server is all consumed.

3.3 PPA Mode

This mode performs deduplication work after data is temporary written to the disk in a server. The server saves the data from the client at first, then it send all data to deduplication engine for doing deduplication processing. After eliminating the duplication data, the server saves only non-duplicated data blocks to storage. The PPA usually consumes the system resource of a server while reducing the system resource of the client because all the deduplication work is processed on the server side. However, it needs additional disk storage for saving data on a disk. Moreover, DR-Ready time also increase because there are the gap between backup time and data deduplication.

4 Architecture of the Proposed System

To provide the context for presenting our methods for supporting the multi-mode and file deduplication, this section describes the architecture of deduplication system and explains the internal of each mode.

In figure 2, we show the main modules of the proposed system. The deduplication system are composed of several modules: File Accessing, Data Deduplication, Data Transfer, File Deduplication, Data Receiving, Data Scheduling, Data Deduplication, Compressing and DBMS.

File accessing module performs file grouping for efficient file deduplication by considering disk buffer cache size. If there are too many files to deduplication, the buffer cache will be busy for handling disk I/O. Therefore, we only treat enough files by grouping it within the size of disk buffer. With this approach we can avoid heavy disk I/O traffic. In file deduplication module, duplicated files are eliminated by using file hash. In client, it creates file hash for the file group and passes the file hash list to file deduplication module on the server. In

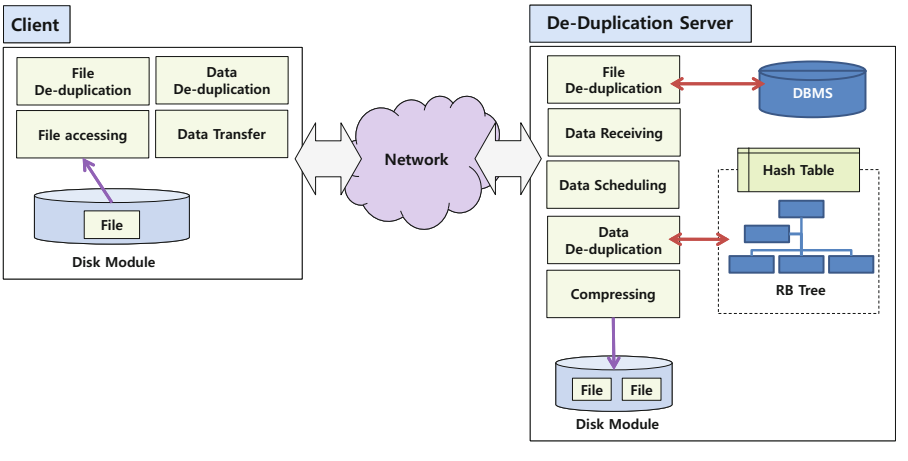


Fig. 2. The Architecture of the Proposed System

server, it checks duplicated files by comparing existing file hashes on a DBMS. The server sends only duplicated file hash list to the client. With this approach, we can prevent duplicated files are transferring to the server. In data deduplication module, block-level data deduplication is processed. The system divides data stream into blocks with chunking function. And then, we can get each data block hash using hash function such as SHA1, MD5 and SHA256. Generally, a chunking function can be dived into a fixed length chunking method and variable length chunking method. In our work, we adapted a fixed length chunking method because it is much more simple and easy to implement. The chunking size of data block is varying from 4Kbyte to several Mega Byte. In our work, we fixed 4Kbyte chunking size for increasing the performance of data deduplication. By choosing small chunking block, we can increase the possibility of finding

Algorithm 1. Multi Mode Algorithm

```
begin
    modeselection ← receivemode();
    filededuplication();
    if modeselection = SBA then
        while eof ≠ null do
            hash ← receivehash();
            check ← checkhash(hash);
            sendcheck(check);
            if check ≠ false then
                continue;
            else
                receivedata();
            end
            savemeta();
        end
    end
    if modeselection = ILA then
        while eof ≠ (data data ← receivedata() ) do
            hash ← makehash(data);
            check ← checkhash(hash);
            if check ≠ false then
                continue;
            else
                savedata();
            end
            savemeta();
        end
    end
    if modeselection = PPA then
        data ← receivedata();
        meta ← todisk(data);
        dataschedule(meta);
    end
end
```

duplicated block. Hash retrieval also very important because it causes frequent comparison, insert and delete operation. So, we adapted a red-black tree data structure for high performance hash operation. To process file and block deduplication, all the metadata have to be efficiently managed in a database module. The metadata includes file and block information, file and block hash data, file and block location, etc. Moreover, each file is composed of several blocks with and without deduplicated blocks. To build file from blocks, we have to carefully manage each block index.

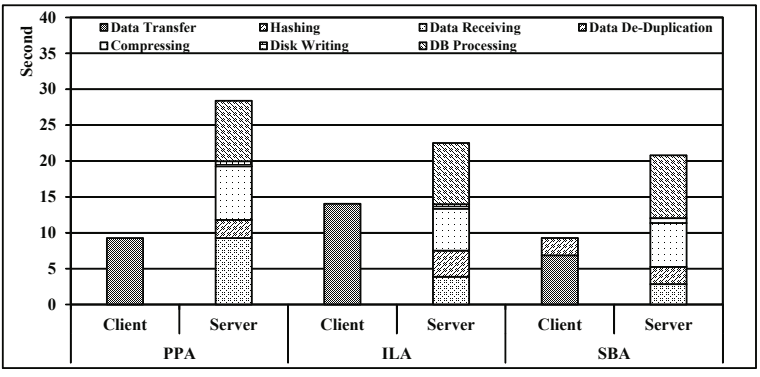
4.1 Multi-mode Algorithm for Deduplication

Algorithm 1 shows how to support multi-mode deduplication. First, the server receives mode information from the client. Second, the server performs file deduplication work, which is a common work regardless of each mode. Third, if the mode is set to SBA, then the client chunks the data and gets the block set and creates hash for each block before sending data blocks to the server. After verifying that the hash file is in the server, if server has the hash file then the server skips the data transferring step. Otherwise the server receives the data from the client and saves it in the disk. Fourth, if the mode is set to ILA, the server receives data from client and makes a data hash file by splitting data blocks. Similar to the SBA mode, it verifies if the hash file is in the server. If the server doesn't have the same hash value, the server saves the hash and metadata to the disk. The server repeats this routine until it receives EOF control message from the client. Finally, if the mode is set to PPA, the server receives file metadata from the client until all the data blocks are transferred and it saves the blocks temporarily into the disk. The server eliminates duplicated files and blocks when it scheduled to do deduplication work. Algorithm 1 shows the overall steps of the proposed system.

5 Performance Evaluation

In this work, we implemented the proposed system on Fedora core 9 operating systems. The hardware platform is equipped with Pentium 4 3.0 GHz CPU, 1024MB RAM, and WD-1600JS(7200/8MB) hard disk. All experiments are performed in the computer with Fedora Core 9. The Client and deduplication server is under the same hardware. In our experiment, data stream is composed of data blocks which duplicated rate varying 0, 40, and 80 percentages with 100 MByte size. 0 percentages means that there is no duplication in data blocks and 100 percentages mean that all of the data are duplicated. We measured the execution time, CPU resource and network bandwidth on the client and the server. All experiment is performed 10 times with SBA, ILA and PPA mode.

Figure 3 shows the execution time measurement for each mode. For the client, PPA is the fast because the client only transfers data to the server without deduplication work. And the execution time is proportional to the size of data. ILA shows much more delayed for processing data stream because ILA must



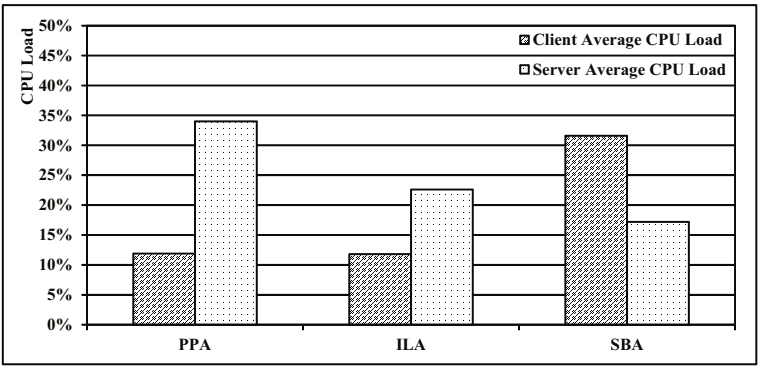
unit : ms

Duplication Ratio	De-dup Scheme	Send Data	Hashing	Recv File	Data Dedup	Compress	Disk Write	DB Process
0%	PPA	9172	0	9162	2601	10242	967	11029
	ILA	16447	0	1395	3815	10147	1006	11062
	SBA	10804	2346	2817	2339	9564	1105	11175
40%	PPA	9262	0	9251	2541	7469	708	8427
	ILA	14044	0	3872	3640	5848	607	8538
	SBA	6861	2411	2832	2413	6109	681	8759
80%	PPA	9077	0	9055	2431	4729	426	6464
	ILA	10688	0	5114	3144	1992	324	6756
	SBA	2432	2395	2616	2395	2136	343	6383

Fig. 3. Execution Time Analysis for Each Mode

wait until the server finishes deduplication work. We can minimize waiting time by increasing buffer size in the server. However, spacious buffer size can result in system performance degrade because physical memory is critical resource handling large hash index data. The execution time of SBA mode is depending on duplication rate. If duplication rate is high, then overall execution time will be decrease because the data transfer time is decreased. For the server, SBA mode needs small portion of system resource and PPA mode needs much more system resource. The system with SBA mode process all the hash work in the client side, therefore the server only store non-duplicated data blocks. However PPA mode has to do all the deduplication work in the server side, which makes the system very busy.

In figure 4, top graph show the average CPU load when duplication rate is 40 percentage and bottom table shows that average CPU load when duplication rate is 0, 40, 80 percentage, respectively. CPU load of the PPA and ILA show almost same value, 11 percentages, regardless of duplication rate. While SBA mode shows 31 percentage CPU loads. The reason why CPU load of SBA is higher than other mode is that file chunking and hashing work is done in the client mode. For the server, the mode that require higher CPU load is PPA because PPA mode requires deduplication work and additional disk I/O for processing temporal data storing. With this experiment, we conclude that if a user want to fast file deduplication in the client side without delay, then PPA is preferable. However, if a user wants to minimize CPU load for both the client and server,



Duplication Ratio	De-dup Scheme	Client Time (ms)	Client CPU Avg(%)	Server Time (ms)	Server CPU Avg(%)
0%	PPA	9172	11.8	34001	40
	ILA	16447	11.9	27425	32
	SBA	13150	31.6	27000	20.2
40%	PPA	9262	11.9	28396	34
	ILA	14044	11.8	22505	22.6
	SBA	9272	31.6	20794	17.2
80%	PPA	9077	11.9	23105	28.7
	ILA	10688	13.1	17330	20
	SBA	4727	30.9	13873	9.3

Fig. 4. Comparison of Average CPU Load

then ILA mode is a good solution. And if a user want to fast file transfer and low CPU load for the server, then SBA is the best choice.

6 Conclusion

In this paper, we presented data deduplication system supporting multi-mode which can dynamically change deduplication mode. Our key idea is to provide suitable deduplication mode considering system resource and user preference. We have designed and implemented the multi-mode deduplication server on the Linux system. We analyzed the performance of SBA, ILA and PPA modes, Experiment results show that PPA can minimize execution time for handling data deduplication and CPU load for the client. Our result shows that if a user wants to fast processing for file deduplication with low CPU workload in the client side, then PPA is preferable. However, if we want to minimize the system overhead on the server side, we have to use SBA mode. For future work, we will study how to convert the optimized method to dynamically considering CPU, I/O and network bandwidth. Moreover, we will design the optimized deduplication module providing QoS(quality of service) concept by continuously monitoring server and client behavior. Also, we believe that energy efficient data deduplication is useful for smartphone environment in the future.

References

1. Tan, Y., Jiang, H., Feng, D., Tian, L., Yan, Z., Zhou, G.: SAM: A Semantic-Aware Multi-tiered Source De-duplication Framework for Cloud Backup. In: 39th International Conference on Parallel Processing (2010)
2. Quinlan, S., Dorward, S.: Venti: a new approach to archival storage. In: Proceedings of the 1st USENIX Conference on File and Storage Technologies, FAST (2002)
3. Muthitacharoen, A., Chen, B., Mazieres, D.: A Low-Bandwidth Network File System. In: Proceedings of the Symposium on Operating Systems Principles (SOSP 2001) (2001)
4. Rabin, M.O.: Fingerprinting by random polynomials:Technical Report TR-15-81, Center for Research in Computing Technology, Harvard University (1981)
5. Zhu, B., Li, K., Patterson, H.: Avoiding the disk bottleneck in the data domain deduplication file system. In: Proceedings of the 6th USENIX Conference on File and Storage Technologies, FAST (2008)
6. Broder, A., Mitzenmacher, M.: Network Applications of Bloom Filters: A Survey. In: Internet Mathematics (2002)
7. Lillibridge, M., Eshghi, K., Bhagwat, D., Deolalikar, V., Trezise, G., Campbell, P.: Sparse Indexing, Large Scale, Inline Deduplication Using Sampling and Locality. In: Proceedings of the 7th USENIX Conference on File and Storage Technologies, FAST (2009)
8. Clements, A., Ahmad, I., Vilayannur, M., Li, J.: Decentralized Deduplication in SAN Cluster File Systems. In: Proceedings of 2009 USENIX Technical Conference (2009)
9. Dubnicki, C., Gryz, L., Heldt, L., Kaczmarczyk, M., Kilian, W., Strzelczak, P., Szczepkowski, J., Ungureanu, C., Welnicki, M.: HYDRAstor: a Scalable Secondary Storage. In: Proceedings of the 7th USENIX Conference on File and Storage Technologies, FAST (2009)
10. Ungureanu, C., Atkin, B., Aranya, A., Salil Gokhale, S.R., Calkowski, G., Dubnicki, C., Bohra, A.: HydraFS: a High-Throughput File System for the HYDRAstor Content-Addressable Storage System. In: Proceedings of the 8th USENIX Conference on File and Storage Technologies, FAST (2010)

On the Maximality of Secret Data Ratio in CPTE Schemes

Trung Huy Phan[1] and Hai Thanh Nguyen[2]

[1] Hanoi University of Science and Technology
huyfr2002@yahoo.com
[2] Ministry of Education and Training
nhthanh@moet.gov.vn

Abstract. Based on the ring of integers modulo 2^r, Chen-Pan-Tseng (2000) introduced a block-based scheme (CPT scheme) which permits in each block F of size $m \times n$ of a given binary image B to embed $r = \lfloor \log_2(k+1) \rfloor$ secret bits by changing at most two entries of F, where $k=mn$. As shown, the highest number of embedded secret bits for at most two bits to be changed in each block of k positions of F in any CPT-based schemes is $rmax=\lfloor \log_2(1+k \ (k+1)/2) \rfloor$, approximately $2r$-1, twice as much as r asymptotically, and this can reached approximately in our CPTE1 scheme by using modules on the ring $\mathbf{Z}_2$ of integers modulo 2. A new modified scheme-CPTE2 to control the quality of the embedded blocks, in the same way as Tseng-Pan's method (2001), is established. Approximately, CPTE2 scheme gives $2r$-2 embedded bits in F, twice as much as r-1 bits given by Tseng-Pan' scheme, while the quality is the same.

Keywords: Maximality, secret data ratio, binary image, steganography, CPTE scheme.

1 Introduction

In the area of steganography, one of the most challenging problems is that we need to hide secret data into binary images with the high ratio of secret data and with the less distortion of the images. In case ones need to use some schemes like CPT [1] to prevent steganalysis, especially to histogram-based attacks (see for examples some interesting analysis in [6], if the alpha ratio of the number of changed pixels to the number of total pixels of a given palette image is lower than 0.1, it is very difficult to guess if the image contains hidden data), our schemes can provide a higher quality of stego palette images since by applying these schemes to a concrete palette, ones can gain a small alpha ratio while the amount of total hidden bits for real problems is large enough. In block-based approach (see references [1,2,3,4,5]), each binary image is partitioned into binary blocks of the same size $m \times n$, each block can be seen as a binary matrix of size $m \times n$. In such a block F of size $m \times n$, by taking WL scheme [2] one can embed one bit by changing at most one bit of F. From CPT scheme proposed by Chen-Pan-Tseng (2000)[1], in F having $q=mn$ entries one can embed $r=\lfloor \log_2(q+1) \rfloor$ bits by changing at most two bits of F. As shown in the subsection **3.1**

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 88–99, 2011.

of this paper, the highest number of secret bits for at most two bits to be changed in such a block F- in any CPT-based scheme (CPTE scheme - for short)- is $rmax=\lfloor \log_2((1+q(q+1)/2)\rfloor$, approximately $2r$-1. Based on modules over the ring Z_2 of integers modulo 2, in our CPTE1 scheme, this ratio can be reached approximately, as shown by Corollary 1. Hence, asymptotically this provides twice as much ratio as r which is gained by CPT scheme. This result allows us to establish a new modified scheme-CPTE2 from CPTE1 to control the quality of the embedded blocks F to improve the invisibility of the whole embedded image, in the same way as in Tseng-Pan's modified method [4] (2001) from CPT one. Approximately, in CPTE2 scheme $2r$-2 secret bits can be embedded which has twice as much as r-1 bits gained by Tseng-Pan' scheme (also twice as much as r-1 bits gained in the scheme modified by Hirohisa (2003)[5] later), while the quality is the same.

The paper is divided into 5 sections. Following the introduction section, section 2 gives a brief representation of the CPT scheme and the maximal secret data ratio (MSDR) of secret bits embedded in each block F of size $m\times n$ of pixels in binary images or palette images. Section 3 is provided as the introduction of CPTE1 scheme and section 4 to CPTE2 scheme. The last section 5- provides results of the evaluation, comparison between MSDR with CPT, CPTE1, CPTE2 schemes, Tseng-Pan's modified CPT scheme (MCPT) in table 1 and conclusion.

2 CPT Scheme

Given a binary image B which is partitioned into p blocks F_t as binary matrices with the same size $m\times n$, $1\le t\le p$. Combined with these blocks of the image are two matrices K, W of the same size $m\times n$: K is a binary key matrix whose elements are randomly chosen, W is a weight matrix whose elements are integers chosen randomly with the restriction as shown in Theorem below.

For details, in each such a concrete block F, one can embed $r=\lfloor \log_2(mn+1)\rfloor$ secret bits. Each entry F_{ij} has a value 0 or 1. Changing (or inverting) F_{ij} in F means that F_{ij} is changed to (F_{ij} XOR 1), i.e. we take $F_{ij}:=F_{ij}$ XOR 1. The operation $F\oplus K$ is taken by the bitwise exclusive-OR on two equal-size binary matrices F and K. The following operation $\otimes$ computes the sum which is obtained by taking pairwise multiplications on two equal-size interger matrices. We define

$$T=F\oplus K,\ \ \mathrm{SUM}[T\otimes W]=\sum_{1\le i\le m}\sum_{1\le j\le n} T_{ij}\times W_{ij} \bmod 2^r.$$

The correctness of the CPT scheme is shown the following.

Theorem. Given arbitrary $m\times n$ binary matrices F, K and an arbitrary $m\times n$ integer weight matrix W satisfying W_{ij}, $1\le i\le m$, $1\le j\le n\}=\{1, 2,\ldots, 2^r\text{-}1\}$ with $r=\lfloor \log_2(m.n+1)\rfloor$. Let $b=b_1b_2\ldots b_r$ be a bit stream. We can invert at most two entries of F to get $b=\mathrm{SUM}[(F\oplus K)\otimes W]$.

3 Maximal Secret Data Ratio of the Embedded Bits and the CPTE1 Scheme

3.1 Maximal Secret Data Ratio of Embedded Bits

In this part, without loss of generality, we consider only one fixed matrix F of size $m \times n$ of pixels in an image G which is considered as a set of pixels. In F each entry F_{ij} *(or the pair (i,j))* can be understood as a pixel, also F_{ij} can be to as the color of this pixel whenever its value is used. Put $q = mn$. For a given integer $k>1$, changing the pixel F_{ij} means that the color F_{ij} is changed to a new F_{ij}' by one of the k other ways, so that F_{ij} is similar to F_{ij} by some selected color distance.

In the case of binary image, $k=2$, but in general, k can be larger than 2 for color images. We consider here *CPTE schemes* which mean secret bits can be embedded in each matrix F by changing at most two entries. Together with F, each new matrix F' after changing entries of F is called a *configuration*. Hence, the number of configurations after changing one entry in total is $(k\text{-}1).q$ at most, since each entry has k-1 ways to change. If we change two entries in F, $(k\text{-}1).q.(q\text{-}1)/2$ ways can be taken.

Hence, in general, in any CPTE scheme we have at most $1+(k\text{-}1).q+(k\text{-}1).q.(q\text{-}1)/2$ configurations. This means that we can hide at most

$$R=\lfloor \log_2(1+(k\text{-}1).q+(k\text{-}1).q.(q\text{-}1)/2)\rfloor \text{ secret bits in } F.$$

We call R the *Maximal Secret Data Ratio* (MSDR) of CPTE schemes.

Specially, for the case of binary image, $k=2$, $R = \lfloor \log_2(1+q+q.(q\text{-}1)/2)\rfloor = \lfloor \log_2(1+q.(q+1)/2)\rfloor$ secret bits can be embedded in F.

3.2 Module Approach in Hiding Secret Data

Each (right) module M over the ring $\mathbf{Z}_q$ is an additive abelian group M with zero 0 together with a scalar multiplication "." to assign each couple $(m,\boldsymbol{k})$ in $M \times \mathbf{Z}_q$ to an element $m.\boldsymbol{k}$ in M. Let $\mathbf{Z}_q=\{\mathbf{0},\mathbf{1},..,\boldsymbol{q}\text{-}\mathbf{1}\}$. Some following basic properties are used usefully in the sequent:

P1) $m.\mathbf{0} = 0$; $m.\mathbf{1}=m$;
P2) $m+n = n+m$ for all m,n in M.
P3) $m.(\boldsymbol{k}+\boldsymbol{l}) = m.\boldsymbol{k} + m.\boldsymbol{l}$ for all m in M, $\boldsymbol{k}, \boldsymbol{l}$ in $\mathbf{Z}_q$.

Given an image G, denote by C_G the set of colors given by G, $C_G= \{C_p: p \text{ in } G\}$ where each C_p is the color of pixel p. Suppose that we can find a function Val: $C_G \rightarrow \mathbf{Z}_q$ and a *color changing mapping* Next: $C_G \rightarrow C_G$ satisfying the condition:

$$\forall\, c \in C_G,\ \text{Val}(\text{Next}(c)) = \text{Val}(c)+1. \tag{1}$$

For the palette image case we claim one extra condition:

$$\forall\, c \in C_G,\ c' = \text{Next}(c) \text{ is a color "similar" to } c. \tag{2}$$

Given any set $S=\{p_1,p_2,..,p_N\}$ of N pixels in G, $N \geq |M|$, we define a surjective mapping

$$h: S \to M\text{-}\{0\} \text{ from } S \text{ onto } M\text{-}\{0\}, \quad (3)$$

h is called a *weight mapping* and for each p in S, $m=h(p)$ is called *the weight of p.*

Consider a set of secret data $D=\{d_m : m\in M\}$ so that each item d_m can be extracted easily whenever m is given. In the two subsections below, we propose a method to hide any element $d\in D$ into S by changing colors of at most one element in S.

3.2.1 Hiding the Secret Item d into S

Step 1) Computing $m = \sum_{p\in S} h(p).\text{Val}(C_p)$ in the right module M.
Step 2) Case $d_m = d$: keep S intact;

Case $d \neq d_m$: suppose $d=d_s$, for some $s \in M$.

i) Find any $po \in S$ such that $h(po)= s\text{-}m$.
ii) Changing the color C_{po} of po into C_{po}'= Next(C_{po}).

3.2.2 Extracting the Secret Item d from S

Step 1) Computing $u= \sum_{p\in S} h(C_p).\text{Val}(C_p)$
Step 2) Using m, recover $d=d_u$.

3.2.3 Correctness of the Method

Theorem 1. *The item d_u extracted in step 1 of the extracting stage above is exactly the secret item d needed to be hidden into S.*

Proof. We need consider only the case $d=d_s \neq d_m$ and prove that $u=s$.

Indeed, in the step 2i) of **3.2.1** an element po can be selected since h is surjective. By $C_{po'}$ = Next(C_{po}) in the part **3.2.1** step (2), using conditions one deduces $\text{Val}(C_{po'})= \text{Val}(\text{Next}(C_{po}))=\text{Val}(C_{po})+1$, then by the property (P1) of modules, we have

$u= \sum_{p\in S}\ h(C_p).\text{Val}(C_p) = \sum_{po\neq p\in S}\ h(p).\text{Val}(C_p) + h(po).\text{Val}(C_{po'})$

$u =\sum_{po\neq p\in S}\ h(p).\text{Val}(C_p)+ h(po).(\text{Val}(C_{po})+1)$, and by property (P3),

$u= \sum_{po\neq p\in S} h(p).\text{Val}(C_p)+ h(po).\text{Val}(C_{po})+h(po).1$. Then by **3.2.1.** step 2(i),

$u= \sum_{p\in S} h(C_p).\text{Val}(C_p) + (s\text{-}m).1$, implying $u = m+(s\text{-}m)=s$ by (P1) and properties of module. This implies $d=d_s=d_u$.‖

3.2.4 Hiding Secret Data in Binary Images

For binary images, applying $q=2$ the addition in $\mathbf{Z}_2$ can be seen as the operation exclusive -OR on bits, and $M=\mathbf{Z}_2 \times \mathbf{Z}_2\times..\times \mathbf{Z}_2$ is the n-fold cartesian product of $\mathbf{Z}_2$ which can be seen as a (right) $\mathbf{Z}_2$-module, each element $x=(x_1,x_2,..,x_n)$ in M can be presented as an n-bit stream $x=x_1x_2..x_n$, with operations defined by:

D1) For any $x=x_1x_2..x_n$, $y=y_1..y_n$ in M, $\boldsymbol{k}$ in $\mathbf{Z}_2$, $x+y = z_1z_2..z_n$ where $z_i=x_i +y_i$, $i=1,..,n$ can be computed by bitwise XOR.

D2) $x.\boldsymbol{k}= z_1z_2..z_n$ where $z_i=x_i.\boldsymbol{k}$ $(= x_i$ AND $\boldsymbol{k})$.

Given a binary image G, we set $C_G= \mathbf{Z}_2 =\{\mathbf{0},\mathbf{1}\}$ and Val is the identical function on $\mathbf{Z}_2$, Val(c)=c for all c in $\mathbf{Z}_2$. The function Next: $\mathbf{Z}_2 \to \mathbf{Z}_2$ is defined by

$$\text{Next}(c)=c+1, \text{ for all } c \text{ in } \mathbf{Z}_2 \tag{4}$$

and *changing a color* c is done by replacing c with c'=Next(c)=c+1.

For any set S=$\{p_0,p_1,..,p_N\}$ of $N+1$ pixels in G, $N+1=|S|\geq 2^n -1$, we can hide a secret n-bit stream $b= b_1b_2..b_n$ by changing color of at most one pixel in S as follows.

3.2.4.1 Hiding the Secret Item b into S. Step 0) Choose a secret set K=$\{k_p \in \mathbf{Z}_2$: p in S$\}$ of |S| key bits k_p. Change the color C of each p $\in$ S into a new color C_p*=C_p+k_p (in $\mathbf{Z}_2$). For the new colors of pixels in S, apply the steps (1),(2) in **3.2.1** by:

Step 1) Computing $m = \sum_{p\in S} h(p).C_p$* in the $\mathbf{Z}_2$ –module M.
Step 2) Case $m = b$: keep S intact;

Case m $\neq b$: find any $px \in S$ such that $h(px)= b-m$, changing the color C_{px} of px into C_{px}'= Next(C_{px})=C_{px}+1.

(Then the new color of px is C_{px}* = C_{px}'+k_{px} = C_{px}+1+k_{px}=C_{px}+k_{px}+1=C_{px}*+1).

Therefore only one pixel p in S changes its color, after hiding b into S.

3.2.4.2 Extracting the Secret Item d from S. Step 0) Using the secret set K, change the color C_p of each p $\in$ S into a new color C_p*=C_p+k_p . For the new colors of pixels in S, apply steps 1), 2) in **3.2.2** as follows:

Step 1) Computing $u = \sum_{p\in S} h(p).C_p$* in the $\mathbf{Z}_2$ –module M.
Step 2) Return $b=u$.

3.2.4.3 Correctness of the Method. Since N+2=$\lfloor$Card(S)+1$\rfloor \geq 2^n$, applying **3.2.1**, **3.2.2** to each pixel p with the new color C_p* in steps (1), (2) of hiding secret data and extracting secret data, and using the properties of $\mathbf{Z}_2$, we deduce

Theorem 2. Given a binary image G, changing color of at most one pixel in any subset S of G by CPTE1 scheme, we can hide $\lfloor$Card(S)+1$\rfloor$ secret bits by **3.2.4.1** and extract exactly these bits by **3.2.4.2.**

Proof. Consider $u = \sum_{p\in S} h(p).C_p$* as in **3.2.4.2**,

If $b=m$ in step 2) in **3.2.4.1**, we have $b = m= \sum_{p\in S} h(p).C_p* = u$;

If $b\neq m$: we have $u = \sum_{p\in S} h(p).C_p* = \sum_{px\neq p\in S} h(p).C_p* + h(px).C_{px}'*$ by step 2(ii) in **3.2.4.1** and by $C_{px}'*=C_{px}*+1$,

$u = \sum_{px\neq p\in S} h(p).C_p* + h(px).(C_{px}*+1) = \sum_{p\in S} h(p).C_p* + h(px).1 = m + h(px)=m+b-m=b$.

Therefore in any cases, $u=b$ as claimed in the return step (3).‖

3.3.1 Parameters for CPTE1 Scheme

We consider now one binary matrix F of size $m\times n$ of a binary image G in which each entry F_{ij} of F presents the pixel (i,j) and its color $F_{ij} \in C_G = \mathbf{Z}_2 = \{\mathbf{0},\mathbf{1}\}$. Put $p = mn+2$. Suppose that p has a binary presentation $p=b_t b_{t-1}..b_0$ with $b_t = 1$, $t \geq 1$. We can split in a secret way F into two parts, say S_1 and S_2, such that (3.7) S_1 has at least 2^α -1 pixels, S_2 has at least 2^β -1 elements where α, β are defined as below

$$\begin{cases} \alpha = t\text{-}1, \beta = t \text{ if } b_{t-1}=\mathbf{1} \\ \alpha = t\text{-}1 = \beta \text{ if } b_{t-1}=\mathbf{0}. \end{cases} \quad (5)$$

Set

$$m(p) = \alpha + \beta. \quad (6)$$

Applying Theorem 2 above, one can hide any α secret bits into S_1 by changing colors of at most one pixel in S_1, β secret bits into S_2 by changing colors of at most one pixel in S_2, therefore $m(p) = \alpha + \beta$ secret bits into F by changing colors of at most two pixels in F.

Remark 1. In practical, we present each element b in M as a bit stream $b=b_{\alpha+\beta}b_{\alpha+\beta-1}..b_1$ of $\alpha+\beta$ bits, and the addition + on M is nothing but bitwise exclusive OR (XOR) on these streams. The result $d = b.c$ for all c in $\mathbf{Z}_2$ can be presented as

$$d=d_{\alpha+\beta}\, d_{\alpha+\beta-}..d_2 d_1 \text{ where } d_j = b_j \,.\, c = b_j \text{ AND } c,\ j = 1,..\,,\alpha+\beta.$$

Remark 2. From now on, for simplicity, we present elements b,d in M as natural numbers except that on them the operation $b \oplus d$ is treated by bitwise XOR on $m(p)$-bit streams, and the product $b.c$, $c\in\mathbf{Z}_2$, is treated as in the Remark 1.

We can select one binary key matrix $K=(K_{ij})$ of size $m\times n$ for both S_1 and S_2: each k_{ij} such that F_{ij} in S_1 is used for S_1, and F_{ij} in S_2 then k_{ij} is used for S_2 vise versa. Besides, the set M_1 given by the following equation (7) is assigned as the set of weights of elements in S_1:

$$M_1=\{b_{\alpha+\beta}b_{\alpha+\beta-1}..b_2b_1 \in M: b_\beta, b_{\beta-1},..,b_2, b_1=0\} - \{0\} \quad (7)$$

The set M_2 given by the following equation (8) is assigned as the set of weights of elements in S_2:

$$M_2 =\{b=b_{\alpha+\beta}b_{\alpha+\beta-1}..b_2b_1 \in M: b_{\alpha+\beta}, b_{\alpha+\beta-1},..,b_{\alpha+1} = 0\} - \{0\}. \quad (8)$$

+) The weight functions from S_1, S_2 to M_1, M_2 are presented by a weight matrix $W=(W_{ij})$ of size $m\times n$ satisfying conditions:

$$\{W_{ij}: i=1,..,m, j=1,..,n\} = M-\{0\};\ \{W_{ij}: F_{ij}\in S_1\} = M_1-\{0\};\ \{W_{ij}: F_{ij}\in S_2\} = M_2-\{0\}. \quad (9)$$

3.3.2 Hiding and Extracting Secret Bits by CPTE1 Scheme

3.3.2.1 Hiding secret bits. Suppose one need to hide a stream d of $\alpha+\beta$ bits in F, $d = d_{\alpha+\beta}d_{\alpha+\beta-1}..d_2d_1$.

Put $u = d_{\alpha+\beta}d_{\alpha+\beta-1}..d_{\beta+1}0..0$ and $v = 0..0d_{\beta}..d_2d_1$ so that $d = u \oplus v$, $u \in M_1$ and $v \in M_2$.

Step 1) Compute $T=F \oplus K$ as in CPT scheme. Each pixel (i,j), $i=1,..,m$; $j=1,..,m$ has a color $F_{i,j}$ and a new color T_{ij}, T is considered as the matrix of new colors of pixels in F.

Step 2) Compute $s = \sum_{i=1,..,m;\ j=1,..,n} W_{ij}.T_{ij}$. This sum is denoted by $[W.T]$;

Present $s = s_1 \oplus s_2$ where $s_1 = s_{\alpha+\beta}s_{\alpha+\beta-1}..s_{\beta+1}0..0 \in M_1$ and $s_2 = 0..0s_{\beta}..s_2s_1 \in M_2$;

Step 3) Consider s_1 and s_2. For s_1, there are two cases:

a) $s_1 = u$: keep S_1 intact;

b) $s_1 \neq u$: compute $d = u - s_1$ $(= u + s_1)$ in $\mathbf{Z}_2$, find any pixel $p=(i,j)$ in S_1 such that d is its weight, changing color $F_{i,j}$ to $F_{i,j}' = F_{i,j} + 1$ in $\mathbf{Z}_2$.

For s_2, there are two cases:

c) $s_2 = v$: keep S_2 intact;

d) $s_2 \neq v$: compute $e = v - s_2$ $(= v + s_2)$ in $\mathbf{Z}_2$, find any pixel $p=(i,j)$ in S_2 such that e is its weight, changing color $F_{i,j}$ to $F_{i,j}' = F_{i,j} + 1$ in $\mathbf{Z}_2$.

3.3.2.2 Extracting Secret Bits. Given F as the matrix which the secret bit stream d is embedded in.

Step 1) Compute $T = F \oplus K$;

Step 2) Compute $s = [W.T]$;

Step 3) Return $d = s$ as the secret bit stream which has been embedded in F.

3.3.2.3 Correctness of CPTE1 Scheme. This comes from the results in **3.2.4.3** above.

Corollary 1. The number of embedded bits $m(p)$ in F by CPTE1 (defined by (3.4)-(3.5)) approximates the MSDR in general.

Proof. Consider $q=mn$, $p=mn+2=q+2$. From (3.4)-(3.5) we deduce $m(p) \geq \log_2((p/2)^2)$ since that if p has a binary presentation $p=b_tb_{t-1}..b_1b_0$ with $b_{t-1}=1$ then $m(p)=2t-1$ and $\log_2((p/2)^2)=2t-2$, and with $b_{t-1}=1$, $m(p)=\log_2((p/2)^2)=2t-2$.

By maximality, MSDR$\geq m(p)$, hence $\log_2(1+q(q+1)/2) \geq m(p) \geq \log_2((q+2)/2)^2$.

From obvious inequality $(1+q(q+1)/2) < 2((q+2)/2)^2$ for any $q > 0$,

$\log_2[(1+q(q+1)/2) / ((q+2)/2)^2] < \log_2 2 = 1$. Based on these inequalities by putting $\log_2((1+q(q+1)/2) = h+\varepsilon$ with $h=\lfloor \log_2(1+q(q+1)/2) \rfloor$ and some real number ε, $0 \leq \varepsilon < 1$, one gets $\log_2((q+2)/2)^2) = h-1+\varepsilon'$ for some real number ε' such that $\varepsilon \leq \varepsilon' < 1+\varepsilon$. Hence

MSDR $-m(p) \leq \lfloor \log_2(1+q(q+1)/2) \rfloor - \lfloor \log_2((q+2)/2)^2) \rfloor = h - \lfloor h-1+\varepsilon' \rfloor = 1 - \lfloor \varepsilon' \rfloor \leq 1$.

Let us remark that $m(p)$ can be equal to MSDR. For example, with any q satisfying $q \geq 8$, $q+1=2^t$ and $p=2^t+1$, or $q+1=2^t-1$ and $p=2^t$, then MCDR$= \lfloor \log_2(1+q(q+1)/2) \rfloor = 2t-2 = m(p)$. The proof is completed.||

3.3.3 Examples for CPTE1 Scheme

Given 3 x 3 binary matrices F, K , a 3×3 weight matrix W, and $d = d_4d_3d_2d_1$ a 4 bit stream ($p=9+2=11=8+2+1$ has the binary presentation 1011, by (3.5) $\alpha = \beta = 2$, $m(p) = 4$).

In details, S_1={F_{11},F_{12},F_{13},F_{21},F_{22}}; S_2={F_{23},F_{31},F_{32},F_{33}}, M_1={0,12,8,4}, M_2={0,3,2,1} or in binary presentation, M_1={0000,1100,1000,0100}, M_2={0000,0011,0010,0001}. For example,

$$F=\begin{bmatrix}1&0&1\\0&0&1\\1&0&1\end{bmatrix}\quad K=\begin{bmatrix}0&1&1\\1&0&1\\0&1&1\end{bmatrix}\quad W=\begin{bmatrix}8&12&4\\4&8&3\\1&2&3\end{bmatrix}\quad T=F\oplus K=\begin{bmatrix}1&1&0\\1&0&0\\1&1&0\end{bmatrix}$$

$$s=[W.T]=8.\mathbf{1}+12.\mathbf{1}+4.\mathbf{1}+1.\mathbf{1}+2.\mathbf{1}=1000\oplus1100\oplus0100\oplus0001\oplus0010=0011.$$

We present $s=s_1\oplus s_2$, s_1=0000, s_2=0011.

a) With d=0011, we have $d=s$, F is not changed.

b) With d=0000, we write $d=u\oplus v$, u=0000, v=0000. Since $u=s_1$, we keep S_1 intact. By $v\neq s_2$, we need to change S_2 by one entry: Putting $a=v\text{-}s_2=v\oplus s_2$ = 0000 ⊕0011= 0011, we need to choose W_{33} =0011 in S_2, or the corresponding F_{33} has to be changed to $F_{33}\oplus1=0$, T_{33} to $T_{33}\oplus1=1$, hence the new sum $s'=[W.T\,new]=s\oplus W_{33}=0000=d$.

c) With d=1110, u=1100,v=0010, $u\neq s_1$ and $v\neq s_2$ so we need to change one entry in S_1 and another in S_2. In S_1, computing $u\oplus s_1=u=W_{12}$ implies that F_{12} is changed to $F_{12}\oplus1=1$. In S_2, computing $v\oplus s_2=0001=W_{31}$ implies that F_{31} is changed to $F_{31}\oplus1=0$. Hence T has two new entries $T'_{12}:=T_{12}\oplus1=0$;$T'_{31}:=T_{31}\oplus1=0$. The new sum $s'=[W.T\,new]=s\oplus W_{12}\oplus W_{31}=0011\oplus1100\oplus0001=1110$=d as claimed.

4 Modified Scheme CPTE2

To improve the invisibility of the whole embedded binary image, in [4] Tseng-Pan(2001) introduced a modified scheme from CPT scheme (MCPT scheme for short - see [1], [4], [5]) to control the high quality of embedded binary images. Here we introduce a new modified CPTE2 scheme based on our CPTE1 scheme by a method similar to Tseng-Pan's method in [4] to improve the quality of embedded binary images. Given an $m\times n$ binary matrix F. Put $p=mn$ and $p=b_tb_{t-1}..b_0$ as the binary presentation of p. We split in a secret way F into two parts S_1 and S_2, such that S_1 has at least 2^α pixels, S_2 has at least 2^β elements due to (3.5),(3.6),(3.7). In this scheme, in the stream of $m(p)=\alpha+\beta$ bits $d=d_{m(p)}d_{m(p)-1}..d_{m(p)-\alpha+1}\,d_\beta d_{\beta-1}..d_2d_1$, ones hide the α - bit stream $d_{m(p)}d_{m(p)-1}..d_{m(p)-\alpha+1}$ of real secret data in S_1 by changing at most one pixel as before, in S_2 ones hide β - bit stream $d_\beta d_{\beta-1}..d_2d_1$ in which only (β-1) -bit stream $d_\beta d_{\beta-1}..d_2$ is considered as real secret data, the rest d_1 is used as a control bit for the hiding process: d_1=1 means that the hiding process in F is failed (the β - bit stream $d_\beta d_{\beta-1}..d_2d_1$ is odd) and we need to hide the secret bits in next blocks F'of the image G.

4.1 Modification for Quality Control

We present following parameters in details:

+) Let F,K,W be matrices of the same size $m\times n$ defined as in the part **3.3.1** above,

$W = (W_{ij})_{m\times n}$ satisfies (3.8),(3.9),(3.10):

$\{W_{ij}: i=1,..,m, j=1,..,n\} = M\text{-}\{0\}$ and $\{W_{ij}: F_{ij} \in S_1\} = M_1$; $\{W_{ij}: F_{ij} \in S_2\} = M_2$.

+) For F we define the distance matrix $d(F)$ of the same size $m\times n$ by

$$d(F)_{ij} = \min_{\forall x,y}\{\sqrt{(i-x)^2+(j-y)^2} \mid F_{ij} = 1 - F_{xy}\}.$$

i.e. $d(F)_{ij}$ is the smallest distance from F_{ij} to some entry having its value as the completion of F_{ij} . This matrix is used to check for safety whether each entry F_{ij} can be changed in cases we need or not. For example, we consider

$$F = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \quad d(F) = \begin{bmatrix} 2 & 1 & 1 & 2 & 3 \\ \sqrt{2} & 1 & 1 & 2 & 3 \\ 1 & 1 & \sqrt{2} & \sqrt{5} & \sqrt{10} \\ 1 & 1 & 2 & \sqrt{8} & \sqrt{13} \\ 1 & 1 & 2 & 3 & 4 \end{bmatrix}$$

In applications, to avoid taking roots, instead of $d(F_{ij})$ we use $(d(F)_{ij}^2)$.

+) We need to set $d=d_{m(p)}d_{m(p)-1}..d_1 = [W.(F' \oplus K)]$, with $d_1 = 0$ if F is changed successfully to F', or $d_1=1$ otherwise.

+) Extra conditions put on W:

Condition 1. (odd condition- as an extension of Tseng-Pan' method): in each square Q of four points of the form $W_{i,j}$; $W_{i,j+1}$;$W_{i+1,j}$,$W_{i+1,j+1}$ in W, if Q has at least two points to be weights of two corresponding elements in S_2 then Q has at least one odd element.

Condition 2. (skip condition) Denote by $\sim K$ the completion matrix of K given by taking completion on each entry of K. It should be

$$[W.\ K]_{S_2} \mod 2 = 1 \text{ (the sum taking restrictly only on } W_{ij}.K_{ij} \text{ with all } F_{ij} \text{ in } S_2). \quad (10)$$

$$[W.\ \sim K]_{S_2} \mod 2 = 1 \text{ (the sum taking restrictly on } W_{ij}.(1\text{-}K_{ij}) \text{ with all } F_{ij} \text{ in } S_2). \quad (11)$$

Remark 3. a) By the condition 1, if S_2 is not mono-value then there exists a square Q having one point to be an odd weight of some element in S_2. This property will be used to prove lemma 1 below.

b) It should be satisfied the skip condition since that without this condition, if S_2 is mono-value (all its entries have the same values 1 or 0), before or after inversion, ones need to check whether S_2 is mono-value or not, since this leads to verify whether F is failed or successful for hiding of secret data. Hence, using the condition 2 ones can speed up the algorithm, but this is not an essential condition.

4.2 Algorithm for Quality Control

We assume that the conditions 1, 2 in **4.1** are satisfied.

Step 1) Prepare the $m(p)$- bit tream $d' = d_{m(p)}d_{m(p)-1}..d_2 0$ to embed in F from given secret $(m(p)\text{-}1)$ - bits stream $d = d_{m(p)}d_{m(p)-1}..d_2$.

Step 1) If S_2 is mono-value, keep F intact and exit, (we can try to hide d in the next block F' and so on...) otherwise, compute $T = F \oplus K$ and go to the next step.

Step 2) Compute $u = [W.\ T]$ and present u as an $m(p)$ - bit stream $u = u_{m(p)}u_{m(p)-1..}u_2u_1$.

Step 3) Try to embed d' into F by CPTE1 scheme by changing at most one pixel $F_{i,j}$ in S_1 satisfying the extra condition $d(F_{i,j})^2 \leq 2$ and at most one $F_{k,l}$ in S_2 satisfying the extra condition $d(F_{k,l})^2 \leq 2$, in case it is possible. Otherwise, go to next step.

Step 3) Try to mark F as a failure.

One of the two following cases can be happen:

+) $u_1 = 1$: this is the fact we need: F is kept intact.

+) $u_1 = 0$: try to invert at most one entry of S_2 so that u becomes odd. This can be done by setting $H = \{F_{ij}$ in $S_2 | W_{ij}$ is odd, $d(F)_{ij}^2 \leq 2\}$. It implies that S_2 is not mono-value (if S_2 is mono-value, then u is odd by the skip condition 2, a contradiction), therefore H is not empty by the condition 1 (see Remark 3.a and Lemma 1 below). Hence we can randomly changing arbitrary F_{ij} in H so that u is changeded to the odd value $u' = u \oplus W_{ij}$, as we need: F is marked as a failure, then go to the next step.

Step 8) End.

4.3 Algorithm Extracting Secret Data

Step 1) Compute $T = F \oplus K$.

Step 2) Compute $u = [W.\ T]$,

- if u is odd: return conclusion "F is a failed block to hide secret data" and exit.
- if u is even: (u has the form $u = u_{m(p)}u_{m(p)-1..}u_20$) go to next step 3.

Step 3) Return the secret $(m(p)-1)$ - bit stream $d = u_{m(p)}u_{m(p)-1..}u_2$.

The correctness of the algorithms **4.2**, **4.3** is deduced from following theorem whose proof is similar to the proof of Theorem 2 with the help of lemmata 1,2 below.

Theorem 3. The algorithm **4.2** always stops after taking a finite steps and returns the block F so that, by running the algorithm **4.3** on F we can verify whether F is a failure or success. In case F is a success we can extract back exactly secret data embedded in.

The proof of this theorem in it's turn is completed by using the obvious lemmata:

Lemma 1. Suppose the $m \times n$ weight matrix W satisfies:

$$\{W_{ij}: i=1,..,m, j=1,..,n\} = M-\{0\};\ \{W_{ij}: F_{ij} \in S_1\} = M_1;\ \{W_{ij}: F_{ij} \in S_2\} = M_2$$

and the conditions 1 in **4.1**. If S_2 is not mono-value, the set

$H = \{F_{ij}$ in $S_2 | W_{ij}$ is odd, $d(F)_{ij}^2 \leq 2\}$ is not empty.

Lemma 2. Suppose the condition 2 in **4.1** is satisfied. If S_2 before or after changed by algorithm **4.2** is mono-value, then F is failed to hide the secret data $b = b_1b_2\ldots b_r$.

5 Experimental Results

We build a program to check CPT, CPTE1, CPTE2 schemes for images of BMP 1pp and 24 bpp formats. For binary images our experimental results show that by CPTE1,

CPTE2 schemes we gain a higher ratio of embedded bit than that of CPT and MCPT schemes, while the quality of embedded images is the same. A comparison between the numbers of secret bits embedded in a block F for each scheme: MSDR, CPT, CPTE1, MCPT, CPTE2 schemes respectively is given by following table.

Table 1. Comparison of MSDR with other schemes

Size of F in pixels	MSDR	Number secret bits by CPT	Number of secret bits by CPTE1	Number of secret bits by MCPT	Number of secret bits by CPTE2
5	4	2	3	1	2
6	4	2	4	1	3
12	6	3	6	2	4
30	8	4	8	3	7
46	10	5	9	4	8
63	10	6	10	5	9
64	11	6	10	5	9
94	12	6	11	5	10

6 Conclusion

1) Experimental results show that in general, the number of total embedded bits taken in CPTE1 scheme is much more lager than in CPT scheme and approximates to MSDR, a comparison of CPTE2 and MCPT gives the same conclusion. Taking account of secrecy of key and weight matrices and the size of these, some analysis of intractable attacks can be taken as the same ways in [1,4].

2) Other applications: CPTE1 scheme can easily be modified to palette images like GIF, 8bpp BMP image format…in case ones need to prevent from steganalysis, especially to histogram-based attacks (see for example analysis in [6]: if the alpha ratio of the number of changed pixels to the number of total pixels of a given palette image G is lower than 0.1, it is very difficult to guess if G contains hidden data or not). Applying CPTE1 scheme to each palette, ones can gain a small alpha ratio while the amount of total hidden bits is large enough for real applications. In cases each color in the palette has k "similar colors" for some k>2, using properties in **3.1**, **3.2** ones can hide more secret bits in each block F of the image. This problem will be developed in future work.

References

1. Chen, Y., Pan, H., Tseng, Y.: A secure of data hiding scheme for two-color images. In: IEEE Symposium on Computers and Communications (2000)
2. Wu, M.Y., Lee, J.H.: Anovel data embedding method for two-color fascimile images. In: Proceedings of International Symposium on Multimedia Information Processing, Chung-Li, Taiwan, R.O.C. (1998)

3. Mirsattari, N.S., Haghani, P., Jamzad, M.: Feature Watermarking in Digital Documents For Retrieval and Authentication. In: 11th International CSI Computer Conference, CSICC 2006, Iran (2006)
4. Tseng, Y.-C., Pan, H.-K.: Secure and Invisible Data Hiding in 2-Color Images. In: Proceedings of INFOCOM 2001, pp. 887–896 (2001)
5. Hioki, H.: A modified CPT scheme for embedding data into binary images. In: Proc. of Pacific Rim Workshop on Digital Steganography 2003, pp. 32–44 (July 2003)
6. Zhang, X., Wang, S.: Analysis of Parity Assignment Steganography in palette Images. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3683, pp. 1025–1031. Springer, Heidelberg (2005)

A Comparative Analysis of Managing XML Data in Relational Database

Kamsuriah Ahmad

Faculty of Information Science and Technology
National University of Malaysia
kam@ftsm.ukm.my

Abstract. The eXtensible Markup Language (XML) has recently emerged as a standard for data representation and interchange on the web. Based on its popularity used in most application, the critical issues are to store and to query XML data to exploit the full power of this technology. Since relational database is widely used technology for storing and querying, therefore replacing it with pure XML database is not a good choice and very expensive process. It is thus crucial to map XML data into relational data and this process is one that occurs frequently. Many existing methods exist in the literature, and defining what the best mapping method is explicitly important. The intention of this paper is to the existing mapping methods in terms of generating good relational schema. At the end a new mapping method is developed to overcome the limitations the limitations and shows that it is efficient in terms of removing relation redundancy.

Keywords: Mapping method, comparative analysis, XML, relational database.

1 Introduction

XML has emerged as a standard data representation and interchange on the web. However, relational database is often used to store the data based on its popularity and reliability. Some mechanism is needed to map XML data to relational data. Until today there is no fast, easy, automated, and most important, free solution existed on the mapping from XML to relational. Currently, there are several approaches for storing XML data, ranging from using files to full-fledged database management systems such as relational, object-relational, object-oriented or native XML database systems. The methods can be classified according to the following categories:

i.Files: This approach takes an entire XML document as a single object in a file and directly mapped and stored as one big text columns such as CLOB (Character Large Object), in much the same way as to store an image file. The advantages of storing XML as CLOB are: there is no preprocessing and the originality of XML document is preserved. However the disadvantage of this approach is that we cannot get help from the database management system: not able to search effectively on the document contents, or to do queries since the data has to be parsed, loaded into memory, processed and, for update operations, we need to dump back to disk. It might cause a

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 100–108, 2011.

large number of nested tables and redundant data in the relations. Therefore the processes are not efficient and impractical.

ii.Native Database System: In this approach a native XML database was developed to directly support XML data model and queried by special purpose engines. The advantages of these approaches are that XML data can be stored and retrieved in their original formats and no additional mappings or translations are needed. Furthermore most native XML databases have the ability to optimize the query techniques. The disadvantages of this approach are that, due to the document-centric nature of XML database, complex searches or aggregation might be cumbersome. On the other hand since XML may contain lots of redundant data then the update operations are not efficient. As reported, the current state of native XML database is still unsettled and does not have any firm standard on its structure [19], this makes it unsuitable to manage huge XML data.

iii.RDBMS: In this approach, XML data is mapped into rows and columns of a relational table, which is known by the term shredding. The queries posed in semi-structured query languages are translated into SQL queries. The results of these queries are later translated back to XML, where all the processes are done internally. Therefore, the processes of mapping XML to relational tables occur frequently. Database vendors such as IBM, Microsoft, Oracle and Sybase are currently building tools to assist in mapping XML documents into relational tables. These vendors are competing against one another and a standard for the XML data type and the methods of mapping is yet to be defined [3].

Different approaches, either files, native database, or DBMS have their own strength and weaknesses [3]. It is still unclear which of these three approaches is going to find a wide-spread acceptance. In theory, native database systems should work best, but it is going to take a long time before such systems are mature and scale well for large amounts of data. Furthermore it is argued that XML can be effectively used as a database language. This is because XML language is best in supporting other applications, such as user-defined tagging documents, cross-referencing between documents and in fact it just stands as a database that store trees. Thus, XML will never be an ideal database language [15]. Therefore a more efficient database language for XML is needed, and relational database language is the best alternatives. Relational database systems are mature and scale very well, and they have the additional features that XML data can co-exist, making it possible to build applications with little extra effort. To optimize the use of relational database systems for managing XML data, recent work has concentrated on models and methods to map from XML to relational. It is believed that the effort on finding the best mapping algorithm will still continue in the future. Currently many existing methods exist in the literature on the mapping process. This gives rise to the following problem: Are some mappings ‘better’ than the others? To approach this problem, the classical relational database design through normalization technique is referred. The technique is based on the functional dependency that exists in the semantics of the data. Therefore in this study we made an approach that the mapping method is considered to be in good criteria if it can produce relational schema without relation redundancies that may lead to anomaly problems. This paper will study and make comparison of the existence approaches and discuss why the existence approaches are

still insufficient for reducing relation redundancy and failed to achieve the optimal relational schema for XML. This paper is organized as follows: section 2 will discuss different ways to map XML to relations, section 3 will discuss the proposed mapping method (XtoR), section 4 provides a motivating example and compares the relational schema generated by XtoR and the existing methods. At the end the conclusion and future enhancement are presented.

2 Managing XML Data in Relational Database

The mapping from XML to relational is not an easy task to accomplish because the data model of an XML document is fundamentally different from that of a relational database. Especially, the structure of an XML document is in hierarchy, and the XML elements may be nested and repeated, while relational model is a flat representation of data with tables and columns. During the data exchange, XML might come with or without a schema (DTD or XML Schema). The existence or the absence of a schema greatly influences the mapping procedure. When the schema of XML data is not available, a generic mapping is used. XML document can be seen as a tree model and the mapping is based on the relationship between the nodes and edges of XML model. But when a schema is available structural constraint information of an XML document from a schema is used to guide the mapping design. Recently, studies in the context of integrity constraint for XML paying particular attention to the class of keys and functional dependencies [15] as renewed interest to adopt these constraints in the mapping framework. It is believed that if the mapping considers the presence of semantic constraints then the relational schema generated will be good. Therefore, the mapping approach can be discussed into three different categories: (i) model-based approach, (i) structural-based approach, and (iii) semantic-based approach.

2.1 Model-Based Approach

This mapping is based on path expression in the XML tree in the absence of schema type. Basically this approach will traverse the tree and store the path for every node visited in a table. Even though the main idea is the same but various strategies have been proposed that improved from the previous one. Among the approaches that fall under this category are Edge [4], XRel [18] and XPev [14]. However these three approaches are in the absence of schema type, therefore relational table is used to store the path information. The information is based on the structure of XML tree and ignores totally the semantics aspect of XML. The problem with relational schema generated by these methods is that the information is split into small pieces that may end up increasing the storage size of the database. In fact most of the targetIDs are zeroes and lots of data duplication in ordinal column as in Edge approach. Furthermore, in order to process query faster, an index (tag, data) is mandatory and more tables need to be joint during the query processing. Since the methods do not consider the semantics constraints, the criteria of generating good relational schema that reduced redundancy is not achieved.

2.2 Structural-Based Approach

This approach is based on the existence of type definition such as XML DTD or XML Schema, which conforms to XML document. By analyzing the structural properties of the schema, it then automatically converts a DTD or XML Schema into relational schemas. DTDs may contain arbitrary regular expressions, such as recursive, disjunction and set value, which need further analysis. The approaches that can be classified under these categories are Inlining [16], LegoDB [9] and CPI [10]. However the semantic keys are not used efficiently where system generated ID, parentID and parentCODE are used widely. Studies as shown that [11] these IDs will generate data and relation redundancy with respect to the constraints exists in XML documents. This approach do not take into account the information about semantic dependencies, therefore it is not efficient in reducing the redundancy of data that may exist in XML.

2.3 Semantics-Based Approach

Since studies on integrity constraints for of XML data started to emerge, there have been efforts considering capturing semantics of XML for the mapping. These semantics constraints are provided along with XML documents and become an input to the system. The constraints information guides the design of relational schema during the mapping processed. The constraints used are keys, foreign keys, and functional dependencies both in the absence or presence of the schema. As been proved in relational database, when semantics are given, redundancy can be reduced with respect to the constraints through normalization process. The mapping under this category can be divided into two: key based approach (X2R [6], Davidson [7], Liu [12]) and functional dependency based approach (Lv&Yan [13], RRXS [3]).

Table 1 shows a comparison for the mapping approaches. The mapping methods under semantic-based approach used semantic constraints that come with the XML documents. X2R mapping method used the information in key and foreign key constraints and used these constraints to guide the schema design. Even though this method able to generate a reduce redundancy relational schema with respect to the semantics in keys and keyrefs but the general concept of functional dependencies (that able to express dependency constraints among the attributes) is ignored, hence they failed to produce good relational schema for XML with respect to the constraints. Furthermore, the study of XFD implication is avoided; therefore it cannot infer other constraints that may exist, given the existing constraints. As the result, other semantics relationships of XML document cannot be identified. System generated ID is used as an ID to the relation and parentID is used to express the relationship between parent and child. Hence this method will generate relational table with data and relation redundancy. Study has shown that if the system generated ID is used when designing the relational schema instead of the given value-based keys, it will decrease the query performance [12]. Davidson's method performs the mapping in the presence of key constraints. The relational schema is mapped from a set of minimum covers propagated from key constraints. But again this approach does not consider functional dependencies therefore it cannot remove redundancies that may exist in XML. Hence this redundancy cannot be captured and it cannot produce a

good relational schema. In relational, functional dependencies constraint is useful for generating optimal schema decomposition for XML thru its normalization step. Ironically, this constraint is the most neglected aspect by many researchers as the approach in model-based and structural-based. The model-based approach as in Edge, XRel and XPev claims to be efficient in processing the queries in the absence of schema definition but this approach ignores totally the semantic aspects in XML. Therefore this approach will produce redundancies in the schema, hence the criteria of producing an optimal relational schema for XML is not achieved.

Table 1. Comparison of the existing XML-Relational Mapping Techniques

	Schema	Model-based	Structural-based	Semantics-based
Edge	No	Path-based	No	No
XRel	No	Path-based	No	No
XPev	No	Path-based	No	No
Stored	Yes, DTD	No	Yes	No
Inlining	Yes, DTD	No	Yes	No
LegoDB	Yes, DTD	No	Yes	No
CPI	Yes, DTD	No	Yes	No
XtoR	Yes, XML-Schema	No	Yes	Yes
Davidson	No	No	No	Yes
Liu	Yes, DTD	No	No	Yes

The structural-based approach as in Inlining, LegoDB and CPI claims to be efficient in processing the queries. In the presence of type definition such as DTD or XML Schema, they need to deal with these type definitions that may contain arbitrary regular expressions, such as recursive, disjunction, set value and also to deals with null values and incomplete relations. Two evaluations studies [9], [21] on alternative storage strategies indicate that the shared-inlining algorithm [20] outperforms other strategies in data representation and performance across different datasets and different queries, when DTDs are available. The studies also indicate that the presence of DTD during the mapping is vital to achieve good performance and compact data representation of XML in relational settings across different datasets and different queries. This is one of the reasons DTD is included in our studies, and the XFDs is defined with respect to this schema. But most of the mapping approaches ignore the existence of semantics as expressed in functional dependencies, therefore the resulted relational schema may contain redundancy in the relations. It would be helpful if tools exist to facilitate the general problem of mapping between different data formats, taking the semantics of data into account. To facilitate such mapping we need a language in which to express transformations and constraints, and the ability to reason about the correctness of the transformations with respect to the constraints.

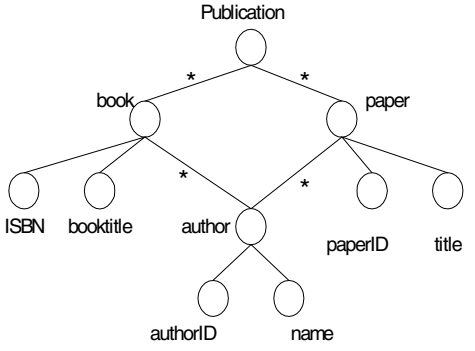
3 X2R: A New Method for Mapping XML to Relations

In this study, a transformation language that is able to extract the semantics information in XML and preserve it during the transformation is developed. However, we have to deal with DTDs that may contain arbitrary regular expressions, and also we have to deals with null values and incomplete relations. Functional dependency is used to define that X determine Y or X->Y. However functional dependency for XML (XFD) is more complicated since we need to deal with the hierarchical structure of XML and the path expression that can be used to express XML. Till now there is no standard definition for XFD, therefore a lot of attempts done to define ones. Studies had shown that different definition of XFD will have different expressive power [1]. The XFD that we adopt is an expression of the form: (C, Q : X -> Y), where C is the downward context path which is defined by an XPath expression from the root of the XML document, Q is a target path, X is an LHS (Left-Hand-Side) and Y is an RHS (Right-Hand-Side). Functional dependencies are used to specify constraints in the XML and use this constraint to infer a non-redundant relational schema for XML. The strategy adopted in this study, is to produce a relational design, which preserves structural and semantic constraints of the XML data while reduced redundancies. First, the structural of XML data is captured by the DTD and generate the DTD schema, which is the formal description of XML. Using the constraint-preserving algorithm, the redundant path is removed. Finally, by mapping paths in XFDs to relational attributes, a set of relational functional dependencies and a relational storage for the XML data is produced, which preserves the content and the structure information of the original XML document. The generated relational schema is able to remove redundancy as indicated by the XFDs, and enforced efficiently using relational primary key constraints.

4 Motivating Example

Publication dataset [16] as in Fig. 1 is used to illustrate the effectiveness of the proposed method. This dataset describes the publication that has many books and papers. Each books and papers contains information about the authors who published either books or papers. To evaluate the effectiveness of the proposed method (X2R), an experiment is conducted where the relational schema generated by X2R, RRXS, Lv&Yan methods is compared. RRXS and Lv&Yan methods which are under semantic-based approach are used in the comparison because they consider XFD in the mapping. Given the DTD graph, its schema and corresponding XFD that may exist in the Publication dataset, the relational schema generated by the three methods shown as in Fig.2.

From the generated schema, X2R method generates a good relational schema for Publication dataset when compares with the schema generates by RRXS and Lv&Yan method. The relational schema generates by RRXS in Fig. 2(ii) will produce two equivalent tables that belong to the same concept which are Author and Author1 table. The reason for this redundant creation is that the algorithm did not check the existence of already created table for the same concept of object (Author),



<!ELEMENT publication(book*, paper*)>
<!ELEMENT book (ISBN, booktitle, author*)>
<!ELEMENT paper(paperID, title, author*)>
<!ELEMENT author (name)
<!ATTLIST author authorID

XFD1 //book: ISBN -> ε
XFD2 //paper: paperID -> ε
XFD3 //author: authorID -> name

Fig. 1. DTD graph, its schema and corresponding XFD

Book (ISBN, booktitle)
Author (authorID, name)
Paper(paperID, title)
Book-author(ISBN, aID)
Paper-author(paperID, authorID)

i.Schema generated by X2R

Book (ISBN, booktitle)
Author (authorID, name)
Author1(ID, authorID, name)
Paper (paperID, title)
Book-author(ISBN, aID)
paper-author1(ID, paperID, author1.ID)

ii.Schema generated by RRXS

Publication(ID)
Book(ID, ISBN, booktitle, author.ID)
Paper(ID, paperID, title, author.ID)
Author(ID, authorID, name)
Publication-book (publication.ID, book.ID)
Publication-paper(publication.ID, paper.ID)
Book-author(book.ID, author.ID)
Paper-author(paper.ID, author.ID)

F_1(publication.book.ISBN, publication.book.ID)
F_2(publication.paper.paperID, publication.paper.ID)
F_3(publication.paper.author.authorID, publication.paper.author.name)

iii.Schema generated by Lv&Yan

Fig. 2. Relational Schema generated by X2R, RRXS and Lv&Yan

it blindly created a new one. These redundant tables may lead to the update anomaly problems. The method by Lv&Yan will create two types of tables: based on structural DTD and based on semantics presented by XFD. Basically three steps involved in this algorithm: i) using structural DTD, a separate relation will be created for each non-leaf vertex and leaf, ii) for each parent-child relation between two vertexes connected by a * operator, a separate relation is created, and ii) for each XFD defined over DTD, a separate relation is created. Based on this algorithm the resulted relational schema is shown as in Fig. 2(iii). As observed, the F_3 table and the Author table are redundant. These tables are used to describe the same concept of object (Author), therefore this will lead to update anomaly problems. For

instance if a new author is added to the publication, both the relations F_3 and author need to be updated for the database to satisfy the constraints. The relational schema generated by X2R overcomes the limitation of these methods by producing a good relational schema design for XML with no redundant relations. In fact the schema produced is correct with respect to keys and functional dependencies.

5 Conclusion

We have investigated the problem of how to design a good relational schema for XML data with no redundant relation. A new method has been developed which given functional dependencies and DTD, redundancy in XML document can be detected and used this information for mapping to relations which can reduce relation redundancy and at the same time preserve the constraints as expressed in functional dependencies. This method can be efficiently operated, automated and eliminates unnecessary ID. As an immediate task, we would like to find efficient algorithm for mapping from relations to XML that based on functional dependencies which may appear in XML. Through this study it is hope that it will give contributions to the database community.

References

1. Ahmad, K., Mamat, A., Ibrahim, H., Mohd Noah, S.A.: Defining functional dependency for XML. Journal for Information Systems Research and Practices (2008)
2. Amer-Yahia, S.A., Du, F., Freire, J.: A Comprehensive Solution to the XML to Relational Mapping Problem. In: Proceedings of the 6th Annual ACM International Workshop on Web Information and Data Management, pp. 31–38 (2004)
3. Atay, M., Chebotko, A., Liu, D., Lu, S., Fotouhi, F.: Efficient Schema-based XML-to-Relational Data Mapping. Journal of Information System 32, 458–476 (2007)
4. Bohannon, P., Freire, J., Roy, P., Simeon, J.: From XML Schema to Relations: A Cost-Based Approach to XML Storage. In: Proceedings of the 18th International Conference on Data Engineering, pp. 64–74 (2002)
5. Chen, Y., Davidson, S., Hara, C., Zheng, Y.: RRXS: Redundancy Reducing XML Storage in Relations. In: Proceedings of 29th International Conference on Very Large Data Base, pp. 189–200 (2003)
6. Chen, Y., Davidson, S., Zheng, Y.: Constraint Preserving XML Storage in Relations. In: Proceeding of the 9th International Conference of Database Theory, pp. 7–12 (2002)
7. Davidson, S., Fan, W., Hara, C., Qin, J.: Propagating XML Constraints to Relations. In: Proceedings of the 19th International Conference on Data Engineering, pp. 543–554 (2003)
8. Fan, W.: XML Constraints: Specification, Analysis, and Application. In: Proceedings of the 16th International Workshop on Database and Expert Systems Applications, pp. 805–809 (2005)
9. Florescu, D., Kossman, D.: A Performance Evaluation of Alternative Mapping Schemes for Storing XML Data in A Relational Database. In: Proceedings of the VLDB (1999)
10. Lee, D., Chu, W.W.: CPI- Constraint-Preserving Inlining Algorithms For Mapping XML DTD to Relational Schema. Journal of Data Knowledge and Engeering 39(1), 3–25 (2001)

11. Lee, Q., Bressan, S.,Rahayu, W.: XShreX:Maintaining Integrity Constraints in the Mapping of XML Schema to Relational. Proceedings of the 17th International Conference on Database and Expert Systems Application, pp.492-496 (2006)
12. Liu, C., Vincent, M., Liu, J.: Constraint Preserving Transformation from Relational Schema to XML Schema. Journal of World Wide Web 9(1), 93–110 (2006)
13. Lv, T., Yan, P.: Mapping DTDs to Relational Schemas with Semantic Constraints. Journal of Information and Software Technology 48(4), 245–252 (2006)
14. Qin, J., Zhao, S., Yang, S., Dou, W.: XPEV: A Storage Model for Well-Formed XML Documents. In: Wang, L., Jin, Y. (eds.) FSKD 2005. LNCS (LNAI), vol. 3613, pp. 360–369. Springer, Heidelberg (2005)
15. Schewe, K.: Redundancy, Dependencies and Normal Forms for XML Databases. In: Sixteenth Australasian Database Conference, vol. 39 (2005)
16. Shanmugasundaram, J.: Relational Databases for Querying XML Documents: Limitations and Opportunities. In: Proceedings of the 25th VLDB Conference, pp. 302–314 (1999)
17. Tian, F., DeWitt, J., Chen, J., Zhang, C.: The Design and Performance Evaluation of Alternative XML Storage Strategies. SIGMOD Record 31(1), 5–10 (2002)
18. Yoshikawa, M., Amagasa, T., Shimura, T.: XRel: A Path-based Approach To Storage and Retrieval of XML Documents Using Relational Database. ACM Transactions on Internet Technology 1, 110–141 (2001)
19. Zhanga, S., Gana, J., Xua, J., Lva, G.: Study On Native XML Database Based GML Storage Model. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Beijing, vol. XXXVII (2008)

B^{ob}-Tree: An Efficient B^{+}-Tree Based Index Structure for Geographic-Aware Obfuscation

Quoc Cuong To[1], Tran Khanh Dang[1], and Josef Küng[2]

[1] Faculty of Computer Science & Engineering, HCM University of Technology, Vietnam
{qcuong, khanh}@cse.hcmut.edu.vn
[2] FAW Institute, Johannes Kepler University Linz, Austria/Europe
josef.kueng@faw.jku.at

Abstract. The privacy protection of personal location information increasingly gains special attention in the field of location-based services, and obfuscation is the most popular technique aiming at protecting this sensitive information. However, all of the conventional obfuscation techniques are geometry-based and separated from the database level. Thus, the query processing has two time-consuming phases due to the number of disk accesses required to retrieve the user's exact location, and the location obfuscation. Also, since these techniques are geometry-based, they cannot assure location privacy when the adversary has knowledge about the geography of the obfuscated region. We address these problems by proposing B^{ob}-tree, an index structure that is based on B^{dual}-tree and contains geographic-aware information on its nodes. Experiments show that B^{ob}-tree provides a significant improvement over the algorithm separated from the database level for query processing time and location privacy protection.

Keywords: LBS, obfuscation, privacy-preserving, spatio-temporal indexing.

1 Introduction

With the rapid development of mobile technologies, there are more than 4.5 billion mobile users by the year 2009 and the number is expected to increase more. Among various services for mobile phone, the location-based service (LBS) is the most promising one since it supplies users with many value-added services. In order to benefit from these services, users, however, have to reveal their sensitive information such as their current location. Such novel services pose many challenges because users are not willing to reveal their sensitive information but still want to benefit from these useful services. We consider location privacy as an enabling technology for the proliferation of LBS, and so must balance the privacy and service quality.

To solve this privacy-preserving problem, many techniques have been suggested and the most popular one is obfuscation [1,2,3,4]. The general idea of this technique is to degrade the quality of user's location information but still allow them to use services with acceptable quality. However, this technique has two major limitations. First, all of the proposed obfuscation algorithms are geometry-based. In other words, they do not consider the geographic feature constituting the obfuscated region (e.g., a lake in an obfuscated area). Based on the knowledge of the region geography, an

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 109–118, 2011.

adversary can increase the inference probability of a user's exact location. Second, these algorithms are separated from the database level, making the algorithms go through two time-consuming phases due to the number of disk access required to (1) retrieve user's exact location on the database level, and (2) obfuscate this information on the algorithm level. Also, it is prone to privacy violation and more deployment complexity because both phases, together with the communication channel between them, must be trusted.

Motivated by these reasons, in this work, we create a new geographic-aware obfuscation technique and propose B^{ob}-tree, a new spatio-temporal index structure based on B^{dual}-tree [8]. By taking into account the geographic feature inside the obfuscated region, our new technique ensures a higher privacy protection degree than that of the geometry-based obfuscation techniques in [1,2,3,4]. Furthermore, because B^{ob}-tree embeds geographic-aware region information on its nodes, the process of calculating the obfuscated region can be done in only one phase: traversing the index structure to retrieve the appropriate *obfuscated* region that contains a user's exact location. This one-phase process can reduce the processing time considerably comparing to the two-phase process mentioned above.

The rest of this paper is organized as follows. Section 2 briefly reviews related work. Section 3 presents the new geographic-aware obfuscation technique. Section 4 introduces B^{ob}-tree. Section 5 gives privacy and performance analyses. Section 6 presents experimental results, and section 7 gives concluding remarks.

2 Related Work

2.1 Location Obfuscation

Among the most popular techniques to protect user's location privacy, obfuscation based techniques have gained much interest due to its intuition and implementation simplicity. Location obfuscation aims at hiding user's exact location by decreasing the quality of user's location information. In [1], Ardagna et al. propose obfuscation techniques by enlarging the area containing user's real location. However, these techniques just deal with geometry of the obfuscated region, not concerning about what is included inside (i.e., the geographic feature). Of late, the semantic-aware obfuscation technique introduced in [13,14] considers sensitive feature types inside an obfuscated region. But, this technique does not concern about how big the area of the obfuscated region is. It focuses only on the probability that a user is located in a sensitive place. In various LBS, however, an indispensable requirement is that the area of any obfuscated region must be big enough to protect user's location privacy.

With obfuscation techniques, the bigger the area, the harder an attacker can infer the user's exact location. If the area, however, is too big, it can affect the quality of location-based services. So, it is the responsibility of users to decide what location accuracy degree to be revealed to which service providers. Inspired by this, in [9], Dang et al. introduce an architecture to classify the service providers depending on the user's trust. This architecture inherits the property of mandatory access control to label each service provider so that users only reveal their locations on an appropriate level based on the labels assigned to the service providers. Similar to this idea, our

proposed approach classifies service providers in the way that the more reliable the service providers, the smaller area of the obfuscated region they can obtain.

2.2 Spatio-temporal Structures for Indexing Moving Objects

A number of recent researches focus on indexing the present and future positions of moving objects [5], and the two most dominant popular methods are *parametric spatial access* and *space-filling-curve transformation*. With the former, the main idea is that the bounding rectangle is a temporal function, and thus can enclose moving objects. The most popular access method in this category, TPR-tree [6], inherits the idea of parametric bounding rectangles in R-tree [15] to create time-parameterized bounding rectangles (TPBR). However, the TPBR bear two crucial limitations that dramatically affect the performance of TPR-tree: overlapping and high storage cost. The latter overcomes these two limitations by using the space filling curves (e.g., Peano/z-order, Hilbert) to transform object locations from multi-dimension to one-dimension space. Then, these one-dimensional values are indexed by a B^+-tree, which is the typical one-dimensional index. Two most popular access methods in this category is B^x-tree [7] and B^{dual}-tree [8]. The B^x-tree outperforms the TPR-tree by factors of as much as 10 but it fails to consider object velocity, and thus the query processing with B^x-tree retrieves a large number of false hits, which seriously affects its performance. B^{dual}-tree overcomes this limitation by capturing also the velocity information. By using the partitioning grid that divides the data space into cells, B^{dual}-tree can effectively answer progressive spatio-temporal queries which are poorly supported by B^x-tree.

Despite the existence of several indexing techniques for present and future positions, to the best of our knowledge, no moving-object index has yet been reported in the literature that achieves the goal of obfuscating the geographic-aware region.

2.3 Access Methods for Privacy-Preserving

All of existing privacy-preserving algorithms are separated from the database level [1,2,3,4,13]. This separation, as mentioned above, makes the two-phase query processing time-consuming. Motivated by this, Atluri et al. [10] create S^{STP}-tree, a unified index structure that embeds users' profile vectors directly into its nodes, to support profile conditions. The limitation of this access method is that it only allows or denies the access request of subjects, but does not concern about obfuscating the user's location. In other words, the access request evaluation has only two levels of result: reject or accept. Our proposed index structure, however, has multi-level form of result as evaluating an access request, based on the user's trust in service providers.

Very recently, the OST-tree [11] embeds the user's privacy policy into its nodes and obfuscates spatio-temporal data. But, since OST-tree is based on TPR-tree and concerns only with geometry-based obfuscation, it has high storage cost and quite low privacy protection.

It is evident from the above discussions that currently there does not exist any spatio-temporal index structure that can effectively handle geographic-aware obfuscation. Again, all of them are based on TPR-tree which is much less efficient than B^{dual}-tree in terms of storage cost and query processing time [8]. Towards this

goal, in this paper, we propose the B^{ob}-tree, a structure originally based on B^{dual}-tree, but with essential modifications to support geographic-aware obfuscation.

3 Geographic-Aware Obfuscation

As discussed above, although there exist a variety of research activities in spatial obfuscation, none of the proposed techniques concern with the geographic features. This can leave a backdoor to privacy open as the adversary has the geography knowledge of the obfuscated region. To address this problem, in this section, we present a new geographic-aware obfuscation technique that takes into account both the area of and the geographic feature inside the obfuscated region. This newly proposed technique not only ensures the same quality of service as others as in [1,2,3, 4] (because the obfuscated regions produced by these techniques have the same area), but also has better user's location privacy protection (as proved in section 5.2).

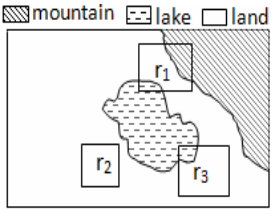


Fig. 1. Example of unapproachable region

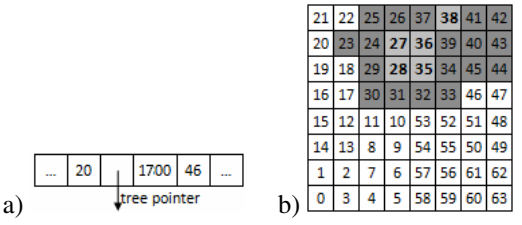


Fig. 2. A part of internal node of B^{ob}-tree and its projection on coordinate space

In our proposed technique, the region is divided into two geographic features: *approachable* and *unapproachable* parts. The unapproachable parts represent places where users, because of some reasons, cannot enter. In contrast, users can enter approachable parts. For example, the lake and mountain are the unapproachable parts of the region in Fig. 1 because no boats are allowed on the lake and user cannot climb the mountain. Our proposed obfuscation technique returns to service providers the obfuscated regions that not only have the same area as that of geometry-based techniques, but also contain only approachable features inside it.

Adversary model: Using the external knowledge of the geographic feature inside the obfuscated region, the adversary tries to eliminate the unapproachable parts of the obfuscated region, leaving only the approachable parts. In this way, the area of the original obfuscated region created by the previously proposed algorithms, e.g., as in [1,2,3,4], will be reduced since the returned region, in this case, includes both the approachable and unapproachable parts. As a result, the probability that an adversary, with his external knowledge, can infer the user's exact location within the obfuscated region is higher. The adversary, however, cannot reduce the area of the region created by our newly proposed geographic-aware obfuscation technique because this returned region includes only the approachable parts. Thus, for the two techniques, although the areas of the two regions are the same, the region created by our proposed

technique achieves better location privacy protection. For example, in Fig. 1, since the region r_1 contains two unapproachable parts (a mountain and a lake), the adversary can reduce r_1 to a smaller region by eliminating the intersection of r_1 with the lake and mountain. The region r_2, however, does not intersect with the lake or mountain, and so it is impossible for the adversary to reduce this region.

4 Index Structure

The base structure of the B^{ob}-tree is originated from that of the B^+-tree which indexes the one-dimensional values. Similar to B^{dual}-tree [8], a d-dimensional moving point o in our index structure with a reference timestamp $o.t_{ref}$, d coordinates o[1],…,o[d], and d velocities o.v[1],…,o.v[d] has its dual in the 2d-dimensional vector as follows:

o^{dual}=(o[1](T_{ref}),…,o[d](T_{ref}),o.v[1],...,o.v[d]), where o[i](T_{ref}) is the i-th coordinate of o at time T_{ref} and is given by: o[i](T_{ref})=o[i] + o.v[i]*(T_{ref}-$o.t_{ref}$)

This 2d-dimensional point in a dual space is mapped to an one-dimensional value using Hilbert curve, and then this value is indexed by B^+-tree. However, in order to specify the geographic-aware region, the node structure is modified to attach this information. Specifically, beside the one-dimensional Hilbert value transformed from the corresponding multi-dimensional point, each internal node contains the area of the approachable regions corresponding to the Hilbert range of the node.

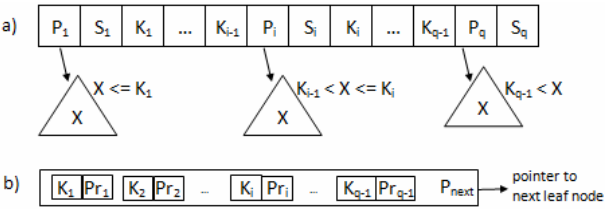


Fig. 3. B^{ob}-tree

Fig. 3 illustrates the structure of the B^{ob}-tree. Each internal node is of the form <P_1,S_1,K_1,P_2,S_2,K_2,…P_{q-1},S_{q-1},K_{q-1},P_q,S_q> where P_i is the tree pointer, S_i is the area of the approachable regions associated with a Hilbert interval [K_{i-1},K_i], where K_i is the search key value. Each leaf node is of the form <<K_1,Pr_1>, <K_2,Pr_2>, …<K_{q-1},Pr_{q-1}>, P_{next}> where Pr_i is a data pointer, and P_{next} points to the next leaf node of the B^{ob}-tree.

In B^{dual}-tree, each internal node e is implicitly accompanied by an interval [$e.h^l$, $e.h^u$], where $e.h^l$ and $e.h^u$ are Hilbert values of the starting and ending cells of the region represented by node e, and thus e is associated with $e.h^u$-$e.h^l$ cells. The area of a region associated with each internal node is then calculated by multiplying the total number of cells of each internal node with the area of the projection of each cell into the coordinate space. However, the region associated with $e.h^u$-$e.h^l$ cells includes both approachable and unapproachable regions. Thus, in order to increase the privacy degree of the region, we must filter out all unapproachable portions in this region.

Assume that the projection result of a region associated with an internal node e (associated with $e.h^u$-$e.h^l$ cells in the 2d-dimensional space) into the coordinate

space is the region consisting of $e_1.h^u$-$e_1.h^l$ cells (in the 1d-dimensional space), and there are x unapproachable cells within this 1d-dimensional region. Obviously, the number of approachable cells associated with e is $e_1.h^u$-$e_1.h^l$-x. Thus, the area of approachable regions associated with e is ($e_1.h^u$-$e_1.h^l$-x)S_c, where S_c is the area of each cell in the 1d-dimensional space. For example, Fig. 2a shows an internal node and its associated Hilbert value transformed from the 4-dimensional space. Fig. 2b is the projection of this node into the 2-dimensional coordinate space. The five gray cells 27, 28, 35, 36, and 38 are unapproachable. Assume that area of each cell is $100m^2$, the area of approachable regions associated with this internal node is (45-23-5)x100=$1700m^2$, where two values 23 and 45 are the Hilbert values of projection into the 2-dimensional space of the two cells 20, 46 in the 4-dimensional space.

The authorization α used in our approach is a 3-tuple $<id_{sp}, id_{user}, \Delta s>$ where id_{sp} is the identity of the service provider, id_{user} is the user's identity, and Δs is the area of the approachable region. The meaning of an authorization is that a user id_{user} only allows the service provider id_{sp} to access his/her sensitive personal location information with an accuracy degree of Δs. For example, the user #U232 is willing to reveal his position in an approachable region, with the area of $600m^2$, to the advertising service #S101. This authorization can be expressed as α_1 = <#S101, #U232, $600m^2$>. If the user's exact position is at coordinate $<x_0, y_0>$, the result returned to the service provider is an approachable region of $600m^2$, containing the coordinate $<x_0, y_0>$.

The area of an obfuscated region associated with each node in a B^{ob}-tree is hierarchical because the interval [$e_1.h^l$,$e_1.h^u$] is smaller when traversing from the root to the leaf nodes. Therefore, when traversing from the root down in a B^{ob}-tree, the accuracy of user's position increases because the area of the obfuscated region is smaller and vice versa. Based on this basic property, if a service provider that has a low trust level from a user wants to retrieve the user's location, the search process can stop at some internal nodes that may be close to the root.

Search, Insertion, Deletion and Update with B^{ob}-tree. The following algorithm outlines the procedure to search for a record in B^{ob}-tree.

```
Algorithm Search
Input: a dual vector of a moving point o^dual, area of an
obfuscated region S.
Output: the region that its area equals S and contains
a moving point with a dual vector o^dual.
Transform o^dual into the Hilbert value h
while (n is not a leaf node) do
    Search node n for an entry i such that K_{i-1} < h ≤ K_i
    if S = S_i then
        return the region corresponding to the Hilbert
        interval [K_{i-1};K_i]
    else if S > S_i then
        return ExtendCell(K_{i-1},K_i,S)
    else
        n ← n.P_i //the i-th tree pointer in node n
Search leaf node n for an entry (K_i,Pr_i) with h=K_i
if found then retrieve the user's exact location
else the search value h is not in the database
```

The algorithm *ExtendCell(K_i,K_j,S)* extends the region corresponding to the Hilbert interval [K_i,K_j] by adding more *approachable* cells until the area of the extended region equals S. This ensures that the obfuscated region produced by our technique has the same area as that of the geometry-based techniques and achieves better location privacy protection because it contains only approachable features.

In this search algorithm, if the area S is big (e.g., the service provider gets a low trust level from the user, and thus can only obtain a big region containing the user's exact location), the search process can stop at some internal node near the root. In this case, the disk access number is reduced significantly. So, we do not have to traverse to the leaf node to find the user's exact position as in the previously proposed algorithms, which are separated from the database level. With our approach, only when users are willing to reveal their exact location to service providers, the search process must traverse to leaf nodes.

The insert, delete and update operations of B^{ob}-tree are similar to those of B^+-tree. However, since these operations change the key value in each node, the area of the obfuscated region associated with each node needs to be re-calculated.

5 Privacy and Performance Analyses

5.1 Privacy Analysis

For obfuscation techniques, the *relevance* [1] is used to measure the location privacy protection. The lower the relevance, the higher the location privacy protection is, and thus the lower the probability an adversary can infer the user's real location.

$$R_s = \frac{(A_i \cap A_f)^2}{A_i . A_f} \tag{1}$$

where A_i is location measurement [1] and depends completely on the used positioning technology, and A_f is the obfuscated region area. With the algorithms separated from the database level, because A_f includes both the accessible and inaccessible regions, an adversary can eliminate the inaccessible part of A_f (cf. the adversary model in section 3). We call A_{fa} the accessible region of A_f after eliminating the inaccessible (e.g $A_{fa} \leq A_f$). The relevance of A_{fa} is calculated as follows:

$$R_{sa} = \frac{(A_i \cap A_{fa})^2}{A_i . A_{fa}} \tag{2}$$

In B^{ob}-tree, since A_f includes only the accessible region, an adversary cannot reduce A_f to a smaller region. Thus, the relevance of our proposed approach is still R_s. Since $A_{fa} \leq A_f$, from (1) and (2) we have $R_{sa} \geq R_s$. This means that the location privacy protection of our proposed approach is better than that of the introduced algorithms. More specifically, by considering the geographic feature inside the obfuscated region, we reduce the probability that an adversary can infer the user's exact location.

Similar to TPR-tree, the adversary can eliminate the inaccessible parts of A_f in relevance of OST-tree. However, the temporal obfuscation [11] in this relevance compensates the inaccessible part of A_f as follows:

$$R_{st} = \frac{(A_i \cap A_f)^2}{A_i.A_f} \cdot \frac{1}{\Delta t} \tag{3}$$

Therefore, if the area of the inaccessible parts of A_f or the temporal obfuscation is small, the relevance of B^{ob}-tree is still smaller than that of OST-tree (e.g., $R_s \leq R_{st}$.).

5.2 Performance Analysis

In this section, we compare the performance between the TPR-tree, OST-tree and B^{ob}-tree in terms of the tree height and number of disk accesses in the query processing. Let m, m', q, q' r, r' denote the average number of entries (i.e., the fan-out) at the root, internal nodes, and leaf nodes; R, R' be the total number of records being indexed; and d, d' be the depth of the B^{ob}-tree and TPR-tree, respectively. Then we have:

$$R = (m+1)(q+1)^{d-1} r \Rightarrow d = log_{q+1}\left(\frac{R}{(m+1)r}\right) + 1 \tag{4}$$

Similarly, we have:

$$d' = log_{q+1}\left(\frac{R}{(m'+1)r'}\right) + 1 \tag{5}$$

In B^{ob}-tree, since each node contains only the integers representing the search key values and areas of the approachable regions, the storage cost for each entry is low, and thus the node fan-out is high. On the contrary, in TPR-tree, each node contains the TPBR that require high storage cost; hence the fan-out is low. In other words, averagely each internal node of TPR-tree contains fewer entries than B^{ob}-tree ($m'<m$, $q'<q$, $r'<r$). So, with the same number of records ($R=R'$), from (4) and (5) we can see that the height of B^{ob}-tree is lower than that of TPR-tree ($d'>d$), resulting in the fewer number of disk accesses in the query processing of B^{ob}-tree than that of TPR-tree. Even more, since TPBR of TPR-tree may overlap, each query processing in TPR-tree may traverse multiple paths which require many disk accesses. It is not the case for B^{ob}-tree, which is based on B^+-tree and does not incur the curse of dimensionality problem [16]. Furthermore, since $d'>d$ and the TPBR requires higher storage cost than single values in B^{ob}-tree, the storage cost of TPR-tree is higher than that of B^{ob}-tree. Also, as shown in [11], OST-tree has the higher height and requires more disk accesses in the query processing than TPR-tree since OST-tree nodes have to reserve the space to contain authorizations. Overall, B^{ob}-tree is the best one among the three.

6 Performance Experiments

To conduct the experiment, we use the open source library called SaIL (A Spatial Index Library for Efficient Application Integration) [12]. The TPR-tree, OST-tree and B^{ob}-tree are all implemented in C++, and all experiments are conducted on a Core 2 Duo PC, running Windows 7 Professional with 1GB of RAM, 160GB of HDD, and the disk page size of 4KB. For all experiments, we use uniform datasets, where object positions are randomly generated and the moving speed ranging from 0.25 to 1.66 is chosen at random. The fill factor is usually close to 70%. The index node and leaf

node capacity are 20 with 4KB page size. The maximum update interval, number of query, and Horizon time are set to 20, 35, and 20, individually.

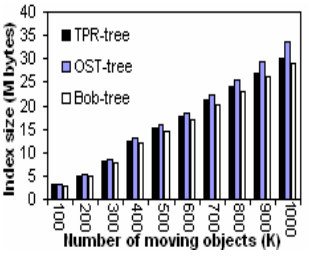


Fig. 4. Storage cost

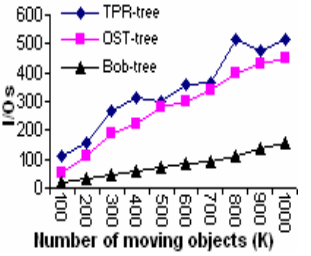


Fig. 5. Query cost (I/Os)

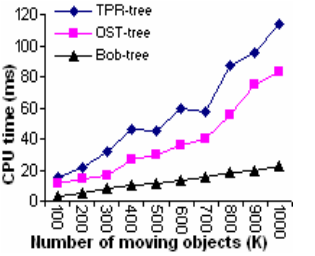


Fig. 6. Query cost (CPU time)

Beside the TPBR, OST-tree's nodes also contain authorization information, thus the number of records in each of OST-tree's node is smaller than that of the TPR-tree. So, the OST-tree requires more storage space than TPR-tree. B^{ob}-tree requires less storage space than TPR-tree since its fan-out is higher (Fig. 4). As the number of moving objects increases, the TPBR in TPR-tree and OST-tree have higher probabilities of overlapping, and the height of B^{ob}-tree is lower than that of TPR-tree and OST-tree. The query cost of B^{ob}-tree is lowest (Fig. 5 and Fig. 6).

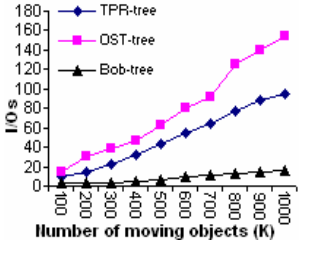


Fig. 7. Update cost (I/Os)

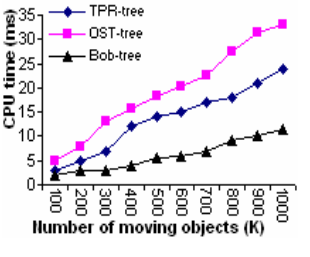


Fig. 8. Update cost(CPU time)

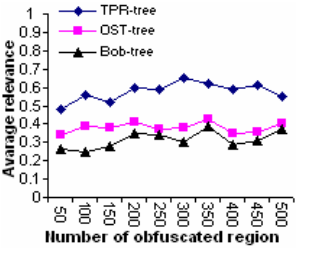


Fig. 9. Relevance

Fig. 7 and Fig. 8 show that the update cost in B^{ob}-tree achieves considerable improvement over TPR-tree and OST-tree because in B^{ob}-tree, given the key, an update needs to traverse only one path. On the contrary, in TPR-tree and OST-tree, an update may traverse multiple paths due to the overlaps among TPBR. As the dataset grows, more overlap happens and thus results in a higher update cost. In Fig. 9, by considering the geographic features inside the obfuscated region, the average relevance of B^{ob}-tree (R_s) is much smaller than that of the geometry-based obfuscation algorithms (R_{sa}). The relevance of B^{ob}-tree is slightly smaller than that of OST-tree (R_{st}) since temporal obfuscation compensates the inaccessible parts.

7 Conclusion and Future Work

In this work, we have created the B^{ob}-tree, based on the B^{+}-tree and B^{dual}-tree, that is capable of obfuscating the geographic-aware regions. Our novel index structure is much more efficient than both OST-tree and TPR-tree in terms of storage space, query processing time, update and privacy protection.

In the future, we will consider the use of the B^{ob}-tree to support other privacy-preserving techniques in LBS (e.g., k-anonymity), and will address the quality of LBS problem due to the different shapes of the returned regions wrt. the B^{ob}-tree and other access methods. Another research problem of interest is to extend authorization from $<id_{sp}, id_{user}, \Delta s>$ to $<id_{sp}, id_{user}, \Delta s, \Delta t, \Delta p>$, where Δt and Δp represent the accuracy degree of time and profile, in order to support temporal and profile obfuscation.

References

1. Ardagna, C.A., Cremonini, M., Vimercati, S.D.C., Samarati, P.: An Obfuscation-Based Approach for Protecting Location Privacy. TDSC 8(1), 13–27 (2009)
2. Anh, T.T., Chi, T.Q., Dang, T.K.: An Adaptive Grid-Based Approach to Location Privacy Preservation. In: ACIIDS, Hue, Vietnam, pp. 133–144 (2010)
3. Jafarian, J.H., Amini, M., Jalili, R.: Protecting Location Privacy through a Graph-based Location Representation and a Robust Obfuscation Technique. In: Lee, P.J., Cheon, J.H. (eds.) ICISC 2008. LNCS, vol. 5461, pp. 116–133. Springer, Heidelberg (2009)
4. Mohamed, F.M.: Privacy in Location-based Services: State-of-the-art and Research Directions. Tutorial, MDM, Germany (2007)
5. Dinh, L.V.N., Aref, W.G., Mokbel, M.F.: Spatio-temporal Access Methods-Part 2. IEEE Data Engineering Bulletin (2010)
6. Saltenis, S., Jensen, C.S., Leutenegger, S.T., Lopez, M.A.: Indexing the Positions of Continuously Moving Objects. In: ACM SIGMOD, USA, pp. 331–342 (2000)
7. Jensen, C.S., Lin, D., Ooi, B.C.: Query and Update Efficient B^{+}-tree based Indexing of Moving Objects. In: VLDB, Canada, pp. 768–779 (2004)
8. Yiu, M.L., Tao, Y., Mamoulis, N.: The B^{dual}-Tree: Indexing Moving Objects by Space-Filling Curves in the Dual Space. VLDB Journal 17(3), 379–400 (2008)
9. Dang, T.K., To, Q.C.: An Extensible and Pragmatic Hybrid Indexing Scheme for MAC-based LBS Privacy-Preserving in Commercial DBMSs. In: ACOMP, Ho Chi Minh City, Vietnam, pp. 58–67 (2010)
10. Atluri, V., Shin, H.: Efficient Security Policy Enforcement in a Location Based Service Environment. In: DBSEC, USA, pp. 61–76 (2007)
11. To, Q.C., Dang, T.K., Küng, J.: OST-tree: An Access Method for Obfuscating Spatio-temporal Data in Location-based Services. In: NTMS, France (to appear, 2011)
12. Hadjieleftheriou, M., Hoel, E., Tsotras, V.J.: SaIL: A Spatial Index Library for Efficient Application Integration. Geoinformatica 9(4), 367–389 (2005)
13. Damiani, M., Bertino, E., Silvestri, C.: Protecting Location Privacy through Semantics-aware Obfuscation Techniques. In: IFIPTM, Norway, pp. 231–245 (2008)
14. Damiani, M., Bertino, E., Silvestri, C.: PROBE: an Obfuscation System for the Protection of Sensitive Location Information in LBS. TR2001-145, CERIAS (2008)
15. Guttman, A.: R-trees: A Dynamic Index Structure for Spatial Searching. In: ACM SIGMOD, USA, pp. 47–57 (1984)
16. Dang, T.K., Küng, J., Wagner, R.: The SH-tree: A Super Hybrid Index Structure for Multidimensional Data. In: Mayr, H.C., Lazanský, J., Quirchmayr, G., Vogel, P. (eds.) DEXA 2001. LNCS, vol. 2113, pp. 340–349. Springer, Heidelberg (2001)

A Mutual and Pseudo Inverse Matrix – Based Authentication Mechanism for Outsourcing Service

Hue T.B. Pham[1], Thuc D. Nguyen[1], Van H. Dang[1], Isao Echizen[2], and Thuy T.B. Dong[1]

[1] Faculty of Information Technology, University of Science, HCMC, Vietnam
227 Nguyen Van Cu, District 5, Ho Chi Minh City, Vietnam
{ptbhue,ndthuc,dhvan,dtbthuy}@fit.hcmus.edu.vn
[2] National Institute of Informatics, 2-1-1 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
iechizen@nii.ac.jp

Abstract. Database outsourcing is becoming popular in which the data owners ship their data to external service provider. Such a model provides organizations advantages such as cost savings and service benefits. However, the delegation of database management to service provider, which is not fully trusted, introduces many significant security and privacy issues. They can be referred to as authentication, data confidentiality and integrity, data privacy, secure auditing. Among them, authentication takes an important role and is the first defence to prevent an unauthorized user from accessing to the outsourced data. In this paper, we first propose a novel public key encryption scheme with keyword search based on pseudo inverse matrix, named PEKS-PM. We prove that PEKS-PM is secure and more efficient than the public key encryption scheme with keyword search based on the Decisional Diffie-Hellman (DDH) which is the best searchable encryption scheme known to date. Based on PEKS-PM, we propose a mutual authentication mechanism which can be used to authenticate the user and the server mutually to establish an intended connection but the server learns nothing about the user's login information. Our proposed authentication mechanism can prevent man-in-the-middle, session high-jacking and replay attacks.

Keywords: Authentication mechanism, untrusted server, outsourcing service.

1 Introduction

The amount of data held by organizations is increasing quickly and it often contains sensitive information. The management and protection of such data are expensive. An emerging solution to this problem is called *database as a service* (DAS), in which, an organization's database is stored at an external service provider. There are mainly four entities in the DAS scenario (Fig. 1): (1) Data owner - individual or organization that is the subject of data to be made available for controlled external use (2) User - individual that requests data from the system (3) Client - front–end that transforms the user queries into a form that can be executed over the encrypted data which is stored

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 119–128, 2011.

on the server (4) Server - an organization that receives the data sent from data owners and makes it available for distribution to clients.

The advantages of DAS are cost savings and service benefits. However, sensitive data, which is now stored on a site that is not under the direct control of the data owner, can be put at risk. Many security and privacy issues need to be taken into account. They can be referred to as authentication, data confidentiality and integrity, data and user privacy, secure auditing, etc. It means that we need multiple defences to ensure these security and privacy issues in DAS. In many contexts, data confidentiality and integrity are managed by means of encryption [9]. Hacigümüs et al. proposed an encryption scheme that uses one key to encrypt the whole database and stores together with the encrypted database the additional index information [9]. This encryption scheme enables the server to execute queries without the need of decrypting the data. There are four steps to process a user's query Q posed from the client (1) the query is firstly translated to its server-side representation Q^S (2) Q^S is then sent to the server and is executed over the encrypted database (3) the result is sent to the client, client decrypts it and filters out those tuples not satisfying the query condition (4) client sends the final result to the user.

To ensure data privacy, several access control approaches combined cryptographic protection and authorization access control and enforce access control via selective encryption, which means users can decrypt only the data they are authorized to access ([10], [11], [12], [13], [14]).

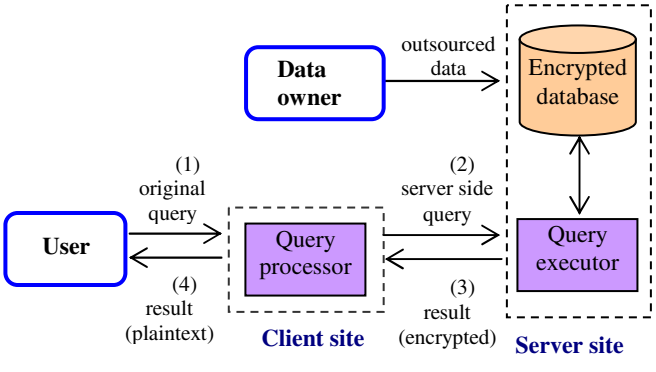


Fig. 1. DAS model

To address the user privacy issue, protocols based on private information retrieval (PIR) were proposed ([15], [16], [17]). Auditing takes an important role in any secure system. It needs a searchable encryption scheme and a protocol for ensuring both accountability and privacy requirements ([2], [3], [18], [19], [20], [21]).

In this paper, we focus on authentication issue. Authentication mechanism takes an indispensable role in DAS scenario. Without an authentication mechanism, the server cannot recognize whether a user is valid one or not and it has to process all the data requests even when they are raised from the invalid users. A user is valid if he is

granted permission to access data by the data owner and he has to pay to the data owner for the fee of retrieving data.

The contribution of this paper is fourfold. First, in section 2, we propose a novel searchable encryption scheme based on the pseudo inverse matrix, named PEKS-PM, which enables the server to test whether the login information sent from a user is valid or not but the server learns nothing about the login information to ensure user privacy.

Second, in section 3, based on PEKS-PM, we propose a mutual authentication mechanism in which, a user has to prove to the server that he is a valid user to log in the system and the untrusted server has to prove to the user that that it is not an imposter to establish a connection, or the session is not highjacked. We prove that our authentication mechanism can prevent man-in-the-middle, session high-jacking and replay attacks.

Third, in section 4, we investigate the cases when a user is inserted or deleted into/ from the system or user's password is changed.

Four, in section 5, we prove that our proposed authentication mechanism, based mainly on PEKS-PM, is secure and efficient by evaluating PEKS-PM, and compare with a searchable encryption scheme based on DDH assumption, which is the best scheme known to date. Section 6 concludes the paper.

2 A Novel Searchabe Encryption Scheme

2.1 General Searchable Encryption

We consider the general searchable encryption of Boneh et al. 3. In this scheme, a sender B wants to send a secret message to a recipient A via an untrusted server C. The scheme is briefly described as the following:

- The sender B encrypts his message M. He then appends to the resulting cipher-text a public key encryption with keyword search (PEKS) for each keyword. B sends encrypted message E(M) with keywords $W_1,\ldots,W_p$ to the server C: $E(M)||PEKS(A_{pub}, W_1)||\ldots||PEKS(A_{pub}, W_p)$, where A_{pub} is A's public key and E, an encryption function.
- The recipient A gives the third party C a certain trapdoor T_W that enables C to test whether one of the keywords associated with the encrypted message is equal to the word W of A's choice: given the encrypted values $PEKS(A_{pub}, W')$ and T_W, C can test if exists W' that satisfies W' = W.

Definition 1. A public key encryption with keyword search scheme is a group of probabilistic polynomial time algorithms (KeyGen; PEKS; Trapdoor; Test), such that:

- KeyGen(s): takes a security parameter s and returns a pair of keys (A_{pub}, A_{priv}).
- PEKS(A_{pub}, W): produces a searchable encryption of W from A_{pub} and W.
- Trapdoor(A_{priv}, W): returns a trapdoor T_W corresponding to A_{priv} and W.
- Test(A_{pub}, S, T_W): given a public-key A_{pub}, a searchable encryption S = PEKS(A_{pub}, W'), and a trapdoor T_W = Trapdoor(A_{priv}, W), outputs if W = W'.

2.2 Preliminaries

2.2.1 Pseudo-inverse Matrix

Definition 2. Consider the linear equation system: $Ax = b$ (2.1), where $A \in R^{m \times n}$, $x \in R^n$, and $b \in R^m$. Moore [4] and Penrose [5] showed that there is a general solution to (2.1) $x = A^+b$. The matrix A^+ is called pseudo-inverse matrix of the matrix A.

The authors proved that this matrix is the unique matrix that satisfies all of the following four criteria: (1). $AA^+A = A$ (2). $A^+AA^+ = A^+$ (3). $(AA^+)^T = AA^+$ (4). $(A^+A)^T = A^+A$ where M^T is the transpose matrix of M: $M = (m_{ij})$, $M^T = (m_{ji})$.

2.2.2 Properties of Pseudo-inverse Matrices

Properties of pseudo-inverse matrices will be presented as propositions. We can find their proofs in [6], and [7].

Proposition 1. [6]. (1). If A is invertible, then $A^+ = A^{-1}$ (2). If O is a zero matrix, then $O^+ = O^T$ (3). $(A^+)^+ = A$ (4). $(A^T)^+ = (A^+)^T$ (5). $(A^+)^T = (A^T)^+$.

Proposition 2. [7]. If the columns of A are linearly independent, then A^TA is invertible. In this case, $A^+ = (A^TA)^{-1}A^T$. If the rows of A are linearly independent, then AA^T is invertible. In this case, $A^+ = A^T(AA^T)^{-1}$.

2.2.3 Generating Pseudo-inverse Matrices on Z2

Algorithm 1 generates a non-singular matrix on Z_2. We work on Z_2 because on this field, matrix operations will be executed efficiently by using logical operations instead.

Algorithm 1. MatrixGen(W, n) (Generating a random non-singular matrix)

(1) Generates a non-singular lower-triangle matrix on Z_2, $L \in \{0,1\}^{n \times n}$, using bit-string W and an arbitrary hash function.
(2) Generates a non-singular upper-triangle matrix on Z_2, $U \in \{0,1\}^{n \times n}$, using bit-string W and an arbitrary hash function.
(3) Returns A = LU.

L and U are two non-singular matrices, the matrix A = LU is also a non-singular one. Based on proposition 2, algorithm 2, PseudoInverseMatrixGen, generates a pseudo-invertible matrix and its pseudo-inverse matrix**,** suppose that $m < n$.

Algorithm 2. PseudoInverseMatrixGen(m, n) (Generating a pseudo-inverse matrix)

(1) Generates a non-singular matrix $Z \in \{0,1\}^{m \times m}$ using MatrixGen algorithm.
(2) Generates randomly a matrix $W \in \{0,1\}^{m \times (n-m)}$
(3) Returns $A = [Z \parallel W]^T \in \{0,1\}^{m \times n}$ and $A^+ = (A^TA)^{-1}A^T \in \{0,1\}^{m \times n}$ (where $\parallel$ is the concatenation operator. If $j \leq m$ then A[i,j] = Z[i,j] else A[i,j] = W[i,j]. By proposition 2, A is pseudo-invertible).

2.3 An Effective Scheme for Searchable Encryption

In this section, we implement general PEKS scheme of Boneh et al. [5] using the pseudo-inverse matrices instead of using DDH assumption and bilinear maps as the implementation of Golle et al. [4]. Our proposed searchable encryption scheme is named PEKS-PM.

Definition 3. (definition of PEKS-PM) Let H_1: $\{0,1\}^* \rightarrow \{0,1\}^m$, H_2: $\{0,1\}^* \rightarrow \{0,1\}^n$, and H:$\{0,1\}^* \rightarrow \{0,1\}^m$ be three one-way hash functions, where m, n $\in$ N\{0}. Our public key encryption with keyword search, PEKS-PM, is defined as, PEKS-PM = <KeyGen, PEKS, Trapdoor, Test>, where:

- KeyGen(m, n): returns A_{pub}= XX^+ and A_{priv} = X^+ using PseudoInverseMatrixGen algorithm, for example.
- PEKS(A_{pub}, W): generates a non-singular matrix Q $\in$ $\{0,1\}^{m\times m}$ using MatrixGen(W, n); returns S = XX^+Q.
- Trapdoor(A_{priv}, W): computes V = $H_2(H_1(W)X)$; generates Q $\in$ $\{0,1\}^{m\times m}$ using MatrixGen(W,n); returns T_W =(C, D), where C = VX^+ and D = $H(VX^+Q)$.
- Test(A_{pub}, S, T_W): Let T_W = (C, D). If H(CS) = D then returns true; false otherwise.

3 A Mutual Authentication Mechanism

Authentication is defined as *"[…] the process of verifying that the identity claimed by a user is his true identity*" [1]. There are several reasons for authenticating users: for preventing an invalid user from logging in the system, for catching the user identity which is a parameter in access control decisions, for recording user identity when logging security-relevant events in the audit trail, etc. In in-house database storage model, database is stored at organization's premises and the server is considered to be trusted to manage the data. Therefore, the authentication process is necessary only at the server site to authenticate the validity of the user. In DAS scenario, the server is considered to be trusted to maintain the outsourced data or to authenticate the user but it is assumed not to be trusted with the confidentiality of database content and some metadata such as user's login information.

The data owner maintains a list of valid users. Each user is identified by a username. The data owner should not maintain the password of users to ensure user privacy. Every user sets and holds his/ her own password and no other entity knows user's password even the data owner or server's operator.

With the participation of three participants, the users, the data owner and the server, the user authentication mechanism based on PEKS-PM is performed according to the following protocol (Fig. 2):

1. The data owner generates a username un_i for each valid user i.
2. Each user i generates a pair of keys (A_{pub}^i, A_{priv}^i), keeps A_{priv}^i private and sends A_{pub}^i to the data owner.
3. The data owner maintains the relation DO_USERS (P_UN, E_UN) where P_UN and E_UN respectively store the plaintext form of each username and the

searchable encryption form of it using PEKS-PM and the key A_{pub}. Each row of relation DO_USER is of the form (un_i, PEKS(A_{pub}^i, un_i)).

4. Each user i then sets his/her password pw_i and sends the pair of values (PEKS(A_{pub}^i, un_i), h(pw_i)) to the server, where h(pw_i) is the hash value of the password pw_i, h is a hash function, for example, MD5 or SHA1.
5. The data owner also sends part of the relation that he maintains $\prod_{E\_UN}$(DO_USERS), which is the list of searchable encryption values of usernames of all the valid users. It is of the form PEKS(A_{pub}^i, un_i) for user i. The server needs to match each value PEKS(A_{pub}, un) sent from a user with that sent from the server to test for valid user and stores the password correctly. A username sent from a user (in the form PEKS(A_{pub}, un)) is valid if the server also receives that value from the data owner.
6. The server holds the relation S_USERS (E_UN, H_PW) where E_UN stores the searchable encryption form of all the usernames, and H_PW stores the hash values of the corresponding passwords of all the users. The server stores these values for authenticating users but it learns nothing about users' usernames or passwords.

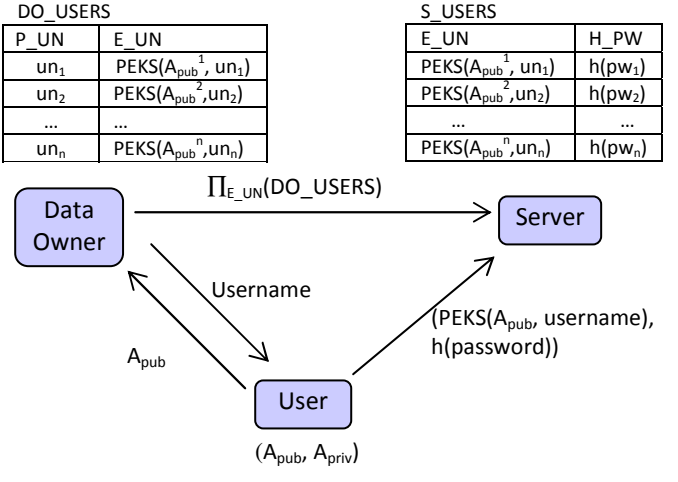


Fig. 2. Data needed for authentication

Our mutual authentication mechanism consists of five steps:

Step 1: On the login form at the client site, user U inputs username *un* and password *pw*. For security reason, client should not send the plaintext form of the username *un* and the password *pw* to the server. Based on the private key A_{priv} of U, the client computes T_W = Trapdoor (A_{priv}, *un*) and sends T_W to server S.

Step 2: For preventing man-in-the-middle attacks, server S generates an integer N randomly and uses N in the authentication process for each time of receiving a login request from the client. Based on our proposed PEKS-PM, via the trapdoor T_W = Trapdoor (A_{priv}, *un*), the server tests whether the username *un* exists in the valid user list which was stored at the server before. If yes, the server knows that user who is

making this login request having inputted the correct username. The server gets the corresponding hash value of the password and computes H = h(h(*pw*) || N), then sends H and N to the user U. In the next step, the server S has to prove to the user U that it is not an imposter or the session is not highjacked. Note that || is the concatenation operator, h is a hash function. They are the same at the client side and the server side.

Step 3: Based on the password of user U and the integer N received from server S, U computes H_1 = h(h(*pw*) || N). If H = H_1, U can be sure that S is not an imposter and he is establishing the connection with the server that he wants to make the connection, because only the server S can find the correct username (via the trapdoor value) and the corresponding password and computes the value H exactly (except U). U then computes H_2 = h(h(*pw*) || (N +1)) and sends H_2 to server S.

Step 4: Based on the hash value of the password of user U that was stored on server S before and the integer N, S computes H_3 = h(h(*pw*) || (N +1)). If H_2 = H_3, server S can be sure that user U is a valid one, because he inputs the correct username and password; otherwise, U is an invalid user.

Step 5: Server S sends the acknowledgment to user U in the case U is a valid user, or the rejection otherwise. Fig. 3 illustrates the authentication process.

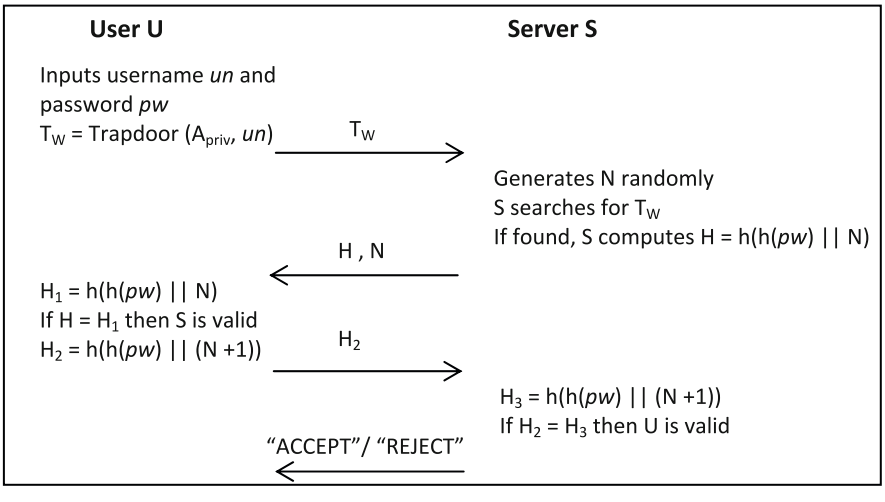


Fig. 3. Mutual authentication protocol

In the case a middle man catches the trapdoor of *un*, he cannot log in the system successfully because he cannot compute the correct value of H_2.

4 User Management

In this section, we investigate the cases when a user is deleted or inserted from/into the system or the case when a user changes his/ her password.

When a user U with username *un* is deleted from the system, the data owner sends PEKS(A_{pub}, *un*) to the server and asks the server delete from the table S_USERS the row with the value at attribute E_UN is PEKS(A_{pub}, *un*). From then on, U cannot log in the system.

When a new user U is inserted into the system, U is given by the data owner a username *un*. U then generates a pair of keys (A_{pub}, A_{priv}) and distributes A_{pub} to the data owner. Using A_{pub} sent from U and PEKS-PM, the data owner encrypts the username *un* of U and sends encryption value PEKS(A_{pub}, *un*) to the server. For security reason, U is required to set his password. U sets password *pw* and sends the pair of values (PEKS(A_{pub}, *un*), h(*pw*)) to the server, where h(*pw*) is the hash value of the password *pw*, h can be MD5, or SHA1.

When a user U wants to change his password, he simply sends the encryption value of his username PEKS(A_{pub}, *un*) and the hash value of his password h(*pw*) to the server. The server will update the hash value of his password.

In the case user U has lost his private key, he should reset his keys. U generates a new pair of keys (A_{pub}, A_{priv}), then sends A_{pub} to the data owner, and sends the encryption value of his username PEKS(A_{pub}, *un*) and the hash value of his password h(*pw*) to the server.

5 Performance and Comparison

Our authentication mechanism is based mainly on our proposed searchable encryption scheme PEKS-PM. So we need an evaluation of PEKS-PM.

5.1 Complexity of Proposed Searchable Encryption Scheme

For performance evaluation of PEKS-PM, we will analyze PEKS-PM in terms of memory cost, computational cost, security and availability.

1) Memory and Transmission Cost

In PEKS-PM, the searchable encryption $S = XX^{+}Q$ which is transferred and stored at the untrusted server needs m^2 bits.

To search an encrypted keyword, there are 2m bits of the trapdoor $T_w = (VX^{+},H(VX^{+}Q))$ transferred to the untrusted server. Therefore, the maximum of transmission bits is m^2 (m>1).

2) Computational Cost

PEKS-PM scheme uses mainly linear matrix operations. The complexity of the matrix operations is very low. Therefore, the cost to generate a searchable encryption S of the algorithm PEKS-PM and the cost to generate the trapdoor T_w of a given keyword W are low.

The complexity of test algorithm is thus low because it needs one linear matrix operation, one hash operation, and one vector comparison operation.

3) Security and Availability

We prove that, having known the messages $XX^{+}Q$ and VX^{+}, it is impossible to recover the secrets X^{+}, Q.

Given $XX^{+}Q$, suppose that rank(X) = r, and we further assume that:

$$X^{+}X = \begin{bmatrix} I_{r\times r} & O \\ O & O \end{bmatrix} \Rightarrow XX^{+}Q = \begin{bmatrix} Q_{r\times r} & O \\ O & O \end{bmatrix}$$

where $I_{r\times r} \in \{0,1\}_{r\times r}$ is an identity matrix of order r×r, and $Q_{r\times r} \in \{0,1\}_{r\times r}$ the left-upper sub-matrix of the matrix Q. Then the probability of determining the correct Q is $2^{-(n-r)m}$.

Given VX^+, even if V is completely known, the probability of determining the correct value of X^+ is very small because the probability of determining each element of X^+ would be ½, and X is generated randomly.

Based on these analyses, that the probability of successful cracking of PEKS-PM is $2^{-(n-r)m}$. Thus, the security of the PEKS-PM is reasonably high. Note that the parameters must be chosen carefully. To ensure $2^{(n-r)m}$ to be a large number, n must be considerably larger than r, and this can be guaranteed by ensuring that m < n.

5.2 Comparison

In this section, we present a comparison between proposed searchable encryption, PEKS-PM, and the searchable encryption scheme based on DDH assumption, PEKS-DDH 2. DDH assumption was based on the difficulty of the discrete logarithm problem. The average computational complexity of the discrete logarithm problem using the best method known to date [10] is $O(\exp(1.923+O(1))(\log_2 p)^{1/3}(\log_2 \log_2 p)^{2/3}))$ bit operations. To achieve a security level complexity of $2^{49.3}$, PEKS-DDH needs 200 bits and therefore, 200 transmission bits. On the other hand, to achieve a similar level of security of 2^{50}, PEKS-PM needs about 100 bits (m = 5, n =10). In order to have the security level of $2^{74.4}$, PEKS-DDH needs 500 transmission bits. Meanwhile, to obtain the security level of 2^{75}, PEKS-PM needs only 225 bits (m = 15, n = 20). The comparison is showed in Table 1.

Table 1. Number of transmission bits required for PEKS-DDH and PEKS-PM

Security level	PEKS-DDH	PEKS-PM
$2^{49.3} \approx 2^{50}$	200	100
$2^{74.4} \approx 2^{75}$	500	225

6 Conclusion

Based on the general scheme of public key encryption with keyword search, which was proposed by Boneh et al. [3], and theory of pseudo-inverse matrix, we have developed a novel searchable encryption scheme named PEKS-PM. PEKS-PM uses mainly the linear operations on the matrix, therefore the complexity is low. Specially, the storage and transmission costs of PEKS-PM are very low while the security is still assured, comparing with a scheme based on the Decisional Diffie-Hellman assumption.

Using PEKS-PM, we have proposed a mutual authentication mechanism which is very useful in data outsourcing service. In this authentication mechanism, the untrusted server has to prove to the user that it is not an imposter to establish the connection or the session is not highjacked and the user has to prove to the server that he is a valid user to log in the system. Our proposed authentication mechanism can prevent man-in-the-middle, session high-jacking and replay attacks.

References

1. Georg, G., Ray, I., France, R.: Using aspects to design a secure system. In: Proc. of 8th Int. Conf. on Engineering of Complex Computer Systems, pp. 117–126 (2002); ISSN 1054-4729
2. Golle, P., Staddon, J., Waters, B.: Secure Conjunctive Keyword Search over Encrypted Data. In: Jakobsson, M., Yung, M., Zhou, J. (eds.) ACNS 2004. LNCS, vol. 3089, pp. 31–45. Springer, Heidelberg (2004)
3. Boneh, D., Crescenzo, G.D., Ostrovsky, R., Persiano, G.: Public-key encryption with keyword search. In: Cachin, C. (ed.) EUROCRYPT 2004. LNCS, vol. 3027, pp. 506–522. Springer, Heidelberg (2004)
4. Moore, E.H.: On the reciprocal of the general algebraic matrix. Bulletin of the American Mathematical Society 26, 394–395 (1920)
5. Penrose, R.: A generalized inverse for matrices. Proc. of the Cambridge Philosophical Society 51, 406–413 (1955)
6. Goul, G.H., Charles, F.V.L.: Matrix computations, 3rd edn., pp. 257–258. Johns Hopkins, Baltimore (1996)
7. Ben-Israel, A., Thomas, N.E.G.: Generalized Inverses. Springer, Heidelberg (2003)
8. Menezes, A.J., Oorschot, P.C.V., Vanstone, S.A.: Handbook of Applied Cryptography. CRC Press, Boca Raton (1997)
9. Hacigümüs, H., Iyer, B.R., Li, C., Mehrotra, S.: Executing SQL over encrypted data in the database-service-provider model. In: SIGMOD, pp. 216–227 (2002)
10. Damiani, E., De Capitani di Vimercati, S., Foresti, S., Jajodia, S., Paraboschi, S., Samarati, P.: Key Management for Multi-User Encrypted Databases. In: Proc. of the 2005 ACM Workshop on Storage Security and Survivability, pp. 74–83 (2005)
11. De Capitani di Vimercati, S., Foresti, S., Jajodia, S., Paraboschi, S., Samarati, P.: Over-encryption: Management of Access Control Evolution on Outsourced Data. In: VLDB, pp. 123–134 (2007)
12. El-khoury, V., Bennani, N., Ouksel, A.M.: Distributed Key Management in Dynamic Outsourced Databases: a Trie-based Approach. In: First Int. Conf. on Advances in Databases, Knowledge and Data Applications, pp. 56–61 (2009)
13. Sandhu, R.S.: Cryptographic implementation of a Tree Hierarchy for access control, pp. 95–98. Elsevier, Amsterdam (1988)
14. Zych, A., Petkovic, M., Jonker, W.: Efficient key management for cryptographically enforced access control, pp. 410–417. Elsevier Science, Amsterdam (2008)
15. Asonov, D.: Private information retrieval: An overview and current trends. In: ECDPvA Workshop, pp. 889–894 (2001)
16. Chor, B., Goldreich, O., Kushilevitz, E., Sudan, M.: Private information retrieval. In: FOCS, pp. 41–50 (1995)
17. Lin, P., Candan, K.S.: Hiding traversal of tree structured data from untrusted data stores. In: Proc. of the 2nd Int. Workshop on Security in Information Systems, Portugal, pp. 314–323 (2004)
18. Song, D.X., Wagner, D., Perrig, A.: Practical techniques for searches on encrypted data. In: IEEE Symposium on Security and Privacy, pp. 44–55 (2000)
19. Waters, B.R., Balfanz, D., Durfee, G., Smetters, D.K.: Building an encrypted and searchable audit log. In: 11th Annual Network and Distributed System Security Symposium (2004)
20. Chang, Y.C., Mitzenmacher, M.: Privacy preserving keyword searches on remote encrypted data, Cryptology ePrint Archive, Report 2004/051 (2004), `http://eprint.iacr.org/2004/051/`
21. Thuc, D.N., Hue, T.B.P., Van, H.D.: An Efficient Pseudo Inverse Matrix-Based Solution for Secure Auditing. IEEE-RIVF, 7–12 (2010); ISBN: 978-1-4244-8072-2

Anonymizing Shortest Paths on Social Network Graphs

Shyue-Liang Wang[1], Zheng-Ze Tsai[1], Tzung-Pei Hong[2], and I-Hsien Ting[1]

[1] Department of Information Management
[2] Department of Computer Science and Information Engineering
National University of Kaohsiung
Kaohsiung, Taiwan 81148

Abstract. Social networking is gaining enormous popularity in the past few years. However, the popularity may also bring unexpected consequences for users regarding safety and privacy concerns. To prevent privacy being breached and modeling a social network as a weighted graph, many effective anonymization techniques have been proposed. In this work, we consider the edge weight anonymity problem. In particular, to protect the weight privacy of the shortest path between two vertices on a weighted graph, we present a new concept called *k-anonymous path privacy*. A published social network graph with *k-anonymous path privacy* has at least k indistinguishable shortest paths between the source and destination vertices. Greedy-based modification algorithms and experimental results showing the feasibility and characteristics of the proposed approach are presented.

Keywords: Social networks, privacy preserving, edge weight, shortest path, k-anonymity.

1 Introduction

Privacy preserving data mining, privacy preserving data publishing, and privacy preserving network publishing have attracted considerable attention in recent years because of the concern of breaching privacy from published data. Social network applications, such as MySpace and Facebook and other online communities, collaboration networks, telecommunication networks, have become very popular for sharing information. There are millions of registered users associated with others through friendships, hobbies, professional association, and so on. These user information and relationship can be modeled as vertices, edges, and edge weights in complex graphs and are of significant importance in various application domains such as marketing, psychology, epidemiology and homeland security. As a result, companies and institutions hosting the data are interested and expect to be beneficial in releasing portions of the graphs so that research communities can analyze the data. However, these social network graphs may contain sensitive information. In order to protect the privacy of users against different types of attacks, graphs should be anonymized before they are published.

Some current practices to protect user privacy from published data include removing all identifiable personal information such as names and social security numbers,

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 129–136, 2011.

limiting access, "fuzzing" the data, eliminating unnecessary groupings, augmenting with additional data, etc. However, it is still easy for an attacker to identify the target by performing different structural and non-structural queries. Let's consider the following examples of re-identification attack on relational data, transaction data, and graph data.

For published relational data, given a public voter registration data and a private microdata such as the de-identified (name and social security number removed) patient data of Massachusetts's state employees, a simple "linking" attack by joining the two datasets can re-identify the identity and medical history of the state's governor. According to one study, approximately 87% of the population of the United States can be uniquely identified on the basis of their 5-digit zip code, sex, and date of birth [1, 13, 14].

For published transaction data, America Online (AOL) released a large portion of its search engine query logs for research purposes in August 2006. The dataset contained 20 million queries posed by 650,000 AOL users over a 3 month period. Before releasing the data, AOL replaces each user's name by a random identifier. However, by examining unique query terms, the New York Times [2] demonstrated that the searcher No. 4417749 was traced back to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Georgia. Despite a query does not contain address or name, a searcher may still be re-identified from combination of query terms that are unique enough about the searcher.

For published graph data, even when a network is published without any identity information, it is still possible to locate the target with high probability based on some structural information around the target [5, 17]. Similar to the quasi-identifiers in relational or set-valued data that can be used as background knowledge for re-identification; any topological structure of the network can be utilized to identify the target in a released network. There have been four types of structural attacks in this environment [5, 7, 16]: degree-attack, subgraph attack, 1-neighborhood attack, and hub-fingerprint-attack. It is also possible that an attacker can also launch a query based on non-structural information (such as vertex label) to identify the target.

There are basically three types of sensitive information that one may want to keep private and may be under attack in a social network environment: node information, link information and edge weight information [3, 8, 9]. The node information is the information attached to a vertex. For example, the emails sent by an individual, the personal information such as age, sex, zip code, and transaction data such as purchased items [6, 10-12]. The link information is about the relationships among the individuals which may be considered sensitive. Links can be used to represent financial exchanges, friend relationships, conflict likelihood, sexual relations, disease transmission [9]. Depending on the application, the edge weight information can semantically represent "degree of friendship", "trustworthiness", and "behavior" etc. If considering routing problem, (for information spread and marketing), edge weights may correspond to the cost of information propagation [4]. To protect node information, many generalization and suppression-based k-anonymity techniques for relational and set-valued data have been proposed. To protect link information, there are some studies such as k-degree, k-automorphism, k-isomorphism privacy models addressing various types of structural attacks. To protect edge weight privacy, perturbation-based approaches to preserve linear property such as shortest paths by anonymizing the edge weights have been proposed recently [4, 8, 9]. In this work, we consider the problem of anonymizing

the shortest path by minimally modifying edge weights such that the published social network graph reveals at least k shortest paths between source and destination vertices. We define a new concept called *k-anonymous path privacy*. Greedy-based modification algorithms and experimental results showing the feasibility and characteristics of the proposed approach are presented.

The rest of the paper is organized as follows. Section 2 gives the problem description. Section 3 describes the proposed algorithms. Section 4 reports the numerical experiments. Section 5 concludes the paper.

2 Problem Description

Recent studies in privacy preserving social networks have proposed many novel models and anonymization approaches. Most of them model the social networks by un-weighted graphs. It then perturbed the graphs before the publication in order to conceal the identities of vertices or link relationships among group of vertices.

However, weighted graphs can be used for analyzing the formation of communities within the network, business transaction networks, viral and targeted marketing and advertising, modeling the structure and dynamics such as opinion formation, and for analysis of the network for maximizing the spread of information through the social links [4]. Depending on the applications, the edge weights could be used to represent "degree of friendship", "trustworthiness", and "business transaction", etc.

In order to protect the privacy of these sensitive information (sensitive edges), current works concentrate on preserving the shortest paths characteristic between pairs of vertices [4, 9] and k-anonymous weight privacy [8]. To preserve the shortest paths between pairs of vertices, Gaussian randomization perturbation and greedy perturbation techniques that minimally modify the edge weights without adding or deleting any vertices and edges have been proposed. A linear programming abstract model that can preserve linear properties of edge weights (including shortest paths) after anonymization is presented in [4]. To eliminate the distinguishability between edge weights, the k-anonymous weight privacy is defined as[8]: the edge $(i \rightarrow j)$ is k-anonymous if and only if there exist at least k edges in $\Phi(i)$ whose weights $w_{i,tl}$, $l = 1, ..., c$, and $c \geq k$, satisfy $\| w_{i,j} - w_{i,tl} \| \leqq \mu$, l=1,..., c. Here, μ is a predefined positive parameter to control the degree of privacy and $\Phi(i)$ is the adjacent edge set in which all edges come from the i-th node.

In this work, we consider a different type of privacy, *k-anonymous path privacy*, that hiders adversary to infer the sensitive relationship between two entities (vertices) in a social network. The basic idea is to hide the true sensitive information, e.g. the shortest path, by obfuscating it with at least *k-1* other paths so that the true path will not be revealed. Figure 1 shows an undirected weighted graph with six vertices. Assuming the relationship represented by the shortest path between vertices v_1 and v_6 is sensitive, $\{(v_1, v_5), (v_5, v_4), (v_4, v_6)\}$, and expected to be hidden. One possible technique is to perturb minimal number of edge weights so that there will be k shortest paths between the two vertices. Figure 2 shows the anonymized graph with two shortest paths $\{\{(v_1, v_5), (v_5, v_4), (v_4, v_6)\}, \{(v_1, v_5), (v_5, v_6)\}\}$, where the weight of edge $e_{5,6}$ is modified to two. Therefore, given a graph G, a set of source and destination nodes H, and privacy level k, the objective of *k-anonymous path privacy* is to minimally modify the graph

such that there exists k shortest paths between each given pair of nodes specified in H, without adding or deleting any vertices or edges. The privacy level k is treated as the number of shortest paths between the specified source and destination vertices in this work.

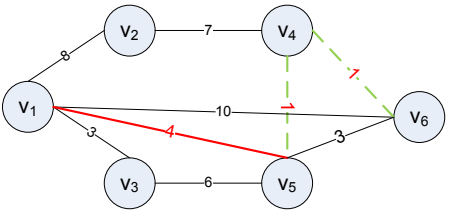


Fig. 1. The Original Network

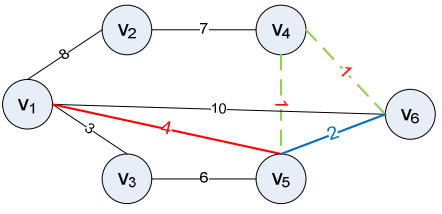


Fig. 2. The Anonymized Network with k=2

3 Proposed Algorithms

For one given pair of source and destination vertices, the objective of *k-anonymous path privacy* is to minimally modify the graph so that k shortest paths can be achieved. We propose a greedy-based approach and modify the edge weights in the top-k shortest paths so that they all possess the same path length. The proposed algorithm first finds the second shortest path and reduces proportionally the edge weights of non-overlapping edges between the second shortest and the shortest paths. For example, in Figure 1, the second shortest path between v_1 and v_6 is $p_{1,6} = \{(v_1, v_5), (v_5, v_6)\}$. The non-overlapping edge is $e_{5,6}$ which has weight three. The non-overlapping edges in the shortest path are $e_{5,4}$ and $e_{4,6}$ which has total weight of two. Therefore the edge weight of $e_{5,6}$ is reduced to two so that both paths will have the same path length. The process repeats itself after all top-k shortest paths are modified. The proposed algorithms for anonymizing single pair of source and destination vertices (K-Single Path anonymization algorithm; KSP) and multiple pairs of source and destination vertices (K-Multiple Path anonymization algorithm; MSP) are given in the following. For simplicity, we assume that edge weights can be modified only once in this work.

The following notations will be used:
v_i: vertex i;
$e_{i,j}$: edge between vertices v_i and v_j;

$w_{i,j}$: weight of edge $e_{i,j}$;
$p_{i,j}$: path between vertices v_i and v_j;
$d_{i,j}$: length of path $p_{i,j}$;
SPL: Shortest Path List;
TSPL: Temporary Shortest Path List;

Given a graph G, a source vertex and a destination vertex (v_i, v_j), and privacy level k, the objective is to minimally modify the graph such that there exists k shortest paths between the given pair of vertices, without adding or deleting any vertices or edges.

K-Single Path Anonymization Algorithm (KSP)
Input: W, weighted adjacency matrix of a given graph G,
The source and destination vertices for which the shortest path is to be anonymized,
K, number of shortest path between each pair of source and destination vertices,
Output: anonymized weighted adjacency matrix W*,

1. Find the shortest path $p_{i,j}$ & its length $d_{i,j}$;
2. For (j = 2 to k)
3. {Find the *j-th* shortest path $p'_{i,j}$ & its length $d'_{i,j}$;
4. For (each edge $e'_{p,q}$ on $p'_{i,j}$ that is non-overlap with top *(j-1)* shortest paths)
5. $\{w'_{pq} = w'_{pq} + \frac{w'_{pq}}{\sum w'_{pq}} \times (d'_{ij} - d_{ij})$; //summation over non-overlapping edges
6. Update the adjacency matrix;
7. }; // end of for each edge
8. }; // end of for j = 2 to k

For a set of source and destination vertices H, a given privacy level k, the *k-anonymous path privacy* problem is to minimally modify the graph G such that there exists k shortest paths between each given pair of vertices specified in H, without adding or deleting any vertices or edges. The following algorithm further assumed that anonymized paths cannot be modified again when anonymizing other sets of paths (for different pairs of source and destination vertices).

K-Multiple Paths Anonymization Algorithm (KMP)
Input: W, weighted adjacency matrix of a given graph G,
H, the set of source and destination vertices for which the shortest paths are to be anonymized,
K, number of shortest path between each pair of source and destination vertices,
Output: anonymized weighted adjacency matrix W*,

1. Initialize $SPL = \phi$; //shortest path list
2. while ($H \neq \phi$)
3. {for (each pair of vertices (v_i, v_j) in H)

find its shortest path $p_{i,j}$ and length $d_{i,j}$;
$d_{r,s} := \min_H d_{i,j}$; //minimum of all shortest paths
$H := H - \{(v_r, v_s)\}$;
$TSPL := \{p_{r,s}\}$; //the shortest path for (v_r, v_s)
while ($|TSPL| < k$) //anonymizing k-1 paths
{find next shortest path $p'_{r,s}$ and its length $d'_{r,s}$;
if ($d'_{r,s} = d_{r,s}$) //same length
{$TSPL := TSPL + p'_{r,s}$; // add to anonymized list
continue;} //find next shortest path
else // different length
{let $diff := d'_{r,s} - d_{r,s}$;
$p''_{r,s} := p'_{r,s} - \{$edges in SPL and $TSPL\}$;
If ($p''_{r,s} \neq \phi$ and $d''_{r,s} > diff$) //available edges
for (each edge $e''_{i,j}$ on the path $p''_{r,s}$)
$\{w''_{ij} = w''_{ij} + \frac{w''_{ij}}{\sum w''_{ij}} \times (d'_{rs} - diff)$;
update the adjacency matrix;
$TSPL := TSPL + p'_{r,s}$;
}; // end of for each edge $e''_{i,j}$
}; // end of if/else
}; // end of while ($|TSPL| < k$)
$SPL := SPL + TSPL$;
}; // end of while ($H \neq \phi$)

4 Numerical Experiments

To evaluate the characteristics of the proposed algorithms, we run simulations on a real data set, EIES (Electronic Information Exchange System) Acquaintanceship at time 2 collected in [15] and can be downloaded from International Network of Social Network Analysis website. The EIES at time 2 is a network of researchers who participated in an early study on the impact of a computer conference on the formation of interpersonal ties among scientists. The measure of acquaintanceship between users has four levels, ranging from 1 (do not know the other) to 4 (very good friendships). The data set contains 48 users and 830 acquaintanceships and is modeled as a weighted graph.

All experiments reported in this section were performed on an Intel Core 2 Duo P8700 CPU, 2.53 GHz machine with 4 GB main memory, running Microsoft Windows 7 operating system. All the methods were implemented using Java programming language.

Figure 3 shows the preliminary results of ratios of perturbed edges. The ratio is the number of modified edges over the total number of edges on the k shortest paths. It can be observed that the percentage of perturbed edges remain quite stable for different privacy level k and for multiple pairs of source and destination vertices. For anonymizing one pair of source and destination vertices (*H1*), the average ratio of five randomly selected pairs is about 40%. The average running time is 0.61 seconds. For anonymizing two pairs (*H2*) and three pairs (*H3*) of source and destination vertices, the

average ratios of five randomly selected pairs are about 39% and 36% respectively. The average running times are 2.77 and 4.49 seconds respectively. Figure 4 shows the running times of anonymization of k shortest paths for different privacy level k and for multiple pairs of source and destination vertices. It can be observed that anonymizing multiple pairs of source and destination vertices require relatively more running time when k increases to 10. This is due to the fact that it takes longer time to search for extra paths to be anonymized. However, the required level of privacy k would depend on applications and is usually not very large.

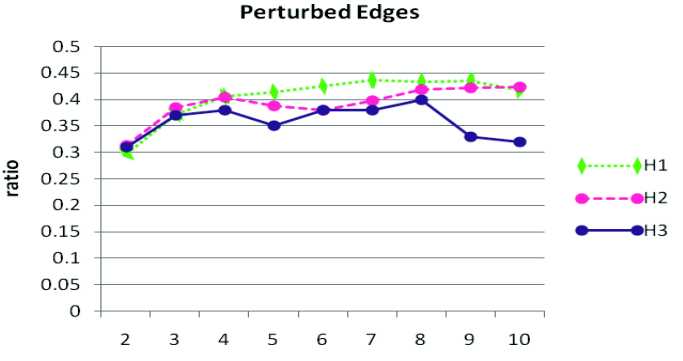


Fig. 3. Ratios of perturbed edges for different k

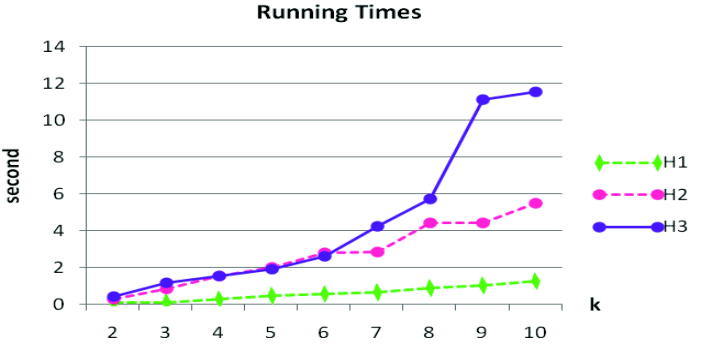


Fig. 4. Running times for different k

5 Conclusions

In this work, we have studied the problem of preserving sensitive paths in social networks. We proposed a new concept called *k-anonymous path privacy* and algorithms that minimally perturbed the edge weights to achieve the path anonymity. Examples illustrating the approach and numerical experiments showing the characteristics of the proposed approach were given. It demonstrates that the proposed technique is feasible to achieve the *k-anonymous path privacy*. In the future, we will consider overlapped

shortest paths and preserving other types of sensitive characteristics and privacy such as minimal cost spanning trees and others.

Acknowledgments. This work was supported in part by the National Science Council, Taiwan, under grant NSC-99-2221-E-390-033.

References

1. Backstrom, L., Huttenlocher, D.P., Kleinberg, J.M., Lan, X.: Group formation in large social networks: membership, growth, and evolution. In: KDD, pp. 44–54 (2006)
2. Barbaro, M., Zeller Jr., T.: A face is exposed for AOL searcher no. 4417749. New York Times (August 2006)
3. Cheng, J., Fu, A., Liu, J.: K-isomorphism: privacy preserving network publication against structural attacks. In: SIGMOD Conference, pp. 459–470 (2010)
4. Das, S., Egecioglu, O., Abbadi, A.E.: Anonymizing weighted social network graphs. In: ICDE, pp. 904–907 (2010)
5. Hay, M., Miklau, G., Jensen, D., Towsley, D.F., Weis, P.: Resisting structural re-identification in anonymized social networks. PVLDB 1(1), 102–114 (2008)
6. He, Y., Naughton, J.F.: Anonymization of set-valued data via top-down, local generalization. In: VLDB (2009)
7. Liu, K., Terzi, E.: Towards identity anonymization on graphs. In: SIGMOD Conference, pp. 93–106 (2008)
8. Liu, L., Liu, J., Zhang, J.: Privacy preservation of affinities in social networks. In: ICIS (2010)
9. Liu, L., Wang, J., Liu, J., Zhang, J.: Privacy preservation in social networks with sensitive edge weights. In: SDM, pp. 954–965 (2009)
10. Meyerson, A., Williams, R.: On the complexity of optimal k-anonymity. In: Proc. of PODS (2004)
11. Motwani, R., Nabar, S.U.: Anonymizing unstructured data, arXiv: 0810.5582v2, [cs.DB] (2008)
12. Park, H., Shim, K.: Approximate algorithms for k-anonymity. In: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, pp. 67–78 (2007)
13. Samarati, P., Sweeny, L.: Generalizing data to provide anonymity when disclosing information. In: Proc. of ACM Symposium on Principles of Database Systems, p. 188 (1998)
14. Sweeny, L.: k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10(5), 557–570 (2002)
15. Wasserman, S., Faust, K.: Social network analysis: methods and applications. Cambridge University Press, New York (1994)
16. Zhou, B., Pei, J.: Preserving privacy in social networks against neighborhood attacks. In: ICDE, pp. 506–515 (2008)
17. Zou, L., Chen, L., Ozsu, M.T.: K-automorphism: A general framework for privacy preserving network publication. In: VLDB (2009)

Mining Latent Sources of Causal Time Series Using Nonlinear State Space Modeling

Wei-Shing Chen and Fong-Jung Yu

Department of Industrial Engineering and Technology Management,
Da-Yeh University, 168, University Rd. Dacun, Changhua, 51591, Taiwan, R.O.C.
weishing@mail.dyu.edu.tw

Abstract. Data mining refers to use of new methods for the intelligent analysis of large data sets. This paper applies one of nonlinear state space modeling (NSSM) techniques named nonlinear dynamical factor analysis (NDFA) to mine the latent factors which are the original sources for producing the observations of causal time series. The purpose of mining indirect sources rather than the time series observation is that much better results can be obtained from the latent sources, for example, economics data driven by an "explanatory variables" like inflation, unobserved trends and fluctuations. The effectiveness of NDFA is evaluated by a simulated time series data set. Our empirical study indicates the performance of NDFA is better than the independent component analysis in exploring the latent sources of Taiwan unemployment rate time series.

Keywords: Data mining, latent sources, time series, nonlinear state space modeling, nonlinear dynamical factor analysis.

1 Introduction

Starting from 1980s, modeling time series has become a popular theme [1] for understanding the behaviors of the dynamical process by observing responses. A time series is sequence of regularly sampled quantities from an observed process which is frequently driven by a rather small, typically unobservable set of influences. For example, when there are a multivariate time series that represents the quantity sold and price associated with demand for a product or service, it is often desirable to decide the influences of price and the demand. Such the latent time series describe the underlying data-generating process of a dependent time series using some latent sources (explanatory variables). The developments of this paper were motivated by the need for exploring the driven processes for the given observations. These driven processes contain quantities that cannot be measured direct. Instead, only a portion of noise-corrupted observation is available. Such an effort is needed because the measurement process is either more expensive or more impossible for the latent time series [2]. The present work is therefore interested in the problem of reverse engineering of a time series where one uses a set of given time series x_t to get another hidden set of time series s_t.

In this paper, we used the state space model (SSM) as the data mining technique to explore the latent sources. SSM is a model comprised of two parts: states and

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 137–148, 2011.

observations. The SSM is one of the most powerful methods for modeling a dynamic system and has been widely employed for engineering control systems[3]. The remaining of the paper is organized as follows. In Section 2 we briefly introduce how an SSM can be used to learn dynamical input-output mappings. In Section 3 we discuss the issue of mining hidden sources from multivariate time series by using nonlinear dynamical factor analysis (NDFA). The methods for deciding the parameters used in NDFA are given in Section 4. In Section 5, an application study is carried out and the results are compared with the FastICA method. The paper is concluded in Section 6.

2 State-Space Models

An SSM describes evolving two time series running in parallel, one referred to as the state process (s_t) and the other as the observation process (x_t). An SSM assumes that an observation vector x_t is generated by its latent sources s_t through an unknown transformation matrix Z and additive observation disturbance n_t:

$$x_t = Zs_t + n_t. \tag{1}$$

In (1), the transformation matrix $\boldsymbol{Z}$ is a parameterized mapping from one state space to an observation space. In a dynamical process, a current source s_t can also be generated through another transformation matrix $\boldsymbol{T}$ from the sources s_{t-1} at the previous time instant as follows

$$s_t = Ts_{t-1} + m_t . \tag{2}$$

The system noise m_t and the observational noise n_t noise can be Gaussian or non-Gaussian. The observation equation (1) and state equation (2) form a linear state-space representation for the dynamic behavior of $\boldsymbol{x}$. An SSM provides an important body of techniques for analyzing time series, only the observations $\boldsymbol{x}$ are known earlier, and both the states $\boldsymbol{s}$ and the mappings $\boldsymbol{Z}$ and $\boldsymbol{T}$ are learned from the data.

Many statistical methods [4] have been developed to find latent sources from the observation data. Three types of inference, including filtering, smoothing and prediction, for the state and the model from the observations $\boldsymbol{x}$ are commonly discussed in the literature [5]. For linear SSMs with Gaussian process and observation noise, the Kalman filter [6] and Kalman smoothing [7] are the well-known methods of choice for the consistent estimation of the indirectly observed or unobserved states. For many real-world data, the affect of the desired sources to the observed data is, however, not linear. Therefore, an SSM supports nonlinear state space models (NSSM) in the form (3) and (4) are proposed [8] where $\boldsymbol{\theta}$ is a vector containing the model parameters and time $\mathbf{t}$ is discrete.

$$x_t = Z(s_t, \theta_z) + n_t. \tag{3}$$

$$s_t = T(s_{t-1}, \theta_t) + m_t . \tag{4}$$

An SSM is a model to describe how $\boldsymbol{s_t}$ generates or "causes" $\boldsymbol{x_t}$ and $\boldsymbol{s_{t+1}}$. The goal of mining in this paper is to invert this mapping, that is, to mine $\boldsymbol{s_{1:t}}$ from the given $\boldsymbol{x_{1:t}}$ .Methods have been proposed for learning a linear SSM [9], but estimating a nonlinear SSM is inherently more difficult. Indeed, it is almost impossible to find any simple

closed form solution for the distribution of the sources in (3) and (4). Because in an NSSM, the statistical learning problem is no longer solved in the closed form as opposed to the linear SSM and this raises several computational difficulties. Quach et al. [10] used an unscented Kalman filtering (UKF) technique to tackle nonlinearities by deriving NSSMs from ordinary differential equations (ODEs). The nonlinear mapping in (3) and (4) can be modeled by a multilayer perceptron (MLP) networks or a radial basis function (RBF) network which suit well to modeling both strong and mild nonlinearities [11]. In this paper, we apply a method called nonlinear dynamical factor analysis (NDFA), which is based on variational Bayesian learning [12, 13], for learning form of (3) and (4) in an NSSM of a dynamic process. Variational Bayesian methods, also called ensemble learning[14], are a group of methods to approximate intractable integrals arising in Bayesian inference and machine learning.

3 Nonlinear Dynamical Factor Analysis

This section briefly introduces the NDFA model and its learning algorithm. A comprehensive discussion can be found in [12, 13]. The NDFA model is a dynamical extension of nonlinear factor analysis (NFA) [11]. The function of NDFA is to find dynamic sources which explain nonlinearity of the observed data[8]. In NDFA, the mapping from sources to observations is modeled by a multilayer perceptrons (MLP) to represent the nonlinearity. Like in figure 1, the sources are on the top layer and observations in the bottom layer. The middle layer consists of hidden neurons of sigmoidal nonlinearities each of which estimates a nonlinear function of the inputs as

$$x_t = Z(s_t, \theta_Z)+n_t = B\phi(As+a)+b+n_t = B\tanh(As+a)+b+ n_t . \quad (5)$$

where the matrices A and B are the weights of the hidden and output layer and a and b are the corresponding biases. The *tanh* is a common nonlinear activation function ϕ being used in the MLPs. The function T has a similar mapping structure except the MLP network is used to model only the change in the state values.

$$T(s)=s+D\phi(Cs+c)+d=s+D\tanh(Cs+c)+d . \quad (6)$$

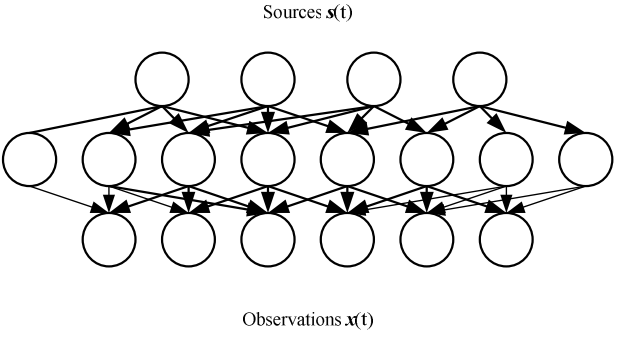


Fig. 1. Mapping structure from sources to observations using a MLP network

All the assumptions of NDFA are expressed in the form of the density model $\boldsymbol{p}(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\theta})$. The parameters $\boldsymbol{\theta}$ consist of the parameters of the MLPs in (5) and (6), the variance parameters of the noise terms $\boldsymbol{n}_t$ and $\boldsymbol{m}_t$ as well as hyperparameters. To estimate both the unknown nonlinear function $\boldsymbol{Z}$ and $\boldsymbol{T}$ and the unknown sources $\boldsymbol{s}_t$, the general approach is to estimate the posterior probability distribution of the unknown parts of the model by using the Bayesian approach where all the assumptions made in the model are expressed in the form of the joint distribution of the observations $\boldsymbol{X}$, states $\boldsymbol{S}$ and parameters $\boldsymbol{\theta}$ of the model

$$p(x,s,\theta)=p(x|s,\theta)p(s|\theta)p(\theta). \tag{7}$$

Gaussian distributions are used to describe the weights of the MLPs for computational tractability. Based on the Bayesian approach, once the joint distribution (7) is defined and the set of observations is obtained, all the relevant information about the unknown parameters is contained in the posterior

$$p(s,\theta|x)=p(s,\theta,x)/p(x). \tag{8}$$

In general, one would find point estimates of the unknowns by maximizing the posterior (8), which yields the well-known maximum a posteriori (MAP) estimate. However, this approach would easily cause over-fitting problems especially in this ill-posed estimation problem. So, Giannakopoulos and Valpola [15] proposed an ensemble learning technique to approximate the actual posterior probability $p(s,\theta|x)$ (8) by an approximating distribution $q(s,\theta)$ over $\boldsymbol{S}$ and $\boldsymbol{\theta}$. The structure of the model and the learning scheme were optimized based on the value of a cost function which is measured by the Kullback-Leibler divergence:

$$C_{KL}=\int q(s,\theta)log(q(s,\theta)/p(s,q|x)d\theta ds. \tag{9}$$

A more comprehensive explanation of the algorithm is given in [12, 16], and an MATLAB software implementation for the NDFA technique is available at [17].

4 Setting the Parameters for NDFA Algorithm

To apply the NDFA algorithm, a number of parameters including the number of sources, the number of hidden neurons, type of activation function, and an approximation method need to be determined. For the practicable applying the NDFA algorithm in data mining of latent sources, we outline the guidelines for setting up the parameters step-by-step as follows:

Step 1: Convert a one-dimensional time series to a state space with embedded dimensions for modeling the dynamics of the causal time series.

Nonlinear factor analysis is used to extract the undying state from a sequence of observations. Phase space embedding techniques are the standard methods to analyze nonlinear dynamical systems. For a one-dimensional time series data, we need to convert it to an embedded data vector $x(t)=[x^T(t)x^T(t-\tau)x^T(t-2\tau)\ldots x^T(t-m\tau)]$ by using the Takens' embedding theory [18]. The delay parameter (τ) can be determined as the first minimum of the mutual information (MI) function [19] or the first zero of the autocorrelation function. An embedding dimension (m) can be decided by using the

mutual information. To determine the proper embedding dimension m, the false nearest neighbors (FNN) method [20] can be used.

Step 2: Determine number of latent sources and set up the initial values for the sources.

According to the Fig 1, mining latent sources using NDFA requires prior knowledge of the number of sources before the ensemble learning can be performed. Here, we applied ICA approach and used a measure of the number of sources based on the normalized determinant value of the global matrix (G) to determine the number of sources in the given observations [21]. The global matrix G is the product of estimated the mixing matrix and unmixing matrix. It can be stated that if |G| is near to zero, then it points out there are some dependent sources. So, the number of sources should be fewer than the number of variables of observations. Otherwise, if |G| is near to one, the number of sources can be assumed to be equal to the number of observational variables. FastICA [22] can be applied to obtain the mixing matrix and unmixing matrix. Another alternative to determine the number of latent sources is based on the eigenvalue spectrum of the data covariance matrix. Everson & Roberts [23] proposed a method of inferring the true eigenvalue spectrum from the sample spectrum.

Because of the flexibility of the MLP network and the gradient based learning algorithm, the NDFA is sensitive to the initialization. Honkela and Valpola suggested to use linear PCA for initialization of the means of the sources S [16].

Step 3: Determine number of hidden neurons in the mappings Z and T in (3) and (4) which are modeled by an MLP network.

There is no magic formula for selecting the best number of hidden neurons. Santos et al. [24] proposed a novel technique to estimate the number of hidden neurons of an MLP. The proposed approach consists in the post-training application of SVD/PCA to the back-propagated error and local gradient matrices associated with the hidden neurons. The number of hidden neurons is then set to the number of relevant singular values or eigenvalues of the involved matrices. Some thumb rules are also available for calculating number of hidden neurons such as the following:

1. The number of hidden neurons should be between the size of the input layer and the size of the output layer.
2. The number of hidden neurons should be 2/3 the size of the input layer, plus the size of the output layer.
3. The number of hidden neurons should be fewer than twice the size of the input layer.

Step 4: Decide nonlinear activation function to use in the MLP. Typical choices for ϕ include *tanh* or the logistic *sigmoid* function.

In case of the inputs of the hidden neurons are not Gaussian, use the standard logistic sigmoid activation function in the form of *tanh* which is more common and faster to evaluate numerically.

Step 5: Decide the approximation method for the nonlinearity.

Since it is impossible to find any closed form for the distribution of the sources, we should use a fixed form approximation where the form is fixed and only the parameters are optimized by minimizing the cost function. Two natural choices for the functional form of the approximation for the nonlinearity are Taylor approximation for low

variance inputs and Gauss-Hermite quadrature approximation for high variance inputs and adaptive approach based on the combination of Gauss-Hermite quadrature formula [25]. The Taylor approximation method fails in case of high input variance because it relies on information of the activation function at a single point.

5 Application to Unemployment Time Series

This section takes advantage the NDFA technique to detect presence of latent sources for characterizing the evolution of the unemployment transition in Taiwan. Unemployment and labor market variability typically shows a complex behavior and it is difficult to identify specific patterns in the long run economic cycle. The time series is the monthly registered UR in Taiwan, as published by the Directorate General of Budget, Accounting and Statistics (DGBAS) of Executive Yuan. The series covers the period between January 1978 and June 2010, for 390 values. Figure 2 provides a visual representation of the time series with the y-axis defined as the unemployment rate and the x-axis as the time index. From an eyeball inspection of the plotted series, it seems obvious that this series is nonstationary and seasonal. Table 1 shows the descriptive statistics of the data. The Jarque-Bera test is a goodness-of-fit measure of departure from normality, based on the sample kurtosis and skewness. The wide vertical spaces in the 2000s until the 2010s show high frustration rates throughout the period. A detail discussion of the nonstationarity of the Taiwan unemployment rate can be found in [26].

Table 1. Basic descriptive statistics of the data

Mean	Std Dev	Skew	Kurtosis	C(6)	Max value	Min value	Jarque-Bera
2.79	1.35	0.674	-0.667	-1.330	6.13	0.860	36.6860

The data has enough values for us to apply the techniques developed while it also covers a period of time that is sufficiently homogeneous for us to be able to adequately analyze and characterize the evolution in Taiwan unemployment. According to statistic in Table 1, if the UR random variable sequences obey independent same normal distribution, its Skewness should equal zero and Kurtosis is three. If the random variable sequences obey normal distribution, the Jarque-Bera statistic should obey a chi-square distribution with freedom 2, standard value of which is 9.21 and 5.99 under 1% and 5% significance level respectively. From Table 1, we see the distribution of UR far deviates from normal distribution and the shape is heavy-tailed. The result of remarkable deviation suggests that UR time series may have nonlinear dynamic structure.

The analysis is conducted in the Matlab environment using the NDFA package[17]. Before conducting the analysis, several parameters needed to be set up. First, we convert the time series into an embedded data vector with proper dimensions and delays to model the dynamics of the time series. As shown in Fig 3, the MI function $I(\tau)$ exhibits a local minimum at $\tau=6$ time steps under four cases of different number of boxes for partition. Thus, we should consider $\tau=6$ to be the best delay time in this study. Fig 4 shows the application of FNN method by maximum norm distance yields an embedding dimension m = 5. The embedded data vector was $x(t)=[x^T(t)x^T(t-6)x^T(t-12)x^T(t-18)x^T(t-24)]$.

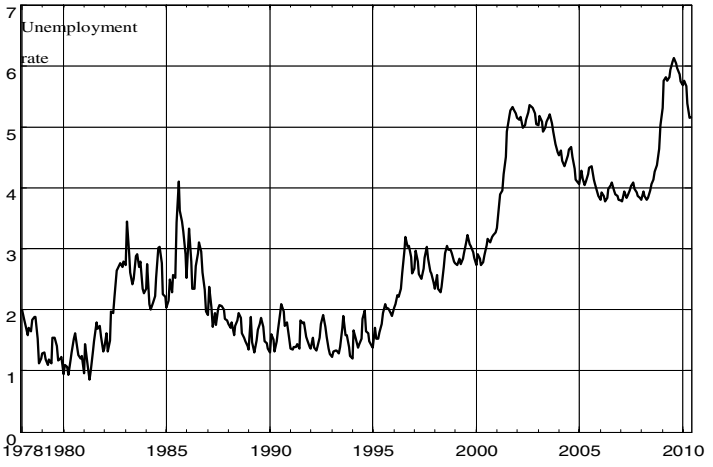


Fig. 2. The time-series representing the historical monthly movement of The UR: Jan 1978 - June 2010

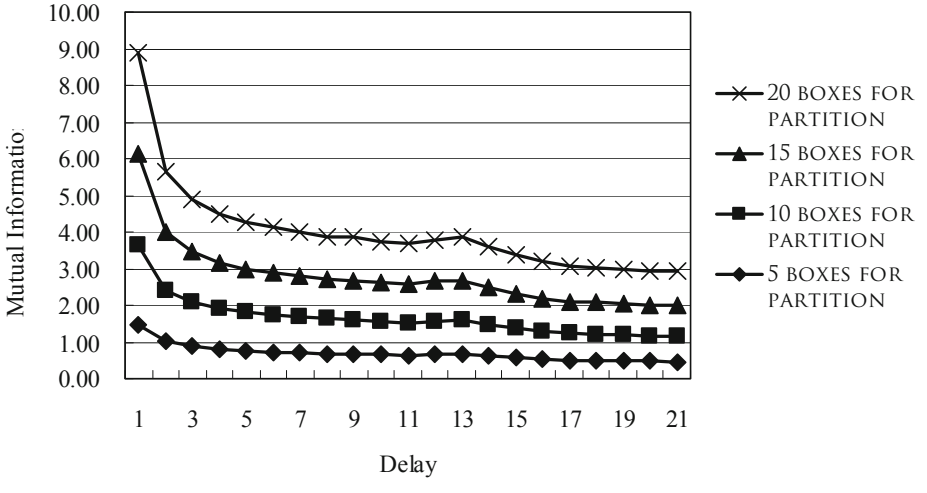


Fig. 3. Mutual information versus delay time

We then estimate the number of sources for searching. Based on the Naik and Kumar [21] method, the determinant of global matrix parameters of ICA is used to estimate the number of search sources. We calculated $|G|=2.6557e\text{-}035\approx0$ which shows several dependent sources are presented. Observing the plot of eigenvalues of covariance matrix in Figure 6, we eliminated three signals from the five observations and set two latent sources for searching.

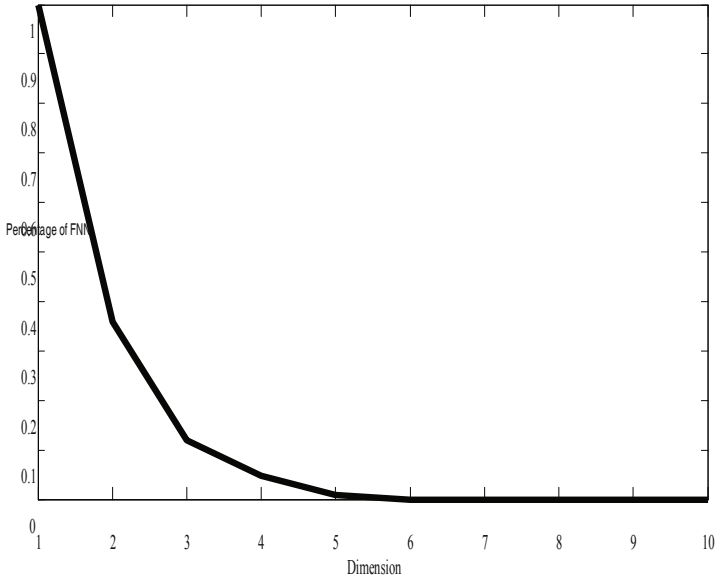


Fig. 4. Percentage of FNN vs m

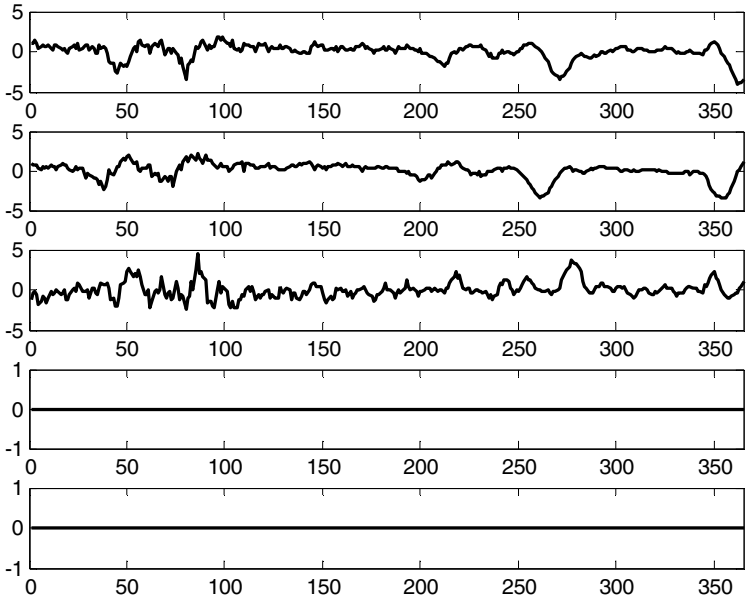


Fig. 5. Sources generated from FastICA

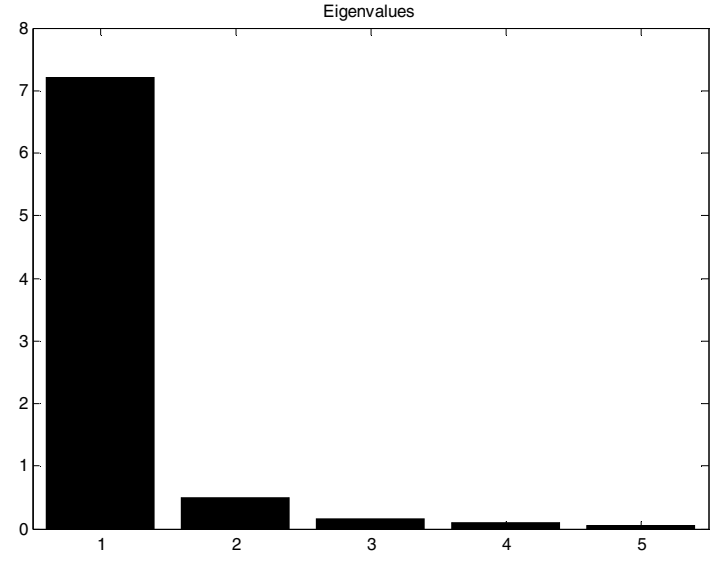


Fig. 6. From the largest to the smallest eigenvalues of covariance matrix

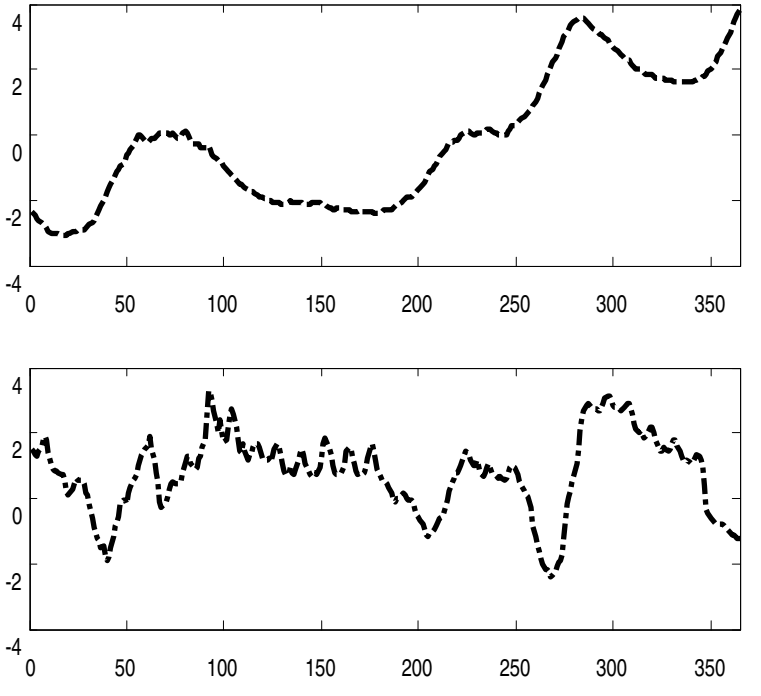


Fig. 7. The two latent sources generated from NDFA

Given two number of search sources, the sources were initialized using PCA. Following the second rule-of-thumb method of the step 3, we set number of both layer hidden neurons to seven. The nonlinear activation is the tanh function. The Gauss-Hermite quadrature approximation is selected as the approximating method for the nonlinearity. We set 500 iterations to run the algorithm. Figure 7 shows the two latent sources of controlling the evolution of Taiwan unemployment rate. The sources have different interpretations. The first latent source drives the unemployment rate in an increasing trend. The second latent source forces the employment market up and down because of some positive and negative forces which include both economic and political issues. Figure 8 shows two latent sources produced from independent component analysis by using FastICA package [22]. These two latent sources reflect similar variation pattern and provide less knowledge for inferring the causality of the time series. Comparing Figure 7 with Figure 8, we found the NDFA can extract one latent trend source and one variation source while the FastICA would extract both two variation sources.

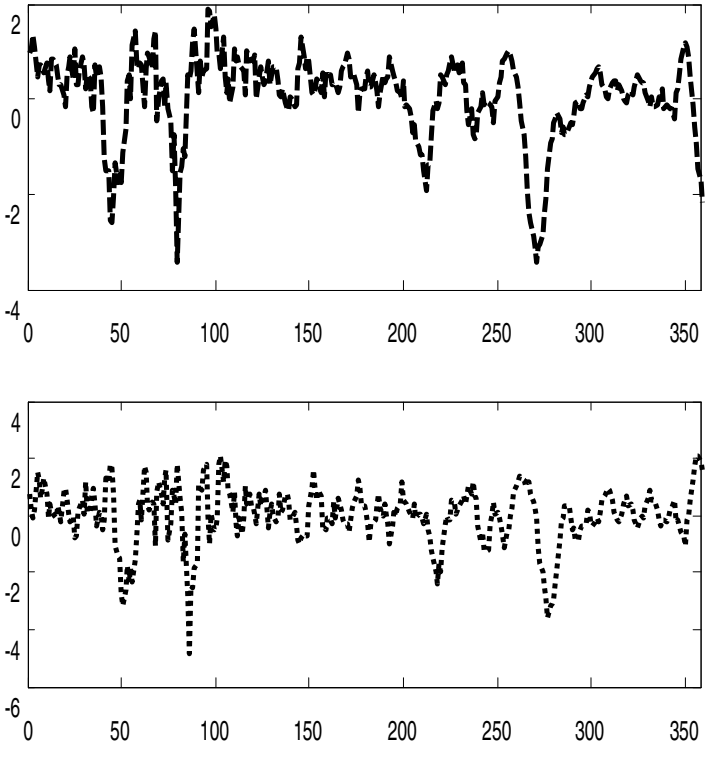


Fig. 8. The two latent sources generated from FastICA

6 Conclusions

It is a tempting alternative to try NDFA on the unemployment time series data. To assume having some underlying independent components and linear transformation in

this specific application may be unrealistic. This paper presents a unified data mining framework for jointly defining process dynamics models and measurements taken on the process. The framework is a state-space model where the process is modeled by the state process and measurements are modeled by the observation process. Parameter estimation and estimation of state process variables can be conducted using MLP with ensemble learning procedures.

Acknowledgements. This work was supported in part by the National Science Council of Republic of China under the NSC-99-2221-E-212-010.

References

1. Makridakis, S.: Time series prediction: Forecasting the future and understanding the past. In: Weigend, A.S., Gershenfeld, N.A. (eds.), p. 643. Addison-Wesley Publishing Company, Reading (1993), ISBN 0-201-62; International Journal of Forecasting 10, 463–466 (1994)
2. Hu, X., Xu, P., Wu, S., Asgari, S., Bergsneider, M.: A data mining framework for time series estimation. Journal of Biomedical Informatics 43, 190–199 (2010)
3. Chen, C.T.: Linear System Theory and Design, 3rd edn. Oxford University Press, New York (1999)
4. Everitt, B.S., Dunn, G.: Applied Multivariate Data Analysis. Oxford University Press, New York (1992)
5. West, M., Harrison, J.: Bayesian Forecasting and Dynamic Models. Springer, New York (1990)
6. De Jong, P.: The diffuse Kalman filter Annals of Statistics 19 (1991)
7. Anderson, B.D.D., Moore, J.B.: Optimal filtering. Prentice-Hall, Englewood Cliffs (1979)
8. Ilin, A., Valpola, H., Oja, E.: Nonlinear dynamical factor analysis for state change detection. IEEE Transactions on Neural Networks 15, 559–575 (2004)
9. Overschee, P.v., Moor, B.D.: Subspace Identification for Linear Systems: Theory, Implementation Applications. Springer, Heidelberg (1996)
10. Quach, M., Brunel, N., d'Alché-Buc, F.: Estimating parameters and hidden variables in nonlinear state-space models based on ODEs for biological networks inference. Bioinformatics (2007)
11. Lappalainen, H., Honkela, A.: Bayesian Nonlinear Independent Component Analysis by Multi-Layer Perceptrons. In: Girolami, M. (ed.) Advances in Independent Component Analysis, pp. 93–121. Springer, Heidelberg (2000)
12. Valpola, H., Karhunen, J.: An unsupervised ensemble learning method for nonlinear dynamic state-space models. Neural Comput. 14, 2647–2692 (2002)
13. Giannakopoulos, X., Valpola, H.: Nonlinear dynamical factor analysis. In: Bayesian Inference And Maximum Entropy Methods in Science And Engineering: 20th International Workshop. AIP Conference Proceedings, vol. 568 (2001)
14. Barber, D., Bishop, C. (eds.): Ensemble learning in Bayesian neural networks. Springer, Berlin (1998)
15. Giannakopoulos, X., Valpola, H.: Nonlinear dynamical factor analysis. In: AIP Conference Proceedings, vol. 568, p. 305 (2001)
16. Honkela, A., Valpola, H.: Unsupervised variational Bayesian learning of nonlinear models. In: Saul, L.K., Weis, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems (NIPS 2004), vol. 17, pp. 593–600 (2005)

17. Valpola, H., Honkela, A., Giannakopoulos, X.: Matlab Codes for the NFA and NDFA Algorithms (2002), http://www.cis.hut.fi/projects/bayes/
18. Takens, F.: Detecting strange attractors in turbulence. LNM, vol. 898, pp. 366–381. Springer, Heidelberg (1981)
19. Fraser, A.M., Swinney, H.L.: Independent coordinates for strange attractors from mutual information. Physical Review A 33, 1134 (1986)
20. Sprott, J.C.: Chaos and Time Series Analysis, vol. 507. Oxford University Press, Oxford (2003)
21. Naik, G.R., Kumar, D.K.: Determining Number of Independent Sources in Undercomplete Mixture. EURASIP Journal on Advances in Signal Processing 5, Article ID 694850 (2009), doi:10.1155/2009/694850
22. Gävert, H., Hurri, J., Särelä, J., Hyvärinen, A.: FastICA Package (2005), http://www.cis.hut.fi/projects/ica/fastica/code/dlcode.shtml
23. Everson, R., Roberts, S.: Inferring the eigenvalues of covariance matrices from limited, noisy data. IEEE Transactions on Signal Processing 48, 2083–2091 (2000)
24. Santos, J.e.D.A., Barreto, G.A., Medeiros, C.a.M.S.: Estimating the Number of Hidden Neurons of the MLP Using Singular Value Decomposition and Principal Components Analysis: A Novel Approach. In: 2010 Eleventh Brazilian Symposium on Neural Networks, pp. 19–24 (2010)
25. Honkela, A.: Approximating Nonlinear Transformations of Probability Distributions for Nonlinear Independent Component Analysis. In: Proceedings of the 2004 IEEE International Joint Conference on Neural Networks (IJCNN 2004), Budapest, Hungary, pp. 2169–2174 (2004)
26. Chen, W.-S.: Use of recurrence plot and recurrence quantification analysis in Taiwan unemployment rate time series. Physica A: Statistical Mechanics and its Applications (in Press, 2011), doi:10.1016/j.physa.2010.12.020

Time Series Subsequence Matching Based on a Combination of PIP and Clipping

Thanh Son Nguyen and Tuan Anh Duong

Faculty of Computer Science and Engineering,
Ho Chi Minh City University of Technology
dtanh@cse.hcmut.edu.vn

Abstract. Subsequence matching is a non-trivial task in time series data mining. In this paper, we introduce our proposed approach for solving subsequence matching which is based on IPIP, our new method for time series dimensionality reduction. The IPIP method is a combination of PIP (Perceptually Important Points) method and clipping technique in order that the new method not only satisfies the lower bounding condition, but also provides a bit level representation for time series. Furthermore, we can make IPIP indexable by showing that a time series compressed by IPIP can be indexed with the support of Skyline index. Our experiments show that our IPIP method is better than PAA in terms of tightness of lower bound and pruning power, and in subsequence matching, IPIP with Skyline index can perform faster than PAA based on traditional R*- tree.

1 Introduction

A basic task in time series data mining which has received an increasing amount of attention lately is the problem of similarity search in time series databases. The similarity search problem is classified into two classes: whole sequence matching and subsequence matching. In the whole sequence matching, it is supposed that the time series to be compared have the same length. In the subsequence matching, the result of this problem is a consecutive subsequence within the longer time series data that best matches the query sequence.

Many solution methods have been proposed for time series similarity search problem. The most promising one is the method that carries out similarity search in two steps: reducing on the dimensionality of time series data, then indexing the reduced data with a multidimensional index structure.

Some typical methods for dimensionality reduction in time series include the Discrete Fourier Transform (DFT) ([5]), Piecewise Aggregate Approximation (PAA) [7], and Adaptive Piecewise Constant Approximation (APCA) [8]. In addition, a special group of time series dimensionality reduction techniques which are based on important points have also been developed. This group includes the method based on Landmark points, proposed by Perng et al., 2000 [10], Extrema points by Fink et al., 2002 ([4]) and Perceptually Important Points (PIP) by Chung et al., 2001 ([2],[3]). These techniques have some interesting features. First, these techniques can retain the important points (i.e. salient points) in the original time series even under a high

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 149–158, 2011.

compression ratio. Secondly, they can deal with comparing time series sequences with different lengths in the database. Thirdly, these time series compression methods are effective to be applied to real-life applications, especially in financial time series. Through our initial study, among the three important points methods, PIP is the most effective and easiest to implement. However, all the three important points methods have not been proved that they can conform to the lower bounding condition which guarantees a time series dimensionality reduction method not incur false dismissals. Besides, these methods do not have indexing mechanisms to support similarity queries, which prevent them from efficiently searching over very large time series databases.

In this paper, we introduce a subsequence matching approach which is based on IPIP, our new method for time series dimensionality reduction. The IPIP method is a combination of PIP and clipping technique in order that the new method not only satisfies the lower bounding condition for time series dimensionality reduction methods but also provides a bit level representation for time series that allows the user to choose compression ratio. Furthermore, we can make IPIP indexable by showing that a time series subsequence compressed by IPIP can be indexed with a multidimensional index structure based on Skyline index. Our experiments show that in subsequence matching, our IPIP method is better than PAA in terms of tightness of lower bound and pruning power, and IPIP with the support of Skyline index can perform faster than PAA with R*-tree.

2 Preliminaries

2.1 Lower Bound Condition

A time series C of length n can be considered a vector or point in the n-dimensional space. Many techniques exist for indexing such data. A general solution described by Faloutsos et al., 1994 [5] is to extract a low-dimensional feature from each time series, and to index the feature space. An important result given in [5] is the proof that in order to guarantee no false dismissals, the distance measure used in the feature space must lower bound the true distance measure. This condition is called the *lower bounding* lemma.

2.2 Index Structures

The popular multidimensional index structures are R-tree and its variants ([1], [6]). In a multidimensional index structure (e.g., R-tree or R^*-tree), each node is associated with a minimum bounding rectangle (MBR). A MBR at a node is the minimum bounding box of the MBRs of its child nodes. A potential weakness in the method using MBR is that MBRs in index nodes can overlap. Overlapping rectangles could have negative effect on the search performance. Besides, another problem in the method using MBR is that summarizing data in MBRs, the sequence nature of time series is not captured.

Skyline Index, another elegant paradigm for indexing time series data which uses another kind of minimum bounding regions, is proposed by Li et al., 2004 [9]. Skyline Index adopts new Skyline Bounding Regions (SBR) to approximate and

represent a group of time series data according to their collective shape. An SBR is defined in the same *time-value* space where time series data are defined. Therefore, SBRs can capture the sequential nature of time series. SBRs allow us to define a distance function that tightly lower-bounds the distance between a query and a group of time series data. SBRs are free of internal overlaps. Hence using the same amount of space in an index node, SBR defined a better bounding region. For k-nearest-neighbor (KNN) queries, Skyline index approach can be coupled with some well-known dimensionality reduction technique such as APCA and improve its performance by up to a factor of 3 ([9]).

2.3 Perpetually Important Point (PIP) Method

The PIP method is an important point method for time series compression, proposed by Chung et al., 2001 [2]. The PIPs extracted from a time series are identified by the following algorithm.

Let a time series $C = \{c_1, \ldots, c_n\}$. The first two PIPs are c_1 and c_n. The third PIP is the point in C with maximum distance to the first two PIPs. The next PIP is the point in C with maximum distance to its two adjacent PIPs and so on. The PIPs identified in the earlier iterations are considered to be more important than those points identified later.

Figure 1.a illustrates the process of identification of 5 PIPs. The distance metric used in the PIP identifying process is the *vertical distance* between the test point and the line connecting the two adjacent PIPs. To determine the maximum distance to the two adjacent PIPs (x_1, y_1) and (x_2, y_2), as described in Figure 1.b, is the vertical distance (VD) between the test point p_3 and the line connecting the two adjacent PIPs, i.e.,

$$VD(p_3, p_c) = |y_c - y_3| = |(y_1 + (y_2 - y_1)(x_c - x_1)/(x_2 - x_1)) - y_3|$$

where $x_c = x_3$.

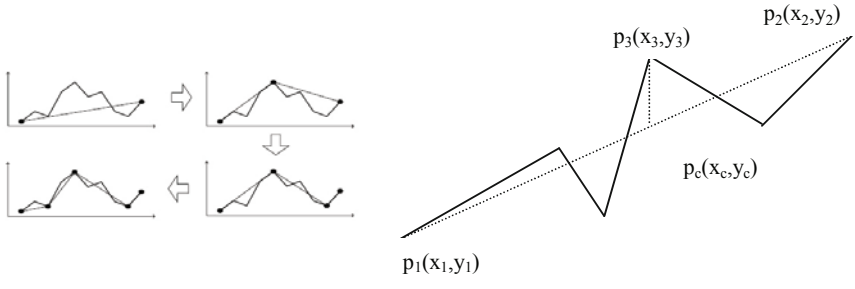


Fig. 1. (a) PIP identification process (b) Vertical distance measure

Storing and Retrieving PIPs in Subsequences Based on SB-tree. In subsequence matching, we have to slide a window of length w across a long time series and identify the necessary PIPs in each subsequence formed during this process. However, this task is time consuming. To get a fast retrieval of PIPs from subsequences, an approximate approach based on a Specialized Binary (SB) tree is proposed by Fu et al. [3] in which PIPs of subsequences can be retrieved according to their data point importance from the SB-tree of the corresponding time series without any re-calculation. With the starting and ending points of a subsequence and the number of PIPs needed to retrieve are given, we can easily retrieve the approximate PIPs of the subsequence based on their positions in the SB-tree by accessing the tree. The PIP retrieving algorithm in subsequences can be seen in [3].

3 IPIP Representation

3.1 IPIP – Compression and Clipping

Given a time series subsequence C and a query Q, without loss of generality, we assume C and Q are n units long. C is divided into segments. We use the PIPs identifying algorithm to choose l PIPs in each segment of C. Next, these PIPs are transformed into a sequence of bits, where 1 represents above the segment average and 0 represents below, i.e., if μ is the mean of segment C, then

$$c_t = \begin{cases} 1 & \text{if } c_t > \mu \\ 0 & \text{otherwise} \end{cases}$$

Figure 2 shows the intuition behind this technique, with l = 5. In this case, the sequence of bit 00100 and the μ value are recorded.

Note that this kind of bit level time series representation has been first considered by Ratanamahatana et al., 2005 [11]. This bit level representation for time series has four main advantages: (i) the clipped distance is less than or equal to the Euclidean distance, (ii) the bit level representation can be efficiently stored and manipulated, (iii) the representation allows time series clustering to scale to much larger datasets, and (iv) the bit level representation allows us to use many available algorithms which

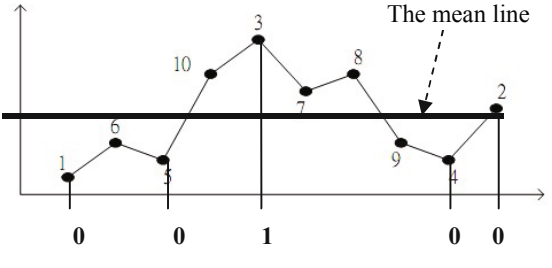


Fig. 2. An illustration of IPIP method

are applicable to binary data only. But clipping has one disadvantage: the data itself dictates the compression ratio and the user has no choice to make.

In order to match the query Q with a time series C in the database, first Q is transformed into the same feature space as C. But the PIPs of Q are not transformed into a sequence of bits since the clipped representation allows raw data to be directly compared to the clipped representation. Then we shift the segment mean lines in Q to meet those in C so that we can compare the similarity in shape between Q and C. The computation of the distance between Q and C', the IPIP representation of C, will be described later.

3.2 Similarity Measure Defined for IPIP

In order to guarantee no false dismissals we must produce a distance measure in the reduced space, D_{IPIP} which is less than or equal to the distance measure in the original space.

Definition 1 (IPIP Similarity Measure). Given a query Q and a subsequence C (of length n) in raw data. Both C and Q are divided into N segments ($N << n$). Suppose each segment has the length of w. Let C' be an IPIP representation of C. The distance measure between Q and C' in IPIP space, $D_{IPIP}(Q, C')$, is computed as follows.

$$D_{IPIP}(Q,C') = \sqrt{D_1(Q,C') + D_2(Q,C')} \tag{1}$$

$D_1(Q, C')$ and $D_2(Q, C')$ are defined as

$$D_1(Q,C') = \sum_{i=1}^{N} w(q\mu_i - c\mu_i)^2 \tag{2}$$

$$D_2(Q,C') = \sum_{j=1}^{N} \sum_{i=1}^{l} (d(q_i, bc_i))^2 \tag{3}$$

where
$q\mu_i$ is the mean value of the i-th segment in Q, $c\mu_i$ is the mean value of the i-th segment in C, bc_i is binary representation of c_i. $d(q_i, bc_i)$ is computed by the following formula:

$$d(q_i, bc_i) = \begin{cases} q_i' & \text{if } (q_i' > 0 \text{ and } bc_i = 0) \\ & \text{or} \\ & (q_i' \le 0 \text{ and } bc_i = 1) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

q_i' is defined as $q_i' = q_i - q\mu_k$, where q_i belongs to the k^{th} segment in Q.

Lemma 1. If $D(Q, C)$ is the Euclidean distance between query Q and time series C, then $D_{IPIP}(Q, C') \le D(Q, C)$.

The proof of Lemma 1 can be seen in our previous paper [12].

4 A Skyline Index for IPIP

In this section, we describe how we can adopt Skyline index [9] for IPIP time series compression method. First, we introduce the concept of the IPIP Bounding Region (IBR). Then, we describe the use of IBRs for indexing and searching time series data.

4.1 IPIP Bounding Regions

In traditional multidimensional index structure such as R*-tree, minimum bounding rectangles (MBRs) are used to group time series data which are mapped into points in a low dimensional feature-space. If a MBR is defined in the two-dimensional space in which a time series data exists, the overlap between MBRs will be large. So by using the ideas from Skyline index, we can represent more accurately the collective shape of a group of time series data with tighter bounding regions. To attain this aim, we use IPIP bounding regions (IBRs) for bounding a group of time series data.

Definition 2 (IPIP Bounding Region). Given a group C' consisting of k IPIP sequences in a N-dimensional feature space. The IBR R of C', is defined as

$$R = (C'_{max}, C'_{min})$$

where $C'_{max} = \{c'_{1max}, c'_{2max},\ldots,c'_{Nmax}\}$, $C'_{min} = \{c'_{1min}, c'_{2min},\ldots,c'_{Nmin}\}$
and, for $1 \le i \le N$, $c'_{imax} = \max\{c'_{i1}, \ldots, c'_{ik}\}$ and $c'_{imin} = \min\{c'_{i1}, \ldots, c'_{ik}\}$
where c'_{ij} is the i^{th} mean value of the j^{th} IPIP sequence in C'.

Figure 3 illustrates an example of IBR. In this example, BC_i is a bit sequence of time series C_i and the number of PIPs in each segment is three.

4.2 Time Series Indexing Based on IBRs

Once the Skyline Index for IPIP has been built, we have to define the distance function $D_{region}(Q, R)$ of the query Q from the IBR R associated with a node in the index structure such that it satisfies the group lower-bound condition $D_{region}(Q, R) \le D(Q, C)$,

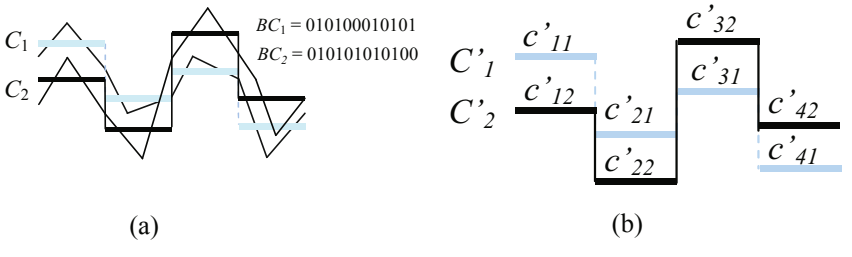


Fig. 3. An example of IBR. (a)Two time series C_1, C_2 and their approximate IPIP representations in four dimensional space. (b) The IBR of two IPIP sequences C'_1 and C'_2. $C'_{max} = \{c'_{11}, c'_{21}, c'_{32}, c'_{42}\}$ and $C'_{min} = \{c'_{12}, c'_{22}, c'_{31}, c'_{41}\}$.

for any time series C in the IBR R. The proof of this group lower-bound condition is given in our previous work [12].

We can index the IPIP representation of time series data by first building a Skyline index which based on a spatial index structure such as R^*-tree [1]. Each leaf node in the R^*-tree-based Skyline index contains an IPIP sequence and a pointer refer to an original time series data in the database. The IBR associated with a non-leaf node is the smallest bounding region that spatially contains the IBRs associated with its immediate children.

Two searching problems which we apply in our experiments are ε-range search and KNN search algorithms. These algorithms are similar to those described in the paper [8], by Keogh et al., 2001.

5 Subsequence Matching Based on IPIP and Skyline Index

Figure 4 shows the outline of our subsequence matching process. For simplicity, we assume that the query sequence Q has the same length w of the sliding window. It is necessary to normalize C and Q so that the comparison between sequences in different "amplitude" ranges can be facilitated.

Algorithm Subsequence_Matching

Inputs: time series C, n (number of PIPs to be extracted from C), query sequence Q and tolerance ε

Outputs: all the subsequences in C of which are in ε-match with Q.

Algorithm:

1. Index Building
   - 1.1 Use PIP identification process to identify n PIPs and store them on the SB-tree.
   - 1.2 Use a sliding window of size w to divide the time series C into subsequences of length w from C, extract k PIPs from each subsequence and apply IPIP transformation on each such subsequence. Store the features transformed from all such subsequences in Skyline index.
2. Index searching
   - 2.1 Apply the PIP transformation on query sequence Q.
   - 2.2 Search Skyline index to find the candidate set of the subsequences on C of which are in ε-match with Q.
3. Postprocessing.
   - 3.1 Examine the original subsequences of the time series C which correspond to the candidate set obtained at step 3.2 to discard the false alarms.

Fig. 4. The subsequence matching algorithm

In the algorithm, we deal with the basic case: the length of the query sequence is equal to the window size. As for the two other cases: (1) the query sequence is composed of exactly p (≥ 1) disjoint windows and (2) the query sequence has a remainder when it is divided into p disjoint windows, we apply the ideas from the "*PrefixSearch*" method and "*Multipiece*" methods proposed by Faloutsos et al. in [5].

6 Experimental Results

In this section we report the experimental results on subsequence matching using IPIP dimensionality reduction technique. We compare our proposed technique IPIP using Skyline index to the popular method PAA based on R^*-tree.

We perform all tests over different reduction ratios, different numbers of chosen PIPs and datasets of different lengths. We consider a length of 1024 to be the longest query. Time series datasets for experiments come from various sources publicly available through the Internet and are organized into five separate datasets. The five datasets are Stock-data (37MB - over two million points), Consumer-data (27MB), FederalFund (24MB), Hydrology (30MB), and DiscordAnomaly (20MB). The comparison between IPIP and PAA is based on the tightness of lower bound, the pruning power and the implemented system. Besides, we also compare the index building time of IPIP and PAA.

The Tightness of Lower Bound. The tightness of lower bound (T) is used to evaluate preliminary effect of a dimensionality reduction technique. It is computed as follows

$$T = D_{feature}(Q', C')/D(Q, C)$$

where $D_{feature}(Q', C')$ is the distance between Q' and C' in reduced space and $D(Q, C)$ is the distance between original time series Q and C.

In order to evaluate of the tightness of lower bound, we experiment over different values of reduction ratios: 128, 64, 32, 16, and a range of the number of extracted PIPs from 8 to128. Due to the space limit, we show the results of this experiment on Stock dataset only in Figure 5. In Figure 5, the horizontal axis is the number of chosen PIPs and the vertical axis is the tightness of lower bound. Based on these experimental results, we can see that lower bound of the IPIP technique is higher (i.e. tighter) than that of PAA.

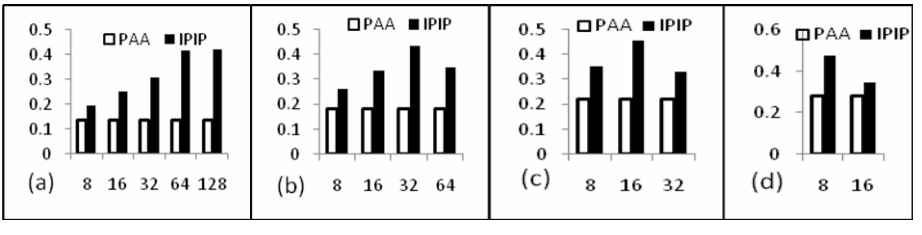


Fig. 5. The experiment results on tightness of lower bound - Stock dataset

Pruning Power. In order to compare the effectiveness of two dimensionality reduction techniques, we need to compare their pruning powers. Pruning power P is the fraction of the database that must be examined before we can guarantee that an ε-match to a query has been found. This ratio is based on the number of times we cannot perform similarity search on the transformed data and have to check directly on the original data to find nearest match.

$$P = \frac{Number\ of\ sequences\ that\ must\ be\ examined}{Number\ of\ sequences\ in\ database}$$

Figure 6 shows the experimental results of *P*. In the charts of Figure 6, the horizontal axis represents the value of reduction ratios and the vertical axis represents the pruning power. When the value of *P* becomes smaller, approaching to 0, the querying method is more effective. Based on these experimental results, we can observe that the pruning power of IPIP technique is better than that of PAA.

Notice that tightness of lower bound and pruning power of a time series dimensionality reduction method are *independent* of the used index structure.

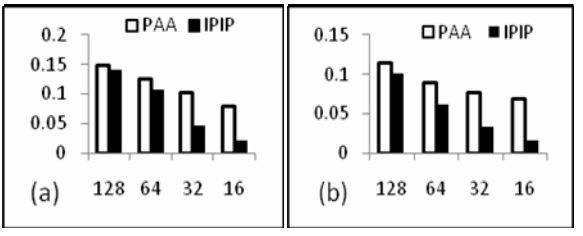


Fig. 6. The pruning powers on Stock data, tested over a range of reduction ratios (16-128) and query lengths (1024 (a), 512 (b))

Implemented System. We also need to compare IPIP to PAA in terms of implemented systems. The implemented system experiment is evaluated on the normalized CPU cost which is the fraction of the average CPU time to perform a query using the index to the average CPU time required to perform a sequential search. The normalized CPU cost of a sequential search is 1.0.

The experiments have been performed over a range of query lengths (256-1024) and values of reduction ratios (8-128). For brevity, we show just two typical results. Figure 7 shows the experiment results with a fixed query length 1024.

Between the two competing techniques, in subsequence matching the IPIP technique using Skyline index performs faster than PAA based on R*-tree.

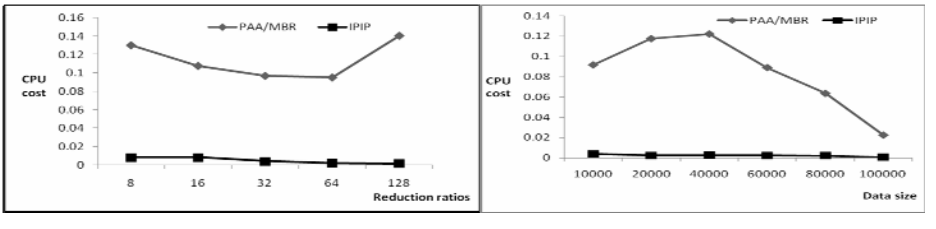


Fig. 7. CPU cost of IPIP and PAA over (a) a range of reduction ratios. (b) a range of data sizes.

Building the Index. We also compare IPIP to PAA in terms of the time taken to build the index. The experimental results over a range of reduction ratios (8-128) show that for low reduction ratios, the index building time of IPIP using Skyline index is shorter than that of PAA. However, for high reduction ratios (e.g. above 128), the index building time of the two approaches are approximately the same.

7 Conclusions

We introduced a subsequence matching approach which is based on IPIP, our new method for time series dimensionality reduction. IPIP is a combination of PIP and clipping technique in order that the new method not only satisfies the lower bounding condition for time series dimensionality reduction but also provides a bit level representation for time series that allows the user to choose compression ratio. In other words, IPIP combines the strengths of PIP and clipping but overcomes the weaknesses of both methods. Besides, we can make IPIP indexable by showing that time series subsequences compressed by IPIP can be indexed with Skyline index. Experimental results demonstrate that our IPIP method is better than PAA in terms of tightness of lower bound and pruning power, and in subsequence matching, IPIP with the support of Skyline index can perform faster than PAA with R*-tree.

References

1. Beckman, N., Kriegel, H.P., Schneider, R., Seeger, B.: The R*-tree: An Efficient and Robust Access Method for Points and Rectangles. In: Proc. of 1990 ACM-SIGMOD Conf., Atlantic City, NJ, pp. 322–331 (May 1990)
2. Chung, F.L., Fu, T.C., Luk, R., Ng, V.: Flexible Time Series Pattern Matching Based on Perceptually Important Points. In: Proc. of Int. Joint Conf. on Artificial Intelligence-Workshop on Learning from Temporal and Spatial Data, pp. 1–7 (2001)
3. Fu, T.C., Chan, H.P., Chung, F.L., Ng, C.M.: Time Series Subsequence Searching in Specialized Binary Tree. In: Wang, L., Jiao, L., Shi, G., Li, X., Liu, J. (eds.) FSKD 2006. LNCS (LNAI), vol. 4223, pp. 568–577. Springer, Heidelberg (2006)
4. Fink, E., Pratt, K.B.: Indexing of Compressed Time Series. In: Last, M., Kandel, A., Bunke, H. (eds.) Data mining in time series Databases, pp. 43–65. World Scientific, Singapore (2003)
5. Faloutsos, C., Ranganathan, M., Manolopoulos, Y.: Fast Subsequence Matching in Time-Series Databases. In: Proc. of ACM SIGMOD Int'l Conf. on Management of Data, Minneapolis, MN, May 24-27, pp. 419–429 (1994)
6. Guttman, A.: R-trees: a Dynamic Index Structure for Spatial Searching. In: Proc. of the ACM SIGMOD Int. Conf. on Management of Data, June 18-21, pp. 47–57 (1984)
7. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. Knowledge and Information Systems 3(3), 263–286 (2000)
8. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases. In: Proc. of ACM SIGMOD Conf. on Management of Data, Santa Barbara, CA, May 21-24, pp. 151–162 (2001)
9. Li, Q., Lopez, I.F.V., Moon, B.: Skyline Index for Time Series Data. IEEE Trans. on Knowledge and Data Engineering 16(6) (2004)
10. Perng, C.S., Wang, H., Zhang, S.R., Parker, D.S.: Landmarks: A New Model for Similarity-Based Pattern Querying in Time Series Databases. In: Proc. of the IEEE Int'l Conf. on Data Engineering, pp. 33–42 (2000)
11. Ratanamahatana, C.A., Keogh, E., Bagnall, A.J., Lonardi, S.: A Novel Bit Level Time Series Representation with Implications for Similarity Search and Clustering. In: Ho, T.-B., Cheung, D., Liu, H. (eds.) PAKDD 2005. LNCS (LNAI), vol. 3518, pp. 771–777. Springer, Heidelberg (2005)
12. Son, N.T., Anh, D.T.: An Improvement of PIP for Time Series Dimensionality Reduction and Its Index Structure. In: Proc. of Int. Conf. on Knowledge and Systems Engineering (KSE 2010), Hanoi, Vietnam, October 7-9, pp. 47–54 (2010)

Cloud Intelligent Services for Calculating Emissions and Costs of Air Pollutants and Greenhouse Gases

Thanh Binh Nguyen, Fabian Wagner, and Wolfgang Schoepp

International Institute for Applied Systems Analysis (IIASA)
Schlossplatz 1
A-2361 Laxenburg, Austria
Tel.: (+43 2236) 71 327
{nguyenb,wagnerf,schoepp}@iiasa.ac.at

Abstract. The GAINS (Greenhouse gas – Air pollution Interactions and Synergies) model quantifies the full DPSIR (demand-pressure-state-impact-response) chain for the emissions of air pollutants and greenhouse gases. To fulfill regional specific requirements of the GAINS model, we have studied and developed a cloud intelligent service system for calculating emissions and costs for reducing emissions at regional as well as global levels. In this paper, first we present a cloud intelligent conceptual model that is used to specify an application framework, namely GAINS cloud intelligent application framework. Using this application framework, first we build a global data warehouse called GAINS DWH World, then a class of regional data warehouses, e.g. GAINS DWH Europe, GAINS DWH Asia, etc, are specified and used for regional data analysis and cost optimization.

Keywords: GAINS, Cloud intelligence, Data Warehouse.

1 Introduction

IIASA's Greenhouse gas – Air Pollution Interactions and Synergies (GAINS) model explores synergies and trade-offs between the control of local and regional air pollution and the mitigation of global greenhouse gas (GHG) emissions [5,7,14]. The GAINS model estimates emissions, mitigation potentials and costs for five air pollutants (SO_2, NO_x, PM, NH_3, VOC) and for the greenhouse gases included in the Kyoto protocol. Historic emissions of air pollutants and GHGs are estimated for each country based on information collected by available international emission inventories and on national information supplied by individual countries.

Geographically the model was initially developed for European countries [7]. In 1994 the World Bank and IIASA started the RAINS Asia project (IIASA Options, 1993) using the application logic for Asian countries as well [5]. When the model became more popular, individual countries became interested in modeling their national pollution scenarios. During 2004 regional models for Italy and the Netherlands were developed. Even with the new versions, the model was still being developed on standalone schemas. This allowed easy modeling of certain regional aspects that had

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 159–168, 2011.

to be embedded into the model, but it also increased the maintenance efforts: Every change in the common structure of the model had to be introduced in each schema separately.

To solve the above challenge, cloud computing, which is a new computing paradigm to provide reliable, customized and QoS guaranteed dynamic computing environments for end-users [15], is considered as an option. Cloud computing delivers infrastructure, platform, and software (application) as services, which are made available as subscription-based services in a pay-as-you-go model to consumers. These services in industry are respectively referred to as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) [16]. According to [8], cloud computing, the long-held dream of computing as a utility, has the potential to transform a large part of the IT industry, making software even more attractive as a service.

In addition, data warehouse and business intelligent (BI) [1,2,4,9,10,11,12] are the processes of gathering enough of the right information in the right manner at the right time, and delivering the right results to the right people for decision-making purposes so that they can continue to yield real business benefits, or have a positive impact on business strategy, tactics, and operations. In our previous paper [1], we developed the business intelligent system called as GAINS-BI (GAINS Business Intelligent) that enables analysis and reporting. In this context, we introduced mathematical models used to calculate emission and cost [14] for a given pollutant, a given GAINS region, and a given year within a given GAINS scenario. These mathematical sound concepts enable us to specify the GAINS-BI conceptual data model as well as to calculate emission and costs in the ETL (Extract-Transform-Load) processes and data cube generation.

In this paper, we present an integrated approach, i.e. a cloud intelligent service system for calculating emissions and costs for reducing emissions to fulfill regional specific as well as global requirements of the GAINS model. In this context, first a GAINS cloud intelligent conceptual model is introduced to specify the application in a very formal manner. Hereafter, the GAINS cloud intelligent system architecture and its main components, i.e. data warehouses, metadata, platform and application services introduced as an application framework. In this top down approach, first, a global data warehouse, namely GAINS World DWH, is developed by integrating data from international emission inventories and national information supplied by individual countries. Then, each new regional data warehouse is built and administrated by using the GAINS data warehouse platform services. The regional data warehouse is then used for regional data analysis and cost optimization by using the GAINS business intelligent services. To illustrate the concepts, some typical examples have been presented. The utilization of the cloud intelligent approach provides a feasible and effective method to improve the ability of building, managing as well as analysis of multi-regional data warehouses.

The rest of this paper is organized as follows: section 2 introduces some approaches and projects related to our work; after an introduction of GAINS cloud data warehouse concepts in section 3, a GAINS cloud data warehouse system architecture and framework are presented and modeled by using UML (Unified Modeling Language) in section 4. This section will also present our implementation results. At last, section 5 gives a summary of what have been achieved and future works.

2 Related Work

The characters of the proposed approach can be rooted in several research areas of new-introduced cloud intelligence research field, including the trends and concepts, the combined use of cloud computing and data warehousing technologies in supporting cloud intelligent systems, as well as its utilization in GAINS. With the amount of data generated in GAINS models increasing continuously, delivering the right and sufficient amount of information at the right time to the right business users has become more complicated and critical [4].

Cloud computing has come up fast with companies such as Amazon, Google and Salesforce.com stealing a march on the information technology infrastructure stalwarts such as HP, IBM, Microsoft, Dell, EMC, Sun and Oracle [6]. The latter are certainly participating, but doing so more behind the scenes notwithstanding some high profile press releases. Providers such as Amazon, Google, Salesforce, IBM, Microsoft, and Sun Microsystems have begun to establish new data centers for hosting Cloud computing application services such as social networking and gaming portals, business applications (e.g., SalesForce.com), media content delivery, and scientific workflows. Actual usage patterns of many real-world application services vary with time, most of the time in unpredictable ways.

More and more enterprise solutions and platforms for Business Intelligence have been developed such as IBM DB2 with Business Intelligence Tools, Microsoft SQL Server, Teradata Warehouse, SAS, iData Analyzer, Oracle, Cognos, Business Objects, etc. [8], have been developed aim to empower businesses by providing direct access to information used to make decisions, create more effective plans and respond more quickly to problems and opportunities [11]. Thus, this approach effectively and efficiently leverages the data resources to satisfy their requirements for analysis, reporting and decision making process.

In the context of the GAINS [5] model, there are several specific questions like "How much would a migration from one technology to another, more effective one, cost and how much emissions would it save?", or "What is the most effective way in terms of use of technologies to save emissions within a given budget?". Questions like this are answered with the help of the GAINS optimization module [14]. Compared to the earlier RAINS model, the GAINS model incorporates now aspects that constitute important interactions with greenhouse gas mitigation strategies. GAINS covers now a wider range of pollutants, and it includes structural changes in the energy systems such as energy efficiency improvements and fuel substitution as means for emission reductions. It models the impacts of emission control measures on multiple pollutants (co-control) for a wide range of mitigation options in the energy and agricultural sectors.

This paper focuses on integrating cloud computing and data warehousing technologies to facilitate the larger reusability of the GAINS data warehousing system architecture. Specifically, we also introduce the modeling of conceptual architectural artifacts and their relationships, which formalization enables the design reusability and consistent development of GAINS cloud data warehouse systems.

3 Cloud-Based GAINS Data Warehousing System

In this section, the two main GAINS cloud data warehousing concepts, i.e. dimensions, decision variables or facts are presented in a formal manner. 7 GAINS dimensions and their domain values are described. Afterwards, decision and derived variables are defined and formulated.

3.1 GAINS Cloud Data Warehousing Conceptual Model

On the basis of mathematical modeling, the architecture of GAINS Cloud Data Warehousing systems (GAINS-CDWH) is formally modeled, with the objectives of setting design alternatives in the context of GAINS-BI previously described. The aim of this conceptual model is to provide an extension of the standard DW architecture used in the literature by cloud-based modeling aspects, and to connect the defined model with GAINS-specific mathematical models. The mathematical model of a GAINS-CDWH can be written in the form:

$$GAINS-CDWH = <\{Dim_n\},\{Fact_m\},\{Calc_o\},\{Service_q\},\{DWH_t\}>$$

where:

- $\{Dim_n\}$=$\{region,pollutant,sector,fuelactivity,technology,year,scenario\}$ is a set of GAINS dimensions.
- $\{Fact_m\}$ is a set of GAINS decision variables.
- $\{Calc_o\}$=$\{\{emissioncal_u\},costcalc\}$ is a set of emission and cost calculation formulas. Emission and cost calculation have been introduced in several publications as [5,7,14].
- $\{Services_q\} = \{platformservices, applicationsevices\}$ is a set of services in the context of GAINS cloud data warehouse framework and used to build, manage as well as analysis data in global and regional data warehouses.
- $\{DWH_t\}$=$\{$GAINS DWH World, GAINS DWH Asia, GAINS DWH Europe,…$\}$ is a set of global and regional data warehouses generated and managed by the GAINS cloud data warehousing system.

In the next sections, we will introduce two main components of GAINS multidimensional data model, i.e. dimension and facts (variables). Descriptions about *GAINS Services* and *Data Warehouses* will be introduced in section 4 in the context of GAINS Cloud Data Warehouse Application Framework.

3.2 Dimensions

In GAINS, *region*,*pollutant*,*sector*,*fuelactivity*,*technology*,*year*, and *scenario* are 7 GAINS dimensions and denoted by $\{Dim_n\}$= {*region*,*pollutant*, *sector*, *fuelactivity*, *technology*,*year*,*scenario*}. Furthermore:

- *I* =dom(*region)* is domain value of dimension *region, i* is a region.
- *P* =dom(*pollutant)* is domain value of dimension *pollutant, p* is a pollutant.
- *S* =dom(*sector)* is domain value of dimension *sector, s* is a sector.

- F =dom(*fuelactivity)* is domain value of dimension *fuelactivity, f* is a fuelactivity.
- T =dom(*technology)* is domain value of dimension *technology, t* is a technology.

3.3 Decision and Derived Variables

Decision variables in GAINS, these variables, denoted by $x_{i,s,f,t}$, describe the level of the activity f in sector s and country i that is controlled by technology Naturally, theses variables can only take non-negative values and the following has to hold: $f \in F_{i,s}$ and also $t \in T_{i,s,f}$. Thus:

$$0 \le x_{i,s,f,t} \text{ , } \forall i \in I \text{ , } \forall s \in S \text{ , } \forall i \in I \text{ , } f \in F_{i,s} \text{ , } t \in T_{i,s,f}$$

Derived variables. There are a number of variables that can be derived from decision variables, which are described as follows:

• *Activity data.* This is the pollutant-specific activity data, and defined as $xp_{i,s,f} = \sum_{t \in T_{s,f,p}} x_{i,s,f,t} \text{ , } \forall p \in P_{s,f}, i \in I, s \in S, f \in F_{i,s}$

• *Application rates/Control strategies*. Having defined the activity data it is possible to derive the application rates $q_{i,s,f,t}$ of individual technologies (the set of application rates of all relevant control technologies is referred to as a 'control strategy') as:

$q_{i,s,f,t} = \frac{x_{i,s,f,t}}{xp_{i,s,f}}$, $\forall i \in I, s \in S, f \in F_{i,s}, t \in T_{s,f}$, so that $0\% \le q_{i,s,f,t} \le 100\%$.

• *Emissions of pollutant p in country i***.** It is relatively easy to calculate the emissions of pollutant p in country i from the decision variables, i.e., the technology specific activity data: $emissions_{i,p} = \left(\sum_{s \in S} \sum_{f \in F_{i,s}} \sum_{t \in T_{s,f,p}} EF^{abated}_{i,s,f,t,p} \bullet x_{i,s,f,t} \right)$

where the abated emission factor $EF^{abated}_{i,s,f,t,p}$ is calculated as:

$$EF^{abated}_{i,s,f,t,p} = EF_{i,s,f,t,p} \bullet (1 - remeff_{i,s,f,t,p})$$

where in turn $EF_{i,s,f,t,p}$ is the unabated emission factor of pollutant p associated with the sector-activity combination (s, f) in country i, and $remeff_{i,s,f,t,p}$ is the removal efficiency for pollutant p associated with technology t.

4 GAINS Cloud Data Warehouse Application Framework

In this section, a cloud-based GAINS data warehousing system architecture is introduced as a framework for specifying based components, including platform and application services, data warehouses, as well as metadata. The framework is then designed by using UML and described in section 4.2.

4.1 GAINS Cloud Data Warehouse System Architecture

Figure 1 shows the GAINS cloud data warehouse system architecture, including four main components, i.e. platform services, data warehouses, application services, and metadata. These components are described as follows:

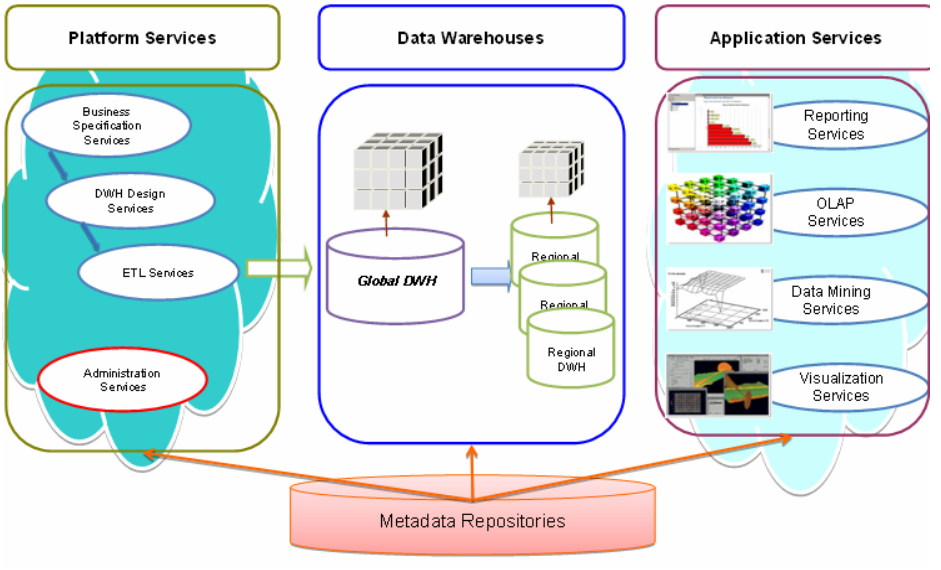


Fig. 1. The GAINS cloud data warehouse system architecture

Global and Regional Data Warehouses. In our approach, first we build a global data warehouse namely GAINS DWH World to hold the global scientific data, i.e. activity, emission and cost data collected by available international emission inventories and calculated in GAINS. Based on the GAINS DWH World, each regional data warehouse will be specified and built as a subset of the GAINS DWH World.

GAINS metadata are developed to contain information of business models, design patterns, and analytical capacities. GAINS metadata contains two following kinds:

- GAINS multidimensional metadata stores basic information about the scientific structure of the GAINS cloud warehouse system. This information includes:
  - Business model contains a name, multidimensional schema, i.e. dimensions, facts, constraints of a regional DWH.
  - Building service metadata holds the design pattern, mapping, and information for the ETL (extract, transformation and loading) process to build a regional data warehouse. This information has to be accessed and referenced by global as well as regional data warehouses as it is the basis for almost every calculation of the system.

  - Administration service metadata is description about a regional GAINS DWH, e.g. privilege, access, log information.

- The GAINS application metadata holds all necessary information to manage analytical functions of the scientific data stored in a regional DWH.

Platform Services are specified for GAINS cloud data warehouse system to provide abilities to build and manage regional data warehouses. There are two main groups of platform services, namely business services and development services. The full descriptions of platform services will be presented in 4.2.

Application services can be seen as report generation, data mining tasks, OLAP, and visualization functions. In the GAINS cloud-DWH system context, they are GAINS SaaS and on-demand services, which will be presented in the next section.

4.2 Design GAINS Cloud Data Warehouse Application Framework with UML

Figure 2 shows an overview about the GAINS cloud data warehouse application framework modeled by using UML. The GAINS infrastructure is developed based on Oracle, Java, and UNIX. Such processing, storage, networks, and other fundamental computing capabilities are not presented in this paper since we would like to focus in the platform and application services, which are specified as follows:

GAINS platform as a Service, provided to a GAINS data warehouse vendor, e.g. a country, is to deploy onto the GAINS cloud data warehouse application framework

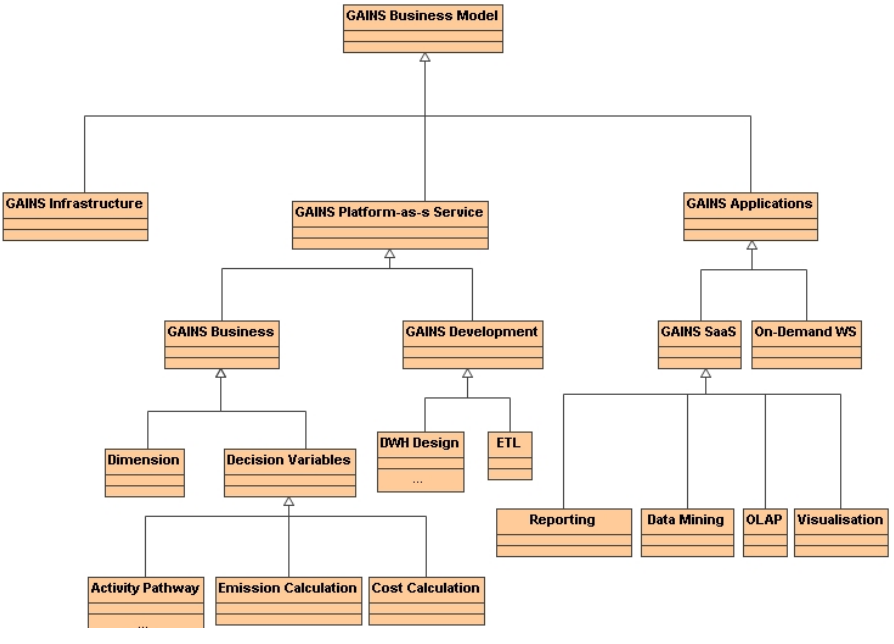


Fig. 2. Modeling GAINS cloud data warehouse application framework with UML

a regional data warehouse by using the GAINS platform services. The following steps show how the GAINS vendor has control over the deployed regional data warehouse and possibly application hosting environment configurations.

- *Business Services.* This step is used to specify multidimensional model, i.e. modeling dimensions, facts, and specifying constraints. This configuration is considered as a multidimensional schema and holds business model and rules of a regional data warehouse.
- *Development services*. Based on the business model, a regional data warehouse will be designed. All detail structural components, e.g. attributes, hierarchies, mappings, as well as transformation and calculation operators are designed or assigned in this step.

GAINS Application Services, provided to GAINS user(s), is to use global or regional data warehouses running on the GAINS cloud data warehouse system. There are two typical services described as follows:

- *GAINS DWH SaaS* is a set of GAINS data warehouse application services, such as reporting, data mining, OLAP, and visualization services, running on the GAINS cloud data warehouse infrastructure.
- *On-demand services* are also known as "software on demand" [16], the GAINS on-demand services allows our vendors to develop, host and operate software for their usage purposes, e.g. GAINS DWH Italia, GAINS DWH Netherland, etc.

4.3 Implementation Results

IIASA's GAINS model explores cost-effective emission control strategies to reduce greenhouse gases and/or improve local and regional air quality. By selecting a smart mix of measures, countries can reduce air pollution control costs as well as cut greenhouse gas emissions.

The above figure shows the GAINS cloud data warehouse portal which provides access to a number of GAINS global and regional data warehouses including:

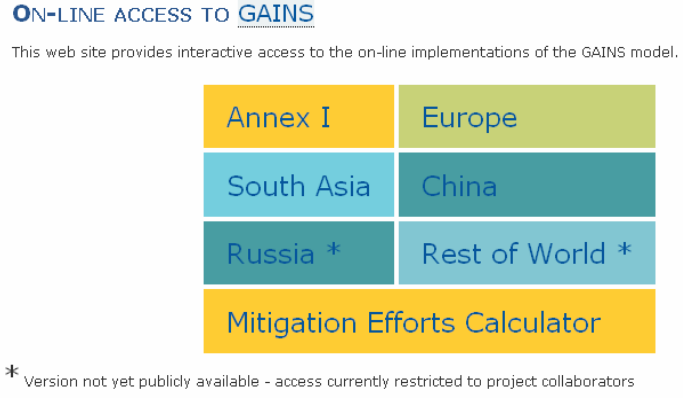
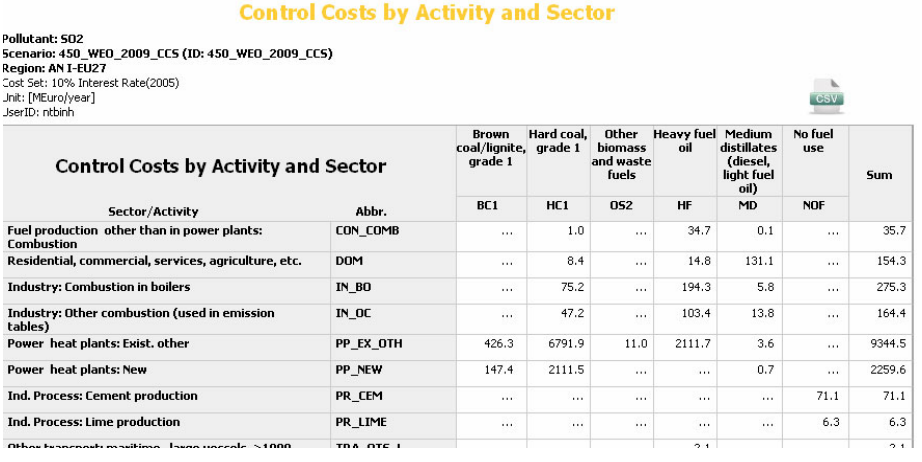


Fig. 3. GAINS cloud data warehouse portal

- The GAINS DWH World is developed by collecting data from available international emission inventories and on national information supplied by individual countries. It is the global data warehouse in our cloud data warehouse application framework and used to configure and generate regional data warehouse(s) as required.
- For Annex I countries, under United Nations Framework Convention on Climate Change (UNFCCC) where we not only provide access to GAINS, but also interact Mitigation Effort Calculator (MEC), and this helps to compare the relative climate change mitigation efforts of all industrialized countries.
- For European countries where GAINS is used extensively by Member States of the EU and the European Commission to develop cost-effective strategies to reduce the environmental impact of air pollutions. For example, figure 4 shows an example of using GAINS Global DWH to calculate Costs by Activity and Sector of Annex I countries.
- For China and South Asia, including India, where GAINS is being used to explore sustainable development pathways for the future.



Control Costs by Activity and Sector

Pollutant: SO2
Scenario: 450_WEO_2009_CCS (ID: 450_WEO_2009_CCS)
Region: AN I-EU27
Cost Set: 10% Interest Rate(2005)
Unit: [MEuro/year]
UserID: ntbinh

CSV

Control Costs by Activity and Sector		Brown coal/lignite, grade 1	Hard coal, grade 1	Other biomass and waste fuels	Heavy fuel oil	Medium distillates (diesel, light fuel oil)	No fuel use	Sum
Sector/Activity	Abbr.	BC1	HC1	OS2	HF	MD	NOF	
Fuel production other than in power plants: Combustion	CON_COMB	...	1.0	...	34.7	0.1	...	35.7
Residential, commercial, services, agriculture, etc.	DOM	...	8.4	...	14.8	131.1	...	154.3
Industry: Combustion in boilers	IN_BO	...	75.2	...	194.3	5.8	...	275.3
Industry: Other combustion (used in emission tables)	IN_OC	...	47.2	...	103.4	13.8	...	164.4
Power heat plants: Exist. other	PP_EX_OTH	426.3	6791.9	11.0	2111.7	3.6	...	9344.5
Power heat plants: New	PP_NEW	147.4	2111.5	...	...	0.7	...	2259.6
Ind. Process: Cement production	PR_CEM	...	...	...	...	...	71.1	71.1
Ind. Process: Lime production	PR_LIME	...	...	...	...	...	6.3	6.3
Other transport: maritime, large vessels >1000	TRA_OTS_1				3.1			3.1

Fig. 4. An example of using GAINS Global DWH to calculate Costs by Activity and Sector of Annex I countries

5 Conclusions and Future Work

This paper introduces a new sound concept in the cloud intelligence research field when modeling the GAINS cloud data warehouse conceptual model in a very formal manner. Furthermore, GAINS dimensions and decision variables are also specified and formulated to present very special features in their structures and constraints. Moreover, this paper also introduces our approach to leverage the sharing multidimensional metadata data and reusing design patterns in building, managing and using GAINS data warehouses globally as well as regionally in the context of the GAINS cloud data warehouse application framework.

However, in order to successfully implement this framework, there are some considerations must be solved. The first consideration is how to defining at global level, i.e. GAINS DWH World, a full set of dimensions and their constraint, which can

effectively infer matching models/pattern of the global data warehouse with respect to a set of regional requirements. The next consideration is the way building interoperability consistent with other system's repositories to help the wrapper access internal system intelligently. The last consideration is how to seamlessly integrate a new artifact with new terms, concepts, constraints and potential conflicts to certain autonomous systems. These challenges could be seen as our future work.

References

1. Binh, N.T., Wagner, F., Schoepp, W.: GAINS-BI: Business Intelligent Approach for Greenhouse Gas and Air Pollution Interactions and Synergies Information System. In: Proc of the International Organization for Information Integration and Web-based Application and Services, IIWAS 2008, Linz (2008)
2. Gangadharan, G.R., Swami, S.N.: Business Intelligence Systems: Design and Implementation Strategies. In: Proc. of the 26th International Conference Information Technology Interfaces, ITI 2004, Croatia, pp. 139–144 (2004)
3. Grant, A.J., Luqi: Intranet Portal Model and Metrics: A Strategic Management Perspective. IT Professional 7, 37–44 (2005)
4. Hugh, J.W., Barbara, H.W.: The Current State of Business Intelligence. Computer 40, 96–99 (2007)
5. Klaassen, G., Amann, M., Berglund, C., Cofala, J., Höglund-Isaksson, L., Heyes, C., Mechler, R., Tohka, A., Schöpp, W., Winiwarter, W.: The Extension of the RAINS Model to Greenhouse Gases. An interim report describing the state of work as of April 2004 (2004), IIASA IR-04-015
6. Lou, A.: Data Warehousing in the Clouds Making Sense of the Cloud Computing Market (2009), White paper http://www.b-eye-network.com/view/8702
7. Makowski, M.P.: Data Cleaning and Performance Tuning in the GAINS Model. Thesis at the Database and Artificial Intelligence Group (DBAI) of the Technical University of Vienna (2008)
8. Michael, A., Armando, F., Rean, G., Anthony, D., Randy, K., Andy, K., Gunho, L., David, P., Ariel, R., Ion, S., Matei, Z.: Above the Clouds: A Berkeley View of Cloud Computing, University of California at Berkeley (2009)
9. Ta'a, A., Bakar, M.S.A., Saleh, A.R.: Academic business intelligence system development using SAS® tools. In: Online Proc. of the SAS Global Forum (2008)
10. Tvrdikova, M.: Support of Decision Making by Business Intelligence Tools. In: Proc. of the 6th International Conference on Computer Information Systems and Industrial Management Applications, pp. 364–368 (2007)
11. Wei, X., Xiaofe, X., Lei, S., Quanlong, L., Hao, L.: Business intelligence based group decision support system. In: Proc of the International Conferences on Info-tech and Info-net ICII 2001, Beijing, China, pp. 295–300 (2001)
12. Zeng, L., Shi, Z., Wang, M., Wu, W.: Techniques, Process, and Enterprise Solutions of Business Intelligence. In: Proc. of the IEEE Conference on Systems, Man and Cybernetics, Taipei, Taiwan, pp. 4722–4726 (2006)
13. Wagner, F., Schoepp, W., Heyes, C.: The RAINS optimization module for the Clean Air For Europe (CAFE) Programme, Interim Report IR-06-029, International Institute for Applied Systems Analysis (IIASA) (September 2006)
14. Wang, L., Laszewski, G., Kunze, M., Tao, J.: Cloud computing: A Perspective study. Grid Computing Environments (GCE) (2008)
15. Wikipedia, http://en.wikipedia.org/wiki/Cloud_computing

Distributed Representation of Word

Jau-Chi Huang[1], Wei-Chen Cheng[1,2], and Cheng-Yuan Liou[1,⋆]

[1] Department of Computer Science and Information Engineering, National Taiwan University, Taiwan, Republic of China
cyliou@csie.ntu.edu.tw
[2] Institute of Statistical Science, Academia Sinica, Taiwan, Republic of China

Abstract. We present a novel method to train the Elman network to learn literal works. This paper reports findings and results during the training process. Both codes and network weights are trained by using this method. The training error can be greatly reduced by iteratively re-encoding all words.

Keywords: Elman network, content addressable memory, semantic indexing, personalized codes, compositional representation.

1 Introduction

In [1], Elman used a simple recurrent network (SRN or called Elman network), see Fig. 1, to model human language processing. During operation of the network, both current input pattern from the input layer and previous state of the hidden layer saved in the context layer activate the hidden layer. The network is trained by using the next input pattern as the desired output. The backpropagation learning algorithm [2] is used to reduce the difference between the output of the network and the desired output. After training, it has the ability to predict the next input patterns. Prediction plays an important role in language processing. Listeners can predict the following words and pay attention to the words which violate expectations. The network autonomously explores the regularities which underlie the temporal order of the characters in words and underlie the word sequences in sentences.

In the simple sentence prediction [1], Elman used *localist representation*. In the representation, each word is represented by a different bit. Therefore, the dimension of input code vectors is the same as the number of different words. All codes are orthogonal to each other and they carry no syntactic or semantic meaning of their words. Elman claimed that since the input representations themselves loaded no information that can be used for prediction, the network should have developed certain internal representations after learning. He averaged the hidden layer outputs of each word after the learning and then fitted those averaged representations of all words into a hierarchical clustering tree. The tree shows that there are several categories of words. These categories possess both syntactic relationships, such as noun v.s. verb, and semantic relationships, such as human v.s. animal.

⋆ Corresponding author.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 169–176, 2011.

The localist representation in [1] has been discussed by Marcus in [3]. Localist representations are costly. The corpus would contain a very limited number of vocabularies. This is because each word occupies a single dimension in an input vector. Localist representations can't provide any form-based or function-based similarity among words. This kind of representation is incapable of supporting real world applications. On the other hand, human infants appear to generalize to novel stimuli, but the networks will not. It's because infants have a far richer perceptual experience than networks, and the composition of perceptual features which are outside the experiment is more powerful than simple representation of each word limited to the network experience [4].

Distributed representation of word may be able to support compositional representation of perceptual features for productive behavior and inference. Elman operated the same simulation with each word encoded as randomly assigned distributed representation and obtained similar results to prove the localist representations are not necessary to the connectionist models [4]. However, the randomly assigned distributed representation of each word is arbitrary and is far from the real world composition of perceptual features that human infants use.

Moreover, the slow convergence obtained from backpropagation learning algorithm stops people using Elman network to process more complex real sentences. This paper rewrites the training method and devises an encoding scheme to renew the distributed representation in each iteration. The training error is significantly reduced without increasing the number of neurons. The representation will possess the form-based and function-based similarity in certain degree. Instead of using the simple artificial sentences [1], this work uses real sentences sampled from a fiction and analyzes their prediction errors. Because it's too complicated to fit all words into a binary tree like Elman did in [1], we change to apply the technique PCA (Principal Component Analysis) to monitor and display the intermediate distributions of codes on $2D$ space during training.

2 Method

Elman network is a powerful tool on linguistic process and has many achievements [1], [5], [6]. It has been proved that the computational capability of the first-order recurrent networks is beyond the Turing limit. In other words, it's a certain kind of super-Turing model [7]. We expect that it is capable of loading and extracting structural information from sequences of words. The proposed training method will adjust the weights to load the structural information autonomously. The iteratively renewed codes will possess similarity information that are useful in many indexing applications. Both code renewing and weight adjusting work together to achieve performance.

2.1 Elman Network

Elman network is a simple recurrent network that has a context layer as an inside self-referenced layer, see Fig. 1. During operation, the output of the hidden layer at time $t-1$ saved in the context layer, together with the current input from the

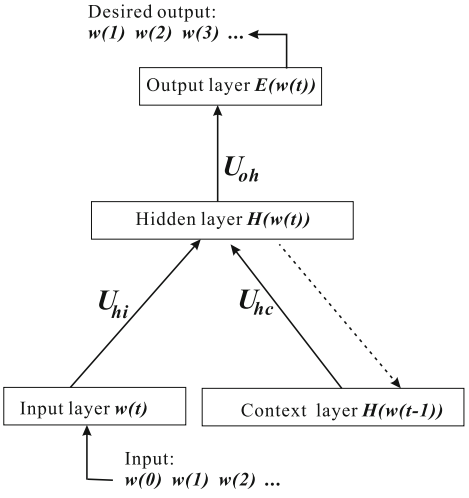


Fig. 1. Illustration of Elman network

input layer, activates the hidden layer at time t. Consider a sequence of words, $\{w(t), t = 0, 1, 2, ...\}$, this sequence can be samples from an article, a book, or a whole corpus of a specific author. Each word is presented one after one along the sentence. The network is trained to predict the following word. For example, the input at time $t = 0$ is $w(0)$, and the desired output at time $t = 0$ will be $w(1)$. The input at time $t = 1$ is $w(1)$, and the desired output at time $t = 1$ will be $w(2)$. The backpropagation learning algorithm [2] is used to adjust the weights.

Let $\{w_n, n = 1 \sim N\}$ be the code set of N different words in a corpus. The code for each word is randomly assigned without repetition. The corpus contains a collection of all given sentences. During training, sentences are fed to the network sequentially, word by word, starting from the first word of the sentence. Let L_o, L_h, L_c, and L_i be the number of neurons in the output layer, the hidden layer, the context layer, and the input layer, respectively. In Elman's network [1], L_h is equal to L_c, that is, $L_h = L_c$. In this work, to predict the following word, the number of neurons in the output layer is the same as the number of neurons in input layer, $L_o = L_i = R$, where R is the dimension of codes. Therefore, $w(t) = w_n$ will be an R by 1 column vector. U_{oh} is an $L_h + 1$ by L_o weight matrix between the hidden layer and the output layer. U_{hi} is an L_i by L_h weight matrix that connects the input layer and the hidden layer, and U_{hc} is an L_c by L_h weight matrix that connects the context layer and the hidden layer. The weight matrix U_{hic} is an $L_i + L_c + 1$ by L_h matrix between the first layer and the hidden layer. The output vector of the hidden layer is denoted as $H(w(t))$ when $w(t)$ is fed to the input layer. $H(w(t))$ is an L_h by 1 column vector with L_h elements. Let $E(w(t))$ be the output vector of the output layer when $w(t)$ is fed to the input layer. $E(w(t))$ is an L_o by 1 column vector.

The function of the hidden layer is $H(w(t)) = \varphi(U_{hic}F(w(t)))$, where $F(w(t))$ is an $L_i + L_c + 1$ by 1 column vector, and φ is a sigmoid activation function,

$\varphi(x) = \frac{2}{1+exp(-0.5x)} - 1$, that operates on each element of a vector [2]. The column vector $F(w(t))$ has the form

$$F(w(t)) = \begin{bmatrix} w(t) \\ H(w(t-1)) \\ 1 \end{bmatrix}.$$

The function of the output layer is $E(w(t)) = \varphi(U_{oh} \begin{bmatrix} H(w(t)) \\ 1 \end{bmatrix})$. The goal of training is to minimize the error between the network's outputs and the desired outputs to meet the prediction $E(w(t)) \approx w(t+1)$.

2.2 Iterative Re-encoding

Instead of using simple generated sentences as corpus, this work plans to process real corpus which contains complex sentences and large amount of vocabularies. This work presents a new approach to accomplish the language task. In the approach, each word initially has a random lexical code, $w_n^{j=0} = [w_{n1}, w_{n2},...,w_{nR}]^T$. After every k training epochs, a renewed code is calculated by

$$w_n^{new} = \frac{1}{|s_n|} \sum_{i=1}^{|s_n|} s_n^i, \quad n = 1 \sim N,$$

where the set s_n contains all predictions for the word w_n,

$$s_n = \{E(w(t)) \mid w(t+1) = w_n\}.$$

$|s_n|$ is the total number of predictions in the set. Note that in [1], Elman averaged all the hidden output vectors for each word w_n, but this work averages all the prediction vectors for it instead.

2.3 Normalization

All renewed codes are normalized before using them in the next iteration. This is because Elman network can reduce the prediction error simply by decreasing the hamming distances among all codes. The worst case is that every word converge to a same code vector. In this case, the network achieves zero prediction error. The normalization contains two steps. The first step is to let each row in the code matrix $W_{R\times N}^{new} = \left[w_1^{new}\ w_2^{new}\ ...\ w_N^{new}\right]$ become zero mean. This can be gotten by

$$W_{R\times N}^{ave} = W_{R\times N}^{new} - \frac{1}{N} W_{R\times N}^{new} \begin{bmatrix} 1 & ... & 1 \\ \vdots & \vdots & \vdots \\ 1 & ... & 1 \end{bmatrix}_{N\times N}.$$

In the second step, we set $w_n^{renew} = \|w_n^{ave}\|^{-1} w_n^{ave}$, where $\|w_n\| = (w_n^T w_n)^{0.5}$. Therefore, for each word, $\|w_n^{renew}\| = 1$. This setting can prevent a diminished solution, $\{\|w_n\| \sim 0, n = 1 \sim N\}$, that is usually derived by the back-propagation algorithm.

3 Experiment

This work uses the fiction "Peter Pan" [8] as our simulation data. To reduce the amount of vocabularies, we apply several simple grammar rules on plurals, regular verbs, and "'s" to simplify the problem. For example, "playing" becomes "play + ING"; "lights" becomes "light + NVs"; "children's" becomes "children + 's"; "turned" becomes "turn + Ved". After preprocessing, there are 3,805 different root words including "ING", "NVs", "'s", "Ved", and the mark that is added to represent the end of a sentence. The total number of sentences is 3,101; the total number of words is 54,999.

The architecture of the network is $L_i = 15$ input neurons, $L_o = 15$ output neurons, $L_h = 30$ hidden layer neurons, and $L_c = 30$ context layer neurons. The initial values of synapse weights U_{hic} and U_{oh} are randomly assigned in the range $[-1, 1]$, and the initial values of the neurons in the context layer are set to zero. The initial coding $w_n^{j=0}$ is randomly assigned under the restriction that different words have different codes and then is normalized through above normalization method. Use the backpropagation algorithm to reduce the prediction error which is $\|E(w(t)) - w(t+1)\|^2$. The learning rate is fixed to 0.01. U_{hic} and U_{oh} are updated after each word presented. We set one epoch, $j = 1$, as when all 54,999 words were presented, and we renew the codes every k epochs.

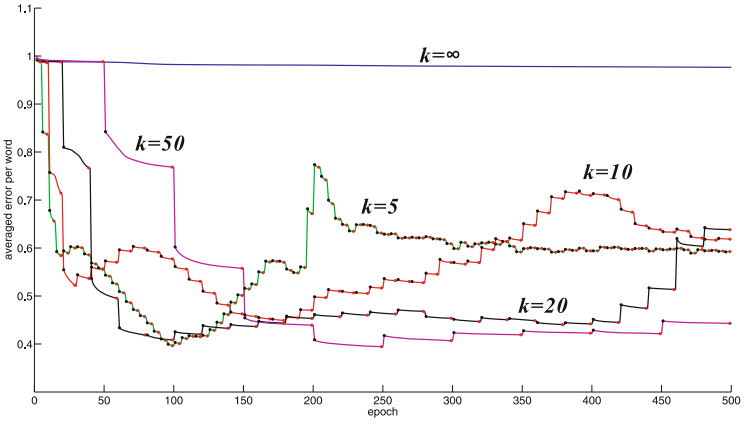


Fig. 2. Error curves under different re-encoding conditions, $k = \infty$, $k = 5$, $k = 10$, $k = 20$, and $k = 50$. The curve from red point to black point displays how error is reduced by re-encoding. The curve from black point to red point displays how error is reduced by weight adjustment.

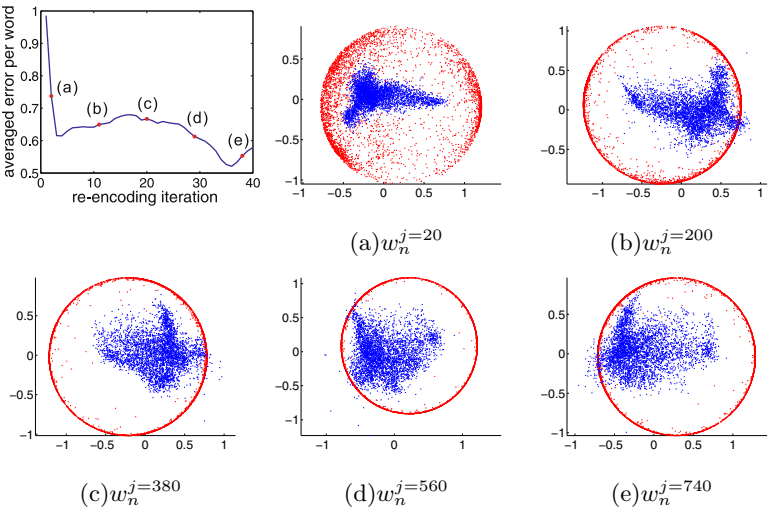


Fig. 3. Using PCA to display the code distributions on $2D$ space after different training epochs. We re-encode the words every 20 epochs. The blue points display w_n^{new}. The red points display the renew codes w_n^{renew} after normalization.

Fig. 2 records the error curves under different numbers of renew epochs, $k = \infty$(codes won't be renewed), 5, 10, 20, and 50. The figure shows that the error decreased sharply when we apply the re-encoding scheme. The performance of weight adjusting is far less than that of code renewing. This figure also shows that intensive re-encoding may cause severe fluctuations.

Fig. 3 displays the code converge under $k = 20$ condition. We can find normalization will force the codes distributed on a circle and not to merge to one point. However, this may change the interior structure of codes and cause the error curve continually fluctuates.

We choose the code $w_n^{j=m+1}$ to analyze, where the m^{th} epoch achieves the minimum error. The hierarchical cluster tree that Elman used to analyze the internal representations of words is restricted to display a limited relationships of words in a small number of artificial vocabularies. We use PCA to map them from the code space ($R = 15$ dimension) to a $2D$ space to display the distributions directly, see Fig. 4. We can find verbs and adverbs are densely crowd to the left part, and nouns and adjectives have no clustering effect.

After learning, we present all $3,101$ sentences to the network again and get averaged prediction error for each word which was used to predict the next word, see Fig. 5(a), and also for each word which was predicted, see Fig. 5(b). From Fig. 5(a), the higher frequency words tend to have lower prediction errors. This implies that higher frequency words can predict the next word more accurately. From Fig. 5(b), We find the words which only appear once in the corpus tend to have lower prediction errors. They may be easy to be predicted.

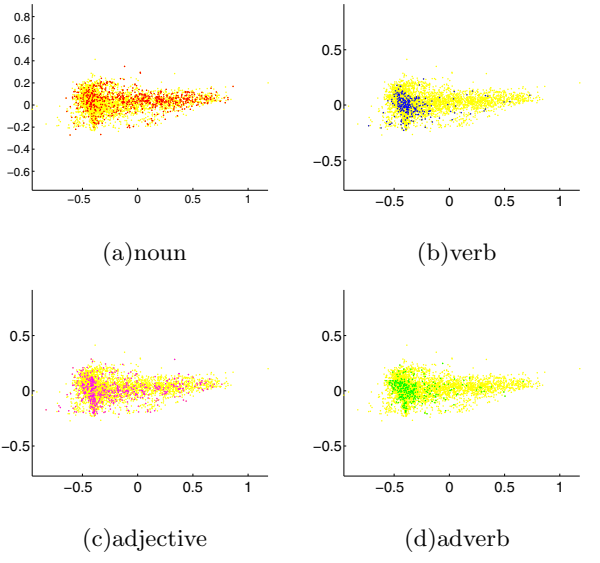


Fig. 4. The code distribution of nouns, verbs, adjectives and adverbs before normalization on $2D$ space

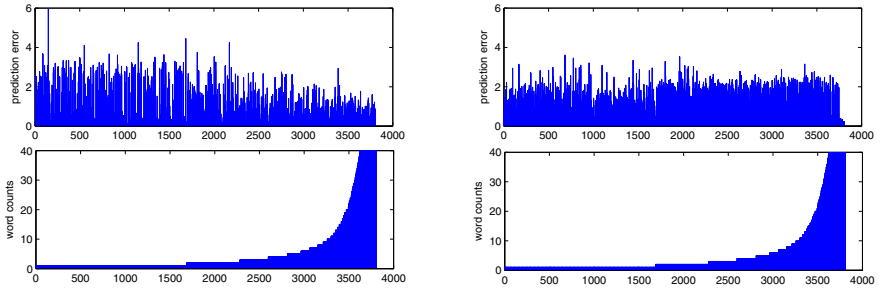


Fig. 5. (a)The prediction errors and the word counts of words which were used to predict the next word. (b)The prediction errors and the word counts of words which were predicted.

4 Summary

The proposed re-encoding method significantly improves performance of Elman network. It allows us to process complex real literal works. The code obtained from our method preserves syntactic relationship. Moreover, we analyze the relationship between the prediction errors and word counts. It shows that the relationship can be extracted automatically by the network. This work displays

the capability of using Elman network with our re-encoding method to process language.

References

1. Elman, J.L.: Finding Structure in Time. Cognitive Science 14, 179–211 (1990)
2. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning Internal Representations by Error Propagation. In: Rumelhart, D.E., McClelland, J.L. (eds.) Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1, pp. 318–362. MIT Press, Cambridge (1986)
3. Marcus, G.: Symposium on Cognitive Architecture: The Algebraic Mind. In: Gernsbacher, M.A., Derry, S. (eds.) Proceedings of the 20th Annual Conference of the Cognitive Science Society. Lawrence Erlbaum Associates, Mahwah (1998)
4. Elman, J.L.: Generalization, Simple Recurrent Networks, and the Emergence of Structure. In: Gernsbacher, M.A., Derry, S. (eds.) Proceedings of the 20th Annual Conference of the Cognitive Science Society. Lawrence Erlbaum Associates, Mahwah (1998)
5. Liou, C.-Y., Huang, J.-C., Yang, W.-C.: Semantic addressable encoding. In: King, I., Wang, J., Chan, L.-W., Wang, D. (eds.) ICONIP 2006. LNCS, vol. 4232, pp. 183–192. Springer, Heidelberg (2006)
6. Liou, C.-Y., Huang, J.-C., Yang, W.-C.: Modeling Word Perception Using the Elman Network. Neurocomputing 71, 3150–3157 (2008)
7. Seigelmann, H.T.: Neural Networks and Analog Computation: Beyond the Turing Limit. Springer, Heidelberg (1999)
8. Peter Pan by J. M. Barrie - Project Gutenberg, http://www.gutenberg.org/ebooks/16

Mining Frequent Itemsets from Multidimensional Databases*

Bay Vo[1], Bac Le[2], and Thang N. Nguyen[3]

[1] Faculty of Information Technology, Ho Chi Minh City University of Technology, Vietnam
vdbay@hcmhutech.edu.vn
[2] Faculty of Information Technology University of Science, Ho Chi Minh, Vietnam
lhbac@fit.hcmus.edu.vn
[3] California State University Long Beach, USA
tnnguyen@csulb.edu

Abstract. Mining frequent itemsets (FIs) has been developing in recent years. However, little attention has been paid to efficient methods for mining in multidimensional databases. In this paper, we propose a new method with a supporting structure called AIO-tree (Attributes Itemset Object identifications – tree) for mining FIs from multidimensional databases. This method need not transform the database into the transaction database, and it is based on the intersections of object identifications for fast computing the support of itemsets. We compare our method to dEclat (after transformation to a transaction database) and indeed claim that they are faster than dEclat.

Keywords: Frequent itemsets, multidimensional databases, equivalence classes, AIO-tree, MARA algorithm, DMARA algorithm.

1 Introduction

Mining FIs is often used for transaction databases (TDB) [1-3, 6, 10-11, 13, 15]. However, we also need mine in multidimensional databases (MDB), which are not fully the TDB. Commonly, we must transform MDB into TDB to mine FIs [12].

Table 1. An example database

OID	A1	A2	A3
1	a1	b1	c1
2	a1	b2	c1
3	a2	b2	c1
4	a3	b3	c1
5	a3	b1	c2
6	a3	b3	c1
7	a1	b3	c2
8	a2	b2	c2

This paper presents a new method for mining FIs from MDB. This method need not transform MDB into TDB. It stores items in AIO-tree for fast generating itemsets and computing the support of them. Transforming databases into TDBs will lose the relationship of items in the same attribute. These items need not be combined in reality because we know that they are not frequent itemsets. Besides, it requires more storage space of transformed databases.

Consider the example database in Table 1 on the left, if we transform into TDB, we have a mapping as shown in Table 2.

* This work was supported by Vietnam's National Foundation for Science and Technology Development (NAFOSTED), project ID: 102.01-2010.02.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 177–186, 2011.

Table 2. The mapping table

OID	A1	Item	A2	item	A3	item
1	a1	A	b1	D	c1	G
2	a1	A	b2	E	c1	G
3	a2	B	b2	E	c1	G
4	a3	C	b3	F	c1	G
5	a3	C	b1	D	c2	H
6	a3	C	b3	F	c1	G
7	a1	A	b3	F	c2	H
8	a2	B	b2	E	c2	H

And output will be a binary database:

Table 3. The transactions database

TID	A	B	C	D	E	F	G	H
1	1			1			1	
2	1				1		1	
3		1			1		1	
4			1			1	1	
5			1	1				1
6			1			1	1	
7	1					1		1
8		1			1			1

Using Apriori algorithm with minSup = 20%, we have results as follows:

L1	
item	Supp
A	3
B	2
C	3
D	2
E	3
F	3
G	5
H	3

⇒

C2		C2 (cont.)	
itemset	Supp	Itemset	Supp
AB	0	CE	0
AC	0	CF	2
AD	1	CG	2
AE	1	CH	1
AF	1	DE	0
AG	2	DF	0
AH	1	DG	1
BC	0	DH	1
BD	0	EF	0
BE	2	EG	2
BF	0	EH	1
BG	1	FG	2
BH	1	FH	1
CD	1	GH	0

L2	
Itemset	Supp
AG	2
BE	2
CF	2
CG	2
EG	2
FG	2

C3	
CFG	2

L3	
CFG	2

Fig. 1. Results of mining FIs using Apriori of the TDB in Table 3

Thus, the number of 2-itemset candidates is 24. We can see that the support of AB, AC, BC, DE, DF, EF, GH are 0 because these 2 items belong to an attribute. In practice, the number of candidates that need to be considered is much fewer if we do not consider these items (17 instead of 24). Similar to the candidates 3, 4,…, k-itemsets.

Thus, the purpose of this paper is mining directly in MDB to take full advantage of this point for not generating non-frequent candidates.

The rest of this paper is as follows: Section 2 presents related work. In section 3, some definitions related to MDB are given. Section 4 proposes an AIO-tree structure to store items of database for mining FIs. Section 5 discusses MARA (Mining Association Rules based on AIO-tree), an algorithm based on AIO-tree structure, to mine FIs. It is based on a hybrid method to mine FIs and based on the intersections of object identifications for fast computing the support. Diffset strategy is also discussed in this section, DMARA algorithm is also proposed. Section 6 shows our experimental results. We conclude our work in section 7.

2 Related Work

In recent years, a number of mining FIs algorithms have been proposed. There are in general three categories [13]:

(1) Test-and-generate [2]: using a level-wise approach to discover FIs. These generate candidates based on apriori property to test whether all subset of a candidate are frequent or not. Disadvantages of this approach includes: (1) multiple scanning database and (2) it is time-consuming for generating and removing candidates.
(2) Divide-and-conquer [3, 4]: using compact data structure to mine FIs. These are based on FP-tree (or some extension of FP-tree) to store database in tree and mine from tree. This approach often scans the database in two times: one for building the f-list and another for building FP-tree.
(3) Hybrid [6, 9, 15]: using both "test-and-generate" and "divide-and-conquer" to mine FIs.

Mining FIs in MDB, to our best knowledge, there is only one paper [12]. The reason is mentioned above, most authors transform the database into TDB and mine from it. In practice, the way of transforming leads to consider many candidates, if we do not transform into transaction, we easily consider that they are not frequent. In [12], authors presented the method for mining FIs in MDB. However, this method still considered itemsets that are in the same attributes. Besides, disadvantage of this method is that it must be based on level-wise approach so that it consumes a lot of time for test-and-generate candidates.

Some applications of mining association rules from MDB are web mining, XML mining [14], classification based on association rule mining [7, 8, 9].

3 Definitions

Let D be the database with n attributes A_1, A_2,…, A_n and |D| rows (cases). Specific values of attribute A_i are denoted by lower case a, respectively.

Definition 3.1: An itemset includes a set of pairs which include an attribute and a specific value for each attribute in the set, denoted < (A_{i1}, a_{i1}), (A_{i2}, a_{i2}),…, (A_{im}, a_{im})>.

Definition 3.2: Support of X, denoted Supp(X), is the number of rows of D that match X.

Definition 3.3: Given X = <(A_{i1},a_{i1}), …, (A_{im}, a_{im})>, Y = <(A_{j1},a_{j1}),…, (A_{jk}, a_{jk})> be two itemsets. We say that X is subset of Y, denoted $X \subseteq Y$, iff $\forall(A_{iv},a_{iv}) \in X$ then $(A_{iv},a_{iv}) \in Y$.

Definition 3.4: A rule r has the form of r: $X \xrightarrow{q,p} Y - X$, where X = <($A_{i1}$,$a_{i1}$), …, ($A_{im}$, a_{im})>, Y = <(A_{j1},a_{j1}), …, (A_{jk}, a_{jk})> are two itemsets, $X \subseteq Y$, q = Supp(Y) is the support of r and $p = \frac{Supp(Y)}{Supp(X)}$ is the confidence of r.

Example: Consider the database in Table 1:
With X = <(A1, a3)> and Y = <(A1, a3), (A2, b3)>, Supp(X) = 3, Supp(Y) = 2.
r: $X \xrightarrow{q,p} Y - X$, where q = Supp(r) = Supp(Y) = 2, p = conf(r) = $\frac{Supp(Y)}{Supp(X)} = \frac{2}{3}$.

Definition 3.5: Obidset (Object identifications set)
Obidset(X) is a set of object identifications in D that matches X.
Example: Consider the database in Table 1

X1 = <(A1,a2)> then Obidset(X1) = {3, 8} or 38
X2 = <(A2, b2)> then Obidset(X2) = 238
X3 = <(A1,a2), (A2, b2)> then Obidset(X3) = Obidset(X1) ∩ Obidset(X2) = 38

We can see that |Obidset(X)| is the support of X. Therefore, based on Obidset, we can get the support of X fast. Besides, if we have Obidset(X) and Obidset(Y), we can compute Obidset(X∪Y) easily by computing the intersection of Obidset(X) and Obidset(Y).

4 AIO-Tree

4.1 Equivalence Class [15]

Let I be a set of items, and $X \subseteq I$. A function p(X, k) = X[1:k] as the k length prefix of X and a prefix-based equivalence relation θ_K on itemsets is defined: $\forall X,Y \subseteq I, X \equiv_{\theta_k} Y \Leftrightarrow p(X,k) = p(Y,k)$. That is, two itemsets are in the same equivalence class if they share a common k length prefix.

4.2 Vertex

Vertex is a triple of attributes (A); itemset (X); and Obidset (O).
The triple is denoted (A, X, O).

Remark: The itemset in this section is different from the itemset defined in section 3. The reason being that we group all attributes contain the itemset into a number to save memory. If the database has n attributes, we can use n bits to store them. For example, the database in Table 1 has 3 attributes, we use 3 bits, bit 1 is 1 if itemset contains the value of attribute A1. Similar to bit 2 for attribute A2, and bit 3 for attribute A3.

Example: Consider the database in Table 1, we have:
X = <(A1,a1)>, because Obidset(X) = {1,2,7} or 127, it can present as (1,a1,127). That means a1 contains in attribute 001 = 1 and in OIDs {1,2,7}.
Y = <(A1,a1), (A2, b1)> can be present as (3,a1b1,1). We can get the Obidset(Y) by computing Obidset(<A1,a1>) ∩ Obidset(<(A2,b1)> = 127 ∩ 15 = 1.

In Fig. 2, {a1b1, a1b2, a1b3, a1c1, a1c2} are in the same equivalence class a1 because they have the same prefix a1. Similarly, {a2b2c1, a2b2c2} are in the same equivalence class a2b2 because they have the same prefix a2b2.

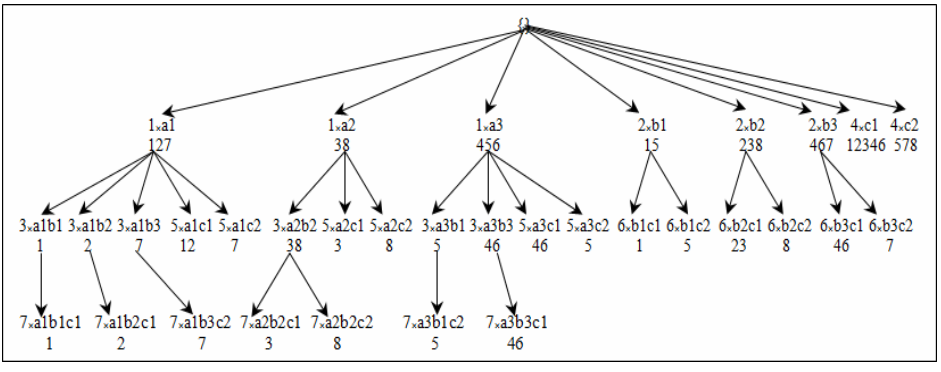


Fig. 2. AIO-tree and equivalence classes

4.3 Arc

An arc connects two vertexes together such that their attributes are in same equivalence class and have parent-child relationship.

5 Mining Frequent Itemsets from MDB

In this section, we present MARA, an efficient algorithm for mining FIs from MDB. It is based on AIO-tree (and a hybrid approach) to mine FIs.

Theorem 5.1: Given two nodes (att_1,$itemset_1$,$Obitset_1$) and (att_2,$itemset_2$,$Obitset_2$), if $att_1 = att_2$ and $itemset_1 \neq itemset_2$ then $Obidset_1 \cap Obidset_2 = \varnothing$.

Proof: Because $att_1 = att_2 \Rightarrow \forall item_1 \in itemset_1$, $\forall item_2 \in itemset_2$, if $item_1 \neq item_2$ and $item_1$ occurs in OID_k then $item_2$ cannot occur in OID_k. Therefore, $\forall OID \in Obidset_1$, because $itemset_1$ occurs in OID so that $itemset_2$ cannot occur in OID $\Rightarrow Obidset_1 \cap Obidset_2 = \varnothing$.

Theorem 5.1 infers that, if two itemsets X and Y have the same attributes, we need not combine them into itemset X∪Y because of Supp(X∪Y) = 0.

For example: consider two nodes (1,a1,127) and (1,a2,38), Obidset(<(1,a1)>) = 127 and Obidset(<(1,a2)>) = 38 ⇒ Obidset(<(1,a1),(1,a2)>) = Obidset(<(1,a1)>) ∩ Obidset(<(1,a2)>) = ∅.

Similarly, Obidset(<(1, a1), (2, b1)>) = 1, Obidset(<(1, a1), (2, b2)>) = 2 $\Rightarrow$ Obidset(<(1, a1), (2, b1)>) $\cap$ Obidset(<(1, a1), (2, b2)>) = $\emptyset$.

5.1 Algorithm for Mining FIs Based AIO-Tree

The equivalence class L_r contains single-item nodes which their supports satisfy minSup (i.e., |Obidset| $\geq$ minSup) (line 1). For example, consider the database in Table 1 with minSup = 20% (itemset occurs in at least two rows), L_r contains the first level in Fig. 1 because all single-items satisfy minSup. After that, the algorithm will sort items in L_r (line 2) in increasing order by their supports (i.e., sort them according to |Obidset|). Finally, it will call **GENERATE_FIs** function (line 4).

Input: MDB D and minSup.
Output: FIs (frequent itemsets) in D that satisfy minSup.
Method:

```
MARA(D, minSup)
 1.  Lr = {(Ai, aj, Obidset(Ai,aj)): |Obidset(Ai,aj)| ≥ minSup}
 2.  Sort(Lr) // Sort AIO nodes increasing by their supports
 3.  FIs = ∅
 4.  GENERATE_FIs(Lr, minSup)

GENERATE_FIs(Lr, minSup)
 5.    for all (Ai, Xi, Oi) ∈ Lr do
 6.        Add Ai, Xi and |Oi| to FIs
 7.        Pi = ∅
 8.        for all (Aj, Xj, Oj) ∈ Lr with j > i do
 9.            if   Ai ≠ Aj   then
10.                A = Ai ∪ Aj
11.                X  = Xi ∪ Xj
12.                O  = Oi ∩ Oj
13.                if   |O| ≥ minSup   then
14.                   Add  (A, X, O) to Pi
15.        GENERATE_FIs(Pi, minSup)
```

Algorithm 1. MARA algorithm

Function **GENERATE_FIs** will traverse all (A_i, X_i, O_i) in L_r (line 5) to generate FIs. With each node (A_i, X_i, O_i), it will add its attributes, its itemset and its support into FIs (line 6). After that, the function will traverse all (A_j, X_j, O_j) in L_r following (A_i, X_i, O_i) (line 8). With each pair {(A_i, X_i, O_i), (A_j, X_j, O_j)}, if their attributes are different ($A_i \neq A_j$, line 9), we compute three information: $A = A_i \cup A_j$, $X = X_i \cup X_j$, $O = O_i \cap O_j$ (line 10 – 12), if the support of X (i.e., |O|, line 13) satisfies minSup then add the AIO node (A, X, O) to P_i (is initialized by $\emptyset$, line 7). After traversing all (A_j, X_j, O_j), it will call **GENERATE_FIs**(P_i, minSup) recursively (line 15).

5.2 Illustrations

Consider the database in Table 1 with minSup = 20%. We have the results in Fig. 3.

First of all, the root node contains nodes which are single items and satisfy minSup.

Consider node (1, a2, 38): Firstly, the algorithm will add (1, a2, 2) to FIs. Because nodes (1, a1, 127) and (1, a3, 456) have the same attribute with (1, a2, 38), the algorithm does not combine them together. With node (2, b1, 15), because O = Obidset(<1,a2>) $\cap$ Obidset(<2,b1>) = 38 $\cap$ 15 = $\varnothing \Rightarrow |O| = 0 <$ minSup. With node (2, b2, 238), because O = Obidset(<1, a2>) $\cap$ Obidset(<2, b2>) = 38 $\cap$ 2385 = 38 $\Rightarrow |O| = 2 \geq$ minSup $\Rightarrow$ add node (3, a2b2, 38) into equivalence class P_i (its attributes is 3 because it is $1 \cup 2 = 001 \cup 010 = 3$). When the algorithm is called recursively with input parameter P_i, because of $|P_i| = 1$, there is no any new equivalence class created.

5.3 Diffset for Fast Computing Support and Saving Memory

Authors Zaki and Gouda in [16] have proposed Diffset strategy for fast computing the support of itemsets and saving memory to store Tidsets. We recognize that it can be used for Obidsets.

$$\text{Diffset}(PX) = \text{Obidset}(P) - \text{Obidset}(X)$$

Assume we have Diffset(PX) and Diffset(PY). Now, we want to get Diffset(PXY). We can get it easily by computing the difference set between Diffset(PY) and Diffset(PX), i.e, Diffset(PXY) = Diffset(PY) – Diffset(PX) [16].

The support of XY will be obtained by: Supp(PXY) = Supp(PX) - |Diffset(PXY)| [16].

For example: Diffset(a2b2) = 38 – 238 = $\varnothing \Rightarrow$ Supp(a2b2) = Supp(a2) - |Diffset(a2b2)| = 2. Similarly, we have Diffset(a3b3) = 5, Diffset(a3c1) = 5

$\Rightarrow$ Diffset(a3b3c1) = Diffset(a3c1) – Diffset(a3b3) = $\varnothing$ and Supp(a3b3c1) = Supp(a3c3) - |Diffset(a3b3c1)| = 2 – 0 = 2.

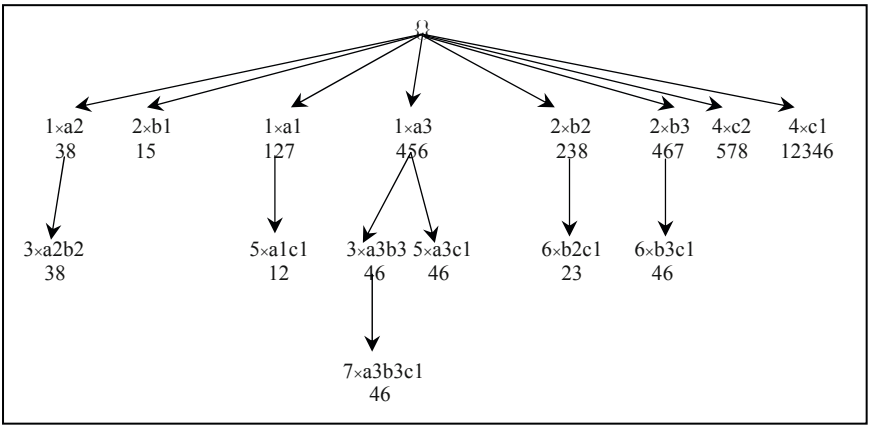


Fig. 3. Results of MARA algorithm with minSup = 20% using AIO-tree

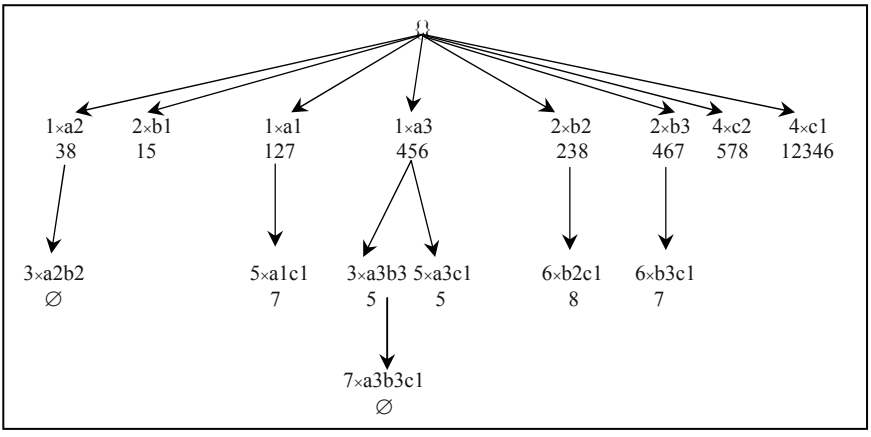


Fig. 4. Results of DMARA algorithm with minSup = 20%

The AIO-tree using Diffset is showed in Fig. 4. We can see that the sum size of Diffsets is often smaller than that of Obidsets. Therefore, the time for computing Diffset will be smaller than that for Obidsets, and the memory for storage Diffsets will be also smaller than that of Obidsets. In Fig. 4, the level 1 still contains Obidsets. From the level 2, Diffset is used. We can see that the sum size of Diffsets is 29, 76.3% compare to the sum size of Obidsets in Fig. 3 (the sum size of Obidsets is 38). The more levels the AIO-tree has, the smaller the scale is.

6 Experimental Results

Algorithms are coded by C# in Visual Studio.NET 2005, Windows XP. The experimental results are tested in databases getting from UCI Machine Learning Repository [5]. The databases are run in centrino core2duo 2×2.0, 1 MB RAM.

Table 4. The experimental databases [5]

Database	#attrs	#distinct items	#records
Breast	12	737	699
German	21	1077	1000
Led7	8	24	3200
Vehicle	19	1434	846
Poker-hand	11	95	1000000

The experimental databases have different features. Breast, German, Vehicle have many attributes and distinctive items (each attribute has many items or values, we count the number of items in all attributes), but the number of objects is few. Led7 has a few attributes, distinctive items and number of objects. Poker-hand (download on 2009) has a few attributes, distinctive items, but the number of objects is large.

Table 5. Results of DMARA compares to dEclat [15]

Database	minSup (%)	#FIs	Time (s)	
			dEclat [15] (mapping to TDB)	DMARA
Breast	1	11391	0.09	0.09
	0.5	19822	0.13	**0.12**
	0.2	67798	0.26	**0.24**
	0.1	996283	1.76	**1.71**
German	5	59356	0.52	**0.51**
	3	182877	1.04	**1.03**
	1	1703187	**4.42**	4.46
	0.5	6780213	**13.33**	14.21
Led7	1	3139	0.11	**0.1**
	0.5	7215	0.15	**0.13**
	0.2	7784	0.15	**0.14**
	0.1	9289	0.16	**0.15**
Vehicle	2	571	0.34	**0.31**
	1	2176	0.75	**0.69**
	0.5	8781	1.31	**1.22**
	0.2	294219	3.43	**3.26**
Poker-hand	6	247	113.06	**101.71**
	4	288	126.23	**112.38**
	2	739	129.58	**112.88**
	1	2699	265.57	**231.06**

As shown in Table 5 below, our algorithm (DMARA – MARA using Diffset) is better than dEclat in many runs (20 of 25 results, here we do not consider the time to transform the database into TDB) because it does not combine itemsets which are in the same attributes. However, it must spend a little time to compute attributes of itemsets (by union two attributes of two itemsets making this itemset). Therefore, DMARA is slight slower than dEclat in some results (German with minSup = 1% and 0.5%).

7 Conclusion and Future Work

In this paper, we have proposed a new method for mining FIs from MDB. The experimental results show that our algorithm is efficient than the other algorithm in the same category (dEclat – using hybrid method). One characteristic of direct mining in MDB is that we do not need to convert frequent itemset to atrributes-itemset because we have stored its attributes in the first level of AIO-tree. Therefore, we can save time when we show frequent itemset to the monitor or file. In AIO-tree, we can use Diffset for computing the supports of itemsets fast and saving more storage memory. When we generate association rules that satisfy minConf, we can use the attributes of itemsets to applicable considerations. Therefore, as future work, we will study more efficient algorithm for mining association rules from MDB. We will also look at using lattice-based approaches [10, 11] to apply them in MDB. This will include mining rules from frequent closed itemsets (FCIs) in MDB.

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD Conference, Washington DC, USA, pp. 207–216 (May 1993)
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: VLDB 1994, pp. 487–499 (1994)
3. Grahne, G., Zhu, J.: Efficiently using prefix-trees in mining frequent itemsets. In: FIMI 2003 Workshop on Frequent Itemset Mining Implementations, pp. 123–132 (2003)
4. Han, J., Kamber, M.: Data mining: concept and techniques, 2nd edn., ch. 5, pp. 234–250. Morgan Kaufmann Publishers, San Francisco (2006)
5. http://mlearn.ics.uci.edu/MLRepository.html (Download on 2007 and 2009)
6. Lee, A.J.T., Wang, C.S., Weng, W.Y., Chen, J.A., Wu, H.W.: An efficient algorithm for mining closed inter-transaction itemsets. Data & Knowl. Eng. 66, 68–91 (2008)
7. Li, W., Han, J., Pei, J.: CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules. In: Proc. of ICDM 2001, pp. 369–376 (2001)
8. Liu, B., Hsu, W., Ma, Y.: Integrating classification and association rule mining. In: Żytkow, J.M. (ed.) PKDD 1998. LNCS, vol. 1510, pp. 80–86. Springer, Heidelberg (1998)
9. Vo, B., Le, B.: A novel classification algorithm based on association rules mining. In: Richards, D., Kang, B.-H. (eds.) PKAW 2008 (Held with PRICAI 2008). LNCS, vol. 5465, pp. 61–75. Springer, Heidelberg (2009)
10. Vo, B., Le, B.: Mining traditional association rules using frequent itemsets lattice. In: The 39th International Conference on Computers & Industrial Engineering, Troyes, France, pp. 1401–1406. IEEE, Los Alamitos (2009)
11. Vo, B., Le, B.: Mining minimal non-redundant association rules using frequent itemsets lattice. Journal of Intelligent Systems Technology and Applications (accepted in March 2010) (to appear)
12. Xu, W., Wang, R.: A novel algorithm of mining multidimensional association rules. In: ICIC 2006. LNCIS, pp. 771–777 (2006)
13. Yahia, S.B., Hamrouni, T., Nguifo, E.M.: Frequent closed itemset based algorithms: A thorough structural and analytical survey. ACM SIGKDD Explorations Newsletter 8(1), 93–104 (2006)
14. Yuliana, O.Y., Chittayasothorn, S.: Deriving conceptual schema from XML databases. In: 1st Asian Conference on Intelligent Information and Database Systems, pp. 40–45 (2009)
15. Zaki, M.J., Hsiao, C.J.: Efficient algorithms for mining closed itemsets and their lattice structure. IEEE Transactions on Knowledge and Data Engineering 17(4), 462–478 (2005)
16. Zaki, M.J., Gouda, K.: Fast vertical mining using diffsets. In: Proc. of Ninth ACM SIGKDD Int'l. Conf. Knowledge Discovery and Data Mining, pp. 326–335 (August 2003)

Hybrid Fuzzy Clustering Using L_P Norms

Tomasz Przybyła[1], Janusz Jeżewski[2], Krzysztof Horoba[2], and Dawid Roj[2]

[1] Silesian University of Technology,
Institute of Electronics,
Akademicka 16, 44-101 Gliwice, Poland
Tomasz.Przybyla@polsl.pl
[2] Institute of Medical Technology and Equipment,
Biomedical Signal Processing Department,
Roosvelta 118, 41-800 Zabrze, Poland

Abstract. The fuzzy clustering methods are useful in the data mining applications. This paper describes a new fuzzy clustering method in which each cluster prototype is calculated as a value that minimizes introducted generalized cost function. The generalized cost function utilizes the L_p norm. The fuzzy meridian is a special case of cluster prototype for $p = 2$ as well as the fuzzy meridian for $p = 1$. A method for the norm selection is proposed. An example illustrating the performance of the proposed method is given.

1 Introduction

The goal of clustering is to find existing subsets in a set of objects $\mathbf{O} = \{o_1, \ldots, o_N\}$. The object set consists of unlabeled data, i.e. labels are not assigned to objects. Objects from one group have a high degree of similarity, while they have a high degree of dissimilarity with objects from other groups. Subsets that are found among the objects of the $\mathbf{O}$ set are called *clusters* [1], [2].

In most cases, each o_i object from the $\mathbf{O}$ object set is represented by an $\mathbf{x}$ vector in the s-dimensional space i.e. $\mathbf{x} \in \mathbb{R}^s$. The set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ is called the object data representation of $\mathbf{O}$. In such a case, the l-th component of the k-th feature vector $\mathbf{x}_k$ gives a measure of l-th feature (e.g. length of flower petal, age of car, weight) of the k-th object o_k.

One of the most popular clustering method is the fuzzy c–means (FCM) method. In this method, cluster prototypes are computed as fuzzy means [2]. However, one of the most important inconvenience of the FCM method is its sensitivity to outliers i.e. there are feature vectors which components have quite different value compared to other feature vectors. There are many modifications for the limitation of the outliers influence. In the first modification, the L_2 norm is replaced by the L_1 norm and by the generalized L_p norm [3]. Another approach has been proposed by Krishnapuram and Keller [4], [5]. This clustering approach is based on possibilistic theory instead of the fuzzy sets theory. One of the most interesting modification has been proposed by Kersten. In this method the L_2 norm is replaced by the L_1 norm, and the cluster prototypes are computed as fuzzy medians [6].

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 187–196, 2011.

On the other hand, clustering method should be robust for data corrupted by outliers or (and) heavy–tailed distributed noise. The heavy tailed distribution is more suitable to model the impulsive noise than the Gaussian distribution [7], [8], [9], [10], [11]. One of the havy tailed distribution is the Cauchy distribution, where the location parameter is called the (sample) myriad [12]. The fuzzy myriads have been used as the cluster prototypes in the fuzzy c–myriads (FCMyr) clustering method [13]. Another example of the heavy tailed distribution is the Meridian distribution proposed by Aysal and Berner [14]. The location parameter for the Meridian distribution is called the (sample) meridian. In the adaptative fuzzy c–meridians (AFCMer) clustering method, the cluster prototypes were computed as fuzzy meridians [16]. The myriad is the maximum likelihood estimator of the location parameter for the Cauchy distribution, so is the meridian for the Meridian distribution. It is important that the Meridian distribution has heavier tails than the Cauchy distribution. Therefore, the Meridian distribution better describes the impulsive noise. The form of the cost function for a sample myriad is very similar to the sample meridian cost function. The L_2 norm is used for the myriad cost function, where for the meridian cost function, the L_1 norm is used. In this paper, the generalized cost function is presented. In the proposed cost function, the L_p norm is used. Assuming $p = 2$ the generalized cost function becomes the myriad cost function, while for $p = 1$ the proposed cost function becomes the meridian cost function. Such a generalized cost function is used to determine the cluster prototypes in the proposed clustering algorithm.

The paper is organized as follows. Section II gives the generalized cost function. The proposed clustering algorithm is introduced in Section III. Section IV illustrates numerical examples. The last section contains some conclusions and ideas for future research.

2 Generalized Cost Function

2.1 Weighted Myriad Cost Function

For the Cauchy distribution, the displacement parameter is called sample myriad. For the given set of N independent and identically distributed (i.i.d.) samples each obeying the Cauchy distribution with common scale parameter, the sample myriad is a value that minimizes the cost function Ψ_K defined as follows [17]:

$$\begin{aligned}\hat{\Theta}_K &= \arg\min_{\Theta\in\mathbb{R}} \Psi_K(\mathbf{x};\Theta) \\ &= \arg\min_{\Theta\in\mathbb{R}} \sum_{k=1}^{N} \log\left[K^2 + (x_k - \Theta)^2\right]\end{aligned}, \tag{1}$$

where: Θ is the location parameter, and K is the scale parameter. By assigning non–negative weights to the input samples, the weighted myriad $\hat{\Theta}_K$ is derived as a generalization of the sample myriad. For the N i.i.d. observations $\{x_k\}_{k=1}^N$ and the $\{u_k\}_{k=1}^N$, the weighted myriad can be computed from the following expression

$$\begin{aligned}\hat{\Theta}_K &= \arg \min_{\Theta \in \mathbb{R}} \Psi_K(\mathbf{x}, \mathbf{u}; \Theta) \\ &= \arg \min_{\Theta \in \mathbb{R}} \sum_{k=1}^{N} \log\left[K^2 + u_k (x_k - \Theta)^2\right] \cdot \\ &= myriad\left\{u_k * x_k|_{k=1}^{N}; K\right\}\end{aligned} \tag{2}$$

The value of weighted myriad depends on the data set $\mathbf{x}$, the assigned weights $\mathbf{u}$ and the scale parameter K. Two interesting cases may occur. First, when the K value tends to infinity (i.e. $K \to \infty$), then the value of weighted myriad converges with the weighted mean, that is

$$\lim_{K \to \infty} \hat{\Theta}_K = \frac{\sum_{k=1}^{N} u_k x_k}{\sum_{k=1}^{N} u_k} \ , \tag{3}$$

where $\hat{\Theta}_K = myriad\left\{u_k * x_k|_{k=1}^{N}; K\right\}_{k=1}^{N}$. This property is called myriad linear property [12], [17].

Second case, called modal property, occurs when the value of K parameter tends to zero (i.e. $K \to 0$). In this case the value of the weighted myriad is always equal to one of most frequent values in the input data set.

2.2 Weighted Meridian Cost Function

The random variable formed as the ratio of two independent zero–mean Laplacian distributed random variables is referred to as the Meridian distribution [14]. For the given set of N i.i.d. samples $\{x_k\}_{k=1}^{N}$ each obeying the Meridian distribution with the common scale parameter δ, the sample meridian $\hat{\beta}_\delta$ is given by [14]:

$$\begin{aligned}\hat{\beta}_\delta &= \arg \min_{\beta \in \mathbb{R}} \phi_\delta(\mathbf{x}; \beta) \\ &= \arg \min_{\beta \in \mathbb{R}} \sum_{k=1}^{N} \log\left[\delta + |x_k - \beta|\right]\end{aligned} \ , \tag{4}$$

where Φ_δ is the sample meridian cost function.

The sample meridian can be generalized to the weighted meridian by assigning non–negative weights to the input samples. So, the weighted meridian is given by

$$\begin{aligned}\hat{\beta}_\delta &= \arg \min_{\beta \in \mathbb{R}} \phi_\delta(\mathbf{x}, \mathbf{u}; \beta) \\ &= \arg \min_{\beta \in \mathbb{R}} \sum_{k=1}^{N} \log\left[\delta + u_k |x_k - \beta|\right] \cdot \\ &= meridian\left\{u_k * x_k|_{k=1}^{N}; \delta\right\}\end{aligned} \tag{5}$$

The behavior of the weighted meridian significantly depends on the value of its medianity parameter δ. Two interesting cases may occur. The first case occurs when the value of the medianity parameter tends to infinity (i.e. $\delta \to \infty$), the weighted meridian is equivalent to the weighted median [14]. For the given data set of N i.i.d. samples $x_1, \cdots, x_N$ and assigned weights $u_1, \cdots, u_N$, the following equation holds true

$$\lim_{\delta \to \infty} \hat{\beta}_\delta = \lim_{\delta \to \infty} meridian \left\{ u_k * x_k |_{k=1}^N ; \delta \right\} = median \left\{ u_k * x_k |_{k=1}^N \right\} \quad . \tag{6}$$

This property is called the median property. The second interesting case, called the modal property, occurs when the medianity parameter δ tends to zero. In this case, the weighted meridian $\hat{\beta}_\delta$ is equal to one of the most repeated values in the input data set.

2.3 Generalized Cost Function

Comparing the properties of the weighted myriad cost function and weighted meridian cost function common features can be found. One of them is the behavior of the both cost function when the K parameter and the δ parameter tend to zero. Then, for the same data set $\mathbf{X}$, the value of weighted myriad is equal to the value of the weighed meridian. Another common feature of both functions is their similar form, but the weighted myriad cost function uses the L_2 norm while the weighted meridian cost function uses the L_1 norm.

Let the L_p norm be defined as follows

$$||\mathbf{z}||_p = \left(\sum_{l=1}^{s} |z_l|^p \right)^{\frac{1}{p}} \quad , \tag{7}$$

where $\mathbf{z}$ is an s–dimensional real vector (i.e. $\mathbf{z} \in \mathbb{R}^s$). Applying the L_p norm to the weighted myriad cost function (2) or weighted meridian cost function (5), the generalized cost function can be expressed in the following form

$$\chi_\gamma^{(p)}(\nu) = \sum_{k=1}^{N} \log \left[\gamma + u_k ||x_k - \nu||_p \right] \quad , \tag{8}$$

where $|| \cdot ||_p$ is the L_p norm to the p power, and parameter γ corresponds to medianity parameter δ for $p = 1$ and corresponds to linearity parameter K for $p = 2$. It should be mentioned, that for $p = 1$ the γ parameter is equal to medianity parameter δ, but for $p = 2$ parameter γ is equal to the square root of the linearity parameter K (i.e. $\gamma = \sqrt{K}$).

For the given data set $\{x_k\}_{k=1}^N$ and the assigned weights $\{u_k\}_{k=1}^N$, let the $\hat{\nu}_\gamma$ be the value minimizing the cost function (8), i.e.

$$\begin{aligned} \hat{\nu}_\gamma &= \arg \min_{\nu \in \mathbb{R}} \chi_\gamma^{(p)}(\nu) \\ &= \arg \min_{\nu \in \mathbb{R}} \sum_{k=1}^{N} \log \left[\gamma + u_k ||x_k - \nu||_p \right] \end{aligned} \quad . \tag{9}$$

Table 1. Properties of the $\hat{\nu}_\gamma$ estimator

γ	$p=1$	$p=2$
$\gamma \to 0$	most frequent value in the input data set	
$0 < \gamma < \infty$	$\hat{\nu}_\gamma = meridian(u_k * x_k\|_{k=1}^N; \gamma)$	$\hat{\nu}_\gamma = myriad(u_k * x_k\|_{k=1}^N; \sqrt{\gamma})$
$\gamma \to \infty$	$\hat{\nu}_\gamma = median(u_k * x_k\|_{k=1}^N)$	$\hat{\nu}_\gamma = mean(u_k * x_k\|_{k=1}^N) = \frac{\sum_{k=1}^N u_k x_k}{\sum_{k=1}^N u_k}$

Properties of the $\hat{\nu}_\gamma$ value are presented in Table 1.

The function $\chi_\gamma^{(p)}(\nu)$ can be regarded as a generalized cost function. For $p=1$ a weighted meridian is a special case of $\hat{\nu}_\gamma$, and for $p=2$ the weighted myriad is a special case of $\hat{\nu}_\gamma$.

Assuming without loss of generality that the weights are in the unit interval (i.e. $u_k \in [0,1]$ where $1 \le k \le N$), the weights can be interpreted as membership degrees. Then, a weighted myriad $\hat{\Theta}_K$ or a weighted meridian $\hat{\beta}_\delta$ can be interpreted as a fuzzy myriad or fuzzy meridian, respectively. In the rest of this paper, the weights will be treated as a membership degrees and the weighted myriad and weighted meridian will be interpreted as fuzzy myriad and fuzzy meridian. Also, the $\hat{\nu}_\gamma$ value will be interpreted as a fuzzy value.

2.4 Selection the L_p Norm

One of the most popular method for the empirical probability function estimation is the Parzen method [15]. For the given data set of N i.i.d. samples $x_1, \cdots, x_N$, the empirical probability density function (PDF) can be computed as follows

$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{h} K\left(\frac{x - x_i}{h}\right), \tag{10}$$

where N is the number of samples, h is the smooth parameter, and $K(\cdot)$ is the kernel function.

Introducting non–negative cost function Ψ as the measure of fit the empirical PDF $\hat{f}$ to the generalized cost function χ, the L_p norm can be determined from

$$\begin{aligned} p &= \arg \min_{p \in \{1,2\}} \min_{x \in X_i} \Psi_p(x) \\ &= \arg \min_{p \in \{1,2\}} \sum_{k=1}^{N} \|\hat{f}(x) - f_p(x;\gamma)\|_2 \end{aligned}, \tag{11}$$

where

$$f_p(x;\gamma) = \begin{cases} \left(\frac{\gamma}{2}\right) \frac{1}{(\gamma + |x|)^2} & \text{if } p=1, \\ \left(\frac{\gamma}{\pi}\right) \frac{1}{\gamma^2 + x^2} & \text{if } p=2. \end{cases}$$

For $p=1$, function $f_p(x;\gamma)$ describes the Meridian distribution and for $p=2$ describes the Cauchy distribution .

The method of the L_p norm determination can be described as follows:

1. For the input data samples $x_1, x_2, \ldots, x_N$, fix the the kernel function $K(\cdot)$, the smooth parameter h, and the γ parameter,
2. For $p = 2$ compute the myriad based on (1) and compute the value of function $\Psi_2(x)$,
3. For $p = 1$ compute the meridian based on (4) and compute the value of function $\Psi_1(x)$,
4. The $L_p = L_2$ norm if $\Psi_2(x) < \Psi_1(x)$; otherwise the $L_p = L_1$ norm.

3 Hybrid Clustering Method

Let us consider a clustering category in which partitions of data set are built on the basis of some performance index, known also as an objective function [2], [18]. The minimization of a certain objective function can be considered as an optimization approach leading to suboptimal configuration of the clusters. The main design challenge is formulating an objective function that is capable of reflecting the nature of the problem so that its minimization reveals a meaningful structure in the data set.

The proposed method is an objective functional based on fuzzy c–partitions of the finite data set [2],[18]. The suggested objective function can be an extension of the classical functional of within–group sum of an absolute error.

The objective function of the chosen method can be described in the following way

$$J_m^{(p)}(\mathbf{U}, \mathbf{V}) = \sum_{i=1}^{c} \sum_{k=1}^{N} \sum_{l=1}^{s} \log\left[\gamma + u_{ik}^m ||x_k(l) - v_i(l)||_p\right] \quad , \tag{12}$$

where c is the number of clusters, N is the number of the data samples, s is the number of features describing the clustered objects. The γ parameter controls the behavior of cluster prototypes, $u_{ik} \in \mathbf{U}$ is the membership degree of the k–sample to the i–th cluster, the $\mathbf{U}$ is the fuzzy partition matrix, $x_k(l)$ represents the l–th feature of the k–th input data from the data set, and m is the fuzzyfing exponent called the fuzzyfier.

The optimization objective function $J_m^{(p)}$ is completed with respect to the partition matrix $\mathbf{U}$ and the prototypes of the clusters $\mathbf{V}$. By minimizing (12) using Lagrangian multipliers, the following new membership u_{ik} update equation

$$u_{ik} = \left(\sum_{j=1}^{c} \left(\frac{||\mathbf{x}_k - \mathbf{v}_i||_p}{||\mathbf{x}_k - \mathbf{v}_j||_p} \right)^{1/(m-1)} \right)^{-1} \quad , \tag{13}$$

can be derived. For the case, where $||\mathbf{x}_k - \mathbf{v}_i||_p = 0$, then $u_{ik} = 1$ and $u_{jk} = 0$ for $j \in \{1 \cdots c\} - \{i\}$.

For the fixed number of clusters c and the partition matrix $\mathbf{U}$ as well as for the exponent m, the prototype values minimizing (12) are the values described as follows

$$v_i(l) = \arg \min_{\nu \in \mathbb{R}} \sum_{k=1}^{N} \log \left[\gamma + u_{ik}^m ||x_k(l) - \nu||_p \right] \quad , \tag{14}$$

where i is the cluster number $1 \leq i \leq c$ and l is the component (feature) number $1 \leq l \leq s$.

3.1 Clustering Data with the Hybrid Clustering Method

The proposed hybrid clustering method can be described as follows:

1. For the given data set $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ where $\mathbf{x}_i \in \mathbb{R}^s$, fix the number of clusters $c \in \{2, \cdots, N\}$, the fuzzyfing exponent $m \in [1, \infty)$ and assume the tolerance limit ε. Initialize randomly the partition matrix $\mathbf{U}$ and fix the value of parameter γ, fix $l = 0$,
2. Select the appropriate L_p norm for each cluster and each feature based on (11),
3. for the obtained L_p norms calculate the prototype values $\mathbf{V}$ for each feature of $\mathbf{v}_i$ based on (14),
4. update the partition matrix $\mathbf{U}$ using (13),
5. if $\left\| \mathbf{U}^{(l+1)} - \mathbf{U}^{(l)} \right\| < \varepsilon$ then STOP the clustering algorithm, otherwise $l = l + 1$ and go to (3).

4 Numerical Experiments

In the numerical experiments the fuzzfing exponent has been fixed to $m = 2$, and the tolerance limit $\varepsilon = 10^{-5}$, and as tke kernel function the Gaussian kernel was chosen. For a computed set of prototype vectors $\mathbf{V}$ the clustering accuracy has been measured as the Frobenius norm distance between the true centers μ and the prototype vectors. The matrix $\mathbf{A}$ is created as $||\mu - \mathbf{V}||_F$, where $||\mathbf{A}||_F$:

$$||\mathbf{A}||_F = \left(\sum_{i,k} A_{i,k}^2 \right)^{1/2} \quad .$$

4.1 Selection the L_p Norm

The purpose of this experiment is to investigate the proposed method of the L_p selection. Two artifical data set have been generated. The first data set includes noise with outliers. Figure 1 shows the data set and the shapes of empirical PDF and $f_p(x; \gamma)$ function. Figure 2 shows the second data set. This data set includes noise without outliers.

For the both data set, the values of the fit function Ψ_p for $\gamma = 3$ are presented in Table 2. It can be seen, that for the data set with outliers, the Meridian distribution better describes the data set than the Cauchy distribution. Generally, this means, that the sample meridian is a more robust estimator than the myriad estimator. This confirms the fact, that the Mridian distribution is better suited for impulsive noise. In the other hand, for data set without outliers, the Cauchy distribution gives better fit to the empirical PDF than the Maridian distribution.

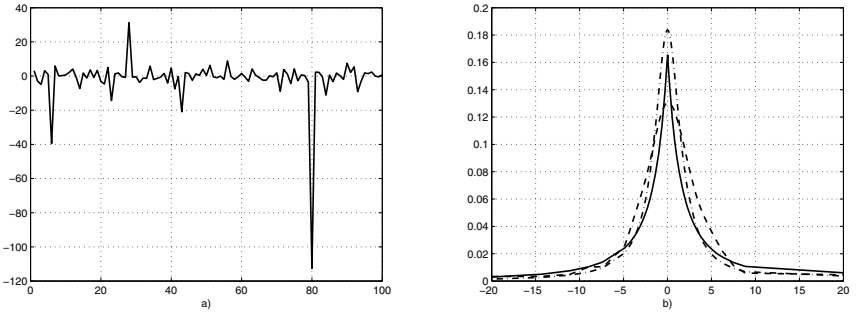


Fig. 1. a) Data set with noise and outliers, b) the Meridian distribution (solid line), the Cauchy distribution (dotted line) and the empirical PDF (dashed line)

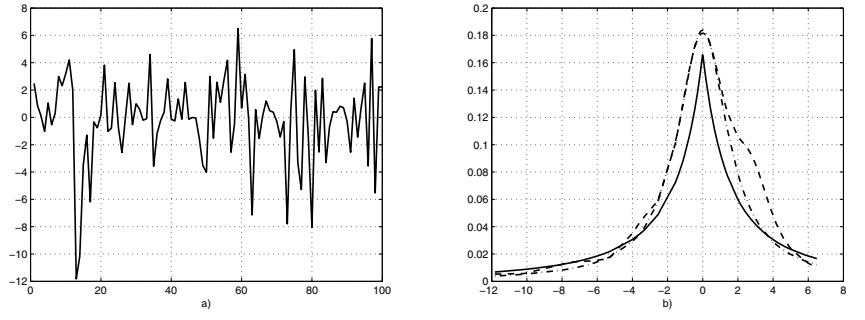


Fig. 2. a) Data set with noise without outliers, b) the Meridian distribution (solid line), the Cauchy distribution (dotted line) and the empirical PDF (dashed line)

Table 2. Values of the $\Psi_p(x)$ function for $\gamma = 3$

data set	$\Psi_1(x)$	$\Psi_2(x)$
with outliers	0.0447	0.0703
without outliers	0.1347	0.0289

4.2 Data Clustering

The example involves three heavy–tailed and overlapped groups of data. The whole data have been generated by a pseudo–random generator. The data set includes 3 groups described by different distributions. The true group centers are: $\mu_1 = [-5, 5]^T$, $\mu_2 = [0, 0]^T$, and $\mu_3 = [10, 0]^T$. The first data set includes overlaping groups generated by Gaussian distributed random generator. In the second data set, the first group data were generated by the Cauchy distributed random generator, while the two others were generated by the Gaussian random generator. In the third data set, the first group data were generated by the Gaussian distributed random generator, while the two other groups were generated

Table 3. The difference among computed cluster centers $\mathbf{V}$ and the true centers μ

	Data set		
γ	I	II	III
1.0	0.6070	2.4653	0.7283
2.0	0.5198	1.0037	0.5852
5.0	0.4792	1.0566	0.5318
10.0	0.4722	1.1900	0.5199
20.0	0.4706	1.4435	0.4953

by the Cauchy distributed random generator. The obtained results for different values of the γ parameter are presented in Table 3.

Small values of parameter γ affect the selectivity of the determination of the cluster prototypes. The selected norm does not influence the results because the for small values of γ the $\hat{\nu}$ value tends to most frequent sample in the data set. For the used data sets, the best results were obtained for $1 < \gamma < 10$.

5 Conclusions

In many cases, the real data are corrupted by noise and outliers. Hence, the clustering methods should be robust for noise and outliers. In this paper the hybrid clustering method is presented. The word *hybrid* stands for different cluster estimation which is dependent on two parameters. The proposed method can be treated as a generalization of two clustering methods: the fuzzy c–means method and the fuzzy c–medians method.

The presented generalization of the cost function allows the application of the L_p norm, where $1 < p < 2$ or $p < 1$. In such cases, it is difficult to interpret and identify the value $\hat{\nu}$.

The current work solves the local minima problem and the performance of the cluster centers estimation for large data sets.

Acknowledgment

This work was partially supported by the Ministry of Science and Higher Education resources in 2010–2012 under Research Project NN518 411138.

References

1. Kaufman, L., Rousseeuw, P.: Finding Groups in Data. Wiley–Interscience, Chichester (1990)
2. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum, New York (1981)
3. Hathaway, R.J., Bezdek, J.C., Hu, Y.: Genalized Fuzzy c-Means Clustering Strategies Using L_p Norm Distances. IEEE Trans. on Fuzzy Sys. 8, 576–582 (2000)
4. Krishnapuram, R., Keller, J.M.: A Possibilistic Approach to Clustering. IEEE Trans. on Fuzzy Sys. 1, 98–110 (1993)

5. Krishnapuram, R., Keller, J.M.: The Possibilistic C–Means Algorithm: Insights and Recomendations. IEEE Trans on Fuzzy Sys. 4, 385–396 (1996)
6. Kersten, P.R.: Fuzzy Order Statistics and Their Application to Fuzzy Clustering. IEEE Trans. on Fuzzy Sys. 7, 708–712 (1999)
7. Huber, P.: Robust statistics. Wiley, New York (1981)
8. Dave, R.N., Krishnapuram, R.: Robust Clustering Methods: A Unified View. IEEE Trans. on Fuzzy System 5, 270–293 (1997)
9. Chatzis, S., Varvarigou, T.: Robust Fuzzy Clustering Using Mixtures of Student's–t Distributions. Pattern Recognition Letters 29, 1901–1905 (2008)
10. Frigui, H., Krishnapuram, R.: A Robust Competitive Clustering Algorithm With Applications in Computer Vision. IEEE Trans. Pattern Analysis and Machine Intelligence 21, 450–465 (1999)
11. Sun, J., Kaban, A., Garibaldi, J.M.: Robust mixture clustering using Pearson type VII distribution. Pattern Recognition Letters 31, 2447–2454
12. Arce, G.R., Kalluri, S.: Fast Algorithm For Weighted Myriad Computation by Fixed Point Search. IEEE Trans. on Signal Proc. 48, 159–171 (2000)
13. Przybyła, T.: Fuzzy c–Myriad Clustering Method. System Modeling Control, 249–254 (2005)
14. Aysal, T.C., Barner, K.E.: Meridian Filtering for Robust Signal Processing. IEEE Trans. on Signal Proc. 55, 3949–3962 (2007)
15. Parzen, E.: On Estimation Of A Probability Density Function And Mode. Ann. Math. Stat. 33, 1065–1076 (1962)
16. Przybyła, T., Jeżewski, J., Horoba, K.: The Adaptive Fuzzy Meridian and Its Application to Fuzzy Clustering. In: Advances in Intelligent and Soft Computing, vol. 57, pp. 247–256. Springer, Heidelberg (2009)
17. Arce, G.R., Kalluri, S.: Robust Frequency–Selective Filtering Using Weighted Myriad Filters admitting Real–Valued Weights. IEEE Trans. on Signal Proc. 49, 2721–2733 (2001)
18. Pedrycz, W.: Konwledge–Based Clustering. Wiley–Interscience, Chichester (2005)

Using Intelligence Techniques to Predict Postoperative Morbidity of Endovascular Aneurysm Repair

Nan-Chen Hsieh[1], Jui-Fa Chen[2], Kuo-Chen Lee[3], and Hsin-Che Tsai[2]

[1] Department of Information Management, National Taipei University of Nursing and Health Sciences, Taiwan, Republic of China
[2] Department of Computer Science and Information Engineering, Tamkang University, Taiwan, Republic of China
[3] Division of Cardiovascular Surgery, Department of Surgery Heart Center, Cheng-Hsin General Hospital, Taiwan, Republic of China

Abstract. Endovascular aneurysm repair (EVAR) is an advanced minimally invasive surgical technology that is helpful for reducing patients' recovery time, postoperative mortality and morbidity. This study proposes an ensemble model to predict postoperative morbidity after EVAR. The ensemble model was developed using a training set of consecutive patients who underwent EVAR between 2000 and 2009. All data required for prediction modeling, including patient demographics, preoperative, co-morbidities, and complication as outcome variables, was collected prospectively and entered into a clinical database. A discretization approach was used to categorize numerical values into informative feature space. The research outcomes consisted of an ensemble model to predict postoperative morbidity, the occurrence of postoperative complications prospectively recorded, and the causal-effect decision rules. The probabilities of complication calculated by the model were compared to the actual occurrence of complications and a receiver operating characteristic (ROC) curve was used to evaluate the accuracy of postoperative morbidity prediction. In this series, the ensemble of Bayesian network (BN), artificial neural network (ANN) and support vector machine (SVM) models offered satisfactory performance in predicting postoperative morbidity after EVAR.

Keywords: Endovascular aneurysm repair (EVAR), postoperative morbidity, ensemble model, machine learning.

1 Introduction

Aortic surgery is a complex surgical operation that is indicated for patients with severe insufficiency in cardiac function. Major cardiac surgical interventions include coronary artery bypass grafting (CABG), repair of congenital heart defects, surgical treatment of atrial fibrillation, heart transplantation, repair or replacement of heart valves, aortic surgery, aneurysm repair or a combination of these surgical procedures. During the operation and the postoperative stay at the ICU and nursing ward, there is considerable morbidity for aortic surgery patients with postoperative complications, which results in increased hospital mortality and postoperative morbidity. Many prediction models for

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 197–206, 2011.

cardiac surgical outcome apply logistic or multivariable regression to assess preoperative risk [1-3]. Most of the risk prediction models in current use were derived for patients undergoing open abdominal aortic aneurysm (AAA) repair and appear to lack utility when applied to EVAR patients. The predicted outcome can be used by surgeons and patients to evaluate whether or not the surgical procedures is likely to be successful. Similarly, postoperative morbidity is a key factor in recovery and through-put of cardiac hospital patients. Prediction of surgical mortality and postoperative morbidity is important in selecting low-risk patients for operation, and in counseling patients about the risks of undergoing surgical operation. The development of a robust prediction model can therefore both assist vascular surgeons in evaluating the expected outcome for a given patient and facilitate counseling and preoperative decision-making. Reliable and accurate prediction of operative mortality and morbidity is an essential criterion for any such risk evaluation models. EVAR is an advanced minimally invasive surgical technology that helps reduce patients' recovery time as well as postoperative mortality and morbidity[4]; it is especially helpful in the treatment of patients judged to be high surgical risk for conventional surgery. EVAR benefits patients with medical co-morbidities, and postoperative complications highly significantly influence longer-term postoperative outcomes in EVAR patients.

Data mining techniques are currently used in medical decision support to increase diagnostic accuracy and to provide additional knowledge to medical stuff. Their increased use provides expanded opportunities for determining the utility of medical decision-making models from retrospective data. Compared to data mining for business applications, medical data mining includes prediction data mining and descriptive data mining two distinct concepts. That is, medical data mining not only requires a prediction model with satisfactory accuracy, but also requires a safety context in which decision-making activities require explanatory support. The main distinction is that predictive data mining requires that the training dataset include an outcome variable, while descriptive data mining uses a global strategy to find the characteristics of each affinity granulation of the data. Both these data mining techniques can produce accurate, predictive and interpretable descriptive models that contribute greatly to handling medical data gathered through systematic use of clinical, laboratory, and hospital information systems. The goal of predictive data mining in clinical surgery is to derive models that can use medical data to predict patient's mortality and morbidity and thereby support clinical surgical decision-making. Predictive data mining can also aid in prognosis, diagnosis and treatment planning for surgical procedures. In contrast, descriptive data mining considers the data as affinity granulations, and aims at finding interpretable patterns and associations among data.

The use of machine learning models has become widely accepted in medical applications. Various machine learning models including BNs, ANNs, and SVMs have been tested in a wide variety of clinical and medical applications[5]. Soft-computing, including fuzzy set and rough set techniques that work well in descriptive data mining, is also a promising technique. BN is a probability-based inference model that has a wide range of applications and is increasingly used medically as a prediction and knowledge representation modeling technique. Verduijn, M., et al. [6] presented the prognostic BN as a new type of prognostic model that builds on the BN methodology and implements a dynamic, process-oriented view of cardiac surgical prognosis. Lin

and Haug [7] proposed BN suitable for exploiting missing clinical data for medical prediction. ANNs have featured in a wide range of medical applications, often with promising results. Eom et al. [8] developed a classifier ensemble-based, including ANNs, DTs, and SVMs, clinical decision support system for predicting cardiovascular disease level. SVMs have been successfully used in a wide variety of medical applications. Polat and Güne [9] used a least square support vector machine to assist breast cancer diagnosis. Babaoĝlu et al. [10] first used principle component analysis method to reduce data features, and acquired an optimum support vector machine model for the diagnosis of coronary artery disease. Choi [11] proposed the detection of valvular heart disorder (VHD) by wavelet packet decomposition and SVM techniques.

This study describes the development of an informative ensemble prediction model consisting of BNs, ANNs and SVMs for the prediction of postoperative morbidity between preoperative variables and complication outcomes in EVAR patients. For a better understanding of our study, Section 2 of this paper begins with an overview of study background and experimental methods in general. Section 3 describes the experimental design and procedures used in this study, including entropy/MDL-based method for the discretization of numerical features, feature selection method, ensemble model for the prediction of postoperative morbidity. Section 4 discusses the experimental findings and offers observations about practical applications and directions for future research.

2 Study Background and Materials

Abdominal aneurysm (AAA) is an enlargement that occurs in a weakened area within the largest artery in the abdomen (http://www.vascularweb.org/patients/NorthPoint/Abdominal_Aortic_Aneurysm.html). If an AAA is not treated in due time, the pressure generated by heartbeats causes the aneurysm to continuously grow larger, and the aortic wall continues to weaken. Finally, rupture occurs and massive internal bleeding occurs. The best way to prevent the high mortality associated with AAA is to find the lesion before rupture occurs. However, patients with aortic diseases are often elderly with severe co-morbidities and sometimes devastating morbidity, making them extremely challenging candidates for surgery. For such patients, EVAR represents a lower risk approach than conventional open surgery and is associated with shorter operating times, shorter hospitalizations, more rapid recovery and improved quality of life during the perioperative period and postoperative follow-up. Although long-term data on the clinical outcomes of patients who received EVAR are not yet available, given its importance, building a prediction of postoperative morbidity after EVAR is critical.

We retrospectively examined 140 consecutive patients who underwent EVAR surgery at Taipei Veteran General Hospital, a teaching center hospital in Taiwan, between 2000 and 2009. The dataset contains preoperative patient characteristics, details of the operative information, and pathological and laboratory findings from the general ward, operating room and intensive care unit (ICU). The dataset also included length of ICU stay, variables that describe postoperative complications that frequently

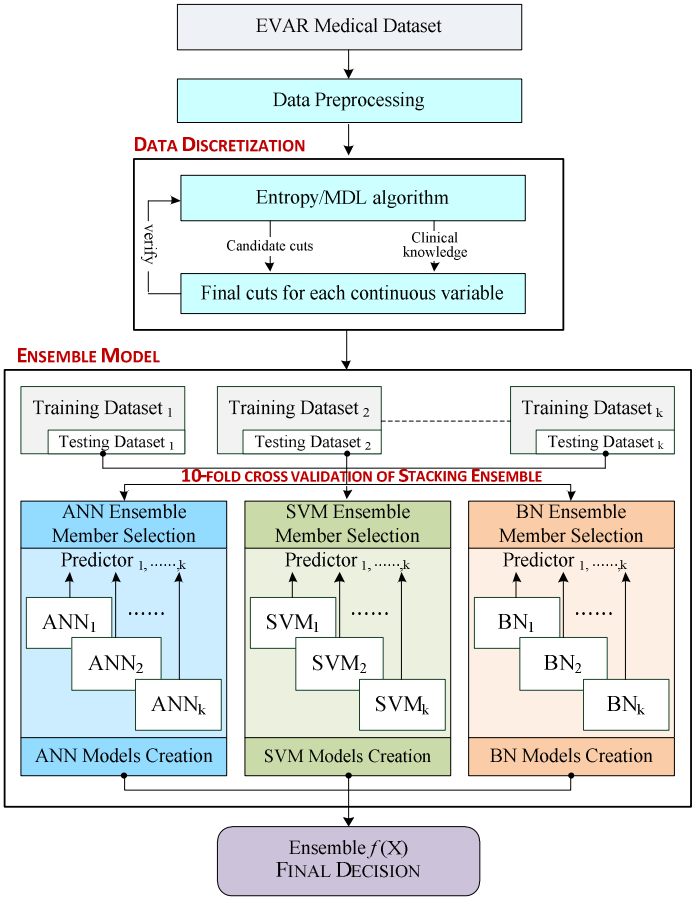


Fig. 1. Proposed architecture for ensemble model development

occur in EVAR surgery, death during hospitalization, and time of death for patients who expired. Postoperative complication was used as the binary outcome variable of the ensemble model, and types of complications were used as subsidiary outcome variables. The original dataset contained 137 variables, but included many missing values. Preliminary inspection of the dataset showed that many variables contained missing values for at least 50% of the patients; these variables were not included in further analysis. In order to identify significant variables for use in the ensemble model, a number of criteria were employed. Variables that were subjective, ambiguous or inadequately defined were excluded; variables that were frequently incomplete were also excluded from subsequent analysis. Sixty-seven of the 140 patients experienced postoperative complications during their stay at hospital. Data collected included preoperative patient characteristics, risk factors, details of the operative information, and physical characteristics of the aneurysm, postoperative physiological and laboratory findings, and postoperative complications as the outcome variable.

The development of this prediction model proceeds as follows Fig.1. First, an informative discretization method for numerical values, which employs a data discretization method to guide the categorized numerical features on the basis of entropy/MDL algorithm and laboratory surgeon' s knowledge, is developed. Through data discretization, numerical values are converted into discrete values. Second, the proposed ensemble-based architecture focuses on fusing three types of models. BNs, ANNs, and SVMs). During the training process, the same training data set is used for all individual models in order to reduce the diversity among individual models, keeping in mind that, in an ensemble model, it is important to construct appropriate training data sets that maintain good balance between accuracy and diversity among individual models. The model selection scheme, designated a stacking scheme, is a mixture of stacking and cross-validation that is chosen in order to improve overall classification by combining models trained on randomly generated subsets of the entire training set.

3 Experimental Design

3.1 Entropy/MDL-Based Method for the Discretization of Numerical Values

Most studies dealing with cardiac surgery prediction models have applied logistic or multivariable regression to assess the preoperative risk. Few studies have utilized machine learning algorithms i.e., decision trees, Bayesian networks or artificial neural networks in analyzing clinical data. These state-of-the-art machine learning algorithms are often informative and can represent more knowledge in the clinical data. Generally, these algorithms require discrete categorical values but clinical datasets usually involve numerical variables. To satisfy the requirements of machine learning algorithms, the employment of a discretization approach is necessary. Discretization is defined as a process that divides numerical values into states of discrete categorical values, leading to informative expressed categorical values. For example, the CART model originally was not designed to handle numerical attributes. During the construction of a CART model, numerical attributes were divided into discrete categorical values. Aside from CART, discretization techniques were frequently adopted in other popular learning paradigms, such as C4.5, BNs, ANNs, and genetic algorithms.

This study employed entropy/MDL to divide numerical domains into intervals. Incorporation investigates data discretization; the intervals are characterized by discrete categorical values. The steps for automatically finding discrete categorical values from a given dataset are described herein. Assuming that the domain of a numerical attribute ranges from v1 to v2, and that {c1, c2,…,ck} denote the k cut-points obtained by the entropy/MDL algorithm, using these k cut-points, k discretize values can be determined. For example, AAA_Size is a numerical attribute obtained from the dataset. The domain of AAA_Size ranges from 5.0 to 9.6. Six cut-points, (5.5, 5.7, 5.8, 7.3, 8.1, 9.6), were computed using the entropy/MDL algorithm. As depicted in Fig. 2, six discrete categorical values were obtained. Within each analysis, the laboratory system should contain an interval that delimits life-compatible values. Therefore, if cut

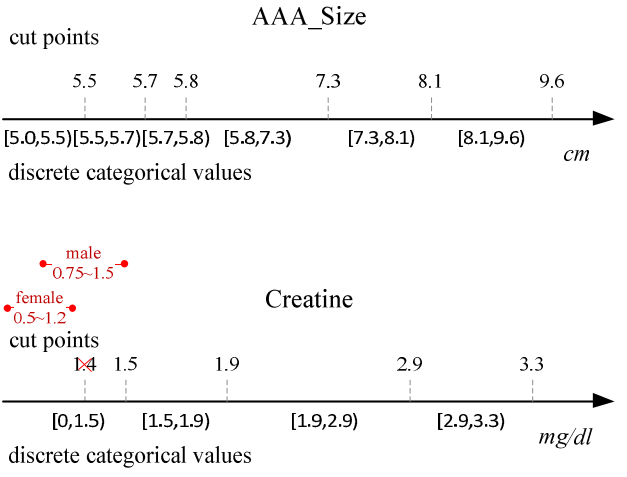


Fig. 2. The discrete categorical values of AAA_Size and Creatine

points fall within the interval, they should be eliminated. For example, one cut point {1.4 mg/dl} of creatine lay within an interval, i.e., [0.75~1.5] for male, [0.5~1.2] for female; therefore the cut point {1.4 mg/dl} was eliminated.

3.2 Significant Attribute Selection

Feature selection methods can be carried out by attribute ranking methods and attribute subset evaluation methods [12]. The attribute ranking methods assess the goodness of individual variables for prediction independently of other attributes and only the high ranked attributes are used to build the prediction model. For practical applications, the former methods can be used as an initial screening to reduce dimensionality in highly dimension datasets. The later methods can be used to find more relevant subset of attributes simultaneously. Hall and Holmes [12] compared six feature selection methods and benchmarking attribute selection techniques for discrete class data mining. The best performing methods according to their study were Information Gain, Recursive Elimination of Features (Relief), Correlation-based Feature Selection (CFS), Consistency-based Subset Evaluation, and Wrapper Subset Evaluation. The former two are attribute ranking methods and the latter three are attribute subset evaluation methods. The results of this benchmark study provide guidelines for the choice of feature selection methods. This search property is in favor of small feature subsets with high class consistency which is more proper for clinical application.

This study used entropy/MDL to discrete numeric attributes. The attributes are then ranked according to their overall contribution to the consistency of the attribute set. As shown in Table 1, we selected attributes by three attribute subset evaluation algorithms. The obtained attribute subsets were employed as major input variables for training ensemble prediction models based on StackingC algorithm.

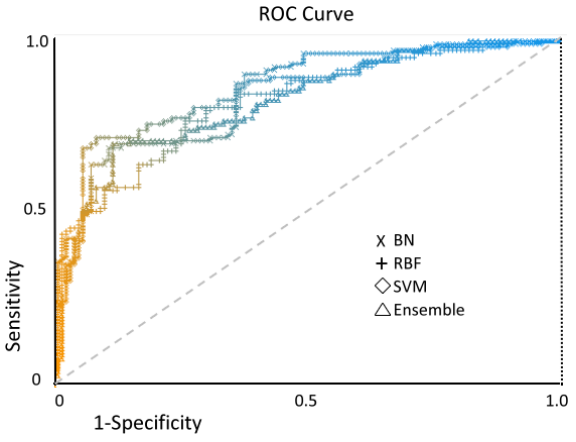
Table 1. Selected attributes by attribute subset evaluation algorithms

Variable	Definition	CFS	Consistency	Wrapper (C4.5)
Gender	Male/female			●
Age	0-100 years			
Smoking	No smoking, smoking, uncertain		●	●
Hypertension	Yes/no		●	
DM	Diabetes mellitus, yes/no	●	●	●
Hyperlipidemia	Yes/no		●	●
COPD	Chronic obstructive pulmonary disease, yes/no		●	●
CVA	Cerebral vascular accident, yes/no	●		
Heart Disease	Yes/no	●	●	
CRI	Chronic renal insufficiency, yes/no	●		
Hgb	Hemoglobin level, g/dl	●	●	●
Hct	Hematocrit percentage, 0-100%	●		
PLT	Platelet count, /CUMM	●	●	●
BUN	Blood urea nitrogen level, mg/dl	●	●	
Creatine	Creatine level, mg/dl	●	●	
AAA size	Size of abdominal aortic aneurysm, mm		●	
AAA site	Site of abdominal aortic aneurysm			●

3.3 Ensemble Model for the Prediction of Postoperative Morbidity

In this study, BNs, ANNs, and SVMs were chosen as based models because they represented different approaches, each of which is simple to implement and has been shown to perform well in medical applications. The rationale of employing these models is that BNs can easily model complex relationships among variables, ANNs are generally superior to conventional statistical models, and SVMs perform reasonably well in most prediction problems and can be used as a benchmark technique. Then, each individual model makes its own prediction estimating probabilities for each class. The final prediction of stacking is computed using multiple-linear regression as a Meta classifier. For the implementation of models, we chose BayesNet, MultilayerPerceptron and SMO as base models in WEKA, and StackingC as ensemble method [13]. All the default parameters in WEKA were used. Besides, we used 10-fold cross-validation to lower the variability of training set.

The model selection scheme is a mixture of stacking [14] and cross-validation that aims to improve the classification by combining models trained on randomly generated subsets of the entire training set. We first applied a cross validation scheme for model selection on each subset; subsequently, for the sake of simplicity and so as not to run



Models	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
10 folds CV- bagging weighted average						
BN	0.736	0.270	0.737	0.736	0.734	0.834
NN	0.675	0.330	0.657	0.657	0.674	0.755
SVM	0.757	0.247	0.758	0.757	0.756	0.755
Ensemble	0.757	0.249	0.760	0.757	0.755	0.813

Fig. 3. The results of the postoperative morbidity prediction

into over-fitting problems, we combined the selected models by stacking approach to reach a final prediction. In cases where specific inside medical knowledge is not available, such a cross validation method can be used to select a classification method empirically, because it seems to be obvious that no classification method is uniformly superior [15]. The result, a heterogeneous ensemble, allows classification methods to be used more effectively. The detailed accuracy of individual models and of the ensemble model is shown in Fig 3. The results of the experiment show that individual models' performance and improvements in performance were achieved by applying the ensemble of models. This indicates that model combination techniques indeed yield stable performance improvements for ensemble models.

4 Discussions

In comparison to open repair, endovascular aneurysm repair of abdominal aortic aneurysms provides a lower risk approach associated with a shorter operating time, shorter hospitalization stay, more rapid recovery time and improved quality of life during the perioperative period and postoperative follow-up. Identification of the reasons for postoperative complications and risk factors for re-intervention during follow-up to maintain aneurysm exclusion remain challenges for surgeons. Most surgeons perform EVAR on sicker patients; however, patients with aneurysm anatomy are usually elderly and might be considered marginal for EVAR repair. These patients are

likely to have a relatively high postoperative morbidity rate with complications, and will highly influence longer-term postoperative outcomes. It is essential to create reliable and satisfactory risk prediction models for postoperative morbidity as an aid to clinical decision-making. Although several risk prediction systems have been proposed for patients undergoing open aneurysm repair, they basically rely on traditional statistical methods and provide scant accuracy and utility when applied to EVAR patients.

We have proposed an ensemble model to predict postoperative morbidity after EVAR and support clinical decision-making. The proposed ensemble model is constructed by incorporating discretization of categorical numerical values; BNs, ANNs, and SVMs were used to augment the ensemble model and the dataset was processed by cross validation, showing moderate performance. The experimental result shows that the proposed ensemble model predicts postoperative morbidity with relatively satisfactory accuracy, even when data is missing and/or sparse, showing its usefulness in support of clinical decision-making. The supplementary nature of multi-models distinguish the proposed model from existing risk scoring systems that are based on conventional statistical methods and from various machine learning models. To summarize, the advantage of using the proposed ensemble model is that it can provide surgeons with practical, relatively accurate aid in their daily diagnostic tasks and enable them to extract meaningful relationships among features of medical datasets through the use of constrained decision rules.

References

1. Barnes, M., Boult, M., Maddern, G., Fitridge, R.: A model to predict outcomes for Endovascular aneurysm repair using preoperative variables. European Journal of Vascular and Endovascular Surgery 35, 571–579 (2008)
2. Bohm, N., Wales, L., Dunckley, M., Morgan, R., Loftus, I., Thompson, M.: Objective risk-scoring systems for repair of abdominal aortic aneurysms: applicability in Endovascular repair. European Journal of Vascular and Endovascular Surgery 36, 172–177 (2008)
3. Stijn, C.W., Wouters, L.N., Freek, W.A.V., Rene, M.H.J.B.: Preoperative prediction of early mortality and morbidity in coronary bypass surgery. Cardiovascular Surgery 10, 500–505 (2002)
4. Bush, R.L., Johnson, M.L., Hedayati, N., Henderson, W.G., Lin, P.H., Lumsden, A.B.: Performance of endovascular aortic aneurysm repair in high-risk patients: Results from the Veterans Affairs National Surgical Quality Improvement Program. Journal of Vascular Surgery 45(2), 227–235 (2007)
5. Roques, F., Michel, P., Goldstone, A.R., Nashef, S.A.: The logistic EuroSCORE. European Heart Journal 24, 1–2 (2003)
6. Bellazzi, R., Zupan, B.: Predictive data mining in clinical medicine: Current issues and guidelines. International Journal of Medical Informatics 77, 81–97 (2008)
7. Verduijn, M., Peek, N., Rosseel, P.M.J., Jonge, E.d., Mol, B.A.J.M.d.: Prognostic Bayesian networks I: Rationale, learning procedure and clinical use. Journal of Biomedical Informatics 40, 609–618 (2007)
8. Lin, J.H., Haug, J.: Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. Journal of Biomedical Informatics 41, 1–14 (2008)

9. Rowan, M., Ryan, T., Hegarty, F., O'Hare, N.: The use of artificial neural networks to stratify the length of stay of cardiac patients based on preoperative and initial postoperative factors. Artificial Intelligence in Medicine 40, 211–221 (2007)
10. Lisboa, P.J., Taktak, A.F.G.: The use of artificial neural networks in decision support in cancer: A systematic review. Neural Networks 19, 408–415 (2006)
11. Dokur, Z.: A unified framework for image compression and segmentation by using an incremental neural network. Expert Systems with Applications 34, 611–619 (2008)
12. Hall, M.A., Holmes, G.: Benchmarking attribute selection techniques for discrete class data mining. IEEE Transactions on Knowledge and Data Engineering 15, 1–16 (2003)
13. Witten, I.H., Frank, E.: Data mining: Practical machine learning tools and techniques. Morgan Kaufmann Publishers, Boston (2005)
14. Seewald, A.: How to make stacking better and faster while also taking care of an unknown weakness. In: Proceedings of the 19th International Conference on Machine Learning, pp. 554–561 (2002)
15. Quinlan, J.R.: Comparing connectionist and symbolic learning methods. In: Computational Learning Theory and Natural Learning Systems, vol. 1. MIT Press, Cambridge (1994)

Using Quick Decision Tree Algorithm to Find Better RBF Networks*

Hyontai Sug

Division of Computer and Information Engineering, Dongseo University,
Busan, 617-716, Korea
hyontai@yahoo.com

Abstract. It is known that generated knowledge models for data mining tasks are dependent upon supplied data sets, so supplying good data sets for target data mining algorithms is important for the success of data mining. Therefore, in order to find better RBF networks of k-means clustering efficiently, we refer to the number of errors that are from decision trees, and use the information to improve training data sets for RBF networks and we also refer to terminal nodes to initialize the k value. Experiments with real world data sets showed good results.

Keywords: decision tree, radial basis function network, classification, large database.

1 Introduction

In the field of data mining tasks there are two challenges that may hinder success for the tasks. The first challenge is the fact that there can be a lot of data that can cause computational complexity problem, and the second challenge is the fact that data may not be complete for target data mining models so that the trained knowledge model might act poorly for future unseen cases. There are many data mining algorithms to deal with the problem [1], and decision trees and artificial neural networks are some of representative algorithms for the problem. In order to cope with the first problem decision trees can be used, since they are especially good for handling large data sets because of relatively shorter training time. On the other hand, for incomplete or imperfect data problem artificial neural networks can be used, since they are known to be good for the problem with increased computational complexity.

For tasks of data mining artificial neural networks like MLPs and radial basis function (RBF) networks are mostly used because of their good performance in many applications [2, 3, 4]. We are especially interested in RBF networks, because the neural networks have been applied successfully for classification tasks of data mining [5]. RBF networks make approximation based on the training data, and Gaussian functions are used mostly as the radial basis function. In order to train RBF networks first we should find appropriate centre and radius of radial basis function. For this

* This work was supported by Dongseo University, "Dongseo Frontier Project" Research Fund of 2010.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 207–217, 2011.

task, we may use some unsupervised learning algorithms like k-means clustering. K-means clustering is one of the mostly used algorithms for clustering [6]. For k-means clustering an appropriate number of clusters has to be given for initialization. For this initialization we usually use domain knowledge to set the number of clusters. So, the task of setting the number of clusters is arbitrary in nature, so is true for the task of initializing the number of clusters of RBF networks. But the RBF networks have different performance depending on the number of clusters and training data sets, so we want to find better RBF networks exploiting decision trees that can be trained more quickly than RBF networks.

In section 2, we provide related work to the research, and in sections 3 we present the procedure. Experiments were run to see the effect of the method in section 4. Finally section 5 presents some conclusions and future work.

2 Related Work

Because it is easy for us to understand the structure of decision trees unless they are not very large, decision trees have been used often as a knowledge model for the task of data mining. Research efforts have been devoted to build better decision trees. Among them C4.5 is some representative decision tree algorithm, because it has been referred often in literature and freely available [7]. C4.5 uses entropy-based measure and generates decision trees in relatively quick time, but the generated tree size is relatively big.

When available data set size is relatively small, artificial neural networks are regarded as good data mining tools [8]. Among many artificial neural networks MLPs and RBF networks are some representative neural networks that have been often referred in literature [9]. A good point of neural networks is robustness to irrelevant features as well as erroneous data. RBF networks are one of the most popular feed-forward networks [10]. Even though RBF networks have three layers including the input layer, hidden layer, and output layer, they differ from MLPs, because the hidden units of RBF networks are constructed based on some clustering algorithms mostly.

There were some efforts to use decision trees to build better RBF networks. Kubat [11] tried to utilize the information of terminal nodes of C4.5 in building RBF networks. The terminal nodes were used as center points for clustering for RBF network. He showed that the RBF networks have better accuracy than decision trees of C4.5 in some data sets. Schwenker et al. also showed that decision trees can be used to initialize three kinds of RBF networks deterministically [12]. But, because the task of generating an optimal decision tree is NP-complete problem, the data space divided by decision tree is one of many possible ways to divide data space, so it is not easy to mention that the RBF networks are optimal.

Because training task of data mining models like neural networks is induction, the behavior of trained data mining models is dependent on the training data set. So, we can infer that the trained knowledge model will be dependent on sample size as well as the composition of data in the samples. Fukunaga and Hayes [13] discussed the effect of sample size for parameter estimates in a family of functions for classifiers. SMOTE method [14] used synthetic data generation method for minor classes, and showed that it is effective for decision trees. In [15] the authors showed that

class imbalance has different effect in neural networks for medical domain data. In previous work [16] experiments with smaller sample sizes from original data sets were tried, and it showed good results. So in this paper we want to expand the work with modified method in biased sampling to cope with class imbalance for larger data sets.

3 The Method

Most target data sets for data mining have some skewed distribution in class values, and this fact can be checked easily by inspecting the terminal nodes of decision trees. Moreover, we may also use the information of the number of terminal nodes in the trees as the initial number of clusters for k-means clustering of RBF networks. If the generated tree is very large, the task of interpreting the structure of generated tree is difficult. So, we want to use the information of the number of terminal nodes of decision trees only to find better RBF networks. The method first builds a decision tree using some fast decision tree generation algorithms like C4.5. Then, we inspect the number of misclassified objects for each class. Then we choose classes that should be sampled more for more balanced training set of samples with respect to class value distribution in the samples.

We use the number of terminal nodes in the decision tress to determine the initial number of clusters for the RBF network. But the initial value for the number of clusters in RBF network might not be the best value for the given data set. So we first try to decrease the number of clusters from the initial value in arithmetical progression, then we also increase the number of clusters from the initial value. But increasing or decreasing the number of clusters sequentially and generating corresponding RBF networks may take a lot of computing time without much improvement in accuracy, so we increment or decrement the number as some multiple of the initial number of clusters. If the accuracy values of RBF networks do not increase within given criteria, the search stops. The following is a brief description of the procedure of the method.

```
procedure (Output)
  /* X, K, D,σ: parameters */
 1. Generate a decision tree;
  2. Inspect the terminal nodes of the decision tree to
      determine further sampling for inferior classes
      and count the number of terminal nodes;
 3. Do sampling of X % more for inferior classes
 4. Initialize the number of clusters of RBFN as C where
     C is the largest number that is less than the number
     of terminal nodes and the multiple of the number of
     classes;
  5. Generate a RBFN /* initial_accuracy = the accuracy
     of the network */
  6. loop_better_accuracy := initial_accuracy;
     global_better_accuracy := initial_accuracy;
```

```
/* check decreasingly */
7. Repeat K times
7.1 Generate a RBFN after decreasing the number
   of clusters by D;
7.2 If the accuracy of RBFN > loop_better_accuracy Then
      loop_better_accuracy := the accuracy of RBFN;
   End if;
8. End repeat
9. If loop_better_accuracy > global_better_accuracy
  Then
     K := K - σ;
     global_better_accuracy:=loop_better_accuracy;
     Go to 7;
  End If;
/* check increasingly */
10. loop_better_accuracy := global_better_accuracy;
11. Repeat K times
11.1 Generate a RBFN after increasing the number of
      clusters by D;
11.2 If the accuracy of RBFN > loop_better_accuracy
   Then
      loop_better_accuracy := the accuracy of RBFN;
   End if;
12. End repeat
13. If loop_better_accuracy > global_better_accuracy
Then
     K := K - σ;
     global_better_accuracy:=loop_better_accuracy;
     Go to 11;
  End If;
End.
```

In the above procedure there are four parameters to be defined, X, K, D, and σ. X represents additional percentage to do more sampling. K represents the number of repeats in generating RBF networks while we increase or decrease the number of clusters by D. Depending on the number of terminal nodes of the decision tree, we set the value of D and K appropriately. σ is for the adjustment of K value for the next round of the loop. In the following experiment X is set to 20%, K is set to five, and σ is set to two, and D is set depending on how many classes we have and how many terminal nodes exist in the generated decision tree. One may give smaller value of D, if he wants more through search. Increasing or decreasing the number of clusters will be stopped, when the accuracies of the generated RBF networks are not improved further. We decrease the number of clusters from starting point first, because training time of RBF networks for smaller number of clusters takes less time.

4 Experimentation

Experiments were run using data sets in UCI machine learning repository [17] called ‘adult’ [18] and 'statlog(Landsat satellite)' [19] to see the effect of the method. The number of instances in adult data set is 48,842, and the number of instances in statlog data set is 6,435. The data sets were selected, because they are relatively large, and adult data set may represent business domain and stalog data set may represent scientific domain. The total number of attributes is 14 and 36, and there are two classes and six classes for adult and statlog data set respectively. There are six continuous attributes for adult data set, and all attributes are continuous attributes for stalog data set. We used RBF network using K-means clustering [20] to train for various number of clusters. Because most applications of RBF network use relatively small-sized data sets, we did sampling of relatively small sizes for the experiment to simulate the situation. For adult data set sample size of 960 and 1,920 are used, and for stalog data set sample size of 480 and 960 are used. All the remaining data are used for testing. For each sample size seven random sample data sets were drawn. Used decision tree algorithm is C4.5 and default pruning parameter of 25% is used.

Before we sample in the above mentioned sample sizes, we sampled the sample size of 800 and 1,600 for adult data set, and the sample size of 400 and 800 for statlog data set to determine which class objects should be sampled more. Table 1 and table 2 shows average misclassification ratio of fourteen random sample sets for each class when we generate decision trees of C4.5. All sample data are used to generate the trees.

Table 1. Average misclassification ratio for each class of adult data set samples

Class	Misclassification ratio
>50K	35.5%
≤50K	4.7%

Table 2. Average misclassification ratio for each class of statlog data set samples

Class	Average misclassification ratio
1	1.2%
2	1.9%
3	1.6%
4	14.1%
5	8.1%
6	3.2%

So, 20% more objects were sampled from the object pool of ‘>50K’ class for adult data set, and 10% more objects were sampled for each class 4 and 5 for statlog data set. The following table 3 to 10 show accuracies RBF networks depending on the number of clusters for adult and statlog data set. In the tables, the first row contains results of decision tree C4.5, and in the second row ‘#cls’ means the number of clusters, and ‘Acc.’ means accuracy in percentage. ‘*’ in the body of the table

Table 3. Results of adult data set with sample size of 960

Sample 1: Accuracy: 81.9761 # terminal nodes: 17		Sample 2: Accuracy: 81.8153 # terminal nodes: 72		Sample 3: Accuracy: 84.0897 # terminal nodes: 40		Sample 4: Accuracy: 80.4662 # terminal nodes: 89	
#cls	Acc.	#cls	Acc.	#cls	Acc.	#cls	Acc.
		32	82.2038			24	82.1435
		40	82.3583	8	82.6904	32	81.9699
4	**83.4506**	48	82.2122	16	82.9933	40	81.7903
8	82.4168	56	82.0576	24	**83.0016**	48	**82.4085**
12	81.8989	64	82.2205	32	82.5296	56	81.8341
*16	82.3542	*72	82.4962	*40	82.8262	64	81.5209
20	82.0555	80	**82.776**	48	82.3918	72	81.8049
24	82.6211	88	81.7986	56	82.4001	80	81.6921
28	81.5605	96	81.1115	64	82.3792	*88	81.3392
32	82.1474	104	80.2098	72	81.8884	96	81.1303
36		112	81.2577	80	82.0242	104	81.3266
		120	80.9444			112	81.9761
		128	80.5393			120	82.0346
		136				128	81.9072

Table 4. Results of adult data set with sample size of 960 (cont.)

Sample 5: Accuracy: 83.3693 # terminal nodes: 30		Sample 6: Accuracy: 79.8793 # terminal nodes: 51		Sample 7: Accuracy: 82.7677 # terminal nodes: 67	
#cls	Acc.	#cls	Acc.	#cls	Acc.
		10	82.6319	26	**82.9431**
		18	**83.2209**	34	82.586
6	83.6678	26	83.1019	42	81.5041
14	**83.7033**	34	83.0413	50	81.4958
22	83.6762	42	83.1541	58	81.7151
*30	83.5091	*50	82.9619	*66	80.863
38	83.2564	58	82.4001	74	81.0656
46	83.342	66	82.2644	82	80.4181
54	83.3274	74	81.6169	90	80.0735
62	83.035	82	82.1913	98	80.0798
70		90	81.5376	106	80.0338

indicates the initial number of clusters. The number is based on the number of terminal nodes in the corresponding decision tree of C4.5. The best accuracy value in the experiment for each sample set is represented in bold numbers. Table 3 and 4 show the result of experiment for adult data set with sample size of 960.

If we give attention to sample 3 in table 3, the accuracy of RBF network is worse than the accuracy of decision tree of C4.5. But RBF networks for other sample sizes are better than C4.5, so we can say that better RBF networks can be found because of the repeated trials with different number of clusters. Table 5 and 6 show the result of experiment for adult data set with sample size of 1,920.

Table 5. Results of adult data set with sample size of 1920

Sample 1: Accuracy: 82.2915 # terminal nodes: 68		Sample 2: Accuracy: 82.6453 # terminal nodes: 94		Sample 3: Accuracy: 81.317 # terminal nodes: 102		Sample 4: Accuracy: 82.5387 # terminal nodes: 98	
#cls	Acc.	#cls	Acc.	#cls	Acc.	#cls	Acc.
4	**83.8388**	2	**83.6214**	6	**84.4717**	2	**83.615**
12	83.7791	10	83.4275	18	82.7565	14	82.8478
20	83.4253	22	83.1781	30	83.1087	26	82.8648
28	82.8158	34	82.9607	42	82.925	38	82.7476
36	82.6794	46	83.2378	54	82.4186	50	82.6389
44	81.959	58	82.6645	66	83.4303	62	82.9714
52	81.9526	70	83.3783	78	83.0475	74	82.7732
60	81.9974	82	82.7028	90	82.4962	86	82.5281
*68	81.9313	*94	82.7881	*102	82.7412	*98	82.5941
76	82.2894	106	82.7966	114	82.2511	110	82.6219
84	82.1849	118	82.0933	126	82.5727	122	82.7476
92	82.0123	130	82.2809	138	83.124	134	82.4577
100	81.5307	142	82.1892	150	83.0781	146	82.4663
104	81.1747	154	82.6943	162	82.7259	158	82.1849

Table 6. Results of adult data set with sample size of 1920 (cont.)

Sample 5: Accuracy: 82.2915 # terminal nodes: 68		Sample 6: Accuracy: 82.6453 # terminal nodes: 94		Sample 7: Accuracy: 81.317 # terminal nodes: 102	
#cls	Acc.	#cls	Acc.	#cls	Acc.
6	**83.875**				
14	83.3337	4	**83.777**	28	**82.7854**
22	82.9884	16	82.8542	44	82.6952
30	83.5042	28	82.2318	60	82.4247
38	82.9948	40	82.6751	76	82.8094
46	83.1419	52	81.633	92	82.4307
54	83.3998	64	82.7646	108	81.8898
62	82.9479	76	81.9761	124	82.016
70	82.786	88	81.9079	140	81.3007
*78	83.2783	*100	81.7203	*156	80.8019
86	83.4253	112	81.3666	114	81.1625
94	83.1078	124	82.0699	126	80.8499
102	82.445	136	81.9036	138	80.5854
108	83.0588	148	82.04	150	80.4532
116	82.9224	160	81.9548	162	80.8059

If we look at table 5 and table 6, we can notice that the best accuracies have been found in smaller number of clusters than the number of terminal nodes of C4.5. From the results of the both sample sizes for adult data set, we can find the fact that more accurate RBF networks could be found with relatively small number of clusters.

The following table 7 to table 10 show results of experiment for statlog data set. Table 7 and 8 show the result of experiment for statlog data set with sample size of 480.

Table 7. Results of statlog data set with sample size of 480

Sample 1: Accuracy: 78.623 # terminal nodes: 35		Sample 2: Accuracy: 80.1175 # terminal nodes: 31		Sample 3: Accuracy: 80.084 # terminal nodes: 36		Sample 4: Accuracy: 79.3115 # terminal nodes: 36	
#cls	Acc.	#cls	Acc.	#cls	Acc.	#cls	Acc.
				6	79.3451	6	78.22
6	79.513	6	78.6566	12	84.0134	12	82.0991
12	82.0823	12	83.7615	18	84.5844	18	83.4761
18	83.6272	18	84.534	24	**84.6348**	24	83.6776
24	84.6683	24	**84.9034**	30	83.9798	30	84.0134
*30	83.8455	*30	84.3493	*36	83.6608	*36	83.8959
36	84.2653	36	83.7783	42	82.2334	42	83.2914
42	83.9295	42	84.2989	48	83.0227	48	83.5264
48	**84.6851**	48	83.2317	54	84.0638	54	84.2149
54	84.534	54	80.5542	60	83.5097	60	**84.4668**
60	83.5936	60	80.7389	66	84.0974	66	83.7615
66	83.3249			72	83.9463	72	84.1814
72	83.2914			78	82.4685	78	83.1906
78	82.1998			84	81.7128	84	83.1234

Table 8. Results of statlog data set with sample size of 480 (cont.)

Sample 5: Accuracy: 77.5819 # terminal nodes: 34		Sample 6: Accuracy: 80.1679 # terminal nodes: 36		Sample 7: Accuracy: 78.4587 # terminal nodes: 42	
#cls	Acc.	#cls	Acc.	#cls	Acc.
		6	78.6566	12	82.3707
6	78.2872	12	82.8212	18	82.7233
12	82.1998	18	84.6348	24	83.2774
18	83.2242	24	85.6591	30	82.6058
24	**83.7615**	30	85.3568	36	83.546
*30	83.5936	*36	**86.2972**	*42	83.865
36	83.0563	42	85.7935	48	82.6897
42	82.6868	48	85.8774	54	83.546
48	83.0898	54	84.9874	60	83.9154
54	83.0898	60	84.8027	66	79.7179
60	82.9891	66	84.5844	72	**84.1336**
				78	83.5628
				84	82.9248
				90	81.4641

If we give attention to sample set 2, 4, and 6 of sample size 480, the accuracy of RBF network with six clusters is worse than the accuracy of decision tree of C4.5. But we can avoid choosing this RBF network because of the repeated trials with different number of clusters. Table 9 and 10 show the result of experiment for statlog data set with sample size of 960.

Table 9. Results of statlog data set with sample size of 960

Sample 1: Accuracy: 81.9543 # terminal nodes: 79		Sample 2: Accuracy: 81.1872 # terminal nodes: 63		Sample 3: Accuracy: 81.9316 # terminal nodes: 76		Sample 4: Accuracy: 80.8584 # terminal nodes: 69	
#cls	Acc.	#cls	Acc.	#cls	Acc.	#cls	Acc.
6	79.2511						
18	84.0731	6	79.9269	12	79.5251	6	79.0685
30	84.7854	12	82.9589	24	**85.4064**	18	85.6073
42	85.6986	24	**86.6849**	36	84.2374	30	85.589
54	**86.8128**	36	85.1142	48	84.1096	42	**86.3379**
66	86.3379	48	85.4429	60	84.347	54	85.2968
*78	86.3014	*60	85.5525	*72	84.1826	*66	84.9315
90	85.2237	72	85.7149	84	82.9589	78	84.6027
102	86.6119	84	86.0091	96	83.7626	90	83.5982
114	85.6986	96	86.3927	108	82.6119	102	82.9406
126	86.3014	108	83.3151	120	84.3105	114	79.8904
138	86.5205	120	84.4018	132	83.8539	126	81.4429

Table 10. Results of statlog data set with sample size of 960 (cont.)

Sample 5: Accuracy: 81.4098 # terminal nodes: 71		Sample 6: Accuracy: 80.2557 # terminal nodes: 69		Sample 7: Accuracy: 82.3014 # terminal nodes: 76	
#cls	Acc.	#cls	Acc.	#cls	Acc.
6	79.5654	6	79.0137	12	83.0502
18	84.2586	18	**86.1005**	24	**86.1005**
30	86.176	30	85.3151	36	85.6256
42	85.1717	42	85.1142	48	85.2603
54	85.7378	54	83.9087	60	85.1142
*66	84.7699	*66	85.7717	*72	84.9315
78	**86.6508**	78	85.8265	84	85.8265
90	86.3769	90	85.1142	96	84.895
102	85.9021	102	85.9726	108	84.6758
114	85.4456	114	85.7169	120	84.968
126	82.9255	126	85.5525	132	84.3836
138	83.9664				
150	80.558				
162	80.9898				

If we look at table 9 and table 10, we can notice that the best accuracies have been found in smaller number of clusters than the number of terminal nodes of C4.5. But there is no such regularity for sample size of 480. Anyway, we can find better RBF networks with the algorithm.

5 Conclusions and Future Works

Decision tree algorithms have good property that makes it easy to cope with large-sized data sets, but the good property of decision trees for large-sized data sets can also be harmful in data mining tasks, because we often may not have complete or perfect data

sets so that fragmenting the data sets could neglect minor classes, even the size of the data sets are large. Other good point of decision trees is understandability, because the structure of decision tree is represented in symbolic form.

RBF networks make approximation based on training data, and Gaussian functions are used mostly as the radial basis function. In order to train RBF networks, we may use some unsupervised learning algorithms like k-means clustering. Since RBF networks have different performance depending on the number of clusters and training data sets, we want to find better RBF networks based on some objective knowledge models like decision trees, where we can understand the structure of the decision trees easily. Most target data sets for data mining have some skewed distribution in class values, and this fact can be checked easily by inspecting the terminal nodes of decision trees. Moreover, we may also use the information of the number of terminal nodes in the trees as the initial number of clusters for k-means clustering of RBF networks.

The proposed procedure uses the class distribution information and the information of the number of terminal nodes in the generated tree for over-samplng and initialization for the number of clusters of RBF networks. Experiments with two real world data sets in business and scientific domain give us the possibility that we may find better RBF networks effectively.

Because oversampling can generate different class distribution in training data set, we can infer that trained RBF networks may have some different performance compared to original data set. Future work will be some detailed analysis for the effect of oversampling to utilize the RBF networks effectively.

References

1. Tan, P., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addison Wesley, Reading (2006)
2. Bishop, C.M.: Neural networks for pattern recognition. Oxford University Press, Oxford (1995)
3. Heaton, J.: Introduction to Neural Networks for C#, 2nd edn. Heaton Research Inc. (2008)
4. Lippmann, R.P.: An Introduction to Computing with Neural Nets. IEEE ASSP Magazine 3(4), 4–22 (1987)
5. Howlett, R.J., Jain, L.C.: Radial Basis Function Networks I: recent developments in theory and applications. Physica-Verlag, Heidelberg (2001)
6. Russel, S., Novig, P.: Artificial Intelligence: a Modern Approach, 2nd edn. Prentice Hall, Englewood Cliffs (2002)
7. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, Inc., San Francisco (1993)
8. Larose, D.T.: Data Mining Methods and Models. Wiley-Interscience, Hoboken (2006)
9. Shenouda, E.A.M.A.: A Quantitative Comparison of Different MLP Activation Functions in Classification. In: Wang, J., Yi, Z., Żurada, J.M., Lu, B.-L., Yin, H. (eds.) ISNN 2006. LNCS, vol. 3971, pp. 849–857. Springer, Heidelberg (2006)
10. Orr, M.J.L.: Introduction to Radial Basis Function Networks, `http://www.anc.ed.ac.uk/~mjo/intro.ps`
11. Kubat, M.: Decision Trees Can Initialize Radial-Basis Function Networks. IEEE Transactions on Neural Networks 9(5), 813–821 (1998)

12. Schwenker, F., Kestler, H.A., Palm, G.: Three learning phases for radial-basis-function networks. Neural Networks 14, 439–458 (2001)
13. Fukunaga, K., Hayes, R.R.: Effects of Sample Size in Classifier Design. IEEE Transactions on Pattern Analysis and Machine Intelligence 11(8), 873–885 (1989)
14. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16, 341–378 (2002)
15. Mazuro, M.A., Habas, P.A., Zurada, J.M., Lo, J.Y., Baker, J.A., Tourassi, G.D.: Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural Networks 21(2-3), 427–436 (2008)
16. Sug, H.: An Objective Method to Find Better RBF Networks in Classification. In: Proceedings of the 5th International Conference on Computer Sciences and Convergence Information Technology, vol. 1, pp. 373–376 (2010)
17. Suncion, A., Newman, D.J.: UCI Machine Learning Repository. University of California, School of Information and Computer Sciences, Irvine, CA (2007), http://www.ics.uci.edu/~mlearn/MLRepository.html
18. Kohavi, R.: Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp. 202–207 (1996)
19. Statlog (Landsat Satellite) Data Set, http://archive.ics.uci.edu/ml/datasets/Statlog+%28Landsat+Satellite%29
20. Witten, I.H., Frank, E.: Data Mining, 2nd edn. Morgan Kaufmann, San Francisco (2005)

To Propose Strategic Suggestions for Companies via IPC Classification and Association Analysis

Tzu-Fu Chiu[1], Chao-Fu Hong[2], and Yu-Ting Chiu[3]

[1] Department of Industrial Management and Enterprise Information, Aletheia University, Taiwan, R.O.C.
chiu@mail.au.edu.tw
[2] Department of Information Management, Aletheia University, Taiwan, R.O.C.
cfhong@mail.au.edu.tw
[3] Department of Information Management, National Central University, Taiwan, R.O.C.
gloria@mgt.ncu.edu.tw

Abstract. Strategic suggestions are essential for companies to facilitate the top management to foresee and review the future directions of their company's research and development investment. Therefore, a research design has been formed for performing the strategic planning on technology where IPC classification was employed to divide the patents into different categories and association analysis was adopted to discover the relations between terms and between clusters. Consequently, the visualized results, crystallized diagrams and integrated map, were generated and the relations between technical topics and companies were observed. Finally, according to the relations, the strategic suggestions on thin-film solar cell for companies were recognized and proposed.

Keywords: strategic suggestion, IPC classification, association analysis, thin-film solar cell, patent data.

1 Introduction

Solar cell, one of green energies, is growing at a fast pace with its long-lasting and non-polluting natures. To understand the situation and trend of this technology, especially the thin-film solar cell, is essential for companies to foresee and review the future directions of their research and development activities. In technological information, up to 80% of the disclosures in patents are never published in any other form [1]. Additionally, patent analysis has been recognized as an important task at the company, industry, and government levels [2]. Apart from those existing analysis methods such as task identification, searching segmentation, abstracting, clustering, and visualization [2], a research design of IPC classification and association analysis for conducting strategic planning will be built for patent analysis in order to propose the strategic suggestions for companies in the thin-film solar cell industry.

2 Related Work

As this study is attempted to propose strategic suggestions for the industry and companies via patent data, a research design is required and can be built via a consideration of IPC

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 218–227, 2011.

classification and association analysis. In order to manipulate the homogeneity and heterogeneity of patent data, IPC classification is employed to divide the patents into different categories. For handling the textual nature of patent data (mainly the abstract, claim, and description fields), association analysis (including data crystallization) is adopted to discover the relations between terms and between clusters. Subsequently, the research design will be applied to the domain of strategic planning on thin-film solar cell. Therefore, the related areas of this study would be strategic planning, thin-film solar cell, IPC classification, and association analysis, which will be described briefly in the following subsections.

2.1 Strategic Planning

Strategic planning (also called strategic management) is an organization's process of defining its strategy, or direction, and making decisions on allocating its resources to pursue this strategy, including its capital and people [3]. Strategic planning processes are: mission definition, objectives setting, external analysis, internal analysis, strategic choice, strategy implementation, and competitive advantages [4]. Among them, the strategic choice (i.e., strategy formulation), following the decision-making process, is to develop and evaluate strategic alternatives and then select strategies that support and complement each other and that allow the organization to best capitalize on its strengths and environmental opportunities [5]. In order to draw the data mining techniques for conducting the strategy formulation via patent data, a research design of IPC classification and association analysis will be formed to propose the strategic suggestions for companies on thin-film solar cell in this study.

2.2 Thin-Film Solar Cell

Solar cell, a sort of green energy, is clean, renewable, and good for protecting our environment. It can be mainly divided into two categories (according to the light absorbing material): crystalline silicon (in a wafer form) and thin films (of other materials) [6]. A thin-film solar cell (TFSC), also called a thin-film photovoltaic cell (TFPV), is made by depositing one or more thin layers (i.e., thin film) of photovoltaic material on a substrate [7]. The most common materials of TFSC are amorphous silicon or polycrystalline materials (such as: CdTe, CIS, and CIGS) [6]. In recent years (2003-2007), total PV production grew in average by almost 50% worldwide, whereas the thin film segment grew in average by over 80% and reached 400 MW or 10% of total PV production in 2007 [8]. Therefore, thin film is the most potential segment with the highest production growth rate in the solar cell industry, and it would be appropriate for academic and practical researchers to contribute efforts to explore and propose strategic suggestions for this technology.

2.3 IPC Classification

Classification, or categorization, is to classify a given data instance into a prespecified set of categories [9]. The classification task can be defined as to approximate an unknown category assignment function $F : D \times C \rightarrow \{0,1\}$, where D is the set of all possible

documents and *C* is the set of predefined categories [9]. IPC (International Patent Classification) provides a hierarchical system of symbols for the classification of patents according to the different areas of technology to which they pertain [10]. IPC classifies technological fields into five hierarchical levels: section, class, subclass, main group and sub-group, containing 70,000 categories [11]. In this study, IPC code will be used to divide the patents into different categories.

2.4 Association Analysis

Association analysis is a useful method for discovering interesting relationships hidden in large data sets. The uncovered relationships can be represented in the form of association rules or co-occurrence graphs [12]. An event map, a sort of co-occurrence graphs, is a two-dimension undirected graph, which consists of event clusters, visible events, and chances [13]. An event cluster is a group of frequent and strongly related events. The occurrence frequency of events and co-occurrence between events within an event cluster are both high. The co-occurrence between two events is measured by the Jaccard coefficient as in Equation (1), where e_i is the *i*th event in a data record (of the data set *D*). The event map is also called as an association diagram in this study.

$$Ja(e_i, e_j) = \frac{Freq(e_i \cap e_j)}{Freq(e_i \cup e_j)} \tag{1}$$

Secondly, data crystallization is employed to add extra data elements into the association diagram for observing the relations between clusters. Data crystallization is a technique to detect the unobservable (but significant) events via inserting these unobservable events as dummy items into the given data set [14]. The unobservable events and their relations with other events are visualized by applying the event map. A generic data crystallization algorithm can be summarized as follows [15]: (a) event identification, (b) clustering, (c) dummy event insertion, (d) co-occurrence calculation, and (e) topology analysis. The co-occurrence between a dummy event and clusters is measured by equation (2), where DE_i is a dummy event and *C* is the specific number of clusters.

$$Co(DE_i, C) = \sum_{j=0}^{|C|-1} \max_{e_k \in c_j} Ja(DE_i, e_k) \tag{2}$$

Data crystallization was originally proposed to deal with unobservable events (i.e., dummy events) so as to emerge the hidden clues from existing circumstances via judging the unknown relations [14]. This method has been modified by the authors to insert extra data elements (e.g., patent-no, assignee, or country fields) as dummy events into the original data set (i.e., the abstract field), so that the relations between the extra data elements and existing clusters can come out and be observed [16].

In this study, the association analysis and modified data crystallization will be adopted to generate association diagrams and crystallized diagrams for strategic planning.

3 A Research Design for Strategic Planning

As this study is attempted to propose the strategic suggestions for thin-film solar cell, a research design, based on the IPC classification and association analysis, has been developed and shown in Fig. 1. It consists of four phases: data preprocessing, patent classification, association analysis, and new findings; and will be described in the following subsections.

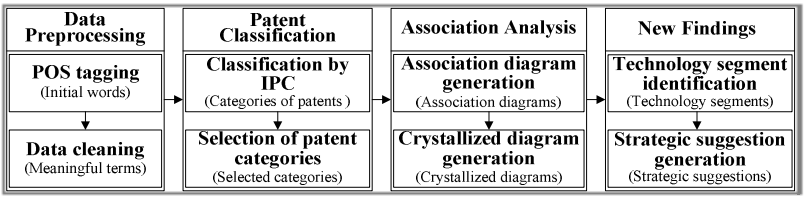


Fig. 1. A research design for strategic planning

3.1 Data Preprocessing

In first phase, the patent data of thin-film solar cell (during a certain period of time) will be downloaded from the USPTO [17]. For considering an essential part to represent a patent document, the abstract, assignee, and country fields are selected as the objects for this study. Afterward, two processes, POS tagging and data cleaning, will be executed to clean up the textual data of the abstract field.

(1) POS Tagging: An English POS tagger (i.e., a Part-Of-Speech tagger for English) from the Stanford Natural Language Processing Group [18] will be employed to perform word segmenting and labeling on the patents (i.e., the abstract field). Then, a list of proper morphological features of words needs to be decided for sifting out the initial words.

(2) Data Cleaning: Upon these initial words, files of n-grams, synonyms, and stop words will be built so as to combine relevant words into compound terms, to aggregate synonymous words, and to eliminate less meaningful words. Consequently, the meaningful terms will be obtained from this process.

3.2 Patent Classification

Second phase is used to conduct the patent classification according to the IPC field. The patents of thin-film solar cell will be classified by IPC code and selected via the number of appearing times of every specific IPC code.

(1) Classification by IPC: Every patent will be assigned to a specific category according to its IPC code and may be assigned to several categories if its IPC field contains more than one code. Meanwhile, the number of patents for an IPC category will be counted for showing its occurrence frequency.

(2) Selection of Patent Categories: Based on a threshold setting (e.g., 15 was set for this study) on the number of patents in an IPC category, a certain number of the

leading IPC categories will be selected and then be utilized to generate the association diagrams and crystallized diagrams.

3.3 Association Analysis

Third phase is designed to perform the association analysis on the meaningful terms of the abstract data for every IPC category using association diagram generation and crystallized diagram generation so as to obtain the technical topics and relations between topics and companies.

(1) Association Diagram Generation: An association diagram will be drawn via the term frequency and co-occurrence from the meaningful terms (of the abstract data) of every IPC category, so that a number of clusters will be generated through the proper thresholds setting of frequency and co-occurrence. These clusters are regarded as technical topics and will be named using the domain knowledge.

(2) Crystallized Diagram Generation: In order to generate a crystallized diagram, a dummy event (i.e., assignee) needs to be inserted into the abstract data. Afterward, using the updated abstract data, the diagram will be drawn to show the clusters, dummy nodes, and links. Secondly, the clusters in crystallized diagrams will also be named and regarded as in the association diagrams. Thirdly, the topics, dummy nodes, and links will be utilized to observe the relations between topics and companies. Finally, these crystallized diagrams will be utilized to form an integrated map in the next phase.

3.4 New Findings

Last phase is intended to identify the technology segments and to recognize the strategic suggestions, based on the crystallized diagrams of IPC categories.

(1) Technology Segment Identification: By combining the crystallized diagrams, an integrated map of technical topics for IPC categories will be constructed via linking the technical topics to the same clusters in the different crystallized diagrams. Subsequently, the technology segments will be identified based on the integrated map.

(2) Strategic Suggestion Generation: According to the above integrated map and technology segments, the relations between technical topics and companies can be observed and the strategic suggestions can be recognized. The strategic suggestions of thin-film solar would be useful for the top management to foresee and review the future directions of their company's research and development activities.

4 Experimental Results and Explanation

The experiment has been implemented according to the research design. The experimental results will be explained in the following four subsections: result of data preprocessing, result of patent classification, result of association analysis, and result of new findings.

4.1 Result of Data Preprocessing

As the aim of this study is to propose the strategic suggestions, the patents of thin-film solar cell are the target data for the experiment. Mainly, the abstract, assignee, and country fields were used in this study. The issued patents (160 records) during year 2000 to 2009 were collected from USPTO, using key words: "'thin film' and ('solar cell' or 'solar cells' or 'photovoltaic cell' or 'photovoltaic cells' or 'PV cell' or 'PV cells')" on "title field or abstract field". Afterward, the POS tagger was triggered and the data cleaning process was executed to do the data preprocessing upon abstract data. Consequently, the abstract data during year 2000 to 2009 were cleaned up and the meaningful terms were obtained.

4.2 Result of Patent Classification

According to the IPC field, the number of IPC categories (down to the fifth level) in 160 patents were 190, as many patents contained more than one IPC code, for example, Patent 06420643 even contained 14 codes. But there were up to 115 categories individually consisting of only one patent. The leading frequent 50 categories were illustrated in Fig. 2, where the first category H01L031-18 consisted of 48 patents and the 49th category C03B033-07 consisted of 2 patents. By setting the threshold of the number of consisting patents to 15, the leading six IPC categories: H01L031-18 (consisting 48 patents), H01L021-02 (27 patents), H01L031-06 (22 patents), H01L031-036 (19 patents), H01L031-00 (18 patents), and H01L021-00 (15 patents), were selected and would be used to generate the association diagrams, crystallized diagrams, as well as the integrated map. Consequently, the strategic suggestions would be drawn up based on these visualized results.

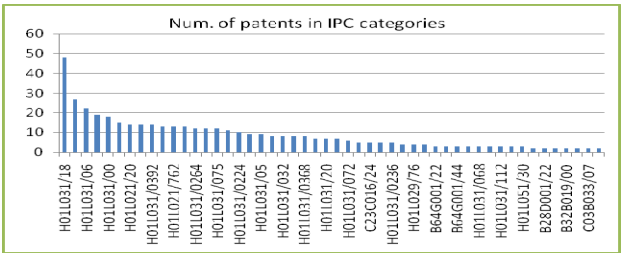


Fig. 2. The number of patents in IPC categories

4.3 Result of Association Analysis

Using the meaningful terms of abstract data (H01L031-18 to H01L021-00), six association diagrams were drawn via 'association diagram generation' successively. Taking the diagram of H01L031-18 as an example, twelve clusters were found while the number of consisting nodes of a cluster was set to no less than four. According to the domain knowledge, the clusters were named as follows: a1-porous-structure, a2-plasma-CVD, a3-accumulated-charge, a4-amorphous-annealing-deposition, a5-encapsulant, a6-absorber-layer, a7-reflective-film, a8-silicon-film, a9-composite-structure, a10-annealing-process, a11-semiconductor-layer, and a12-compound-semiconductor. These

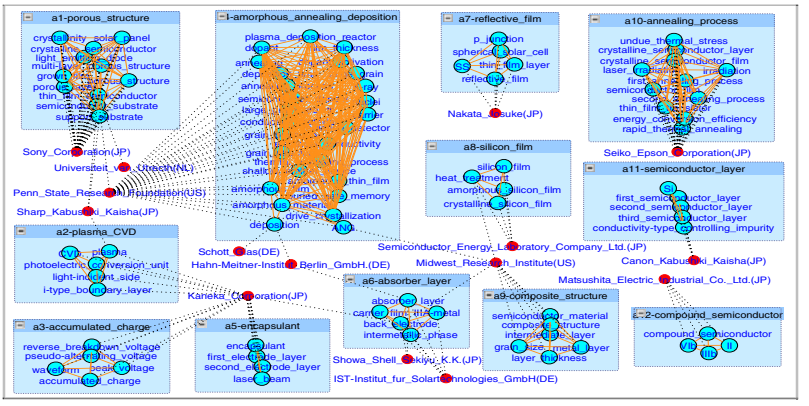


Fig. 3. A crystallized diagram of IPC category "H01L031-18"

named clusters were regarded as technical topics. Afterward, using the same abstract data with inserted dummy event (i.e., company-name), six crystallized diagrams were drawn via 'crystallized diagram generation', showing the relations between topics and companies. Fig. 3 is an example of crystallized diagram of IPC category "H01L031-18".

By combining six 'crystallized diagram of IPC category' (H01L031-18 to H01L021-00), an integrated map of technical topics for IPC categories (in Fig. 4) was constructed via linking every technical topic to the same clusters in the six categories.

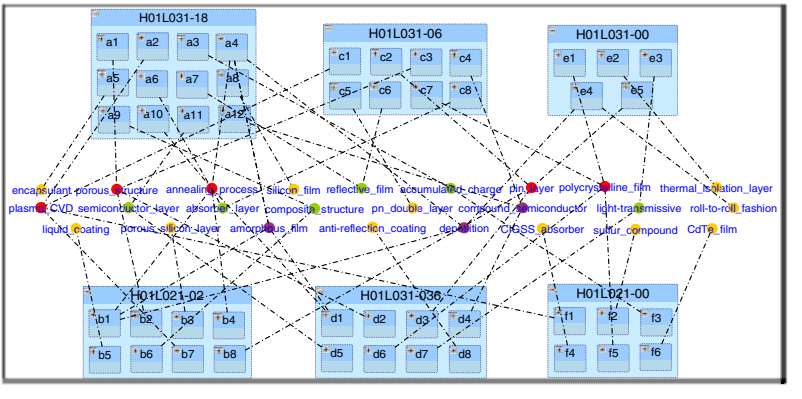


Fig. 4. An integrated map of technical topics for IPC categories (H01L031-18 to H01L021-00)

4.4 Result of New Findings

The integrated map of technical topics for IPC categories will be utilized to identify the technology segments. Afterward, the integrated map and technology segments will be applied to recognize the possible strategic suggestions.

(1) Technology Segment Identification: According to the above integrated map, the relations among technology segments, technical topics, and related companies were

summarized below in Table 1. Depending on the number of clusters which a technical topic was comprised, the technology segments were identified as: significant segment (linking to equal or greater than 4 clusters; the nodes were in red color); regular segment (equal to 3 clusters; in purple color); minor segment (equal to 2 clusters; in green color); and niche segment (equal to 1 cluster; in yellow color). This integrated map and the above crystallized diagrams were then used to recognize the possible strategic suggestions.

Table 1. Technology segments with their technical topics and related companies

Technology segments	Technical topics	H01L0 31-18	H01L0 21-02	H01L0 31-06	H01L0 31-036	H01L0 31-00	H01L0 21-00	Related companies
Significant segment	porous-structure	a1	b2	c3	d2			Sony(JP), U_Utrecth(NL), Penn_State(US), Sharp(JP), National_Institute(JP), Canon(JP)
	annealing-process	a4, a10	b1, b4		d1			Schott(DE), Hahn-Meitner(DE), Seiko_Epson(JP), Penn_State(US), Matsushita(JP)
	plasma-CVD	a2	b7	c1			f1	Sharp(JP), Kaneka(JP), Matsushita(JP)
	pin-layer		b8	c2	d3	e5		Fuji_Electric(JP), Sharp(JP), Kaneka(JP), Angewandte(DE) , U_Utrecth(NL)
	polycrystalline-film			c7	d6	e1	f5	AstroPower(US), Sharp(JP)
Regular segment	compound-semiconductor	a12		c4	d4			Matsushita(JP), Sony(JP)
	amorphous-film	a4	b1		d1			Penn_State(US), Midwest(US), Schott(DE)
	deposition	a4	b1			e4		Schott(DE), Hahn-Meitner(DE), Luch(US), Penn_State(US)
Minor segment	accumulated-charge	a3					f3	Kaneka(JP)
	reflective-film	a7		c6				Nakata(JP)
	semiconductor-layer	a11			d5			Canon(JP)
	absorber-layer	a6		c8				Kaneka(JP), Hahn-Meitner(DE), Midwest(US), IST-Institut(DE), Showa(JP)
	composite-structure	a9	b6					Midwest(US)
	light-transmissive				d7	e3		Sharp(JP)
Niche segment	encapsulant	a5						Kaneka(JP)
	silicon-film	a8						Semiconductor_Energy(JP), Midwest(US)
	liquid-coating		b5					McCandless(US)
	porous-silicon-layer		b3					Canon(JP)
	pn-double-layer			c5				ANTEC_Solar(DE)
	anti-reflection-coating				d8			Pacific_Solar(AU)
	roll-to-roll-fashion					e4		Luch(US)
	thermal-isolation-layer					e2		Industrial_Technology(TW)
	CIGSS-absorber						f4	U_Central_Florida(US)
	sulfur-compound						f2	Honda_Giken(JP)
	CdTe-film						f6	Solar_Systems(IT)

(2) Strategic Suggestions: According to the above table of technology segments, technical topics, and related companies, the relations among three factors could be observed from different viewpoints, including of an industry, technical topics, and companies. The strategic suggestions would be recognized depending on the domain knowledge and described as follows.

Viewpoint of an industry: Referring to Table 1, the most common areas of thin-film solar cell during 2000 till 2009 were H01L031-18, H01L021-02, H01L031-06, H01L031-036, H01L031-00, and H01L021-00, where H01L031 and H01L021 were also indicated as the top two out of the 10 popular areas in the classification search of European Patent Office [19]. It was suggested that most companies should put their efforts in these areas. On the other hand, the less important areas of thin-film solar cell spread widely to 115 categories (each category containing only one patent). It meant that the new directions of this industry emerged rapidly and variedly.

Viewpoint of technical topics: From the above Table 1, the technical topics in the significant segment (i.e., porous-structure, annealing-process, plasma-CVD, pin-layer, and polycrystalline-film) were the most popular and valuable technical items. It was suggested that the government units should pay more attention to these items. The technical topics in the niche segment (i.e., encapsulant, silicon-film, liquid-coating, porous-silicon-layer, pn-double-layer, anti-reflection-coating, roll-to-roll-fashion, thermal-isolation-layer, CIGSS-absorber, sulfur-compound, and CdTe-film) were the emerging technical items. It appeared that some of these items could be potentially new techniques.

Viewpoint of companies: According to Table 1, the companies in the significant segment (e.g., Sony (JP), U-Utrecth (NL), Penn-State (US), Sharp (JP), National-Institute (JP), and Canon (JP) in the "porous-structure" topic) were the more competitive ones. It was suggested that the companies in the same technical topic could cooperate together to form an alliance to increase their strength; but for the strongest company, it might be appropriate to compete with all others to get the leader position. In contrast, the companies in the niche segment were the ones with innovative ideas. It was suggested that these companies should re-examine their R&D plans to decide whether to put more resources into this potential technical topic or to withdraw resources from this unavailable technical topic.

In addition, the focused companies were the ones with higher frequency (equal or greater than 3) in Table 1: Kaneka (JP), Sharp (JP), Midwest (US), Canon (JP), Matsushita (JP), Penn-State (US), and Schott (DE). It would be suitable for these companies to compete for the leader position.

5 Conclusions

The research design of IPC classification and association analysis for strategic planning has been formed and applied to propose the strategic suggestions for thin-film solar cell using patent data. The experiment was performed and the experimental results were obtained. The visualized results: association diagrams, crystallized diagrams, and integrated map, were generated. The leading six IPC categories were: H01L031-18, H01L021-02, H01L031-06, H01L031-036, H01L031-00, and H01L021-00. The technical topics, related companies, and technology segments were identified. The technical topics in the significant segment were: porous-structure, annealing-process, plasma-CVD, pin-layer, and polycrystalline-film. The focused companies were: Kaneka (JP), Sharp (JP), Midwest (US), Canon (JP), Matsushita (JP), Penn-State (US), and Schott (DE). Finally, the strategic suggestions on thin-film solar cell were also recognized and proposed.

In the future work, the other aspects of company information (e.g., the public announcement, open product information, and financial reports) can be included so as to

enhance the validity of research result. Additionally, the patent database can be expanded from USPTO to WIPO or TIPO in order to perform the strategic planning on thin-film solar cell widely.

Acknowledgments. This research was supported by the National Science Council of the Republic of China under the Grants NSC 99-2410-H-156-014.

References

1. Blackman, M.: Provision of Patent Information: A National Patent Office Perspective. World Patent Information 17(2), 115–123 (1995)
2. Tseng, Y., Lin, C., Lin, Y.: Text Mining Techniques for Patent Analysis. Information Processing and Management 43, 1216–1247 (2007)
3. Wikipedia, Strategic planning (October 15, 2010), http://en.wikipedia.org/wiki/Strategic_planning
4. Barney, J.B., Hesterly, W.S.: Strategic Management and Competitive Advantage: Concepts and Cases. Prentice Hall, Englewood Cliffs (2010)
5. Robbins, S., Coulter, M.: Management, 10th edn. Prentice-Hall, Englewood Cliffs (2008)
6. Solarbuzz, Solar Cell Technologies (October 20, 2010), http://www.solarbuzz.com/technologies.htm
7. Wikipedia, Thin film solar cell (October 20, 2010), http://en.wikipedia.org/wiki/Thin_film_solar_cell
8. Jager-Waldau, A.: PV Status Report 2008: Research, Solar Cell Production and Market Implementation of Photovoltaics, JRC Technical Notes (2008)
9. Feldman, R., Sanger, J.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, Cambridge (2007)
10. WIPO, Preface to the International Patent Classification (IPC)(October 30, 2010), http://www.wipo.int/classifications/ipc/en/general/preface.html
11. Sakata, J., Suzuki, K., Hosoya, J.: The analysis of research and development efficiency in Japanese companies in the field of fuel cells using patent data. R&D Management 39(3), 291–304 (2009)
12. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addison Wesley, Reading (2006)
13. Ohsawa, Y., Benson, N.E., Yachida, M.: KeyGraph: Automatic Indexing by Co-Occurrence Graph Based on Building Construction Metaphor. In: Proceedings of the Advanced Digital Library Conference (IEEE ADL 1998), pp. 12–18 (1998)
14. Maeno, Y., Ohsawa, Y.: Human-Computer Interactive Annealing for Discovering Invisible Dark Events. IEEE Transactions on Industrial Electronics 54(2), 1184–1192 (2007)
15. Maeno, Y., Ohsawa, Y.: Stable Deterministic Crystallization for Discovering Hidden Hubs. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, vol. 2, pp. 1393–1398 (2006)
16. Chiu, T.F.: Applying KeyGraph and Data Crystallization to Technology Monitoring on Solar Cell. Journal of Intelligent & Fuzzy Systems 21(3), 209–219 (2010)
17. USPTO. the United States Patent and Trademark Office (October 30, 2010), http://www.uspto.gov/
18. Stanford Natural Language Processing Group, Stanford Log-linear Part-Of-Speech Tagger (October 15, 2010), http://nlp.stanford.edu/software/tagger.shtml
19. European Patent Office, Search the European classification (October 30, 2010), http://v3.espacenet.com/eclasrch?classification=ecla&locale=en_EP

A New Vertex Similarity Metric for Community Discovery: A Distance Neighbor Model

Yueping Li

Shenzhen Graduate School, Harbin Institute of Technology, Xili
Shenzhen, 518055, China
leeyueping@gmail.com

Abstract. The hierarchical clustering methods based on vertex similarity can be employed for community discovery. Vertex similarity metric is the most important part of these methods. However, the existing metrics do not perform well compared with the state-of-the-art algorithms. In this paper, we propose a new vertex similarity metric based on distance neighbor model, called Distance Neighbor Ratio Metric (DNRM), for community discovery. DNRM considers both distance and nearby edge density which are essential measures in community structure. Compared with the existing metrics of vertex similarity, DNRM outperforms substantially in community discovery quality and the computing time. The experiments are designed rigorously involving both well-known social networks in real world and computer generated networks.

Keywords: Hierarchy clustering, vertex similarity, community discovery, modularity, time complexity.

1 Introduction

1.1 Background

Graph mining attracts much attention in both academy and industry. Many systems of current interest can be represented as graphs. Each of these graphs consists of vertices and their connecting edges, where the vertices indicate the individuals and the edges represent the relations. Recent studies [1] reveal that many graphs in society often exhibit hierarchical *community structure*. In addition, the communities correspond to known sets of units dealing with related topics, such as citation networks [2], food webs [3] , and biochemical networks [4,5]. Thus, community discover plays an essential role for the identification and characterization of real networks [6]. Furthermore, uncover hierarchical community structure emerges as a premise task for capturing an in-depth understanding of networks.

In the literature, community discovery algorithms have been well studied. Generally, they can be divided into three categories: graph partitioning, hierarchical clustering, and methods for hyperlink-based network. Graph partition algorithms include the Kernighan-Lin algorithm [7], spectral partitioning

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 228–237, 2011.

[8,9]. Hierarchical algorithms contain two classes: agglomerative methods based on the optimization of modularity or the similarity metrics, and the divisive methods based on betweenness metrics such as Girvan-Newman (GN) algorithm [10], Tyler algorithm [11], and Radicchi algorithm [12]. The methods for detecting hyperlink-based Web communities such as the maximum flow communities (MFC) algorithm [13], the hyperlink- induced topic search (HITS) algorithm [14], the spreading activation energy (SAE) algorithm [15].

There are also many other kinds of methods based on different technologies such as spectral property of graph matrix[16] [17] [18], spin-spin interactions [19], random walks [20] and synchronization [21]. For more details, the reader can refer to the survey article by Fortunato [22].

1.2 Related Metrics of Vertex Similarity

In essence, similarity metric is the most important part of agglomerative methods based on vertex similarity. The idea of these methods is to compute the similarity between each pair of vertices, firstly, no matter whether they are connected by an edge or not. Then, merge the vertex or the (temporary) community into the vertex or community most similar to it.

However, it appears that these methods perform well for specific types of problems , but work poorly in more general cases [23]. The reason is that existing vertex similarity metrics are designed for particular kinds of graphs. Thus, the algorithms based on these metrics cannot tackle a variety of graphs.

Next, we present several classical metrics of vertex similarity for community discovery. Then, we will show their limitations and shortcomings.

One well-known similarity measure is Jaccard Index [24], which is defined as follows. For a vertex $u \in V(G)$, let $\Gamma(u)$ be the the set of neighbor vertices of u. It is natural that two vertices u and v are more likely if they share more common neighbor vertices. In addition, if the shared neighbors take up more proportion of all their neighbors, it also shows more similarity between these two vertices. The formula is defined as follows:

$$s_{JI}(i,j) = \frac{|\Gamma(i) \bigcap \Gamma(j)|}{|\Gamma(i) \bigcup \Gamma(j)|} \tag{1}$$

where $\Gamma(i)$ is the neighbor set of vertex i.

Another metric is the number of independent paths between two vertices, denoted by s_{NIP}. Independent paths do not share any edge (vertex). This metric indicates the maximum flow that can be conveyed between the two vertices. It can be computed under the constraint that each edge can carry only one unit of flow, and the flow value is an integer.

An alternate metric evaluates the number of the paths between two vertices. In this case, the problem is that the total number of paths is huge. But one method solving this problem is to compute a weighted sum of the number of paths, where paths of length l are weighted by the factor α^l, with $\alpha < 1$. It is clear that the paths with longer length can be neglected due to its tiny weight. Denote this weighted sum metric by s_{WS}.

There are several metrics defined by the distance in an n-dimensional Euclidean space, by assigning a position to each vertex of the given graph. Thus, the positions of the vertices have to be determined before computing these metrics. One approach to determine position of each vertex is the spectral bisection method which is based on the properties of the Laplacian matrix [25]. Due to its high computation cost, these metrics are seldom used in community discovery. Thus, they are not considered in our paper.

In brief, the metrics above consider either edge density in neighbors or the number of connecting paths. Since the similarity of one vertex pair is affected by the following features: distance, local edge density and number of disjoint connecting paths in global, these metrics are not well defined shown in Table 1. In addition, the existing metrics are not good for community discovery, for instance, the time complexity and the quality is not satisfied.

Table 1. Features of existing metrics

Metric	distance	edge density	disjoint paths	quality	time complexity
s_{JI}	×	√	×	mediate	$O(\|E(G)\|)$
s_{NIP}	×	partial	√	bad	$O(\|V(G)\|\|E(G)\|)$
s_{WS}	√	partial	partial	mediate	$O(2^{\|V(G)\|})$

In this paper, we propose a new vertex similarity metric for community discovery. The metric is based on distance neighbor model which enable it to evaluate both topological distance and local edge density. Thus, it can describe the similarity between two vertices better.

The rest of this paper is organized as follows. Section 2 formulas our problem, and propose the measure of quality. Section 3 introduces our metric. Section 4 gives our algorithm. Experimental results are presented in Section 5. Finally, in Section 6, we summarize this work and point out the future work.

2 Problem Statement

Our problem is to divide the considered graph into communities in certain application scenario. Unfortunately, community structure has no universal accepted definition [1]. One common used one is that the division of vertices into groups such that there is a higher edge density within groups while less edge density between them.

This paper considers simple graphs only, i.e., the graphs without loop or multi-edges. Given graph G, $V(G)$ and $E(G)$ denote the sets of its vertices and edges respectively. In addition, our paper considers unweighted graphs, that is, all edges are unweighted.

A community structure is a partition $\mathcal{P} = C_1, C_2, \ldots, C_k$ of graph G such that $C_1 \bigcup C_2 \bigcup \ldots \bigcup C_k = V(G)$ and $C_i \bigcap C_j = \emptyset$ for $i \neq j$.

It appears that the number of the partitions of one graph is huge. One measure is necessary for evaluating the quality of one partition with respect to the

community in scenario. One common used quantitative measure is ***modularity*** [26]. Definition given below states that communities in a good partition has high intra-community edge density and less inter-community edge density:

$$Q(\mathcal{P}) = \frac{1}{2m}\Sigma_{ij}(A_{ij} - \frac{k_i k_j}{2m})\delta(C_i, C_j) \tag{2}$$

where A_{ij} is the adjacency matrix, m is the total number of the edges, and k_i is the degree of vertex i. The function δ yields one if vertices i and j are in the same community ($C_i = C_j$), zero otherwise.

We are supposed to find an optimal partition $\mathcal{P}$ which makes the modularity $Q(\mathcal{P})$ maximum. When one partition has modularity larger than 0.3∼0.4, it can be concluded that this partition has community structure. The larger the modularity is, its community structure is more prominent and clear. However, it is well-known that our problem is an NP-hard problem [22]. Thus, there is not polynomial time algorithm for this problem unless P=NP. Most existing methods are approximation algorithms. Furthermore, modularity has limits [22]. Thus, the result community structure is favorable only if it corresponds to the actual structure in real world. If there is no in-advance structure information, the result will recommend a probable community structure but not an optimal one.

3 A New Vertex Similarity Metric Based on Distance Neighbor Model

3.1 Definition and Properties

Our model is designed based on the fact that the probability that two vertices have the same neighbors when they are in the same community is larger than the case they lie in different ones. This motivation arises from the neighbor ratio metric s_{JI}. But we extend this model by allowing not only adjacent neighbors but also neighbors within certain distance, which is a threshold denoted by d. We next present the definition of our distance neighbor model.

Let $G = (V, E)$ be a simple graph. Assume that two vertices i and j are supposed to compute similarity. Let $dis(x, y)$ be the distance between vertices x and y where $x, y \in V(G)$.

The idea of the neighbor ratio metric s_{JI} is that if the proportion of the common neighbor over all neighbors is larger, then these two vertices i and j is more similar. Since we consider the non-adjacent neighbors within the distance d, it is natural that these neighbors' contribution to the similarity is less than the adjacent neighbors. Thus, it can be concluded that the contribution decreases when the distance from the neighbor to i or j increases. Therefore, the metric s_{DNM} of our distance neighbor model is defined as follows:

$$\Gamma(i, d) = \{u | dis(i, u) = d, \forall u \in V(G)\} \tag{3}$$

$$s_{DNM}(i, j) = \sum_{k=1}^{d} \frac{|\Gamma(i,k) \cap \bigcup_{c=1}^{k} \Gamma(j,c)| + |\Gamma(j,k) \cap \bigcup_{c=1}^{k} \Gamma(i,c)|}{k \times \sum_{c=1}^{k} |\Gamma(i,c) \cup \Gamma(j,c)|} \tag{4}$$

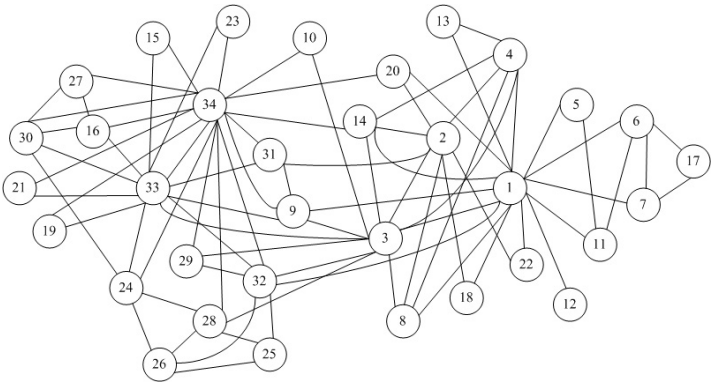


Fig. 1. Karate network

For better understanding of our metric, an illustration example using the well-known dataset karate club network is given. The topology structure is shown in Fig. 1. The statistic data of vertex pair is presented in Table 2.

Table 2. The statistic data of vertex pair $(6, 3)$

d	1	2
$\Gamma(6,d)$	1,7,11,17	3,4,5,8,9,11,12,13,14,18,20,22
$\Gamma(3,d)$	1,2,4,8,9,10,14,28,29,32	5,6,7,11,12,13,20,22,24,25,26,31,33,34
$s_{DNM}(6,3)$	$\frac{1}{13}$	$\frac{1}{13}+\frac{1}{2}\times\frac{11}{28}$
d	3	4
$\Gamma(6,d)$	2,10,28,29,31,32,33,34	15,16,19,21,23,24,25,26,27,30
$\Gamma(3,d)$	15,16,17,19,21,27,30	0
$s_{DNM}(6,3)$	$\frac{1}{13}+\frac{1}{2}\times\frac{11}{28}+\frac{1}{3}\times\frac{9}{34}$	$\frac{1}{13}+\frac{1}{2}\times\frac{11}{28}+\frac{1}{3}\times\frac{9}{34}+\frac{1}{4}\times\frac{10}{34}$

According to the definition, it can be concluded that our metric is symmetric, that is, $s_{DNM}(i,j) = s_{DNM}(j,i)$ for $\forall i,j \in V(G)$.

Next, we propose the relationship between local edge density and $s_{DNM}(i,j)$. Set $d_e(i) = \min\{dis(a,i), dis(b,i)\}$ where the edge $e = (a,b)$ and $a,b,i \in V(G)$. Let $P_{(i,j)}(k)$ be the set of the paths between vertices i and j in which the length of each path is no larger than k. Then we have the following proposition.

Proposition 1. *Let d be the chosen distance threshold, and $E(P_{(i,j)}(k))$ be the edge set of the path set $P_{(i,j)}(k)$ where $i,j \in V(G)$ and k is a positive integer. Set $E^d_{(i,j)}$ to be the edge set satisfying: (I) the distance from i and from j is no larger than d ; (II) lies in $E(P_{(i,j)}(2d))$. Then, we have*

$$\sum_{k=1}^{d} \frac{E^k_{(i,j)}}{k \times \sum_{c=1}^{k} |\Gamma(i,c) \cup \Gamma(j,c)|} \geq s_{DNM}(i,j) \tag{5}$$

3.2 Comparison with Existing Vertex Similarity Metrics

In this subsection, we compare our metric with the existing ones on several pairs of the graph modelled by Zachary's karate club illustrated in Fig. 1. Choose the distance threshold d to be 2. We select a couple of pairs to vertices, and the values are given in Table 3.

Table 3. Metrics comparison in karate club graph

Pairs	(1,12)	(1,6)	(1,2)	(6,7)	(6,2)	(6,30)	(17,30)	(2,12)
s_{JI}	0	2/17	7/16	1/4	0	0	0	0
s_{NIP}	1	4	9	4	4	4	2	1
s_{WS}	1.0	3.2	109.2	3.1	108.3	164.7	56.6	46.3
s_{DNM}	0.581	0.487	0.405	0.610	0.469	0.323	0.164	0.426

Discussion: The metric of neighbor ratio s_{JI} cannot distinguish the similarity of pairs (6,2) and (6,30), since they have no common neighbors. In addition, it cannot describe the topology distance of a pair of vertices. The metric s_{NIP} cannot evaluate the topology distance either, concluded by the values of s_{NIP}(6,2) and s_{NIP}(6,30). The metric s_{WS} outputs wrong evaluations of pairs (1,12) and (2,12). The reason is that there are more paths from vertex 2 to vertex 12 than that from vertex 1 to vertex 12. The values indicate that our distance neighbor model is better in describing the similarity compared with the others.

Finally, we summary the features of our metric in Table 4.

Table 4. Features of our metric s_{DNR}

Metric	distance	edge density	disjoint paths	quality	time complexity
s_{DNR}	$\surd$	$\surd$	$\times$	good	$O(\|E(G)\|)$

4 Algorithm Description

Community Discovery Algorithm Based on Distance Neighbor Model

Input: a simple, undirected and unweighed graph G
Output: a community structure
1. Choose the diameter d.
2. Foreach vertex pair (i, j) where $i \neq j$ and $dis(i, j) \leq 2 \times d$ Do
3. Begin For $k := 1$ to d Do Search $\Gamma(i, k)$ and $\Gamma(j, k)$; //endfor
4. For $k := 1$ to d Do Compute $\Gamma(i, k) \cap \Gamma(j, k)$ and $\Gamma(i, k) \cup \Gamma_j, k)$; //endfor
5. Compute $s_{DNM}(i, j)$;
6. End //foreach
7. Use the classical average linkage method to find the community structure and output it.

Complexity analysis: Steps 3-6 employ a procedure which is a part of breadth-first-search. Thus, the running time is bounded by $|E(G)|$. The step 2 is a loop which repeats for each pair of vertices. Therefore, the computation of metrics needs $O(|V(G)|^2|E(G)|)$ time. It is known that the time complexity of average link methods is $O(|V(G)|^3)$. Hence, the total time complexity of our algorithm is $O(|V(G)|^2|E(G)|)$. It is necessary to mention that the actual running time is much less than this worst time, since if the distance between two vertices is larger than $2d$, no computation is needed.

5 Experimental Results

We implement the algorithm is Section 4 in Java and perform experiments in several well-known datasets. We choose the distance threshold d to be four.

The first data set is Ravasz network [27]. As Ravasz et al. pointed out, conventional network clustering methods are difficult to find the correct community structure of such network. The community structure and the merge sequence (dendrogram) are presented in Fig. 2 and Fig. 3, respectively.

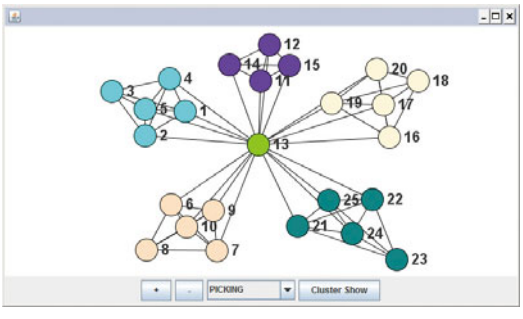


Fig. 2. Community structure using our metric in Ravasz network

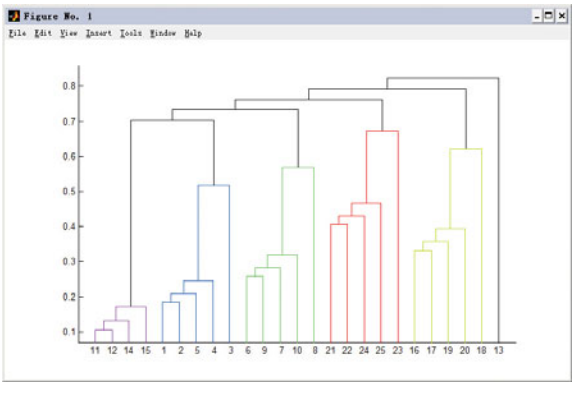


Fig. 3. Merge sequence using our metric in Ravasz network

Table 5. Modularity Results

	Ravasz	karate	football	dolphins	musician
s_{NIP}	0.3452	0.203	0.1300	0.018	0.1450
s_{JI}	0.1308	0.3400	0.6020	0.4280	0.4010
s_{WS}	0	0.2130	0.3230	0.3320	0.3010
s_{DNM}	0.5270	0.3628	0.6042	0.4278	0.4047
GN	0.5509	0.406	0.572	0.52	0.405
CNM	0.3326	0.302	0.402	0.353	0.439

We also test some famous datasets: karate club network [23], US college football league [10], Jazz musician network [28] and dolphin social network [29]. We compare our algorithm with the algorithms based on existing vertex similarity metrics. In addition, we also show the result of the-state-of-the-art algorithms Girvan-Newman algorithm [10] and CNM algorithm presented by Clauset, Newman and Girvan [1].

We propose the experimental results of modularity in Table 5, which shows that our metric is superior to the existing ones. Table 5 indicates that the Girvan-Newman algorithm outperforms in three datasets compared with our algorithm. However, in Ravasz network our metric finds the correct structure shown in Fig. 2, though GN gets high modularity; In dolphins society the modularity of best division is 0.478 ± 0.03 stated by Lusseau [29]. Our metric is better than GN in this dataset. In addition, the time complexity of (improved) Girvan-Newman algorithm $O(|E(G)|^2|V(G)|)$ is higher than our algorithm.

6 Conclusions

In this paper, we have proposed a new vertex similarity metric for community discovery. This metric computes overlapping proportion of neighbors within a distance with respect to considered vertices. Thus, it considers both topological distance and local edge density. The experimental result shows that our metric is better than the existing ones. In addition, it appears that our algorithm based on this metric has several advantage to the Girvan-Newman algorithm and CNM algorithm in some aspects.

Since the computation of our metric is in a local part, distributed or parallel computation is available which enables that our algorithm can tackle large scale graph. In addition, in the light of Clauset's local algorithm [30], our algorithm can be extended to an incremental algorithm of which the running time is reduced dramatically. Furthermore, an online algorithm based on our model can be developed.

Acknowledgements

This research is supported in part by NSFC under grant No.60603066, China National High-tech Program under grants No.2007AA01Z436, and Shenzhen

Science and Technology Program under grants No.NSKJ-200707, 08CXY-44, PCT200805190162A.

References

1. Clauset, A., Moore, C., Newman, M.E.J.: Hierarchical structure and the prediction of missing links in networks. Nature 453, 98–101 (2008)
2. Price, D.: Networks of scientific papers. In: Kochen, M. (ed.) The Growth of Knowledge: Readings on Organization and Retrieval of Information, pp. 145–155. Wiley, Chichester (1965)
3. Dunne, J.A., Williams, R.J., Martinez, N.D.: Foodweb structure and network theory: The role of connectance and size. Proc. Natl. Acad. Sci. USA 99, 12917–12922 (2002)
4. Kauffman, S.A.: Metabolic stability and epigenesis in randomly connected nets. J. Theor. Bio. 22, 437–467 (1969)
5. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y.: A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA 98, 4569–4574 (2001)
6. Sales-Pardo, M., Guimera, R., Moreira, A.A., Amaral, L.A.N.: Module identification in bipartite and directed networks. Proc. Natl. Acad. Sci. USA 104, 15224–15229 (2007)
7. Kernighan, B.W., Lin, S.: An efficient heuristic procedure for partitioning graphs. Bell System Technical Journal 49, 291–308 (1970)
8. Fiedler, M.: Algebraic connectivity of graphs. Czech. Math. J. 23, 298–305 (1973)
9. Pothen, A., Simon, H., Liou, K.P.: Partitioning sparse matrices with eigenvectors of graphs. SIAM J. Matrix Anal. Appl. 11, 430–452 (1990)
10. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99, 7821–7826 (2002)
11. Tyler, J.R., Wilkinson, D.M., Huberman, B.A.: Email as spectroscopy: automated discovery of community structure within organizations. In: Huysman, M.H., Wenger, E., Wulf, V. (eds.) Proceedings of the International Conference on Communities and Technologies, pp. 81–96. Springer, Heidelberg (2003)
12. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and indentifying communities in networks. Proc. Nat. Academy of Science (PNAS) 101(9), 2658–2663 (2004)
13. Flake, G.W., Lawrence, S., Giles, C.L., Coetzee, F.M.: Self-Organization and Identification of Web Communities. Computer 35(3), 66–71 (2002)
14. Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment. J. ACM 46(5), 604–632 (1999)
15. Pirolli, P., Pitkow, J., Rao, R.: Silk from a Sows Ear:Extracting Usable Structures from the Web. In: Proc. ACM Conf. Human Factors in Computing Systems (CHI), pp. 118–125 (1996)
16. Donetti, L., Münoz, M.A.: Detecting network communities: a new systematic and powerful algorithm. J. Stat. Mech., P10012 (2004)
17. Capocci, A., Servedio, V.D.P., Caldarelli, G., Colaiori, F.: The scale-free topology of market investments. Physica A 352, 669 (2005)
18. Alves, N.A.: Unveiling community structures in weighted networks. Phys. Rev. E 76(3), 36101 (2007)

19. Reichardt, J., Bornholdt, S.: Detecting fuzzy community structures in complex networks with a potts model. Phys. Rev. Lett. 93(21), 218701 (2004)
20. Zhou, H.: Distance, dissimilarity index, and network community structure. Phys. Rev. E 67(6), 061901 (2003)
21. Arenas, A., Diaz-Guilera, A., Peerez-Vicente, C.J.: Synchronization reveals topological scalses in complex networks. Phys. Rev. Lett. 96(11), 114102 (2006)
22. Fortunato, S.: Community detection in graphs, arXiv, 0906.0612 (2009)
23. Newman, M.E.J.: Detecting community structure in networks. Eur. Phys. J. B 38, 321–330 (2004)
24. Jaccard, P.: Etude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Socit Vaudoise des Sciences Naturelles 37, 547–579 (1901)
25. Barnes, E.R.: An algorithm for partitioning the nodes of a graph. SIAM Journal for Algorithms and Discrete Methods 3, 541–550 (1982)
26. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Phys. Rev. E 70, 066111 (2004)
27. Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., Barabsi, A.L.: Hierarchical organization of modularity in metabolic networks. Science 297, 1551–1555 (2002)
28. Gleiser, P., Danon, L.: Community structure in Jazz. Adv. Complex Systems 6, 565–573 (2003)
29. Lusseau, D., Schneider, K., Boisseau, O.J., Haase, P., Slooten, E., Dawson, S.M.: The bottlenose dolphin commu-nity of doubful sound features a large problem of long-lasting associations. Behav. Ecol. Sociobiol. 54, 396–405 (2003)
30. Clauset, A.: Finding local community structure in networks. Phys. Rev. E 72, 026132 (2005)

Seat Usage Data Analysis and Its Application for Library Marketing

Toshiro Minami and Eunja Kim

Kyushu Institute of Information Sciences, Japan & Kyushu University Library, Japan and Gwacheon Information Science Library
minami@kiis.ac.jp, gaia55@naver.com
http://www.kiis.ac.jp/~minami/

Abstract. Due to the progress of information and communication technology, our society is changing very quickly. Along with this, people's requirements to information are not only changing vigorously but also have a huge variety in their forms. As a result it becomes more and more difficult for the libraries to provide their patrons with appropriate information services. In this paper, we demonstrate the usefulness of the concept of library marketing through some examples; especially with some analysis methods for seat-usage, or seat-occupation, data in library. We investigate what seats are more preferred than others and try to deduce tips for better seat arrangement, combination of different types of seats, etc. Even though our research is in a very early stage so that we could not infer that good suggestions to better seat arrangements, we believe in its importance and it would propose the best solution in seat design in the future.

Keywords: Library Marketing, Seat Usage/Occupation Data Analysis, Data Mining, Seat Arrangement.

1 Introduction

Due to the progress and popularization of information and communication technology, our society is changing very quickly. Not only it becomes easier and faster to access information but also our way of dealing with and requirements to information has changed thoroughly. Our information access becomes ubiquitous; i.e. at any time from any place.

Libraries have been playing a very important role in information and knowledge finding. Providing the patrons with appropriate information is the major mission for libraries, as was indicated in the Five Laws of Library Science advocated by S.R. Ranganathan in 1960s [6];

(1) Books are for use,
(2) Every reader his book,
(3) Every book his reader,
(4) Save the time of the reader, and
(5) The library is a growing organism.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 238–247, 2011.

About a half century has passed since then and even though the environment of the libraries has changed a lot, the main idea about the mission of library is still very important even now and will continue in the future as well.

In order to keep this major role in our society, libraries have to be changing according to the change of people's requirements about information and knowledge. The big problem is that the spectrum of the people's requirements is very wide now. Thanks to the existence of the Internet and mobile terminals, like cell phones and other wireless equipments for digital communications, people would eager to get information as soon as possible at the site they encounter in their daily lives.

A lot of online information services are provided by quite a lot of (profit-oriented) companies in order to fulfill such requirements. Now we are able to check how many minutes later our intended bus will come to the bus stop where we are waiting. We can find the best route from where we are to our destination building, as well. But such services are not sufficiently enough. A very convenient service that is provided without any charge now may change to be payed service at any time when the company wants to do. It is highly necessary for us to get some kinds of information providing, and educational or research support services, as the essential public services for our society just like we get security services by police and fire departments as public services.

Libraries have been providing with information mainly in the form of printed books and magazines. However, considering the environmental changes relating to information materials, they have to change in order to adapt to the current environment. Of course the librarians recognize it and they put much effort to change themselves. However because that the speed of the change is so fast and the services that people want to get becomes quite a wide spectrum, it is quite difficult to appropriately change the way for the libraries based on the traditional methodology they are taking so far.

Our idea to deal with such a situation is to use the data analysis, or data mining, methodology. It is well known that a convenience store chain utilizes the data. They collect the POS (Point of Sales) data as the customers purchase their goods. The purchase data together with some profile data of the customers will be sent to the central server immediately. The specialists analyze the data and extract the information that will suggest what kind of goods should be delivered to the store for tomorrow. The net-shopping sites can collect the customers behavior data automatically. They analyze them and decide which goods to recommend to each customer according the customer's purchase history and his/her behavioral characters extracted from the data; i.e. Web mining. Such examples suggest that data analysis, or data mining, for library's marketing purpose will be very successful.

However the librarians are very worried about the privacy issue and they do not want to deal with the data the libraries can have for marketing purpose. We believe that they should go one step forward in order to keep being the public service organization that can answer to the patron's requirements more appropriately. They can do their best for protecting the data that may be connected

with the patrons' privacy data from spilling out of the library, by taking the up to date security system. The research results from the privacy preserving data mining (PPDM) [1] will be able to apply to their data analysis. By renaming the patrons' IDs from the real IDs to the tentative IDs that are used for analysis, the risk for privacy issue will be relieved much.

The aim of this paper is to ask for the librarians to start using data analysis for their marketing by showing library data analysis examples. We take up the seat usage data as the analysis target in this paper. Our research has been just started and is just in a very beginning stage. However we are convinced that data analysis methodology has a high potential so that it will provide the important tips and knowledge for improving patron services, creating new services, managing libraries better, and so on.

The rest of this paper is organized as follows: In Section 2, we discuss about library marketing (LM) more precisely. In Section 3, we take up the seat usage data of Kyushu University Library in Japan, and demonstrate the usefulness of the data analysis methodology in understanding the patrons' behavior. Finally in Section 4, we conclude our discussions in this paper.

2 Library Marketing

According to the American Marketing Association (AMA) [2], the concept of marketing used to be defined as: "Marketing is an organizational function and a set of processes for creating, communicating, and delivering value to customers and for managing customer relationships in ways that benefit the organization and its stakeholders." This definition is more profit-oriented than the current definition: "Marketing is the activity, set of institutions, and processes for creating, communicating, delivering, and exchanging offerings that have value for customers, clients, partners, and society at large."

From these two definitions, we recognize that marketing was considered as the activities that benefit the organization (company); which matches with the ordinary people's intuitive image. It is now considered as wider activities that benefit the customers and our society as well. So it is natural to apply marketing to non-profit organizations like libraries including public and university libraries. In this point of view, the aim of marketing activities by libraries (library marketing) is to give better services to their users, or patrons, so that they are able to get better reputations, to be recognized as more reliable organizations, and to get more customer satisfaction (CS)/patron satisfaction (PS) eventually. In addition to this aim it is preferable to perform their jobs more efficiently, and with less cost; which can be another important aim of library marketing.

In this paper we focus on the library marketing methods based on those of analyzing the objective data and extracting useful information and knowledge (see also [5] on this) not only for libraries but also for their patrons. Libraries have many kinds of data including circulation records (borrowing or returning of books and other materials), catalog information, patrons' entrance data, book reservation records and so on. Some libraries also have patrons' exiting time

data, reservation data for study rooms, PCs' session records, etc. However most of these data are not used sufficiently so far. It is really a big waste of potentially very valuable data. We carry out our research on library marketing by dividing the process into four levels [3,4].

(i) Preliminary Investigation
In this level we investigate what information, tips, and knowledge could be obtained by analyzing some kinds of data as case studies. We do not worry much about if we can really get such data or the extracted information is very useful or not. Our aim in this level is to create as many possible ideas as we can imagine which could be and/or may be used for library marketing.

(ii) Real Data Analysis
In this level we apply the methods obtained in the preliminary investigation level. By using the real data, we can evaluate the analysis methods from the practical point of view. If we find out that an analysis method is very useful, then we apply this method to another data. It could happen that we can apply a method to other types of data by modifying it, slightly or largely. Most of the analysis methods presented in this paper can be considered to be those in this level. We will continue our research on this level and try hard to find as many practically useful methods as possible.

(iii) Combination of Methods
Even though one type of data can be useful enough for library marketing, we would be able to extract even more useful information by combining the extracted information/knowledge and combining more than one types of data. We will investigate this type of analysis methods after we investigate the level (ii) sufficiently.

(iv) Development of the Automated Methods
As we have found a very useful analysis method, it should be convenient to apply it by automating the analysis method. This method is a kind of macro procedure so that it is a pack of analysis methods and thus can be considered as one method. As a result, this analysis is easy to use as well as it can be used as a part of more sophisticated automated methods.

In this paper we investigate the seat usage/occupation data as a preliminary investigation (i) and an example analysis of a very small real data (ii).

3 Seat Usage Data Analysis

3.1 Case Study at Kyushu University Library (KUL)

Kyushu University is one of the biggest and high ranked universities in Japan. It consists of 13 faculties, and has about 19,000 students. The main campus is located in Hakozaki area, where the main library of the Kyushu University Library (KUL) system is located. KUL has about 4 million books, about 90 thousand titles of magazines, and about 40 thousand titles of e-journals. Roughly

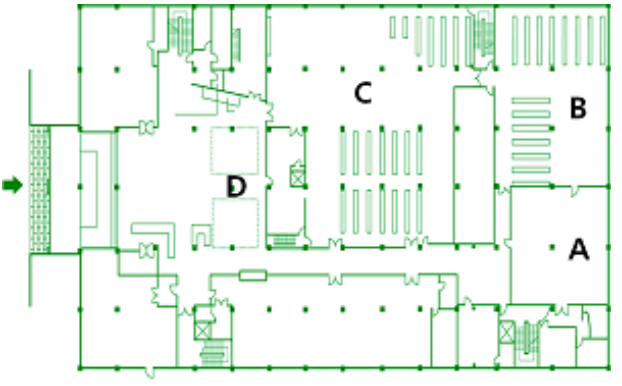


Fig. 1. The 2nd Floor Plan of the Central Library of KUL

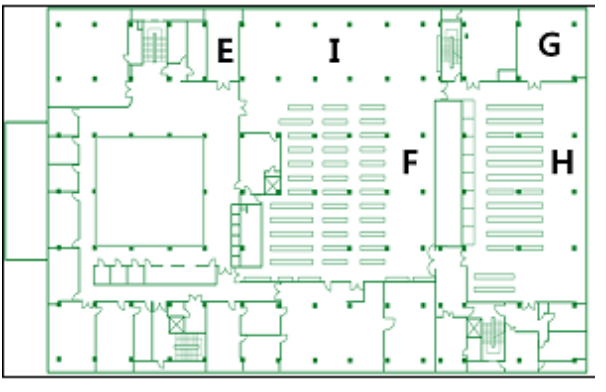


Fig. 2. The 3rd Floor Plan of the Central Library of KUL

340 thousand students, university staff, and others visit the central library per year.

We show the 2nd and 3rd floors of the central library building, which are the main parts of the building to be used by patrons. (2nd Floor: Fig. 1, 3rd Floor: Fig. 2). The 2nd floor is the entrance floor: the entrance gate is located in the left part of the figure. In the lobby, is a PC area (Fig. 3), where 46 PCs are available, which is called the Information Salon 1 (D). In the right area of the lobby as the patrons enter are the counters for circulation and reference services. Rooms marked A and B are the reading rooms. The Room C is now used as a "learning commons" area, which is equipped with many small tables and students are allowed to re-arrange the tables freely and they can talk aloud in a group meeting so some other purposes. It is also used for seminars organized by professors.

In the 3rd floor, are located reading rooms (marked G, H, I, and F) and another PC room (marked E), called the information salon 2 (Fig. 4), which is

Fig. 3. A PC Room: Information Salon 1 (D) in the 2nd Floor

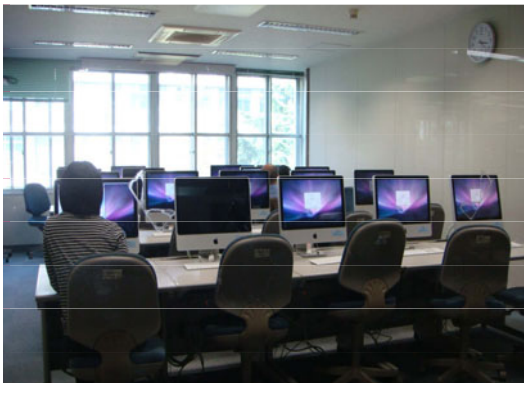


Fig. 4. A PC Room: Information Salon 2 (E) in the 3rd Floor

much smaller than the information salon 1. These information salons are used for lectures occasionally.

3.2 Analysis Example for the Information Salon 2

We created some seat usage data of the central library of KUL. We checked which seats are occupied by patrons by spending several days. Roughly speaking, we recorded every half hour. For example, in the case of Room E, i.e. the Information Salon 2, in the 3rd floor, we collected the data for January 13, April 16, and April 17 in 2009. On January 13, we investigated 7 times during the period from 15:30 to 20:30, and in April 16 and 17, 28 times from 8:00 to 21:30. So we have 63 times of data in total.

The seat arrangement of Room E is illustrated in Fig. 5. As we can see in Fig. 4, there are 5 PCs in a table. There are 4 tables in the room and 2 of them in the front part only have 4 PCs instead of 5 because the central part is used for printers. Therefore there are 18 PCs and seats in Room E.

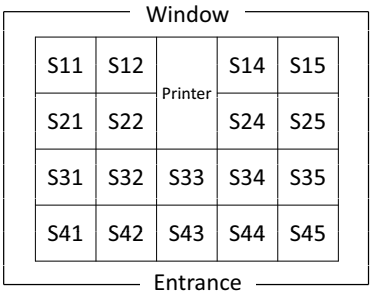


Fig. 5. Seat Arrangement of the Information Salon 2 (E)

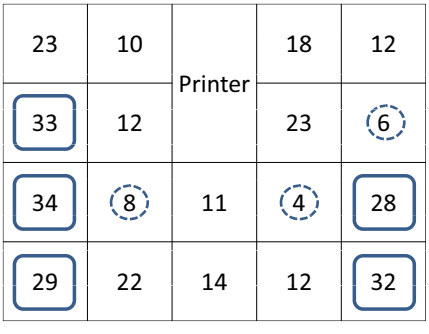


Fig. 6. Seat Usage Frequencies of the Information Salon 2 (E)

Fig. 6 shows the total numbers of the usage of each seat. For example the seat S11 was found to be used 23 times among 63 times; the usage ratio is 37%. The most popularly used seat was found to be S31 (54%), which is followed by S21 (52%) and S45 (51%). On the other hand the most unpopularly use ones are S34 (6%), S25 (10%), and S32 (13%). We put rectangular marks to the most popular seats and put circular marks to the most unpopular seats.

From these data we can say that the people prefer to use the seat located in the edge and at the same time not too much far away from the entrance. So the most popular seats are located in either S*1 and S*5. It is interesting that the left edge, i.e. S*1, is more popular than the right edge, i.e. S*5. Probably in some reason the right side passage is considered to be the main passage area than the left one. So they prefer to use the left one as the more quiet seats than the right ones.

At the same time it seems that they do not want to use the seat next to the one which is already occupied. This describes why S32 and S34 is not popular. It is interesting to see that the seat next to the printer seems to be considered a kind of edge positions. This may be the reason why S14 and S24 is more popular than other inner seats. This also describes why S25 is not popular; it is next to the relatively popular seat S24.

9	5	Printer	4	5
7	3		9	1
8	3	7	3	8
11	7	8	7	10

Fig. 7. Session Frequencies of the Seats in the Information Salon 2 (E)

2.6	2.0	Printer	4.5	2.4
4.7	4.0		2.6	6.0
4.3	2.7	1.6	1.3	3.5
2.6	3.1	1.8	1.7	3.2

Fig. 8. Average Session Times of the Seats in the Information Salon 2 (E)

We take another view on this data. We will define a session of a seat usage if the seat has been used continuously in the data. As an example, let us take the seat S11 in the data for April 17. The seat has been occupied at 11:00, 11:30, 12:00, 12:30, 13:30, 14:00, 14:30, 15:00, 16:00,16:30, 19:00, 19:30. This means it consists of 4 sessions; from 11:00 to 12:30, from 13:30 to 15:00, from 16:00 to 16:30, and from 19:00 to 19:30. The total number of sessions are shown in Fig. 7. As a result we can define the average session time, which is shown in Fig. 8.

The longest average session time is 6.0 at the seat S25, which has only 1 session. So a patron happened to choose this seat and used it for a work that needed relatively long time. Thus this seat usage would be an exceptional case. The seats for the next longest average session times are S21, S14, and S31. Among these 3 seats, S21 and S31 are the seats that have also used for a long time (See Fig. 6). So we can say that these seats are preferably chosen by the patrons for the works that would need to spend a long time. On the other hand the seat S45 has a big frequency while its average session time is not very large; which might mean that this seat is easy to access because it is close to the entrance door and thus this seat is considered to be very useful for relatively shorter works. Similarly the seats S33, S34, S43, and S44 have short session times. Among them the session S33, S43, and S44 have relatively larger usage frequencies. So we can

Fig. 9. A Reading Room (I)

say that these seats are preferably chosen for small works because they are easy to acess.

It is worth pointing out that Uematsu suggests that a table with 6 seats with 3 seats in one side and the rest 3 in another side (see for example in Fig. 9, Reading Room I) is full for two patrons because they will occupy the diagonal seats and nobody will not willing to use the remaining 4 [7]. In a rough analysis so far, it looks true in our data too. We have to analyze in more detail for further investigation.

4 Concluding Remarks

Libraries are supposed to go with society as was described in the five laws of library science [6]. Actually they have been changing itself based on this philosophy. They started using computers in 1980s and started with providing home pages as the Internet and Web become popular in 1990s. However the speed of changes of society is too fast recently for the libraries to catch up with it based on the methodology they have been taking so far. They need a new methodology in the coming information age.

We believe that data collection and data analysis including the data mining technologies have essential importance for the libraries. So we have started with developing data analysis methods for the libraries.

In this paper we put focus on the seat usage data. We have investigated which seats the patrons are using in the central library of Kyushu University. We chose the Information Salon 2 for an example and try to find some tips extracted from the data. Our main concern is to understand how the patrons choose their seats. Then we would like to find some kind of information and knowledge based on the results, which is useful as the librarians plan the seat arrangement in a library.

As we have just started our research based on such a methodology, we do not have much yet. However, we are convinced that we will be able to find more and more useful knowledge as we continue our research in this direction. In order to

do this, we have to collect much more data not only on seat usage data but also other data that will be useful for understanding patrons and their behaviors in and out of the library.

References

1. Aggarwal, C.C., Yu, P.S. (eds.): Privacy-preserving data mining: models and algorithms. Advances in database systems, vol. 34. Springer, New York (2008)
2. American Marketing Association (AMA), http://www.marketingpower.com/
3. Minami, T., Kim, E.: Data Analysis Methods for Library Marketing. In: Lee, Y.-h., Kim, T.-h., Fang, W.-c., Ślęzak, D. (eds.) FGIT 2009. LNCS, vol. 5899, pp. 27–34. Springer, Heidelberg (2009)
4. Minami, T., Kim, E.: Data Analysis Methods for Library Marketing in Order to Provide Advanced Patron Services. International Journal of Database Theory and Application 3(2), 11–20 (2010)
5. Nicholson, S.: The basis for bibliomining: Frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services. Information Processing & Management 42(3), 785–804 (2006)
6. Ranganathan, S.R.: The Five Laws of Library Science, 2nd edn. Asia Publishing House (1957)
7. Uematsu, S.: To Have a Look at Library as Architecture. Bensei Shuppan (1999) (in Japanese)

MDL: Metrics Definition Language*

Jerzy Brzeziński, Dariusz Dwornikowski, Michał Kalewski,
Tomasz Pawlak, and Michał Sajkowski

Institute of Computing Science, Poznań University of Technology
Piotrowo 2, 60-965 Poznań, Poland
{Jerzy.Brzezinski,Dariusz.Dwornikowski,Michal.Kalewski,
Tomasz.Pawlak,Michal.Sajkowski}@cs.put.poznan.pl

Abstract. The paper presents Metrics Definition Language (MDL), a new format for metric definition. MDL is a flexible grammar based language, which clearly separates the idea of a data source from a metric. Hence, MDL provides ways to define complex metrics that can be defined over many data sources or even other metrics. It also supports basic mathematical operations, that can be used to define metric aggregations and transformations. The language is human readable, universal and data format agnostic. It can be easily used with every monitoring system.

1 Introduction

Heterogeneous computer systems tend to grow in terms of quantity of the components involved in the computation, as well as in terms of their overall complexity. Moreover heterogeneous paradigms, such as cloud computing and SOA, introduce additional layers of abstraction.

Services in SOA cooperate to form complex business processes, whereas in cloud computing virtual machines are used as an additional layer over the operating systems. Obviously, the amount of the monitoring information generated by such systems is increasing at even greater pace. With the new abstraction levels new challenges in monitoring appear, one of these challenges is the problem of defining metrics. Commonly used, statically defined metrics may now not be sufficient enough to fully describe complex and dynamic distributed systems. Monitoring and management should provide efficient and easy ways to dynamically define and create new metrics from the existing ones, e.g. for the needs of sophisticated load balancing, resource migration and business process orchestration.

Traditional monitoring systems can be categorised as probe focused. This means that values of gathered data are treated as a separate simple metric, additionally both the data source and the metric are semantically inseparable. The commonly known example is SNMP protocol and its agent. The agent provides many separate metrics, which are defined in a MIB (*Management Information Base*) format. To derive complex metrics,

* This work has been partially supported by the Polish Ministry of Science and Higher Education within the European Regional Development Fund, Grant No. POIG.01.03.01-00-008/08 IT-SOA https://www.soa.edu.pl/

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 248–256, 2011.

off-line analyses has to be carried out in a monitoring system or the agent itself. Its MIB has to be extended, in most of the scenarios by means of programming. Other monitoring systems seem to share this configuration model of defining metrics, which makes it hard to extend them dynamically.

This paper's contribution is a proposal of an universal format for defining metrics — Metrics Definition Language (MDL), which can be used to create dynamically new metrics on the top of monitoring data sources.

The paper is organised as follows. Section 2 presents a motivation for the Metrics Definition Language (MDL). Section 3 presents the Metrics Definition Language, its grammar and example documents with metrics definitions. Section 4 contains conclusions and an outline for a future work.

2 Motivation for Metrics Definition Language

Many monitoring systems use their own configuration system for data sources and metrics definition. Moreover, in most cases, the data source is inseparable from the metric, making them one entity in the context of configuration. This introduces unneeded complexity of the configuration, because every new metric has to be defined separately for the same data source. In Zabbix [10] monitoring system this problem is minimised by introducing so called Host Templates, which provide fairly easy way to define the configuration template for a given host and reuse it in the future for other machines. Nevertheless, this mechanism is Zabbix internal only and is not portable to other monitoring systems. Similar mechanisms can be found in most of the monitoring systems.

Traditional monitoring systems also lack flexibility in metric definition. None of them provides the means to define new complex metrics out of exiting ones. For example, one could need to easily add metric that is a sum of all of the system loads of every host in a network or be an average throughput of all the interfaces in a network. Moreover, such a mechanism should be easy to use and not need system restart or modification. It should also be human readable but at the same time easy to process by a machine.

There exist standards for management, such as WSDM [3,8,9], WS-Management [1], CIM (*Common Information Model*) [4]. However they focus on providing a framework for defining management interfaces and ways for describing managed elements and its interfaces. The substandard of WSDM — MUWS (*Monitoring Using Web Services*) gives some partial support for metrics definition, by extending the *Metric capability* object. This method is only suitable for defining metric metadata and does not tell how to measure the metric and what data to use. Similar approach is in CIM Metrics [5].

The above weaknesses of traditional approaches to metric definitions show a need of new solutions to configuration of monitoring systems.

3 Metrics Definition Language

For defining sources of monitoring data, as well as metrics we propose Metrics Definition Language (MDL) as a universal format. We state MDL's requirements and grammar along with examples of MDL's use. At last we compare the MDL to Event Processing

Language (EPL), that MDL loosely bases its grammar on, highlighting and motivating the differences of these languages in the context of monitoring.

3.1 Requirements for MDL

The requirements for the MDL are directly derived from the motivation outlined in Section 2. These are:

Human and Machine readability. MDL should be easy to understand and to use by a human, but also easy to process by a machine. This requirement opens up the way to full automation of monitoring systems (e.g. autonomic computing), but also is very realistic and assumes that human interventions are common and needed.

Universality. This requirement assumes that MDL should be fairly easy to apply in most of the existing monitoring systems. It can be easily fulfilled by assuming that every data packed in traditional monitoring system can be treated as an event. It might also mean, that on the level of implementation some components should be adjusted or added, but the very syntax of the MDL is universal and does not have to be adjusted.

Clear separation of data sources and metrics. MDL should provide syntax for data source definitions separately from metrics definitions. Metrics should be defined as general measures that are computed over the data acquired from the data sources. These requirement is of a great importance, as it clearly separates the data from the calculation, in result, making it easy to use templates. It also opens the possibility for the next requirement, which is support for complex metrics.

Support for complex metrics. MDL should provide ways for easy and flexible definition of new metrics, that can be freely composed in terms of mathematical operations on the data. Such complex metrics can be created by any mathematical aggregating function, such as sum, average, min or max. This opens also support for global metrics, which are aggregations of many metrics of the same type.

MDL should be data format agnostic. This requirement assumes that our language should not be strictly assigned to one specific format of the event or data packet. The data format is strictly dependent on the implementation.

3.2 The Proposed Approach

To meet tech specified requirements, we propose Metrics Definition Language (MDL) for metric and data source definition. Many modern languages and protocols are created in structured markup languages, like XML. Anyway, we decided not to follow this direction and to create a simple, yet powerful grammar based language, in order to support human readability requirement. We based MDL's grammar loosely on EPL, the language used for defining event processing queries. This step was dictated by two facts. Firstly, EPL is in use and we found its grammar as very intuitive and simple,

yet very powerful. Secondly, this decision is complementary with our approach to treat monitoring data as streams of events. EPL itself is however not suitable for our needs, as it is too complex for this problem's domain and it is strictly event processing oriented.

The MDL has following features:

- grammar based language
- human and machine readable
- intuitive to use
- support for variables
- clear separation of data sources and metrics definitions
- support for metric templates
- complex metrics support
- event oriented (but still universal)

Below, we present MDL's grammar and few examples of MDL documents, which best show the expressiveness of the language.

3.3 Grammar of MDL

One MDL document may contain one or more metric definitions. There are two section types in the document: **STREAMS** and **METRIC**. The **STREAMS** section contains declarations of streams of events used by following metrics, defined in the **METRIC** section(s). It is important that this section must be the first section in the file. Each **METRIC** section contains definition of exactly one metric. There must be at least one **METRIC** section in the document. The general form of document is presented below.

```
STREAMS
  <stream declaration 1>,
    .
  <stream declaration N>
METRIC
  <metric declaration 1>
    .
METRIC
  <metric declaration M>
```

Each declaration of stream contains original stream name (usually the same as the name or the type of events in the stream), optional conditions for selecting events, time window declaration and an alias of the stream. Declarations of conditions and time window can also contain variables. Variable is denoted by preceding and ascending colon ':', e.g. `:time:` – variable named 'time'. Variables cannot contain whitespace characters, moreover first character of its name must be a letter or underscore '_'. In declaration of conditions, variable could be used only as value of the parameter. In time window declaration, variable could be used only as whole time window length declaration.

```
<stream_declaration> : stream_name(<conditions>?).<window> [AS] alias
<conditions> : param1=val1[(AND|OR) param2=val2...[(AND|OR)paramN=valN]]
<window> : last() | win(<window_length>)
<window_length> : [0-9]+ (msec|sec|min|events?)?
```

There are two types of time windows: time-based and event-based windows. Time-based window defines how old (in time) events could be processed when accessing the data stream, on the other hand the event-based window defines how many (newest) events could be processed. Both time and length windows are defined by `win(...)` declaration, type of the window depends on the contents of the window. Moreover, there is a special type of time window `last()` which allows processing only last (the newest) value in the stream. The `last()` window is equal to `win(1 event)`. When no time operator is used, the window length is set to specified number of events.

Each metric declaration contains its name, optional computation interval, equation and optional variable values. If computation interval is not set then metric will be computed every time new event is available in any stream used by particular metric. If computation interval is set to a number, it is interpreted as a number of new events in any of streams. Each variable used in declaration of stream, used in the metric, must be set in definition of the metric. When metric makes use of multiple event streams, then cross join of them is requested.

```
<metric_declaration>    : metric_name [COMPUTE EVERY <interval>]
                <equation> [, var1=value1 [,var2=value2...[, varN=valueN]]]
<interval>              : [0-9]+ (msec|sec|min|events?)?
<equation>              : ( <param> | <func>(<param>) ) [<op> <equation>]
<param>                 : <stream_alias>.<event_param> | -?[0-9]+(.[0-9]+)?
<func>                  : avg | sum | min | max | count
<op>                    : * | / | + | -
```

3.4 Examples

Example 1. Simple metric computing average CPU usage in last 1 minute; metric value is recomputed every 10 seconds.

```
STREAMS
    CPU(global=true).win(1 min) AS CPU_global
METRIC cpu_load60 COMPUTE EVERY 10 sec
    avg(CPU_global.usage)
```

Example 2. MDL document containing two streams, which takes advantages from variables to create set of similar metrics. The *load30* metric utilizes both of streams to compute abstract value called *load*.

```
STREAMS
        CPU(global=true).win(:time:)    CPU_global,
        MEMORY.win(:time:)              mem

METRIC load30 COMPUTE EVERY 15 sec
        avg(CPU_global.usage) * avg(mem.usage), time='30 sec'

METRIC cpu_load30 COMPUTE EVERY 15 sec
        avg(CPU_global.usage), time='30 sec'

METRIC cpu_load60 COMPUTE EVERY 30 sec
        avg(CPU_global.usage), time='60 sec'
```

Example 3. Set of complex metrics used in test system.

```
STREAMS
        CPU(global=true and total=false).win(:time:) AS cpu,
        CPU(cpu=0 and core=:core: AND
                        global=false AND total=false).win(:time:) cpu_core,
        Memory().win(:time:)              AS mem,
        Disk(name=:name:).last()          AS disk,
        Disk(type=:type:).win(5 sec)      AS disk_type,
        Network().win(2000 msec)          AS net,
        Process(name=:name:).last()       AS proc

METRIC cpu_usage15 COMPUTE EVERY 5 sec
        avg(cpu.usage), time = '15 sec'

METRIC cpu_usageMAX60 COMPUTE EVERY 20 sec
        max(cpu.usage), time='60 sec'

METRIC cpu_usage60min COMPUTE EVERY 1 min
        avg(cpu.usage), time='60 min'

METRIC cpu_usageACTUAL
        cpu.usage, time='1 event'

METRIC core0_usage15 COMPUTE EVERY 5 sec
        avg(cpu_core.usage), time='15 sec', core=0

METRIC mem_usage15 COMPUTE EVERY 5 events
        avg(mem.usage), time='15 sec'

METRIC mem_usage60min COMPUTE EVERY 1 min
        avg(mem.usage), time='60 min'

METRIC diskC_usage COMPUTE EVERY 5 sec
        disk.usage, name='C:'

METRIC local_disk_usage COMPUTE EVERY 5 sec
        avg(disk_type.usage), type=3

METRIC total_traffic COMPUTE EVERY 1 sec
        sum(net.totalKBytesPerSec)/2

METRIC cpu_TestApp
        proc.processorUsage, name='TestApp'
```

3.5 Comparison of MDL to EPL

The following comparison of MDL to EPL refers to (N)Esper version of Event Processing Language [6, ch.4-9].

The expected data flow begins in many data sources, which generate streams of events, or tuples. Each stream is passed to the metric engine and a single value is computed for each of the metrics. The EPL data source (the FROM clause) contains streams of events, but its return data is a whole tuple, not as expected in metrics calculations, single value. On the contrary, metric definitions in MDL return single value, which is what is expected by the monitoring system. In that case MDL is simpler than EPL but sufficiently functional for metric definition needs.

Moreover, usage of variables in EPL statement causes creation of statement template, which must be filled by the implementation, or by another EPL statement. Metrics language is supposed to define ready to compute metrics. MDL achieves this by applying

variables to streams only. There are templates of streams filled by each metric using them, everything in one MDL document. In addition templates of streams allow creation of multiple similar metrics, differing only in length of time window or a value of some parameter in condition of event selection.

Single EPL statement can define only one metric, with assumption that it returns exactly one value at once. Creation of EPL statement that returns values of many non-trivial metrics in one query is much more complicated than creation of many equivalent simple EPL statements, so this case is ignored. On the other hand it is possible to create set of metrics, each sharing declaration of streams in a single MDL document.

Finally MDL document is usually shorter and simpler than equivalent set of EPL statements, see comparison of metrics from example 3 with the following set of EPL statements. The code metrics could be found in Table 1. As can be seen, the most significant simplicity gain is obtained by introducing templates of streams in MDL.

```
SELECT  avg(cpu.usage) as cpu_usage15
FROM    CPU(global=true AND total=false).win:time(15 sec) as cpu
OUTPUT EVERY 5 sec

SELECT  max(cpu.usage) as cpu_usageMAX60
FROM    CPU(global=true AND total=false).win:time(60 sec) as cpu
OUTPUT EVERY 20 sec

SELECT  avg(cpu.usage) as cpu_usage60min
FROM    CPU(global=true AND total=false).win:time(60 min) as cpu
OUTPUT EVERY 1 min

SELECT  cpu.usage as cpu_usageACTUAL
FROM    CPU(global=true AND total=false).win:length(1) as cpu

SELECT  avg(cpu_core.usage) as core0_usage15
FROM    CPU(cpu=0 AND core=0  AND
      global=false AND total=false).win:time(15 sec) cpu_core
OUTPUT EVERY 5 sec

SELECT  avg(mem.usage) as mem_usage15
FROM    Memory().win:time(15 sec) as mem
OUTPUT EVERY 5 events

SELECT  avg(mem.usage) as mem_usage60min
FROM    Memory().win:time(60 min) as mem
OUTPUT EVERY 1 min

SELECT  disk.usage as diskC_usage
FROM    Disk(name='C:').win:length(1) as disk
OUTPUT EVERY 5 sec

SELECT  avg(disk_type.usage) as local_disk_usage
FROM    Disk(type=3).win:time(5 sec) as disk_type
OUTPUT EVERY 5 sec

SELECT  sum(net.totalKBytesPerSec)/2 as total_traffic
FROM    Network().win:time(2000 msec) as net
OUTPUT EVERY 1 sec

SELECT  proc.processorUsage as cpu_TestApp
FROM    Process(name='TestApp').win:length(1) as proc
```

Since typical code metrics are defined for object–oriented languages, let us define some simple metrics suitable for declarative languages like EPL or MDL. The names of metrics from Table 1 are very descriptive, so let us discuss the meaning and consequences

of them. The *Lines of code (without empty lines)* is one of the oldest code metrics in the world with, having its advantages and disadvantages, as described in [2,7]. The main advantage of it is natural, intuitive description of length of code. In addition similar metric *Number of characters (without whitespace)* could be used to estimate size of whole MDL document. Since packets in computer networks have limited capacity, e.g. Ethernet frame, smaller size of metric document could fit one frame and consequently speed up distribution of the MDL document over the network. The last metric *Number of declarations of streams* is used to show differences in complexity of code, but not its overall size. In each case simplicity gain is computed as $1 - \frac{metric(MDL)}{metric(EPL)}$.

Table 1. Code complexity comparison between EPL and MDL [7]

Code metric	**EPL**	**MDL**	**Simplicity gain**
Lines of code (without empty lines)	31	30	3.2%
Number of characters (without whitespace)	1070	950	11.2%
Number of declarations of streams	11	7	36.4%

4 Conclusions and Future Work

The paper has presented a new format for defining metrics in monitoring systems, called Metrics Definition Language. MDL is a grammar based language, loosely based on EPL, which supports defining metrics and data streams. MDL is event oriented, but under the assumption, that every data packed is treated as an event, it can be said to be universal and easily used by other monitoring systems. The key features of MDL are: support for complex metrics (aggregations of simple metrics), universality, data format independence and expressiveness. The language is also data format agnostic, it means, that it does not assume any particular format of the data, hence it can be used with existing monitoring systems and protocols, such as SNMP, Nagios, Zabbix etc. In future we plan to implement methods of determining consistent global state in distributed systems for global metrics verification. We would like to extend it to support more scenarios of monitoring and presumably integrate its naming with popular standards, such as CIM or WSDM.

References

1. American Megatrends: About WS-Management (May 2002), `http://www.ami.com/support/doc/AMI_WSMan_Techsheet_v1.1.pdf`
2. Armour, P.G.: Beware of counting loc. Commun. ACM 47(3), 21–24 (2004)
3. Brown, M.: Understand Web Services Distributed Management (WSDM). IBM (July 2005), `https://www6.software.ibm.com/developerworks/education/ws-wsdm/index.html`
4. Distributed Management Task Force, Inc. Common Information Model (CIM) Standards (2009), `http://www.dmtf.org/standards/cim/`
5. DMTF, CIM Metric White Paper, `http://www.dmtf.org/sites/default/files/standards/documents/DSP0141.pdf`

6. EsperTech Inc. Esper Reference Documentation (2009), http://esper.codehaus.org/esper-3.5.0/doc/reference/en/html/:
7. Kan, H., Metrics, S.: Models in Software Quality Engineering. Addison Wesley Professional, Reading (2002)
8. OASIS: Web Services Distributed Management: Management Using Web Services (MUWS 1.1) Part 1 (2006), http://www.oasis-open.org/committees/download.php/20576/wsdm-muws1-1.1-spec-os-01.pdf
9. OASIS: Web Services Distributed Management: Management Using Web Services (MUWS 1.1) Part 2 (2006), http://www.oasis-open.org/committees/download.php/20576/wsdm-muws1-1.1-spec-os-02.pdf
10. Zabbix, S.I.A.: Homepage of Zabbix project, http://www.zabbix.com/:

A Statistical Global Feature Extraction Method for Optical Font Recognition

Bilal Bataineh[1], Siti Norul Huda Sheikh Abdullah[2], and Khairudin Omar[3]

Center for Artificial Intelligence Technology,
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia
[1] bilal_bataineh82@yahoo.com, [2] mimi@ftsm.ukm.my, [3] ko@ftsm.ukm.my

Abstract. The study of optical font recognition has becoming more popular nowadays. In line to that, global analysis approach is extensively used to identify various font type to classify writer identity. Objective of this paper is to propose an enhanced global analysis method. Based on statistical analysis of edge pixels relationships, a novel method in feature extraction for binary images has proposed. We test the proposed method on Arabic calligraphy script image for optical font recognition application. We classify those images using Multilayer Network, Bayes network and Decision Tree classifiers to identify the Arabic calligraphy type. The experiments results shows that our proposed method has boost up the overall performance of the optical font recognition.

Keywords: Arabic calligraphy script, Font recognition, Global feature extraction, Gray level co-occurrence matrix, Statistical feature extraction.

1 Introduction

The feature extraction process is one of the most critical issue in document analysis techniques such as Optical font recognition (OFR). Generally, the feature extraction technique consists of two categories: global analysis [1][2] and local analysis [3][4]. The global analysis approach stresses on a region of text block (two lines or more) in the text context of geometric form. On the other side, the local analysis is an approach that concerns on disconnected parts in document images such as words or characters. The information about font helps in several applications such as reprinting documents, document characterization and classification, document layout analysis and improvement the multi-font optical character recognition (OCR).

Arabic language is one of international language. The Arabic alphabet has been widely adopted into other languages such as Jawi, Persian, Kurdish, Pashto, Urdu, and Hausa. In Arabic alphabet, there are three types of written forms: the printed, handwritten and calligraphy. The Arabic calligraphy is the oldest form, it is has many types, fixed accuracy standards and written by the specialist calligraphers. There are eight main Arabic calligraphy types: Old Kufi, Kufi, Thuluth, Naskh, Roqaa, Diwani, Persian and Andalusi. Fig. 1(a) until (h) show examples of one sentence written by calligraphy types.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 257–267, 2011.

Fig. 1. The (al-khat lesan al-yad) "the calligraphy is a tongue of hand" written in the main Arabic calligraphy types (a) Diwani, (b) Kufi, (c) Thuluth, (d) Persian, (e) Roqaa, (f) Naskh, (g) Andalusi, (h) Old Kufi.

In global analysis approach, usually the texture analysis techniques have used for extracting feature in document image [5][6]. Basically, the texture analysis techniques are designed for high level image such as grayscale and color images. However, the binary document images consist only two level color images and as far as the research concern, OFR is only interested in foreground text body. That may leads to limitation when applying texture analysis techniques in document images. Quite a number of previous OFR researches were focusing on the Latin [1] or Asian language [2][3]. Until now, there is negligence of the benefits of Arabic calligraphy definition such as classifying the documents layout, the purposes of the documents layout, documents library, the documents history classifying, improving the documents preparing and reprinting the documents as the original input image format.

The objective of this work is to propose an enhanced a statistical feature extraction technique for document images. The proposed method are tested on Arabic Calligraphy Script images as the sample case. The proposed method is compared with Gray Level-Co occurrence Matrix proposed by Haralick *et al.* [7]. Then, we evaluate both methods using Multilayer Network, Bayes network and Decision Tree classifiers to obtain the best performance for optical font recognition application. This paper is organized as follows. Section 2 reviews the state of art in previous OFR. Section 3 explains the proposed method and Section 4 presents and analyses the experimental results. Finally, conclusions are presented in Section 5.

2 State of the Art

Feature extraction is a process is between preprocessing and classification phases. These phase have used in several document analysis techniques. Optical font recognition OFR is one of the document analysis that require this stage.

2.1 Previous Work in OFR

Usually, the feature extraction has two categories: local analysis and global analysis, our interest is more on global analysis. The global approaches identify the text font by analyzing generated blocks of text from a document image.

At earlier, Zramdini and Ingold [1] presented a method to identify the weight, size, typeface, and slope of the printed English text image block. It employed a statistical

approach based on global typographical features using vertical and horizontal projection profile, the character main strokes and the connected components by a rectangular shape enveloping the connected components. A multivariate Bayesian classifier was used to recognize in a set of 280 font, size and style. The average accuracy rate of the font recognition was 95.6% after tested on high quality image text blocks. Some recognition process failed with some font types such as Lucida-Sans and Times fonts. Also, it is hardly to recognize the font for a short text.

Zhu *et al.* [2] described a new texture analysis approach to handle the font recognition for Chinese and English fonts. It employs the text block as a texture block where each font has a special texture. The 2-D Gabor filters was used to extract the texture features. The experiment implemented on four Chinese and eight English fonts. They achieved about 99.1% of average accuracy rate with machine documents with high quality. It ignored the scanned documents with low quality. This method obtained less recognition rate when applying a single character for each font.

In general, the local analysis approach suitable for a single dataset. It requires a modifications if other languages are included. It usually does not generalize on different languages. It strongly depends on the segmentation process and the affected noise. The global analysis approach can easily generalize on different languages. It doesn't require any basic modification if any changes or any additional samples are made in the dataset. Furthermore, only few researches focused on global feature extraction approach. Besides that, some research focused on Latin and Chinese languages, whereby other languages have their own significance in our live.

2.2 Global Feature Extraction Approach

The global feature extraction approach is branched into several categories[6,8]. The statistical method is an example of basic category which defines the texture of the images based on spatial distributions of gray level value in an image. It classifies based on statistical orders whereby the first-order static finds the value of each level and extracts the properties based on those values. Whilst, the second-order static finds the value by relating two gray-levels values with some geometrical relationship and indicates them as the important features. Apart from that, the third or higher order statics depend on finding a values of compound properties of the image [9].

The second-order statistics represent by The Gray Level Co-occurrence Matrix (GLCM) [7] and Gray-Level Difference Method (GLDM)[11]. In general, both methods are similar, but the GLCM is the most popular and use statistical method in the texture feature extraction [8,9]. GLCM has been proposed by Haralick *et al.* [7]. It is a matrix gives values explain the occurrences distribution in the image. The occurrences present the number of pair of two pixels of grayscale values are related with specific relationship. If C is a GLCM, I is an n × m image, $(\Delta x,\Delta y)$ denotes the value of the pairs of pixels that have the gray level value of i and j, the formula shown as following:-

$$C_{\Delta x,\Delta y}(i,j) = \sum_{p=1}^{n}\sum_{q=1}^{m}\begin{cases}1, & if\ I(p,q) = i\ and\ I(p+\Delta x, q+\Delta y) = j\\ 0, & otherwise\end{cases} \quad (1)$$

Haralick *et al.* [7] has proposed a several equations use to compute a texture features from the GLCM. The GLCM has used for documents analysis research that

depending on global analysis approach research such as scripts and languages identification [11,12]. It has been used in other domains such as Fingerprint Classification [13] and Chinese Sign Language Recognition [14]. However, the statistical methods provide a statistical description of the image gray-levels. With binary images, we are dealing with two levels only that lead to reduce the effectiveness of the features and to some confusion.

3 The Proposed Method

We divide the framework of Arabic Calligraphy font recognition system into two parts: preprocessing and post-processing modules. In the pre-processing, besides generating texture blocks of the predetermined text, we also include edge deduction process [15]. Whilst the post processing involves two sub processes such as feature extraction using our proposed algorithm based on statistical method and recognition sub processes. Fig. 2 shows the flowchart of the proposed method where it requires the following steps: preprocessing and edge deduction, feature extraction and recognition.

The binary images representing documents or models contain two levels such as black and white levels. Our proposed method concerns all values within closed edge image. The edge image gives a clear contour representation of model type. It represents the simplest representation of any form such as the relationship between angle and adjacent pixels. This approach keeps the information of the shape and removes all the unwanted information that adversely affects the values of features. Unlike the previous techniques such as GLCM, they keep only relationship of the actual pixels value. The proposed statistical method is easy to implement. It can easily generalize on different of image types and less prone to noise.

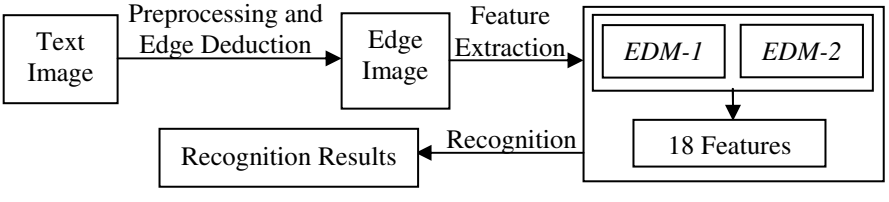
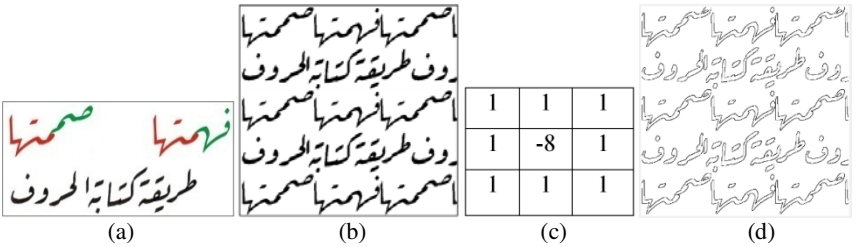


Fig. 2. The flow chart of the proposed Arabic Calligraphy font recognition system

3.1 Preprocessing and Edge Deduction

Usually, the input images have different properties. For that, we apply some methods to overcome this problem. In this stage, we prepare the image using binarization, skew correction, and text normalization subsequently. We obtain the binarization threshold value by using fixed global thresholding method. Then, we use Hough Transform [16] to determine the skew angle in the skew deduction sub-phase. In the text normalization sub-phase, we remove the spaces between words and lines, fill the incomplete lines and prepare the text size and the number of lines to generate the full

text blocks(Fig. 3(b)). Lastly, we apply Laplacian filter with 3×3 kernel matrix (Fig. 3(c)), because it is a powerful technique to deduct edges in all directions and it is also effective to solve salt and pepper noise. Fig. 3(d), shows an image after applying Laplacian filter. It represent the final result of preprocessing, it is a 512×512 edge text block image with 96 DPI.



1	1	1
1	-8	1
1	1	1

(a) (b) (c) (d)

Fig. 3. (a) The original Image, (b) the image results after generating texture blocks, (c) Laplacian Filter value and (d) filtered image by Laplacian

3.2 Features Extraction

We apply an eight neighbouring kernel matrix and associate each pixel according to their two neighbouring pixels. We present a relationship between the Scoped pixel, *S(x, y)* and their neighboring pixels as depicted in Fig. 4(a). We use and transform an encompassing eight pixels into position values as Fig. 4(b). Based on the previous illustration, we introduced our method based on two perspectives: (1) Find the first order relationship, and (2) Find the second order relationship.

(x-1, y-1)	(x, y-1)	(x+1, y-1)
(x-1, y)	*S(x, y)*	(x+1, y)
(x-1, y+1)	(x, y+1)	(x+1, y+1)

(a)

	1	**2**	**3**
1	135°	90°	45°
2	180°	Scoped pixel	0°
3	225°	270°	315°

(b)

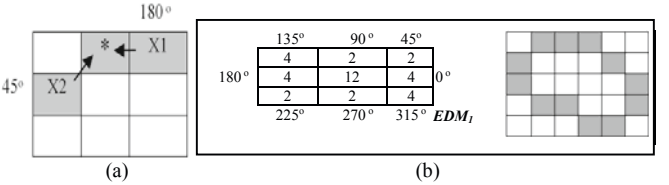
Fig. 4. (a)The eight neighboring pixels, (b) the direction angles of the neighboring pixels, also it represents the edge direction matrix (*EDMS*) and their cells properties

In the first order relationship, we firstly create an 3×3 edge direction matrix (EDM_1) as Fig. 5(b). Each cell in EDM_1 contains a position within 0 until 315 degree value. Secondly, we determine the relationship of the scoped pixel in the edge image $Iedge(x,y)$ by calculating the number of occurrence for each value in EDM_1. The algorithm is as follows:

For each pixel in *Iedge* (x,y)
  If *Iedge (x,y)* is black pixel at center **then**
    Increase number of occurrence at $EDM_1(2,2)$ by 1.
    If *Iedge (x+1,y)* is black pixel at 0° **then**
    Increase number of occurrence at $EDM_1(2,3)$ by 1.

If *Iedge (x+1,y−1)* is black pixel at 45° **then**
Increase number of occurrence at $EDM_1(1,3)$ by 1.
If *Iedge (x,y−1)* is black pixel at 90° **then**
Increase number of occurrence at $EDM_1(1,2)$ by 1.
If *Iedge (x−1,y−1)* is black pixel at 135°**then**
Increase number of occurrence at $EDM_1(1,1)$ by 1.

In the first order relationship, each pixel in the edge image relates with two pixels. For example, as Fig. 5(a) the scoped pixel presents 180° for X1 and 45° for X2. That means each pixel presents a two relationships in (EDM_1).



Order	Angle	Value
1	0°	4
2	180°	4
3	135°	4
4	315°	4
5	45°	2
6	225°	2
7	90°	2
8	270°	2

(c)

Fig. 5. (a) The two neighboring pixels, (b) the edge image and its EDM_1, (c) the order of the angle's importance

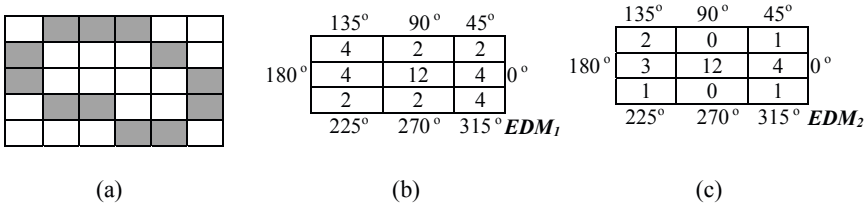
In the second order relationship, each pixel will present by one relationship only. We firstly create an 3×3 edge direction matrix (EDM_2). Secondly, we determine the relationship importance for *Iedge(x, y)* by sorting the values in EDM_1 descendingly as shown in Fig. 5(b) and (c) respectively. We take the most importance relationship of the scoped pixel in *Iedge(x, y)* by calculating the number of occurrence for each value in EDM_2. The relationship orders must follow as below:-

(i) If there are more than one angle which have the same number of occurrence, then the smaller angle, is selected firstly.
(ii) Next, the reversal angle is selected subsequently.

The algorithm of the second order of EDM_2 relationship is as follows:

Step 1: Sort descendingly the relationships in $EDM_1(x, y)$.
Step 2: For each pixel in *Iedge(x, y)*,
Step 3: If *Iedge(x, y)* is a black pixel then
Step 4: Find the available relationships between two neigbouring pixels,
Step 5: Compare the relationship values between two available relationships,
Step 6: Increase number of occurrence at the related cell in $EDM_2(x, y)$.

The results of the first order relationship and the second order relationships presented in EDM_1 and EDM_2 as shown in Fig.6.



(a) (b) (c)

Fig. 6. (a) is a sample image, (b) is its EDM_1 and (c) is its EDM_2 values results

Lastly, we proposed several features from the EDM_1 and EDM_2 values. We summarize 18 features by calculating their homogeneity, pixel regularity, weights, edge direction and edge regularity as the followings:

- **Edges Direction:** This feature represents the main direction of the model, It is calculated by finding the largest value in EDM_1 as follows:

$$Edges\ Direction = \mathrm{Max}(\mathrm{EDM}_1(x, y)) \tag{2}$$

- **Homogeneity:** This feature represent the percentages of each direction to all available directions in the edge image as follows:

$$Homogeneity(\theta) = \mathrm{EDM}_1(x, y)/(\sum \mathrm{EDM}_1(x, y)) \tag{3}$$

- **Weight:** This feature represents the density of the model in the image. It is calculated based on the percentage of edges on the size of the model as follows:

$$Weight = \mathrm{EDM}_1(2,2) / \sum(\mathrm{Iedge}(x, y) = \mathrm{Black}) \tag{4}$$

- **Pixel Regularity:** This feature represent each direction in EDM_1 to the number of scoped pixels in the edge image as follows:

$$Pixel\ Regularity(\theta) = \mathrm{EDM}_1(x, y) / \mathrm{EDM}_1(2,2) \tag{5}$$

- **Edges Regularity:** This feature represent each real direction in EDM_2 to the number of scoped pixels in the edge image as follows:

$$Edges\ Regularity(\theta_*) = \mathrm{EDM}_2(x, y) / \mathrm{EDM}_2(2,2) \tag{6}$$

where θ represents 0°, 45°, 90° and 135°, θ_* presents 0°, 45°, 90°, 135°,180°, 225°, 270° and 135°, and *(x, y)* presents the relative position in EDM_1 and EDM_2.

4 Experiments and Results

The dataset have collected from many resources such as documents, artistic works and calligraphy software products. We have collected 700 samples that consist of 100 samples from each type of Kufi, Diwani, Persian, Roqaa, Thuluth and Naskh. We compare our method with GLCM[7]. In this work, the measures which have been implemented are Contrast, Homogeneity, Angular Second Moment (ASM), Energy, Entropy, Variance and Correlation with 0°, 45°, 90° and 135° angles, that derived to 28 features. Also, we applied Multilayer Network, Bayes network and Decision Tree

classifiers to compare the proposed method performance within different classifiers. The dataset was split into training and testing datasets in percentage between 60% to 70%. Based on our experimental results, the proposed method gives higher performance than the GLCM in all experiments. The experiments performance of each classifier with the different training dataset were convergent, so we chosen the 60% training dataset. Based on the results in Table 1 and Fig. 7, we can note that the proposed method gains the highest performance about 97.85% with the decision tree whereas GLCM method obtains 87.143% with the Multilayer Network.

Table 1. The classification results with 60% training

	Proposed method	**GLCM[7]**
Bayes network	92.473%	77.857%
Multilayer Network	95.341%	87.143%
Decision Tree	97.85%	85.714%

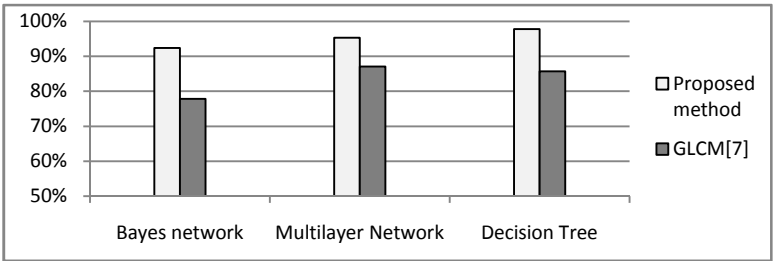


Fig. 7. The classification results of the *GLCM* and proposed method with 60% training

We continued the experiment by analyzing the consistency of the results of the decision tree classifier using the proposed method and the GLCM with a 60% training dataset. We repeated this experiment five times and the results are shown in Table 2.

Table 2. The results of five experiments using a decision tree classifier with 60% training

	Exp # 1	**Exp # 2**	**Exp # 3**	**Exp # 4**	**Exp # 5**
Proposed	97.85%	97.133%	95.699%	94.624%	96.416%
GLCM [7]	85.714%	79.643%	85.714%	83.929%	84.286%

Based on the experiments values of the Table 2, a statistical descriptive provided in Table 3. The mean of values of the proposed method is 96.3444%, which is higher than the GLCM, which is 83.8572%. The proposed method obtains a standard deviation of 1.25201, which is lower than the GLCM, which is 2.49218. That meant the results of the experiments of the proposed methods is most fixed than the results of GLCM. In relation to the above, the proposed method also produces a smaller standard error value of about 0.55992 compared to GLCM, which is 1.11454.

Table 3. The descriptive statistics for the results of the five experiments for the decision tree classifier with 60% training

	Mean	Standard Deviation	StandardError
Proposed	96.3444%	1.25201	0.55992
GLCM [7]	83.8572%	2.49218	1.11454

Based on the results in Table 4, the correction rates for each calligraphy type shows with decision tree and 60% training dataset. The highest accuracy is achieved on Persian and Thuluth. These calligraphy types achieved by 100% accuracy rate. While in the same case, the lowest accuracy is exhibited with the Diwani about 95.5%. In GLCM, the highest accuracy achieved on the Kufi about 95.2%. The lowest accuracy rate are the Persian about 76.2%. In conclusion, the proposed method has given accuracy rate for each calligraphy type higher than the GLCM. Fig. 8 summarizes the results of proposed method and GLCM respectively.

Table 4. Confusion Matrix results of the proposed approach by using Decision Tree Classifier, 60% training dataset

	Proposed	GLCM [7]
Andalusi	97.4%	92.1%
Kufi	97.9%	95.2%
Diwani	95.5%	86.5%
Naskh	97.5%	78.9%
Roqaa	97%	91.9%
Persian	100%	76.2%
Thuluth	100%	80.4%

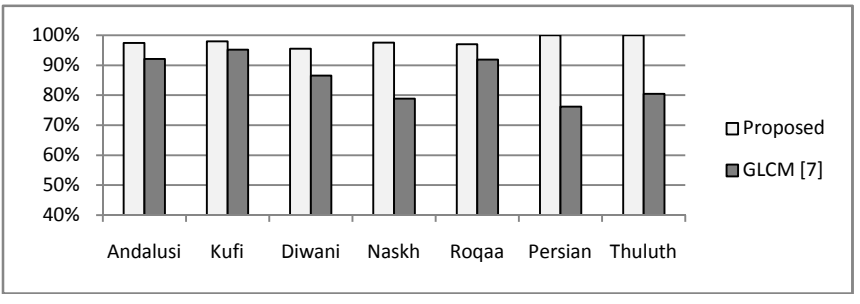


Fig. 8. The correction rate for each class by proposed technique and GLCM features and decision tree, with 60% training dataset

5 Conclusion

In this paper, we propose a global feature extraction method for the binary document images. This method is based on statistical analysis of the relationships between the pixels of an edge image that contains black and white levels as an example. We

applied the proposed method on optical font recognition application to identify the Arabic calligraphy font type. In general, firstly we transform an input image containing texture blocks into closed edge or contour image. Then, we find the values of EDM_1 and EDM_2. In line with that, we obtained 18 features of Arabic calligraphic font type. The proposed method is compared with GLCM on different classifiers. The proposed method outperformed the GLCM method on three chosen classifiers. The highest performance of the proposed method was 97.85% using decision tree on a 60% training dataset. Our method requires simple preprocessing before execution. As conclusion, our proposed global feature extraction method is simple and able to generalize on other image datasets.

Acknowledgments. This research is based on two fundamental research grants from Ministry of Science, Technology and Innovation, Malaysia entitled "Logo and Text Detection for moving object using vision guided" UKM-GGPM-ICT-119-2010 and "Determining adaptive threshold for image segmentation" UKM-TT-03-FRGS0129-2010. We also would like to thank previous the CAIT researchers such as Prof. Dr. Jon Timmis, University of York, UK and Assoc. Prof. Dr. Azuraliza Abu Bakar.

References

1. Zramdini, A., Ingold, R.: Optical Font Recognition Using Typographical Features. IEEE Transactions On Pattern Analysis And Machine Intelligence 20(8), 877–882 (1998)
2. Zhu, Y., Tan, T., Wang, Y.: Font Recognition Based On Global Texture Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(10), 1192–1200 (2001)
3. Sun, H.-M.: Multi-Linguistic Optical Font Recognition Using Stroke Templates. In: The 18th International Conference on Pattern Recognition, Hong Kong, pp. 889–892 (2006)
4. Ding, X., Chen, L., Wu, T.: Character Independent Font Recognition on a Single Chinese Character. IEEE Transactions on Pattern Analysis And Machine Intelligence 29(2), 195–204 (2007)
5. Joshi, G., Garg, S., Sivaswamy, J.: A generalized framework for script identification. International Journal on Document Analysis and Recognition 10(2), 55–68 (2007); ISSN:1433-2833
6. Tuceryan, M., Jain, A.K.: Texture Analysis. In: Chen, C.H., Pau, L.F., Wang, P.S.P. (eds.) The Handbook of Pattern Recognition and Computer Vision, 2nd edn., ch. 2.1, pp. 207–248. World Scientific Publishing Co., Singapore (1998)
7. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural Features for Image Classification. IEEE Trans. Systems, Man and Cybernetics 3(6), 610–621 (1974)
8. Petrou, M., García Sevilla, P.: Image Processing, Dealing with Texture. John Wiley & Sons, Ltd., Chichester (2006)
9. Bian, N.: Evaluation of Texture Features For Analysis of Ovarian Follicular Development. Master Thesis, Department of Computer Science. University of Saskatchewan, Saskatoon, Canada (2005)
10. Conners, R.W., Harlow, C.A.: A Theoretical Comparison of Texture Algorithms. IEEE Transactions on Pattern Analysis And Machine Intelligence PAMI-2(3), 204–222 (1980)
11. Busch, A., Boles, W., Sridharan, S.: Texture for Script Identification. IEEE Transactions on Pattern Analysis And Machine Intelligence 27(11), 1720–1732 (2005)

12. Peake, G., Tan, T.: Script and Language Identification from Document Images. In: Proc.Workshop Document Image Analysis, San Juan, Puerto Rico, vol. 1, pp. 10–17 (1997)
13. Yazdi, M., Yazdi, M., Gheysari, K.: A New Approach for the Fingerprint Classification Based on Gray-Level Co-Occurrence Matrix. Proceedings of World Academy of Science, Engineering and Technology 30 (July 2008)
14. Quan, Y., Jinye, P., Yulong, L.: Chinese Sign Language Recognition Based on Gray-Level Co-Occurrence Matrix and Other Multi-features Fusion. In: 4th IEEE Conference Industrial Electronics and Applications, ICIEA 2009, Xi'an, pp. 1569–1572 (2009)
15. Bataineh, B., Abdullah, S.N.H.S., Omer, K.: Generating an Arabic Calligraphy Text Blocks for Global Texture Analysis. In: International Conference on Advanced Science, Engineering and Information Technology (ICASEIT 2011), Kuala Lumpur, Malaysia (January 2011)
16. Singh, C., Bhatia, N., Kaur, A.: Hough transform based fast skew detection and accurate skew correction methods. Pattern Recognition 41(3), 3528–3546 (2008)

Domain N-Gram Construction and Its Application to Text Editor

Myunggwon Hwang, Dongjin Choi, Hyogap Lee, and Pankoo Kim

814 IT Building, 375 Seoseok-dong, Dong-gu,
501-759, Gwangju, South Korea
{mg.hwang,dongjin.choi84,leoscientist}@gmail.com,
pkkim@chosun.ac.kr

Abstract. Google has published n-gram data which was constructed from huge document set gathered until 2005. However, it is hard to use the data in real world applications due to its huge volume. In this paper, we propose a method to construct domain n-gram data in which a specific domain group is interested and apply the data to text editor for practical efficiency in evaluation. It contains diverse test results according to typing speed level of people and comparison results with other works. The result of this research is conducted through applying to typing only however it has big importance in a point of being capable of expecting its effectiveness because the n-gram data is widely applicable to many fields.

Keywords: n-gram, domain n-gram, Google n-gram, text editor, mobile typing.

1 Introduction

In recent, much of research has been studied in information processing area to grasp context (or semantics) user inputs. To grasp user context is mainly divided into two methods: the first is to understand meaning of word or document which was already completed by a user [1]. It is generally understood based on knowledge base (lexical dictionary such as WordNet). The second is to previously expect a content user may input [2]. It is based on probability calculated through analyzing huge document set. This paper deals with the second method.

Using n-gram data is representative method to expect user content. N-gram is made by analyzing existing huge document set (especially web document) and it allocates value between words (or characters). The value could be called degree of word cohesion and has potentiality to be applicable to various NLP (natural language processing). The n-gram is mainly used for text editor [3], query recommendation for search engine in recent [4], conversational speech recognition [5], and pupil motion tracking based typing [6, 7]. N-gram data for text editor could make an increase of typing speed and reduce of typing wrong word or missing character. Especially it is able to provide big convenience and effectiveness for mobile equipment which has a restricted input device rather than personal computer which has independent keyboard. Among the data, Google n-gram is the most representative data released in

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 268–277, 2011.

recent. However the volume is too huge to be utilized in actual application level due to memory overflow occurrence and storage limit. In the previous work, it has shown a performance of Google trigram [8]. It utilized the trigram data filtered by threshold value however it was still huge for 1.2GB in case of text file. This big data can be possible to be used in a desktop computer but not in mobile device due to its storage capacity. Therefore, in this paper, we propose a method constructing domain n-gram data to reduce its size and to bring work efficiency. In the performance evaluation, we analyze its typing efficiency through a few analyses.

This paper is organized as follows. Explanation of Google n-gram and previous work are described in Section 2. Section 3 explains a method constructing domain n-gram data. Experiments and evaluations for the result are illustrated in the Section 4 and the conclusion and future prospects of the research are mentioned in Section 5.

2 Google N-Gram and Previous Work

The Google n-gram is provided by LDC (Linguistic Data Consortium). The data contains about 1 trillion tokens occurred 40 times or more in web documents which had been gathered by January 2006. Table 1 shows statistics of the n-gram.

Table 1. Statics of Google n-gram

n-gram	Number of tokens
Unigrams	13,588,391
Bigrams	314,843,401
Trigrams	977,069,902
4grams	1,313,818,354
5grams	1,176,470,663

Google is the most representative search engine over the world so that its n-gram is sufficient as fundamental data to be used in various fields. While using n-gram is based on probabilistic method, it can be said that the data implicitly includes semantics because the data is constructed by collective intelligence of the people. Table 2 shows a part of Google 4-gram and 5-gram.

Table 2. Examples of Google 4- and 5gram

4gram		5gram	
Tokens	Occurrence	Tokens	Occurrence
changing the quantity and	522	both SD Memory Card and	128
changing the quantity for	305	both SD and CF cards	85
changing the quantity in	195	both SD and CF expansion	404
changing the quantity of	588	both SD and CF slots	79
changing the quantity please	183	both SD and Compact Flash	57
changing the quantity supplied	70	both SD and HD content	100
changing the quantity to	319	both SD and HD formats	171
…	…	…	…

We know the data is very useful but we should design a method because the volume of the data attains about 86GB (unzipped actual text file size). Therefore, in research [8], a text editor has been implemented based on only small part of data filtered from Google trigram by a few conditions. The small trigram was created through removing tokens containing none English word(s) and choosing tokens occurred 450 times or more. It was just about 1.2GB (5.65%) of the original trigram. The data was applied for checking time efficiency according to typing speed level. It showed the efficiency for the people who type text up to 600 letters per minute but it is not good for human typing more than 600 because the system wastes time for searching next word. From the result, the research has suggested that the data can give efficiency to devices like desktop computers which have independent input unit but be inefficient for mobile devices.

Therefore, in this paper, a method is proposed to construct domain n-gram data in which a specific user group is interested. And in the evaluation, we test how much efficiency the domain n-gram gives to human work.

3 Construction of Domain N-Gram

The total procedure to construct specific domain n-gram using Google data is shown in the figure 1. In this research, 'Computer' is selected as the specific domain. The method to extract domain terms from documents and the method to construct domain n-gram will be described in detail in this section.

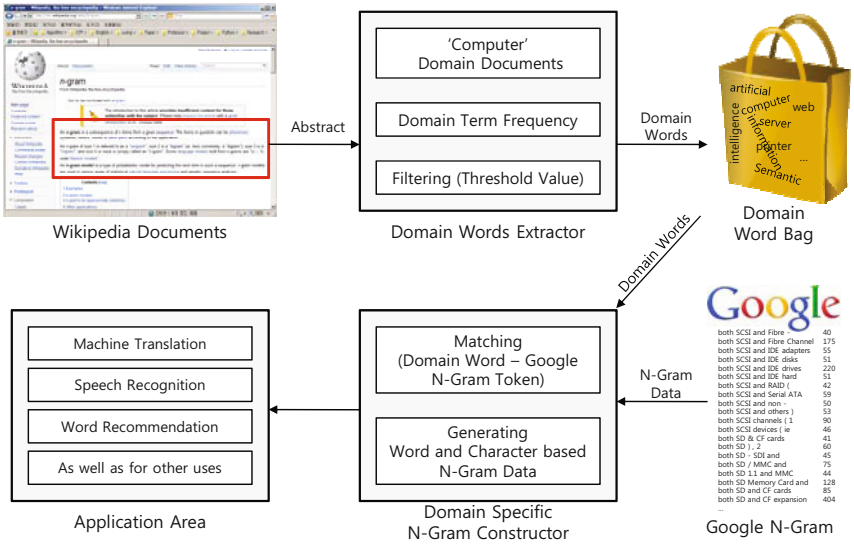


Fig. 1. System Architecture for domain n-gram construction

3.1 Preparation of Domain Word Bag

Domain Specific Wikipedia Documents: Wikipedia document sets will be used to grasp which words can be considered as candidate terms of domain(Domain Word Bag) from the selected domain area('Computer' field is selected in this research). Wikipedia currently contains documents more than 3.35 million and it is still constantly increasing due to the creation of new knowledge such as new technologies, products and trends. Moreover, Wikipedia provides specialized knowledge information retrieval different with existing search engine because Wikipedia contains unlimited diverse information such as history, events, philosophy, culture, arts, sciences, systems, people, technology, animals and so on. The each document in Wikipedia contains single title (subject) and contents deeply related with the title in various forms such as abstract (short and extended), information box, figures, main contents and etc. Especially, the extended abstract describes the core point of the subject so if we use the extended abstract as the corpus, it is possible to reduce the process of noise removal [10]. For this reason, the abstract of Wikipedia will be used to extract specific domain words. The abstract corpus of Wikipedia is an open data resource provided by DBpedia[1] and regularly updated. We only focus on the Wikipedia abstract that contains domain core word 'computer' in DBpedia version 3.4 and following table is briefly describes its statistics.

Table 3. Statistics of computer-domain document

Number of total documents	Number of documents included in computer-domain (%)
2,944,417	36,566 (1.24%)

Extracting domain words: The domain word means a token that appears in the specific domain regardless of types of POS (part-of-speech) or stemming also. The basic criterion to decide which word is suitable for a specific domain is *tf* (term frequency). However, if we only consider *tf* value, the other unexpected word occurred evenly in the all documents will be selected as a domain word. For example, words 'the,' 'a,' and 'to' can appear in any documents. To extract a term appeared in only specific domain, [11] suggested a method though there is a burden to compute approximately 2.9 million documents. To overcome this limitation, the *dw* (domain weight) will be used as a criterion to extract the domain words. The *dw* means a value considering total term frequency regardless of domain and term frequency of specific domain. The value can be considered as the close words to domain if tf_t is high and tf_{ds} is small. The formula to obtain this *dw* is as follows.

$$dw = \frac{tf_{ds}}{tf_t} \tag{1}$$

The maximum and minimum values of the domain weight are 1 and 0 respectively. The following table indicates top fifteen ranked words extracted by using tf_{ds} and *dw* respectively.

[1] http://www.dbpedia.org

Table 4. Top fifteen words extracted by using tf_{ds} and dw respectively

Index	term frequency		Domain weight			
	Word	tf_{ds}	Word	tf_{ds}	tf_t	dw
1	the	429204	computer	44006	44006	1.0000
2	of	231357	computers	8782	8782	1.0000
3	and	216721	computerized	602	406	1.0000
4	a	179436	computer's	448	448	1.0000
5	in	169333	supercomputer	406	406	1.0000
6	to	137901	computer-aided	325	325	1.0000
7	is	107715	computer-based	313	313	1.0000
8	for	73801	microcomputer	305	305	1.0000
9	as	60079	computer-generated	305	305	1.0000
10	was	60025	supercomputers	218	218	1.0000
11	bu	53409	human-computer	193	193	1.0000
12	on	49266	computer-animated	185	185	1.0000
13	with	45152	microcomputers	170	170	1.0000
14	computer	44006	computer-controlled	143	143	1.0000
15	it	41198	minicomputer	142	142	1.0000

As shown in table 4, the result based on domain weight is quite reliable than that based on tf_{ds}. However, there is a drawback when tf_{ds} and tf_t is equal to 1, dw is also equal to 1. This problem is happened because of misspell, web tag, and so on in Wikipedia. This drawback indicates that it is necessary to have threshold value to avoid this situation. Therefore, we make a few conditions for filtering. We only focus on English in this research, so that tokens including numbers, special characters and none English language will be removed. It is need to provide flexibility to satisfy diversity of word recommendation so we make an un-strict threshold value. There is no limit to tf_{ds} but tf_t and dw should be higher than 3 and 0.1 respectively. The reason why tf_t should be higher than 3 is that we can reduce the cases of misspell and dw value make terms dependent to the domain. The following table describes top fifteen ranked word in order of tf_{ds}, tf_t and dw applying threshold value.

Table 5. Domain words extracted through filtering

index	words	tf_{ds}	tf_t	dw
1	computer	44006	44006	1
2	game	16533	136789	0.120865
3	system	15272	129797	0.1176607
4	software	13022	38507	0.3381723
5	systems	11153	50145	0.222415
6	science	10353	73296	0.1412492
7	data	9521	40689	0.2339944
8	computers	8782	8782	1
9	technology	8567	49005	0.1748189
10	information	8160	61197	0.1333399
11	program	7438	73058	0.1018095
12	engineering	6277	32463	0.1933586
13	video	5919	58174	0.1017465
14	digital	4839	26816	0.180452
15	programming	4253	20345	0.209044

After applying this procedure 35,040 domain words are obtained. These domain words will be utilized to extract 'computer' domain n-gram from Google data in next procedure.

3.2 Domain N-Gram Construction and Recommendation Method

It is quite simple to extract n-gram appropriate to computer domain from Google data due to using word based matching. The n-gram tokens contain top six ranked words in table 5 can be extracted as table 6. Also, at the bottom line in the table, the statistics of each extracted n-gram is described with comparison to each Google n-gram.

Table 6. Domain N-Gram data extracted from Google

index	Domain Words (Unigram)	Domain N-Gram data extracted from Google			
		Bigrams	Trigrams	4-grams	5-grams
1	computer	**computer** access	**computer** aided engineering	candidate in **computer** science	Pentium desktop **computer** complete system
2	game	organized **game**	driving adventure **game**	cameras **game** devices pad	perfect your golf **game** at
3	system	operating **system**	reset your **system**	annual home heating **system**	Java installed on your **system**
4	software	solving **software**	resume **software** employment	another blogging **software** based	Java library programming serialization **software**
5	systems	operating **systems**	reset control **system**	another **systems** engineer jobs	jet inks and printing **systems**
6	science	search **science**	resource **science** center	anthropology is the **science**	job market for computer **science**
Total count	35,040	4,586,932	17,201,709	20,003,232	17,726,070
size rate	0.25(%)	1.45(%)	1.76(%)	1.52(%)	1.51(%)

Some rules are necessary to recommend precise words based on user input context for using domain n-gram data described in table 6. First of all, a system has to determine which kinds of n-grams will be used for recommendation. Following recommendation rules are able to cover this matter. For easy understanding of the work, we use a case of typing 'computer science college of engineering.'

Applying unigram information: if user input is the first word of a sentence, or there is no recommendable word in case of applying bigram information,

- Recommend words after receiving two kinds of characters at first.
  Ex 1) user inputs 'co.' The application recommends words such as 'code,' 'computer,' 'copy,' 'courses,' 'communication,' and so on.
  Ex 2) user selects a word from recommended words in example 1 or inputs 'm' character additionally for more detail recommendation (in case of the latter, 'computer,' 'computers,' 'communication,' 'components,' 'computing,' and so on are recommended).

Applying bigram information: if user completed typing a first word or there is no recommendable word in case of applying trigram information,

- words are recommended based on the first word in order of high probability.
  Ex 3) when inputting 'computer', 'science,' 'peripherals,' 'associates,' 'systems,' 'software,' 'technology,' 'hardware,' and so on are provided.
- user selects a word from recommended words in example 3 or user inputs new character(s) for another recommendation.
  Ex 4) if user additionally inputs 's', 'science,' 'systems,' 'software,' 'system,' 'services,' 'skills,' 'security,' and so on are recommended.

Applying trigram information: if user inputs two words or there is no recommendable word in case of applying 4-gram information,

- Words in order of high probability are recommended based on two words.
  Ex 5) user inputs 'computer science' and 'and,' 'department,' 'at,' 'from,' 'memory,' 'or,' 'university,' and so on are recommended.
- user selects a word from the words in example 5 or user inputs character(s).
  Ex 6) in case of input 'c', the application recommends words such as 'colleges,' 'criminal,' 'course,' 'center,' and so on.
  Ex 7) user selects a word in example 6 or if user inputs 'o' character, 'colleges,' 'counselor,' 'conservation,' 'conference,' and so on are recommended.
  Ex 8) user selects a word in example 7 or if user inputs 'l' character, 'colleges,' 'college,' 'colloquium,' and so on are provided.

Applying 4-gram information: if user inputs three words or there is no recommendable word in case of applying 5-gram information,

- words in order of high probability are recommended based on three words.
  Ex 9) user inputs 'computer science college' and 'best,' 'lane,' 'ranking,' 'station,' 'students,' 'a,' 'and,' 'of,' and so on are given.
- user selects a word in example 9 or inputs another character(s).
  Ex 10) user inputs 'o', 'of,' 'on,' and so on are provided.

Applying 5-gram information: if user inputs four words,

- words are recommended based on four words in order of high probability.
  Ex 11) user inputs 'computer science college of' then 'engineering,' 'arts,' 'William,' 'liberal,' 'technology,' and so on are recommended.
- user selects a word in example 11 or user inputs another character(s).
  Ex 12) user inputs 'computer science college of e', the recommended words are 'engineering,' 'education,' and so on.

It is possible to grasp user contexts and expect next following words using domain n-gram through applying above rules. An application and evaluations will be described in section 4.

4 Experiment and Evaluation

This paper has described a method constructing domain n-gram data. Big data can have much information but systems should waste much time to search it. We implemented a program which has a function to check elapsed time for typing. It has two modes; one

is based on n-gram data; the other is based on manual typing. The application has another function to automatically type given text. And the typing speed can be set so that the application measures the time according to typing speed. The domain n-gram was designed especially for mobile device and the authors think the user can type text at most 300 letters. Therefore we made, for the test, speed 60, 120, 180, 240, 300, and 540 letters per minute individually. We collected 20 documents from Wikipedia under conditions of containing 'computer' and 'algorithm' together and of 300 letters or more. The application automatically types the text based on manual[2], the Google trigram [8], domain n-gram (altogether), and domain each-gram according to the speed. Table 7 shows the evaluation result of each elapsed time. In the table, bold typed values mean time reduction in comparison to the manual.

Table 7. Comparison result bed on typing speed and n-gram data type

Typing Method		Typing Speed (keys/min)					
		60	120	180	240	300	540
Manual		23,555	11,696	7,742	5,845	4,655	2,587
Google Trigram		**16,608**	**8,833**	**6,275**	**5,137**	**4,499**	4,083
Domain each-gram	Unigram	**17,700**	**8,848**	**5,855**	**4,432**	**3,530**	**1,966**
	Bigram	**17,824**	**9,117**	**6,203**	**4,795**	**3,948**	**2,526**
	Trigram	**18,436**	**9,483**	**6,417**	**5,072**	**4,256**	3,081
	4gram	**19,139**	**10,221**	**7,209**	5,988	5,070	4,233
	5gram	**19,550**	**10,222**	**7,209**	5,989	5,071	4,489
Domain n-gram (altogether)		**18,660**	**10,217**	**7,206**	5,987	5,070	4,232

Google trigram shows the best performance under typing speed 300 in case of considering only time saving. And, domain unigram, bigram, and trigram could also give time efficiency similar to the trigram. If user can type over speed 180, the domain unigram provides the best efficiency and domain bigram, Google trigram, and domain trigram follow after in order. From the speed 240 or more, it is confirmed that three domain n-grams surpass the Google trigram. In cases of using 4-, 5-, and domain n-gram, the application wastes much of time in retrieving so user will not want to use it. From the time evaluation, we could compare efficiency of each data. However, in the mobile environment, the time is not the most important factor for the performance of data because storage is limited. Therefore, in addition to the time evaluation, we analyzed count of recommended words and count of reduced typing by each data. Table 8 shows the result.

As shown in table 8, the Google trigram recommends the most words. Comparing to this, domain each gram and domain n-gram could not follow the recommendation performance. Especially, domain 5gram is the worst case because it does not have many cases that match to 5 words all. However, the count of recommendation is not representative for mobile efficiency. The average length of recommendation words shows that the Google trigram is the worst length. This means the Google trigram

[2] Manual mode means that the application automatically types given text according to an assigned speed.

Table 8. Counts of recommended words and reduced typing

Typing Method		Count of Rec. words	Typed count	Reduced count	Avg. length of Rec. words (reduced count)/(Rec. count)
Manual		0	23229	0	0
Google Trigram		1429	17926	5303	3.711
Domain each-gram	Unigram	418	20794	2435	5.825
	Bigram	541	20425	2804	5.183
	Trigram	410	21266	1963	4.788
	4gram	215	22220	1009	4.693
	5gram	98	22763	466	4.755
Domain n-gram		730	19768	3461	4.741

recommends general words. In here, the general words mean that people type the words frequently so that people are good at typing the words. In the other hand, it is confirmed that the domain data recommends technical terms involved in specific domain. People should make concentration to type the technical words because it occurs miss-typing often. Moreover, the data size of the Google trigram is about 3.2 times bigger than that of domain trigram but the count of reduced typing merely attains 2.18 times. It means domain trigram has about 146.8% higher efficiency than Google trigram in considering size and reduced typing together.

From the experimental results of elapsed time and typing efficiency, it could give clues for that domain trigram is similar to the Google trigram or is more efficient even though the data is small.

5 Conclusion

This paper has pointed out the limitation of Google n-gram to be used in real world application due to its hugeness and suggested a method constructing domain n-gram which is related to a specific subject. To construct domain n-gram, the work contained a selection method of domain words and a construction method of domain n-gram data containing the domain words. In the evaluation, a few text editors based on manual typing, Google trigram, domain each-gram, and Domain n-gram have been examined respectively. And the results have showed strengths of domain n-gram through analyses in aspects of time efficiency, count of reduced typing, count of recommended typing, and capacity efficiency. Especially, domain unigram, bigram, and trigram can be expected to show high performance however it is not the case for 4- and 5-grams because it could not cover tokens which people would input and it rather wasted much time for searching.

In recent, mobile devices become very famous due to occurrence of smart phone. This work can give much more efficient to them to improve the performance. In this paper, it only contains domain n-gram and each-gram based text editor. In the future, we will confirm performances in cases of mixed two grams or three grams. And we will consider a method expecting next words with lexical semantics.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2009-0064749).

References

1. Budanitsky, A., Hirst, G.: Evaluating WordNet-based Measures of lexical semantic relatedness. Computational Linguistics 32(1), 13–47 (2006)
2. Zukerman, I., Albrecht, D.W.: Predictive Statistical Models for User Modeling. User Modeling and User-Adapted Interaction 11, 5–18 (2004)
3. Cavnar, W.B., Trenkle, J.M.: N-Gram-Based Text Categorization. In: Proceedings of SDAIR 1994, 3rd Annual Symposium on Document Analysis and Information Retrieval, pp. 161–175 (1994)
4. Baeza-Yates, R., Hurtado, C., Mendoza, M.: Query recommendation using query logs in search engines. In: Lindner, W., Fischer, F., Türker, C., Tzitzikas, Y., Vakali, A.I. (eds.) EDBT 2004. LNCS, vol. 3268, pp. 395–397. Springer, Heidelberg (2004)
5. Khudanpur, S., Wu, J.: A Mazimum Entropy Language Model Integrating N-Grams and Topic Dependencies for Conversational Speech Recognition. In: Proceedings of ICASSP 1999, pp. 553–556 (1999)
6. Soon-Beak, K., Soo-Heum, L.: The Pupil Motion Tracking Based on Active Shape Model Using Feature Weight Vector. In: Proceedings of the Korea Institute of Signal Processing and Systems Conference, pp. 205–208 (November 2005)
7. Morimoto, C.H., Koons, D., Amir, A., Flickner, M.: Frame-Rate Pupil Detector and Gaze Tracker. In: ICCV 1999 FRAME-RATE workshop (September 1999)
8. Hwang, M.G., Choi, D.J., Lee, H.G., Kim, P.K.: Text Editor based on Google Trigram and its Usability. In: Proceedings of the UKSim 4th European Modeling Symposium on Computer Modeling and Simulation, pp. 12–15 (2010)
9. Brants, T., Franz, A.: Web 1T 5-gram Corpus Version 1.1 (LDC2006T13) (April 2006)
10. Choi, D.J., Hwang, M.G., Kim, P.K.: Semantic Context Extraction of Wikipedia. In: The Proceedings of the 2010 International Conference on Semantic Web and Web Services (SWWS 2010), pp. 38–41 (2010)
11. Velardi, P., Navigli, R., D'Amadio, P.: Mining the Web to Create Specialized Glossaries. IEEE Intelligent Systems 23(5) (2008)

Grounding Two Notions of Uncertainty in Modal Conditional Statements*

Grzegorz Skorupa and Radosław Katarzyniak

Wroclaw University of Technology
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
{grzegorz.skorupa,radoslaw.katarzyniak}@pwr.wroc.pl

Abstract. We outline a method of grounding natural language conditional statements with modal operators of belief and possibility. Our work extends previous works on grounding natural language statements within cognitive agent. We focus on defining criteria on conditional statement's suitability to agents knowledge. We argue that there are two types of modality based on the positioning of modal operator and that it is often impossible to distinguish between them. We propose formal model of agent's mental state and criteria of choosing 'correct' statements that can be checked against that model. The work sets up a theoretical basis on further analytic and experimental analysis.

1 Introduction

A cooperation of independent intelligent entities is one of key assumptions of multi-agent systems. This cooperation can be modelled between agents themselves or humans and agents. The need for cooperation forces new challenges on agent-human communication. Natural language communication between computer and human is a long searched holy grail of artificial intelligence studies. We analyse how natural language statements are used. Our work can be applied to speech act theory [3] based protocols of communication.

In this work we are discussing on how natural language statements usage can be modelled within autonomous cognitive agent[8]. We are considering an agent, that should summarise observed situation with a chosen natural language statement. We focus on conditional statements that can be said with some dose of doubt. The uncertainty is expressed by possibility and belief modal operators.

Our key problem is to model human natural language communication skills and ie. human ability to pick a statement that is suitable to situation based on his knowledge. We define a formal criteria on when a statement can and can't be used according to agent's mental state of the situation. Similar problems have been widely analysed by philosophers and logicians [4]. The novelty of the approach lies in analysis of conditionals with modal operators and focus on a statement usage, not its logical truth value.

* This paper was supported by Polish Ministry of Science and Higher Education under grant no. N N519 407437.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 278–287, 2011.

The problem described here has its roots in previous work on grounding modal communication language within an multi-agent system [10], where formal criteria have been defined for simple modal statements and modal statements with conjunction and alternative. We are now focusing on various kinds of conditional statements with modal operators. Some progress has already been done in [12,13,14], where we analysed conditional statements with no modal operators and with modal operators applied to a consequent. Within this work we focus on conditional statements with modal operators applied to the whole statement. Such exemplary statement is: "It is possible, that if the sun is shining, there are no clouds".

In further sections we shall describe how do we understand a proper usage of a statement and conditional statement in particular. We shall argue that modal operators must express two types of uncertainty for statements considered here. We will formally model agents' mental state and define constraints that filter out improper natural language statements. These constraints can be checked directly on agents' mental state.

2 Statement Proper Usage

Statement is properly used when it fits the situation well. Statement suitability depends on many factors, some of which are: speaker's background knowledge, speaker's attitude, conversational and situational context, speaker's knowledge about the listener or current aims and tasks. Choosing the best statement requires checking it against many criteria. It is not enough to check statement's classical logical truth value, because a true statement does not have to be suitable. This problem has been widely discussed in literature as: matching the language common [1], statement's implicature [5] or meeting conversational context [2].

Let us assume that someone says: "It is possible, that Paul is rich". We, as listeners, immediately imply that the speaker does not really know whether Paul is rich or not. Telling "It is possible" is the best available option for the speaker. If he knew more he would say that he believes or knows Paul is rich. Let us consider a different example: "If two plus two is five, the sun is a star." This statement is logically true[1] but it makes no sense to tell it. The antecedent has absolutely no impact on the consequent.

Within this work we focus on filtering out conditional statements that aren't consistent with agent's mental state of the world. We favour statements that are most informative based on agent's knowledge. We define formal criteria on when a modal statement can be used by an agent. These criteria are matched against human language behaviour, so that cognitive agent simulates humans in choosing statements. Choosing the right statement is a many step process. We focus only on agent's mental state. We consider agent's knowledge about the environment, but ignore listener's knowledge, complex conversational context or current agent's needs.

[1] According to classical truth table of implication in Boolean logic.

3 Considered Statements

The agent lives in some world and uses a formal language to describe currently observed situation. The formal language allows agent to form many types of natural language statements and is non-extensible. Agent is able to express uncertainty by adding special modal operators of the form: "I believe that" or "It is possible that" called belief and possibility operators, respectively. Many types of the statements have already been formally analysed in [10]. Within this work we are analysing a subgroup of conditional statements.

Conditionals can be categorised in many ways based on their grammatical structure, tenses used and relationship type between antecedent and consequent. The most common classification is into indicative and subjunctive (counterfactual) conditionals. Indicative conditionals refer to realistic situations that according to a speaker really can happen. Subjunctive conditionals describe situations that are only hypothetical or very unlikely to happen. Usually subjunctives have some form of 'would' in a consequent. For example[2]: "If it rains, the match will be cancelled" is an indicative conditional and "If it were to rain, the match would be cancelled" is a subjunctive conditional. In the second statement a speaker is convinced it will not rain. Further we shall assume that all considered conditionals are of indicative form.

Within this work we focus on conditional statements of the forms: "I believe, that if o_1 is q_1, then o_2 is q_2." and "It is possible, that if o_1 is q_1, then o_2 is q_2.". Where o_1, o_2 are objects observed by the agent and q_1, q_2 are properties of these objects. For simplicity we assume that both the antecedent and the consequent refer to current time moment, hence we are dealing with zero conditionals applied to current moment. Time constraint can be omitted with some limitations analogously as in [14]. We assume the modal operator is placed at the beginning of a statement. Statements with modal operator applied to consequent have already been analysed in [13].

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.

3.1 Language Syntax

The alphabet of the language L consist of the following classes of symbols:

- $O = \{o_1, o_2, \ldots, o_M\}$ to represent atomic individuals (objects),
- $Q = \{q_1, q_2, \ldots, q_K\}$ 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 and belief.

Now we have to define proper formulas. The formulas are divided into two classes: atomic formulas (L^A) and complex formulas (L^M). For complex formulas only

[2] Example comes from [2].

a subclass of modal conditional statements is defined, because these will be analysed further in this paper.

Atomic formulas L^A:
Let: $k \in \{1, 2, \ldots, K\}$ and $m \in \{1, 2, \ldots, M\}$. 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^M:
Let $\phi, \psi \in L^A$. Any statement of one of the following forms: $\phi \rightarrow \psi$, $Bel(\phi \rightarrow \psi)$, $Pos(\phi \rightarrow \psi)$ is a proper statement of the language L.

3.2 Intuitive Semantics

The intuitive semantics of the statements are presented in table 1.

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 ϕ, then ψ.
$Bel(\phi \rightarrow \psi)$	I believe, that if ϕ, then ψ.
$Pos(\phi \rightarrow \psi)$	It is possible, that if ϕ, then ψ.

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 [7]. The presented interpretation is not a formal logical definition of modal logic based on Kripke possible worlds [11]. Implications are zero conditionals used to describe current situation.

4 Two Types of Uncertainty in Conditional Statements

In previous works we have analysed conditional statements where modal operator expressing agent's uncertainty was applied to a consequent. Example of such a statement is: "If the apple is green, then I believe that it is not ripe". The modal operator "believe that" clearly stresses that the second part of the statement is uncertain. This means that the speaker knows the implication without modal operator is not true[3]. On the other hand the green color of the apple suggests to the speaker that it rather isn't ripe. It implies that the speaker analyses belief operator for a simple statement "The apple is not ripe" assuming hypothetical situation where apple is green.

Within this work we are considering modal operators applied to whole conditional statement. A statements of the form: $Bel(\phi \rightarrow \psi)$. The questions that

[3] Some green apples can be ripe.

arise are: Isn't that equivalent to telling $\phi \rightarrow Bel(\psi)$? Isn't a belief in a conditional the same as a belief in its consequent? It seems we use the first case more often. Our answer is negative. We claim that the first statement has more general meaning, covering the meaning of the second one. To backup our claim let us consider following example:

Example 1. Albert is a child. He has never heard that water boils at 100 degrees Celsius. He saw a few times a water in a electrical kettle with a thermometer. Always when the temperature has risen to 100 degrees it boiled. His mother bought a new kettle. Albert sees a water being warmed up in a new kettle and he says: "I believe, that if the number changes to 100, then water will boil".

Albert from example one does not connect facts that water boils at 100 degrees and the kettle displays temperature. He has never seen a situation where the temperature has risen to 100 degrees and the water did not boil. On the other hand he has seen only one old kettle in his life. For him the new kettle does not have to behave this way. The situation is somewhat different. He moves the rule that worked for the old kettle to the new one. Because he is not sure it will work, he stresses that the rule does not have to be true. The modal operator does not tell that there is some chance of boiling the water. It tells that Albert is not sure of his reasoning, not sure of the used rule. On the contrary a statement "If the number changes to 100, then I believe that water will boil" means only that water may boil at 100.

Example points out that mental state in case of $Bel(\phi \rightarrow \psi)$ must be different than in case of $\phi \rightarrow Bel(\psi)$. Whereas $\phi \rightarrow Bel(\psi)$ clearly makes the consequent uncertain, $Bel(\phi \rightarrow \psi)$ makes the reasoning uncertain. Uncertainty is at a different level here. Let us call those two types of uncertainty 'rule based uncertainty' and 'consequent based uncertainty' respectively.

We are not neglecting the fact that both statements may lead to simple belief in boiling in the case of 100 degrees. Uncertain reasoning must lead to uncertain results. We mark here that stressing lack of conviction in reasoning is different than stressing possibility of negative result. The difference is very subtle but it has a crucial conclusions to our research.

We argue that $\phi \rightarrow Bel(\psi)$ has a narrower meaning than $Bel(\phi \rightarrow \psi)$. Operator in front of a statement covers both cases where we are uncertain of reasoning and uncertain of consequent. In fact it also covers all cases that are somewhere in between. Cases when we know the rule is not always true, but only in some odd situations. When one tells a statement of the form $Bel(\phi \rightarrow \psi)$ we may not be able to distinguish where exactly the uncertainty is applied. The best we can do is to assume it is somewhere in between.

5 Conditional Statements Meaning

As proven in paragraph 2 the statement told by a 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.

5.1 Statements without Modal Operator

Most of results presented in this sub-paragraph have been already presented in [12] and quoted in [13]. Once more we quote conclusions that are important in the context of the rest of this work.

We are considering conditional statement of the form: "If ϕ, then ψ." told by an agent to describe observed situation. Suppose an agent tells such a statement.

Firstly we claim the speaker does not know whether ϕ holds or not. If the speaker knew that ϕ holds, he would immediately know that ψ holds and would rather tell 'ψ, because of ϕ' or simply 'ψ'. On the other hand, if speaker knew that ϕ does not hold, it would be pointless of him to say such a statement.

Secondly the speaker does not know whether ψ holds or not. If he knew that ψ holds, there would be no point in telling the conditional statement. Similarly the speaker does not know that ψ does not hold. If he knew it doesn't hold he would rather say 'Not ψ, because of not ϕ.' or 'If there were ϕ, there would be ψ'. The understanding of speaker's knowledge is consistent with results presented in [1].

Lastly, the speaker informs, that he has reasoned about both situations (where ϕ holds or not) and found out that ψ is guaranteed to hold only when ϕ holds. In fact the speaker has reasoned about four possible situations and found out that situation where ϕ holds and ψ does not hold is impossible. Hence the speaker is ready to infer ψ, as soon as he finds out that ϕ holds. But as long as he does not know ϕ he is unable to tell much more about ψ.

5.2 Statements with Belief Operator

Let us consider a situation where the statement "I believe, that if ϕ, then ψ." is told. Formal formula for this statement is $Bel(\phi \rightarrow \psi)$. As described in paragraph 4 the phrase 'I believe that' alarms the listener that: Speaker is uncertain whether ϕ really implies ψ (1). Speaker is uncertain whether ψ holds in case ψ holds (2). On the contrary the statement $\phi \rightarrow Bel(\psi)$ means only, that in a case where ϕ holds, the speaker believes ψ holds. Here the speaker has noticed that ϕ changes chance of ψ happening, but it is not that ϕ guarantees ψ [13]. This is analogous to second uncertainty.

We state that the chance of ψ happening has to rise when ϕ holds. When hearing a statement of the form $Bel(\phi \rightarrow \psi)$ one implies that, if not ϕ, ψ is much less probable. We wouldn't say: "I believe, that if the apple is green, then it is ripe.". Most red apples are ripe, but green apples usually aren't. Our rule applies only to considered agent's situation. One can find examples based on complex conversational context, where this rule does not hold.

5.3 Statements with Possibility Operator

Now let us consider statement "I find it possible, that if ϕ, then ψ", formally written as $Pos(\phi \rightarrow \psi)$. The phrase 'I find it possible that' means that: The speaker is very uncertain whether ϕ implies ψ (1). The speaker thinks ψ can hold in case of ϕ but does not have to (2).

On the contrary the statement $\phi \rightarrow Pos(\psi)$ means only that the speaker allows situation where both ϕ and ψ hold and situation where ϕ holds and ψ does not hold is also very probable [13]. This is analogous to second uncertainty.

As in the case of Bel modal operator the chance of ψ happening has to rise when ϕ holds. One wouldn't say: "It is possible, that if he is a smoker, then he doesn't have cancer". This would suggest that smokers are less probable to have cancer then non-smokers. He would rather use a sentence: "It is possible, that if he is a smoker, then he has cancer".

6 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 previous works we defined mental state as a set of possible worlds and respective probabilities of each world. Within this work we are forced to change the model to a more complex one. This is because conditionals considered here require two types of uncertainty: rule and consequent based uncertainty.

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.

6.1 Mental State Model

A mental state model is a set:

$$W = \{(w^{(1)}, \underline{p}^{(1)}, \overline{p}^{(1)}), (w^{(2)}, \underline{p}^{(2)}, \overline{p}^{(2)}), ..., (w^{(S)}, \underline{p}^{(S)}, \overline{p}^{(S)})\} \tag{1}$$

where $w^{(s)}$, $s = 1, 2, ..., S$ is a possible world and $\underline{p}^{(s)}$ is a lower boundary of a chance of this world being an actual, unknown to the agent, world. Analogously $\overline{p}^{(s)}$ is an upper boundary[4]. The real chance $p^{(s)}$ of world being an actual one is unknown to the agent. She can only assume that $p^{(s)} \in [\underline{p}^{(s)}, \overline{p}^{(s)}]$. 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 the actual, real world state.

[4] This is different from mental state model from [13] where we assumed $p^{(s)}$ is exactly known. Now we only assume that $p^{(s)} \in [\underline{p}^{(s)}, \overline{p}^{(s)}]$.

Agent is not omnipotent and does not know which of the worlds is the real one. She can only evaluate the lower and upper probabilities that given world is the actual one.

It is assumed that $\sum_{s=1}^{S} p^{(s)} = 1$ and $p^{(s)} = P(w^{(s)})$, $p^{(s)} \in [\underline{p}^{(s)}, \overline{p}^{(s)}]$ $(s = 1, 2, ..., S)$ defines a probability distribution over W that is unknown to an agent. The greater $p^{(s)}$, the more probable $w^{(s)}$ is. If for some world s, $\overline{p}^{(s)} = 0$ then world is impossible to happen according to the agent. If for some world s, $\underline{p}^{(s)} = 1$, then the agent knows everything.

Each possible world consists of information about objects properties:

$$w^{(s)} = (Q_1^{(s)}, Q_2^{(s)}, ..., Q_K^{(s)}), \quad s = 1, 2, ..., S \tag{2}$$

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 s where $\overline{p}^{(s)} > 0$, then agent is sure object o has property q_k at given time moment. If $o \notin Q_k^{(s)}$ in all possible worlds s where $\overline{p}^{(s)} > 0$, then agent is sure object o does not have property q_k. If there are some worlds with s where $\overline{p}^{(s)} > 0$ 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. Overall probability of o being q_k at given time moment is a sum of probabilities of all worlds where $o \in Q_k^{(s)}$.

7 Grounding Conditional Statements

We are considering statements of the forms: $Bel(\phi \rightarrow \psi)$ and $Pos(\phi \rightarrow \psi)$. 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 [6,9]. We say that a statement is properly grounded if and only if the epistemic relation holds.

Further we will be using probabilistic model, to determine if a given statement can be spoken. We shall use following probabilities:

- $\underline{P}(\phi), \overline{P}(\phi)$ - minimal and maximal probability of the antecedent,
- $\underline{P}(\psi), \overline{P}(\psi)$ - minimal and maximal probability of the consequent,
- $\underline{P}(\psi|\phi), \overline{P}(\psi|\phi)$ - minimal and maximal conditional probability of the consequent assuming antecedent,
- $\underline{P}(\psi|\neg\phi), \overline{P}(\psi|\neg\phi)$ - minimal and maximal conditional probability of the consequent assuming that antecedent does not hold.

Required minimal and maximal probabilities can be directly calculated from mental state model. To calculate them one has to use appropriate lower and upper bounds $\underline{p}^{(s)}, \overline{p}^{(s)}$ of possible worlds from the model.[5]

7.1 Grounding Statements with Belief Operator

We are defining an epistemic relation for a statement of the form: "I believe, that If ψ, then ϕ".

Definition 1. *Epistemic relation* $\models^E Bel(\phi \rightarrow \psi)$ *holds iff all following conditions are met:*

a. $\underline{\alpha} < \overline{P}(\phi)$ *or* $\underline{P}(\phi) < \overline{\alpha}$
b. $\underline{\alpha}_{Bel} \leq \underline{P}(\psi|\phi) < \overline{\alpha}_{Bel}$
c. $(\overline{P}(\psi|\phi) + \underline{P}(\psi|\phi)) > \beta(\overline{P}(\psi|\neg\phi) + \underline{P}(\psi|\neg\phi))$

where $0 < \underline{\alpha} < \overline{\alpha} < 1$, $0 < \underline{\alpha}_{Bel} < \overline{\alpha}_{Bel} < 1$, $\beta > 1$ *are fixed parameters.*

Condition a guarantees that agent does not know whether ϕ holds or not. Parameter $\underline{\alpha}$ should be close to zero, while parameter $\overline{\alpha}$ should be close to one. We don't simply write condition a as $0 < \overline{P}(\phi)$ or $\underline{P}(\phi) < 1$ because we want agent to consider only realistically probable chances of ϕ happening or not.

Condition b guarantees that minimal conditional probability is enough to make agent 'Believe' the conditional statement. The higher the $\underline{\alpha}_{Bel}$ parameter the less prone agent is to tell that she believes something.

Condition c guarantees that ψ is more probable in case of ϕ than the opposite. In other words ϕ must have positive effect on ψ. The β parameter tells how much more probable must ψ be in case of ϕ. For example $\beta = 2$ means that middle of interval $(\underline{P}(\psi|\phi), \overline{P}(\psi|\phi))$ must be two times bigger then middle of interval $(\underline{P}(\psi|\neg\phi), \overline{P}(\psi|\neg\phi))$.

7.2 Grounding Statements with Possibility Operator

We are defining an epistemic relation for a statement of the form: "I find it possible, that if ϕ, then ψ".

Definition 2. *Epistemic relation* $\models^E Pos(\phi \rightarrow \psi)$ *holds iff all following conditions are met:*

a. $\underline{\alpha} < \overline{P}(\phi)$ *or* $\underline{P}(\phi) < \overline{\alpha}$
b. $\underline{\alpha}_{Pos} < \underline{P}(\psi|\phi) < \overline{\alpha}_{Pos}$
c. $(\overline{P}(\psi|\phi) + \underline{P}(\psi|\phi)) > \beta(\overline{P}(\psi|\neg\phi) + \underline{P}(\psi|\neg\phi))$

where $0 < \underline{\alpha} < \overline{\alpha} < 1$, $0 < \underline{\alpha}_{Pos} < \overline{\alpha}_{Pos} < 1$, $\beta > 1$ *are fixed parameters.*

Conditions a and c are the same as in a case of belief operator.

Condition b is almost the same as respective condition for belief operator. It has changed parameters of minimal and maximal allowed probability. To ensure proper understanding of possibility operator, $\underline{\alpha}_{Pos}$ should be close to zero.

[5] Model defined in [13] allowed to calculate exact probabilities. In the new model it is impossible.

8 Summary

We have described that statements $Bel(\phi \rightarrow \psi)$ and $\phi \rightarrow Bel(\psi)$, $Pos(\phi \rightarrow \psi)$ and $\phi \rightarrow Pos(\psi)$ don't have the same meaning, although it is often impossible to the hearer to distinguish them. We claim that $Bel(\rightarrow \psi)$ is more general and usually used to describe uncertainty on the rule level, not only the consequent level. We defined a model of mental state that is based on intervals of probability. Such model is required to analyse rule based uncertainty. Next we defined formal rules that simulate human way of choosing suitable statements.

In the future we wish to prove how $Bel(\phi \rightarrow \psi)$ and $\phi \rightarrow Bel(\psi)$ conditional statements are interrelated. However this requires suiting definitions of epistemic relation from [13] to new mental state model.

References

1. Ajdukiewicz, K.: Conditional sentence and material implication. Studia Logica 4(1), 135–153 (1956)
2. Clark, M.: Ifs and Hooks. Analysis 32(2), 33–39 (1971)
3. Cohen, P., Levesque, H.: Communicative Actions for Artifical Agents. In: Proc. of the 1st International Conference on Multi-agent Systems, San Francisco (1995)
4. Arlo-Costa, H.: The Logic of Conditionals. Entry for the Stanford Encyclopedia of Philosophy (2007)
5. Grice, H.P.: Meaning. Philosophical Review 66, 377–388 (1957)
6. Harnad, S.: The Symbol Grounding Problem. Physica D 42, 335–346 (1990)
7. Hintikka, J.: Knowledge and belief. An introduction to the logic of the two notions. Cornell University Press (1962)
8. Huhns, N., Signh, M.: Cognitive Agents. IEEE Internet Computing 2(6), 87–89 (1998)
9. 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)
10. Katarzyniak, R.: Gruntowanie modalnego języka komunikacji w systemach agentowych. Exit, Warsaw (2007) (in Polish)
11. Kripke, S.: Semantical Analysis of Modal Logic. Zeitschrift für Mathematische Logic und Grundlagen der Mathematik 9, 67–96 (1963)
12. Skorupa, G., Katarzyniak, R.: Extending modal language of agents, communication with modal implications. Information Systems Architecture and Technology, 127–136 (2009)
13. Skorupa, G., Katarzyniak, R.: Applying Possibility and Belief Operators to Conditional Statements. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010. LNCS (LNAI), vol. 6276, pp. 271–280. Springer, Heidelberg (2010)
14. Skorupa, G., Katarzyniak, R.: Conditional Statements grounded in Past, Present and Future. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010. LNCS (LNAI), vol. 6423, pp. 112–121. Springer, Heidelberg (2010)

Developing a Competitive HMM Arabic POS Tagger Using Small Training Corpora

Mohammed Albared, Nazlia Omar, and Mohd. Juzaiddin Ab Aziz

University Kebangsaan Malaysia, Faculty of Information Science and Technology, Department of Computer Science
mohammed_albared@yahoo.com, {no,din}@ftsm.ukm.my
http://www.ukm.my

Abstract. Part Of Speech (POS) tagging is the ability to computationally determine which POS of a word is activated by its use in a particular context. POS is one of the important processing steps for many natural language systems such as information extraction, question answering. This paper presents a study aiming to find out the appropriate strategy to develop a fast and accurate Arabic statistical POS tagger when only a limited amount of training material is available. This is an essential factor when dealing with languages like Arabic for which small annotated resources are scarce and not easily available. Different configurations of a HMM tagger are studied. Namely, bigram and trigram models are tested, as well as different smoothing techniques. In addition, new lexical model has been defined to handle unknown word POS guessing based on the linear interpolation of both word suffix probability and word prefix probability. Several experiments are carried out to determine the performance of the different configurations of HMM with two small training corpora. The first corpus includes about 29300 words from both Modern Standard Arabic and Classical Arabic. The second corpus is the Quranic Arabic Corpus which is consisting of 77,430 words of the Quranic Arabic.

Keywords: Arabic languages, Hidden Markov model, Unknown words.

1 Introduction

Part of speech disambiguation is the ability to computationally determine which POS of a word is activated by its use in a particular context. POS tagging is an important basis of many higher level NLP applications like speech processing and machine translation. Being one of the first processing steps in any such application, the performance of the POS tagger directly impacts the performance of any subsequent text processing steps. Language resources such as annotated corpora are fundamental for research and development in statistical computational linguistics and for the construction of NLP applications. The main challenge involved in constructing any Arabic NLP system for Arabic is amplified by the lack of these language resources [1]. In spite of recent progress, Arabic is still lacking such tools and annotated resources [2].

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 288–296, 2011.

The task of POS tagging is very difficult due to two main reasons. First, many words are ambiguous. For example, English word " past " can be an adjective (e.g. his past performance), an adverb (e.g. its ten past seven), or a noun (e.g. in the past). This ambiguity is exist in all languages. For example, Arabic word "علم" can be a noun (e.g. علم مفيد), or a verb (e.g. (علم الانسان). Moreover, words can be ambiguous to their grammatical properties. For example, Arabic word " كتب " can be a past verb (e.g.كتب الكتاب), or a passive verb (e.g.كتب الكتاب). Second, the existing of unknown words, words that appears in the test data and do not appear in the training data. The problem of unknown words is the main problem in POS tagging [3]. Actually, the size of this problem is proportional to many factors such as: size, genre and the quality of the training data. This Unknown words problem becomes more serious when the training data are small [4]. With small training data, it is very difficult to predict the distribution of the unknown words. The impact of this problem increases in languages which have huge vocabulary and rich morphological system like Arabic.

This work presents a study aiming to find out the appropriate way to develop a competitive Arabic statistical POS tagger when only a limited amount of training material is available. To do so, we compare different smoothing techniques and different order HMMs. In addition, we propose a new lexical model to better handling unknown word in Arabic POS tagging. This new lexical model is based on the linear interpolation of both word's suffix probability and word's prefix probability. Several experiments are conducted to determine the performance of the different configurations using two small training corpora.

The remainder of this paper is organized as follows: First, Section 2 reviews some related works. Then, the training corpora are described in Section 3. Section 4 presents a breif description of the investigated HMM models.We then describe our method for unknown word POS guessing in Sections 5. Section 6 shows the realized experiments and the obtained results. Finally, Section 7 states some conclusions and further work.

2 Related Work on POS Tagging

POS tagging has been largely studied. There are different approaches have been used for POS tagging. There are some machine-learning taggers [5,6,7,8] and some rule based taggers [9]. POS tagging is often considered to be a "solved task ", with published tagging accuracies around 97%. However, In the real-life scenario, Giesbrecht and Evert [10]showed that five state-of-the-art POS taggers are unsuitable for fully automatic processing. Among all these state-of-the-art taggers, HMM taggers are more robust and much faster than other advanced machine learning approaches in the real-life scenario. Previous work on POS tagging has utilized different kind of features to tackle unknown word POS tagging. These features are mainly based on word substring information, word context information and/or global information.

Padr et al. [11] and Ferrndez et al. [12] investigate different configurations of the HMM Spanish POS Tagger when minimum amount of training data is

available. In contrast with our work, they assume the none-existence of unknown words, which is the main problem when the training data is small [3,4]. However, they work with Spanish language which is very different from Arabic language.

Present day Arabic has two literary styles. One is called the Classic Arabic and the other is the Modern standard Arabic (MSA). Classical Arabic is the language of formal writing until nearly the first half of the 20th century and was also the spoken language before the medieval times [13]. MSA is the language of formal modern writing in all Arabic countries.

Recently, several works have been proposed to Arabic POS tagging such as [14,15,16,17,18,19].For more details about Arabic work and also about POS tagging techniques in general,see [20]. Almost all of these taggers are generally developed for Modern Standard Arabic (MSA) and few works are interested in Classical Arabic [21]. In contrast with other Arabic taggers, our POS tagger deals with both MSA and Classic Arabic together. Additionally up to our knowledge, this is the first work which aims to find the appropriate configuration of Arabic POS tagger when small amount of training data is available. Furthermore, our work has defined a new lexical model to handle unknown words in Arabic POS tagging. The proposed lexical model demonstrates to be effective and efficient in handling unknown word even with small training data. An unknown words POS tagging accuracy of 85.3% obtained using the introduced method is as of yet the highest reported in the Arabic POS tagging literature.

3 The Used Data

For our experiments, we used two different and small Arabic corpora: the FUS-HA corpus and The Quranic Arabic Corpus. The FUS-HA corpus is composed of two sub-corpora:

1. MSA corpus: The MSA corpus is composed of journalistic articles discussing general news topics. The news topics cover various subjects of politics, economics and culture. The corpus includes more than 12000 words forms. The This data are 2009-2010 newswire feeds collected from different online Arabic newspaper archives, such as Al-Jazeera and Alsharq Al-Awsat.
2. Classic Arabic corpus: The Classic Arabic corpus is composed of some texts extracted from ALJAHEZ's book entitled "Albayan-wa-tabyin" (255 Hijri). "Albayan-wa-tabyin" "The art of communication and demonstration" is one of the best and earliest Arabic books on Arabic literary theory and literary criticism. The book covers various subjects, such as rhetorical speeches, history and science. The Classic Arabic corpus includes more than 17000 word forms.

We annotated the FUS-HA corpus using two Arabic tag sets. The first one is the Arabic TreeBank tagset, which is consist of 23 tags, used by Diab et al [16]. The second one is quite similar to the first one. We only add some modifications to handle some linguistic limitation. We introduce a tag for the Broken Plural to distinguish between it and the singular noun. Broken Plurals

constitute 10% of any Arabic text and form 40% of the Arabic plurals [22]. The second modification, our tagset does not include NO_FUNC (no solution chosen) tag, which is used as a tag in the above mentioned Arabic TreeBank tagset. Finally, we distinguish between inflected and non inflected verbs. However, the modified tag set consists of 24 tags.

The Quranic Arabic Corpus[23] is an annotated linguistic resource which shows the Arabic grammar, syntax and morphology for each word in the Holy Quran, the religious book of Islam which is written in classical Quranic Arabic (c. 600 CE). The research project is organized at the University of Leeds, and is part of the Arabic language computing research group within the School of Computing. The Quranic Arabic Corpus is consisting of 77,430 words of Quranic Arabic.

4 Hidden Markov Models

In HMM, the POS problem can be defined as the finding the best tag sequence t^n given the word sequence w^n. The label sequence t^n generated by the model is the one which has highest probability among all the possible label sequences for the input word sequence. This is can be formally expressed as:

$$t_1^n = \arg\max_{t_1^n} \prod_{i=1}^{n} p\left(t_i | t_{i-1} \ldots . t_1\right) . p\left(w_i | t_i \ldots . t_1\right) \ . \tag{1}$$

The two models, the state transition probabilities and the emission probabilities, parameters are estimated from annotated corpus by Maximum Likelihood Estimation (MLE), which is derived from the relative frequencies. We will use Hidden Markov Models POS taggers of order two and three. Having computed the state transition probabilities and the emission probabilities and assigning all possible tag sequences to all words in a sentence, now we need an algorithm that can search the tagging sequences and find the most likely sequence of tags. For this purpose we use the Viterbi algorithm [24] which compute the maximized tag sequence with the best score. However, MLE is a bad estimator for statistical inference especially, in NLP application, because data tends to be sparse. In this work, two smoothing methods are used. With the Bigram version, we use the Modified Kneser Ney smoothing technique [25]. In the Trigram version, we use the linear interpolation of unigram, bigram and trigram maximum likelihood estimates [6] in order to estimate the trigram transition probability.

5 Unknown Words Handling

In POS tagging, we frequently encounter words that do not appear in training data. Such words are called unknown words or out-of-vocabulary (OOV) words. The existence of unknown words is the main problem for POS taggers, since the

statistical information of these words are unavailable. It is a non-negligible problem especially where only a limited amount of training material is available. Unknown words are usually handled by an exceptional processing. Accuracy of POS tagging for unknown words is usually much lower than that for known words.

In order to handle the POS tagging for unknown words in Arabic POS tagging, we have defined a new lexical model based on the linear interpolation of both word suffix probability and word prefix probability. It combines together both word suffix information and word prefix information. The main linguistic motivation behind combining affixes information is that in Arabic word sometimes an affix requires or forbids the existence of another affix [13]. Prefix and suffix are the first n and m letters of the word, and are not necessarily morphologically meaningful. In this model, the lexical probabilities are estimated as follows:

1. Given an unknown word w, the lexical probabilities $P(\text{suffix(w)}|t)$ are estimated using the suffix tries as in the following equation:

$$P\left(t|c_{n-i+1},...,c_n\right) = \frac{P(t,c_{n-i+1},...,c_n) + \theta P(t,c_{n-i+2},...,c_n)}{1+\theta} \quad . \qquad (2)$$

$$\theta = \frac{1}{S-1}\sum_{J=1}^{S}\left(P\left(t_j\right)-\overline{P}\right)^2, \overline{P} = \frac{1}{s}\sum_{J=0}^{S}P\left(t_j\right)$$

   where $c_{n-i+1},...,c_n$ represent the last n characters of the word, $S$is the number of tags in the tagsets.
2. Then, the lexical probabilities $P(\text{prefix(w)}|t)$ are estimated using the prefix tries as in Equation 2. But, we reverse the letters in the words before adding them to the new word trie in order to find the prefix probability. Here, the probability distribution for a unknown word prefix is generated from all words in the training set that have the same prefix up to some predefined maximum length.
3. Finally, we use the linear interpolation of both the lexical probabilities obtained from both word's suffix and word's prefix to calculate the lexical probability of the word w as in the following equation:

$$P\left(w|t\right) = \lambda P\left(suffix(w)|t\right) + (1-\lambda)P(prefix(w)|t) \qquad (3)$$

   where λ is an interpolation factor. In addition, the experiments also utilize the following features: the presence of non-alphabetic characters and the existence of foreign characters.

6 Experiments and Results

The main objective of this work is to study the behavior of different configurations for a HMM POS tagger, in order to determine the appropriate way to develop competitive Arabic tagger. To do so, we have carried out several experiments when small amount of training data is available. The data used for

the empirical evaluation come from the above described corpora. The FUS-HA corpus is divided into 78% for training and 22 % for testing. The percentage of unknown words in this test set is 10.7% . Whereas the Quranic Arabic Corpus is divided into 90.1 % for training and 9.9 % for testing. The percentage of unknown words in the Quranic Arabic Corpus test set is 14.9%. This decision was taken because the test data has to be guaranteed as unseen during training [26] and also it is necessary for comparisons to be consistent with previous evaluation works [27].

Results obtained for each HMM tagger configuration are summarized in Table 1 . Results are given both for the FUS-HA corpus and the Quranic Arabic Corpus. We define the tagging accuracy as the ratio of the correctly tagged words to the total number of words. As we can see in Table 1, the experiments show that the best results on all test sets were achieved by the trigram HMM with the linear interpolation of unigram, bigram and trigram smoothing technique.

Table 1. Obtained results for all HMM POS tagger configurations using both corpora

λ	Bigram HMM						Trigram HMM					
	FUSHA 23 tags		FUSHA 24 tags		Quranic		FUSHA 23 tags		FUSHA 24 tags		Quranic	
	Unknown	Overall	Unknown	Overall	Unknown	Overall	Unknown	Overall	Unknown	Overall	Unknown	Overall
0	66.2	94.4	61.5	93.6	72.6	93	64.4	94.6	49.7	92.9	72.2	92.9
0.1	68.6	94.7	62.4	93.7	75	93.5	67.1	94.9	53.3	93.3	76.1	93.5
0.2	68.9	94.7	62.3	93.8	76.7	93.7	69.6	95.2	57	93.7	77.2	93.7
0.3	69.6	94.8	64.8	94	79	94.1	71.4	95.4	60.5	94	80.8	94.3
0.4	71.3	94.9	62.8	93.8	80.7	94.3	70.2	95.2	61.3	94.1	81.9	94.4
0.5	71	94.9	63.1	93.8	80.8	94.3	70.2	95.2	62.6	94.3	84	94.8
0.6	68.9	94.7	61.6	93.6	82.6	94.6	69.3	95.1	62.7	94.3	**85.3**	95
0.7	66.3	94.4	59.9	93.4	82.3	94.6	66.9	94.8	63	94.3	84.6	94.9
0.8	64.4	94.1	58.1	93.3	81.6	94.5	62.8	94.4	60.9	94.1	82.9	94.6
0.9	63.3	94	56.1	93	80.8	94.4	60.8	94.2	59.3	93.9	82.2	94.5
1	61.7	93.8	55.2	92.9	79.3	94.2	56.4	93.7	58.1	93.8	79.4	94.1

Comparing the results for the different order models, we can draw the following conclusions:

- In general, taggers trained the Quranic Arabic Corpus using have higher precision than those taggers trained using the FUS-HA corpus. This demonstrates that increasing the size of the training set has a positive impact on the accuracy rate.
- It is clearly that working with a trigram HMM gives higher precision than working with a bigram one, for both training corpora.

– The most important conclusion is that the proposed lexical model, the linear interpolation of both the suffix probability and the prefix probability, improves the tagging of Arabic text. The linear interpolation model($0 \prec \lambda \prec 1$)improves the POS tagging of unknown words dramatically compared to the suffix model ($\lambda = 1$), which has been used in several previous studies and proofed to be effective for other languages [6],[28], and the prefix model ($\lambda = 0$).

Nevertheless, some important observations can be extracted from these results:

– A competitive HMM taggers may be built using relatively small train sets, which is interesting, especially, with the lack of language resources.
– The linear interpolation, which combine information from both word suffix and word prefix togather, is a good indicator of Arabic unknown word POS.

7 Conclusion

Part of speech tagging is an important tool in many NLP applications. In addition, it accelerates the development of large manually annotated corpora. In this paper, we have studied how competitive Arabic HMM-based POS taggers can be developed using relatively small training corpus. Different configurations of a HMM tagger are investigated. We conducted a series of experiments using two small Arabic training corpora. Results indicate that accurate Arabic taggers can be build provided appropriate lexical model to handle unknown words. Between all configurations studied here, in general the one that gives a higher precision is trigram HMM with the linear interpolation of unigram, bigram and trigram smoothing technique and the linear interpolation of both word's suffix probability and word's prefix probability unknown word guessing algorithm. In the future work, it might also be worthwhile to improve the tagging accuracy of unknown words. This improvement can be done through the integration of highly coverage Arabic morphological analyzer, increasing the size of our training corpus as well as by using specific features of Arabic words, which can be better predictor for Arabic words POS than words suffixes and prefixes. Also, we plan to study their influence on the taggers developed on small corpora.

References

1. Farghaly, A., Shaalan, K.: Arabic Natural Language Processing: Challenges and Solutions. ACM Transactions on Asian Language Information Processing (TALIP), 1–22 (2009), doi:http://doi.acm.org/10.1145/1644879.1644881
2. Maamouri, M., Bies, A., Kulick, S.: Enhanced Annotation and Parsing of the Arabic Treebank. In: INFOS (2008)
3. Fischl, W.: Part of Speech Tagging - A solved problem? Center for Integrative Bioinformatics Vienna, CIBIV (2009) (Unpublished report)
4. Nakagawa, T.: Multilingual word segmentation and part-of-speech tagging: a machine learning approach incorporating diverse features. PhD Thesis, Nara Institute of Science and Technology, Japan (2006)

5. Ratnaparkhi, A.: A maximum entropy part of speech tagger. In: Brill, E., Church, K. (eds.) Conference on Empirical Methods in Natural Language Processing. University of Pennsylvania, Philadelphia (1996)
6. Brants, T.: TnT: A statistical part-of-speech tagger. In: Proceedings of the 6th Conference on applied Natural Language Processing, Seattle, WA, USA (2000)
7. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the International Conference on Machine Learning, MA, USA (2001)
8. Goldwater, S., Griffiths, T.: A fully Bayesian approach to unsupervised part-of-speech tagging. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (2007)
9. Brill, E.: A Corpus-based Approach to Language Learning. PhD thesis, Department of Computer and Information Science. University of Pennsylvania, Philadelphia (1993)
10. Giesbrecht, E., Stefan, E.: Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus. In: Proceedings of the 5th Web as Corpus Workshop (WAC5), Donostia (2009)
11. Padró, M., Padró, L.: Developing Competitive HMM PoS Taggers Using Small Training Corpora. In: Vicedo, J.L., Martínez-Barco, P., Muńoz, R., Saiz Noeda, M. (eds.) EsTAL 2004. LNCS (LNAI), vol. 3230, pp. 127–136. Springer, Heidelberg (2004)
12. Ferrández, S., Peral, J.: Investigating the Best Configuration of HMM Spanish PoS Tagger when Minimum Amount of Training Data Is Available. In: Montoyo, A., Muńoz, R., Métais, E. (eds.) NLDB 2005. LNCS, vol. 3513, pp. 341–344. Springer, Heidelberg (2005)
13. Attia, M.: Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation. PhD thesis, School of Languages, Linguistics and Cultures, Univ. of Manchester, UK (2008)
14. AlGahtani, S., Black, W., McNaught, J.: Arabic Part-Of-Speech Tagging using Transformation-Based Learning. In: Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt (2009)
15. Kulick, S.: Simultaneous Tokenization and Part-of-Speech Tagging for Arabic without a Morphological Analyzer. In: Proceedings of ACL 2010 (2010)
16. Diab, M., Kadri, H., Daniel, J.: Automatic tagging of Arabic text: from raw text to base phrase chunks. In: Proceedings of the 2004 Conference of the North American Chapter of the ACL (2004)
17. Habash, N., Rambow, O.: Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In: Proceedings of the 43rd Annual Meeting on ACL, Ann Arbor, Michigan (2005), doi:10.3115/1219840.1219911
18. Al Shamsi, F., Guessoum, A.: A hidden Markov model-based POS tagger for Arabic. In: Proceeding of the 8th International Conference on the Statistical Analysis of Textual Data, France, pp. 31–42 (2006)
19. Albared, M., Omar, N., Ab Aziz, M., Ahmad Nazri, M.: Automatic Part of Speech Tagging for Arabic: An Experiment Using Bigram Hidden Markov Model. In: Yu, J., Greco, S., Lingras, P., Wang, G., Skowron, A. (eds.) RSKT 2010. LNCS, vol. 6401, pp. 361–370. Springer, Heidelberg (2010), doi:10.1007/978-3-642-16248-0_52
20. Albared, M., Omar, N., Ab Aziz, M.J.: Arabic Part Of Speech Disambiguation: A Survey. International Review on Computers and Software, 517–532 (2009)

21. El Hadj, Y., Al-Sughayeir, I., Al-Ansari, A.: Arabic Part-Of-Speech Tagging using the Sentence Structure. In: Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt (2009)
22. Goweder, A., De Roeck, A.: Assessment of a Significant Arabic Corpus. In: Proc. of Arabic NLP Workshop at ACL/EACL (2001)
23. Dukes, K., Habash, N.: Morphological Annotation of Quranic Arabic. In: Language Resources and Evaluation Conference (LREC), Valletta, Malta (2010)
24. Viterbi, A.J.: Error bounds for convolution codes and an asymptotically optimal decoding algorithm. IEEE Transactions on Information, 260–266 (1967)
25. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, Cambridge(1998)
26. Carrasco, R.M., Gelbukh, A.: Evaluation of TnT Tagger for Spanish. In: Proceedings of the 4th Mexican international Conference on Computer Science. IEEE Computer Society, Washington, DC (2003)
27. Mihalcea, R.: Performance analysis of a part of speech tagging task. In: Gelbukh, A. (ed.) CICLing 2003. LNCS, vol. 2588, pp. 158–167. Springer, Heidelberg (2003)
28. Samuelsson, C.: Handling sparse data by successive abstraction. In: COLING 1996, Copenhagen, Denmark (1996)

Linguistically Informed Mining Lexical Semantic Relations from Wikipedia Structure

Maciej Piasecki, Agnieszka Indyka-Piasecka, and Roman Kurc

Institute of Informatics, Wrocław University of Technology, Poland
{maciej.piasecki,indyka,roman.kurc}@pwr.wroc.pl

Abstract. A method of the extraction of the wordnet lexico-semantic relations from the Polish Wikipedia articles was proposed. The method is based on a set of hand-written set of lexico-morphosyntactic extraction patterns that were developed in less than one man-week of workload. Two kinds of patterns were proposed: processing encyclopaedia articles as text documents, and utilising the information about the structure of the Wikipedia article (including links). Two types of evaluation were applied: manual assessment of the extracted data and on the basis of the application of the extracted data as an additional knowledge source in automatic plWordNet expansion.

1 Motivation, Related Work and Main Ideas

Electronic texts available in huge volumes include large amounts of valuable information and knowledge, but their use by the intelligent systems is problematic due to the limited development of the contemporary human language technology. However, *structured documents* in which text is extended by the additional layers of annotation expressing semantic information (e.g. text categorisation) or the structural organisation of a document, create new possibilities for automated, text-based knowledge extraction. Crowd-sourced encyclopaedias like Wikipedia are of special interest due to their substantial size and free availability for different languages. Moreover, Wikipedia, as being continuously edited by a number of users, reflects potentially contemporary language use.

The potential of Wikipedia as a lexical semantic knowledge base has been widely explored recently. It has been used in NLP tasks like text categorization [4], where article-category links were used for computing semantic relatedness between words in articles, information extraction [15], information retrieval [5], question answering [1], computing semantic relatedness [8,16], or named entity recognition and disambiguation [2]. All these tasks require well-constructed lexical semantic information, which could be acquired from linguistic knowledge bases like WordNet [3]. Ponzetto and Strube [14] combine the Wikipedia categories into a semantic network which served as basis for computing the semantic relatedness of words. The well-developed relatedness measures established for WordNet were used. In the further research they have developed the automatic method for assigning *isa* and *notisa* relations between the categories from Wikipedia [14].

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 297–306, 2011.

In our approach, we perceived Wikipedia as a set of structured documents in the natural language. Our objective is to extract knowledge concerning lexical semantics from them automatically. Similar works are mostly focused on the utilisation of the Wikipedia structure, e.g. calculation of similarity on the basis of the link structure, processing on the category names and their dependencies or mining meta-information. Our goal is to extract pairs of words and multi-word units that are linked by one of the wordnet lexico-semantic relations.

A wordnet is an electronic thesaurus whose construction follows the main lines of Princeton WordNet – the first and still the biggest wordnet. In a wordnet, words are grouped into sets of near synonyms – called *synsets* (basic building blocks). Synsets, but also individual words, are linked by lexico-semantic relations, that belong to a limited, linguistically motivated set, e.g. hypernymy, meronymy or antonymy (among words). Despite several known deficiencies of wordnets, cf e.g. [13], wordnets express the largest sizes among lexical semantics resources, are manually constructed and are useful in many applications. Wordnet development is cheaper than in the case of semantic lexicons of other types, but it is still a laborious process. A semi-automatic, supporting tool can be a substantial advantage, e.g. [10].

Analysis of the Wikipedia structured content gives an opportunity to combine pure text analysis with recognition of different kinds of annotation (e.g. links, categories etc.). In this paper we leave aside the linguistic analysis of the categories, studied in literature, e.g. [7] and instead we concentrate on processing the content of the Polish Wikipedia[1] articles together with the structural information encoded in them. Our goal is to extract from sequences: term – article, lemma[2] pairs that are *instances* of wordnet relations. Polish is an inflectional language and it seems to be more feasible to perform semantic processing on the level of lemmas, than on the level of word forms.

2 Pattern-Based Semantic Relation Extraction

There are two basic approaches to the extraction of lexical semantic knowledge from text corpora: *Distributional Semantics* and *pattern-based methods*. Distributional Semantics is based on measuring semantic relatedness among words on the basis of the similarity of their distributions in the corpus. High values of the relatedness can be correlated with a rich variety of relations. The measure can be a valuable knowledge source, but it is better to construct it on a corpus which is much larger than Polish Wikipedia.

Pattern-based methods originate from the seminal works of Hearst [6] on application of lexico-syntactic patterns to a corpus. Hearst's patterns have the expressive power of the regular expressions and recognise selected words and expressions, but require Noun Phrases boundaries to be identified by a shallow

[1] `http://pl.wikipedia.org/wiki/Strona_główna`

[2] A lemma is a pre-selected basic morphological form representing the whole set of words or multi-word expressions that differ only with respect to values of grammatical categories (like case or gender) but express the same lexical meaning.

parser. Each pattern is aimed at recognition of instances of a particular lexico-semantic relation. Hearst proposed patterns only for the nominal hypernymy, later attempts to apply patterns for the extraction of meronymy did not achieved accuracy on the practical level.

There is no robust shallow parser for Polish, however, patterns based on the language of morpho-syntactic constraints have been proposed and successfully applied to the extraction of data for the semi-automatic expansion of plWordNet, cf [13]. The patterns were expressed in the JOSKIPI language – a language of rules of the morpho-syntactic tagger [12]. JOSKIPI can be used to describe not only token sequences, but also morpho-syntactic relations between pairs of words, e.g. agreement on number, gender and case between a noun and an adjective modifying it. Here this approach will be extended to processing structurally annotated Wikipedia texts.

Wikipedia articles represent a rich variety of forms, but three main parts of a typical article can be distinguished: *term name*, *versions* and *description*. Versions occur in many articles, are enclosed in parentheses, and include information concerning translations, and term synonyms, but also etymology or language register, domain, etc. For instance:
Ciśnienie tętnicze (ang. "blood pressure" – BP) – ciśnienie wywierane przez
Gloss: Blood pressure (...) is the pressure exerted by

Term: *Ciśnienie tętnicze* — **Variants**: (ang. "blood pressure" – BP) — **Description**: ciśnienie ...

Description is generally a free text with links to other articles embedded in it, however, we can observe that a few first sentences are usually directly related to the term classification and, especially, links occurring there directly characterise the term (e.g. in terms of part/whole distinctions). When we go further from the description beginning interpretation of link semantics is becoming less and less predictable and author's intentions behind attachment of link to tokens[3] are much more difficult to be automatically discovered.

Following the article structure, two types of extraction rules were introduced:

- *article text rules* that can be applied to any piece of the description and do not take into account structural annotations,
- *heading rules* that are intended to be applied to the sequence consisting of the article term, versions and the first sentence of the description treated together as a one complex sentence.

2.1 Article Text Extraction Rules

Article text rules were developed on the basis of rules constructed for the general corpus, cf [13]. Information concerning the structure (links in first) is not used in them, as article text rules can be applied to any part of the description, even

[3] Links are not necessarily attached to proper expressions only – they very often encompass only selected words or symbols.

located far from the beginning and related in a remote way to the term or article categories. Each rule defines a scheme of textual context in which if two noun lemmas occur, it is very likely that they are associated by a particular lexico-semantic relation – the context is a marker. A rule describes lemma positions in the context, selected lexical elements and potential morpho-syntactic relations among lemma and context elements. The context is not limited to the token sequence between the two lemma occurrences but can be freely extended to the tokens preceding the first lemma and following the second lemma. Rules (schemes in fact) are next instantiated by a list of lemmas that represent lexical units for which we want to extract semantic information. Constraining the work of the rules to the preselected lemmas filters out information noise created by associations with infrequent Proper Names or their parts.

Three productive article text rules were applied, all focused on the extraction of hypernymic pairs. The example of a rule is given below – by Noun1 and Noun2 we refer to the lemmas instantiating the rule:

```
and(
 not(Noun1 is in the genitive case and preceded by a noun in genitive),
 rlook(1,end,$C, in(base[$C],"i" and,"oraz" and) ),
 in(base[$+1C],"inny" other,"pozostały" the rest of),
 equal(nmb[$+1C],pl),
 only(1,$-1C,$X, adjectives, adverbs, nouns and commas ),
 not( conjunction or punctuation mark on the following position )
 rlook($+2C,end,$Y, in(flex[$Y],noun) ),
 equal(base[$Y],"Noun2"),
 equal(cas[$Y],cas[0]),
 not( Noun2 in genitive and precedes a noun in genitive )
 only($+3C,$-1Y,$Z,in(flex[$Z],adjectives and adverbs)) )
```

The above rule is expressed in JOSKIPI language. For the presentation clarity the exact code was simplified and summarized in some parts. The italic font marks abbreviations and glosses. In the rule, first we test if Noun1 is not an inner element of a sequence of nouns in the genitive case. In such a situation it is very likely that Noun1 is not a noun phrase head, and is not characterised by the relation identified by the rule. Next, we look for one of the particular conjunctions to the right (`rlook`) of Noun1 such that it is in a sequence with one of the particular adjectives in the plural number (`nmb`). Between Noun1 and the found concjunctions only adjectives, adverbs, nouns and commas can occur. Next, starting from the first position after the conjunction–adjective sequence we are looking for a Noun2 occurrence. The Noun2 and Noun1 occurrences must be in same case. Finally we check if only adjectives and adverbs occur between Noun2 and the conjunctive. Besides simple tests presented in the rule, JOSIPI offers also possibility of performing complex tests on morpho-syntactic agreement between pairs of words or across word sequences. Only Noun2 is explicitly referred to in the rule, as it is assumed that rule is run in contexts with Noun1 on the position 0. As the rules are instantiated with preselected lemmas, it is the task of the control mechanism to scan text for lemma occurrences and run

the appropriate rule instances. Henceforth, the above rule will be called R_*IInne*. The other two article text rules are based on two Polish copular constructions:

- R_*Jest* – built around the verb *być* 'to be'
- and R_*To* – built around a predicative word (a quasi-verb) *to* 'to be'.

R_*To* shows a bias towards the identification of synonymic pairs, while R_*Jest* describes a typical *is_a* context. All three rules were applied to the Polish Wikipedia articles extracted in textual form and preprocessed, cf Sec. 4.

2.2 Heading Extraction Rules

Heading rules are focused on the utilisation of the structural information and were designed to be applied for the initial parts of articles. In their construction, it is assumed that Noun1 is equal to the article term and is located on the position 0. As JOSKIPI constraints work within the limits of one sentence, the work of heading rules is limited to a sentence including: term, variants and the first sentence of the description. JOSKIPI works on the level of morphologically annotated text. In order to make the link positions in text visible to JOSKIPI they were marked by the additional symbols: "$LB" and "$LE".

Six heading rules were manually constructed. An example of the rule R_*ToRodzaj_Lnk* extracting pairs: hyponym – hypernym is presented below:

```
and( in(cas[0],nom,acc),
     rlook(1,10,$I,in(orth[$I],"-","to" is,"$LB")),
     not(equal(orth[$I],"$LB")),
     rlook($+1I,15,$R,in(flex[$R],subst,depr,ger)),
     inter(cas[$R],nom,acc),
     in(base[$R],"rodzaj" kind,"typ" type,"podtyp" subtype,
             "dziedzina" domain,"forma" form, "sposób" manner),
     rlook($+1R,$+10R,$N,or( in(flex[$N],subst,depr,ger),
                                  equal(orth[$R],"$LB") )),
     in(flex[$N],subst,depr,ger),
     equal(base[$N],"Noun2") )
```

In the above rule, first, we check if the article term (Noun1) is in the nominative case (or accusative due to possible tagger errors) – a different case can signal the the article has not been written in a typical way. Next we are looking for an occurrence of a dash '–' or predicative quasi-verb *to* 'is' which should be found before the occurrence of the first link. The first link usually classifies the article term. The symbol *$LB* was added during preprocessing. Next, we look for the first noun that comes after the copular predicate (i.e. '–' or *to*). We expect it to be one of the nouns that signal a semantic relation: subordinate – superordinate. A list of such nouns is provided in the rule. The relation marker must be in the nominative (or accusative) case in this language construction. Finally we are looking for Noun2 which should follow the relation marker. Contexts in

which Noun2 occurs as a part of the link are excluded from R_*ToRodzaj*_*Lnk*, as being covered by one of the five other rules (`$LE` represents a link end):

R_Dash_*Lnk*: `Noun1`$_{case\in nom,acc}$ ... '`-`' ... `$LB Noun2`$_{case\in nom,acc}$ ... `$LE`
R_*ToElement*_*Lnk*: `Noun1`$_{case\in nom,acc}$ ... ('`-`' | `to`) ... (`element` *element* | `część` *part* | `fragment` *fragment*) ... `Noun2`
R_Dash_*Noun*: `Noun1`$_{case\in nom,acc}$ ... '`-`' ... `Noun2`$_{case\in nom,acc}$ – there is no beginning of a link before Noun2 and Noun2 is different that triggering words of the rules: R_*ToRodzaj*_*Lnk* and R_*ToElement*_*Lnk* (presented below).
R_After_Parantheses: `Noun1`$_{case\in nom,acc}$ ... '`(`' ... '`)`' ... `Noun2`$_{case\in nom,acc}$ – there is no link beginning before Noun2
R_In_Parantheses: `Noun1`$_{case\in nom,acc}$ '`(`' not(*verbs* and *punctuation marks*) `Noun2`$_{case\in nom,acc}$

R_Dash_*Lnk* reflects a relatively numerous article scheme in which the description starts with a dash after which a link to the superordinate term comes in a close distance. Rules: R_*ToRodzaj*_*Lnk* (presented earlier in details) and R_*ToElement*_*Lnk* describe less frequent, further specifications of this scheme. R_*ToRodzaj*_*Lnk* refers to lemmas directly signalling the hyponymy relation between the article term and the first link, while R_*ToElement*_*Lnk* identifies a narrow group of lemmas marking the meronymy relation (part of).

R_*ToElement*_*Lnk* extracts a limited number of instances, but their accuracy is relatively high. It is worth to be emphasised that the accuracy of rules similar to R_*ToElement*_*Lnk* is low when they are applied to a general corpus.

The other three rules explore information expressed by the article structure, not links, and the fact the first sentence is being processed. R_Dash_*Noun* and R_After_Parantheses are based on the similar assumption like R_Dash_*Lnk*, namely the first noun, if it is in the nominative case, represents a hypernym. All three rules were split, as their accuracy, and thus their reliability as knowledge sources, can be different, cf Sec. 4. R_In_Parantheses explores the fact, that synonyms of the term are often provided the article author in the variants.

3 Semi-automatic Wordnet Expansion

Pattern-based method extract relation instances with relatively high accuracy but limited coverage, cf Sec. 4, while methods of Distributional Semantics produce result for any pair of lemmas but the description is not focused on any single lexico-semantic relation. The idea of combing heterogeneous extraction methods in automated wordnet expansion became the basis for the Algorithm of Activation-area Attachment (henceforth AAA) and the WordnetWeaver system supporting semi-automatic plWordNet expansion. AAA is capable to utilise heterogeneous knowledge sources characterising lemma relations in suggesting senses for new lemmas. Knowledge sources extracted from Wikipedia are potentially valuable for AAA because of their expected high accuracy. AAA was presented in [10] for the description of its latest development see [11] in this volume. AAA is used in one of the two evaluation methods discussed in Sec. 4.

AAA introduces a notion of a *semantic fit* between two lemmas and also between a lemma and a synset (defining a sense). In the later phase synsets that fit the input lemma are grouped into *activation areas* describing the input lemma senses. Fit for a lemma pair depends on how well the given pair is supported by the knowledge sources, e.g. the pair was extracted by several reliable knowledge sources. The function *score* assigns a value to every lemma pair on the basis of weighted voting across all knowledge sources. Reliability of different knowledge sources is estimated by manual evaluation of the accuracy of the extracted pairs. The reliability values are a basis for weighted voting present in fit and *score*.

AAA works in two phases. During the first phase semantic fit between an input lemma x and each synset Y is computed on the basis of: the semantic fit between x and lemmas belonging to the synset Y and and, additionally, the semantic fit between x and synsets linked to Y by the hypernymy or hyponymy relation (up to several links). During the second phase, on the basis of the semantic fit between x and synsets, connected subgraphs of the hypernymic wordnet structure, called *activation areas*, are identified – an activation area includes only synsets for which semantic fit to x is above some threshold; each activation area is assigned its semantic fit values to x which is equal to the maximum of the semantic fit values between x and synsets of the area.

Proposed automatic evaluation checks how well can AAA reconstruct plWordNet in the lower parts of the hypernymy structure, see [11]. Three strategies for evaluating AAA's proposals were proposed: *All* , *One* (the highest-scoring attachment site), $Best_{P\geq 1}$ (one closest attachment site). In case of all suggestions based on strong fit located not further than 2 hypernymic links from the appropriate synset (the range of acceptable errors) the accuracy was 42.87%. For the same range, the accuracy of *One* was 67.99% and $Best_{P\geq 1}$ was 75.02%. Full results can be found in [10]. Almost half of the suggestions based on strong fit were near the correct place. The result for $Best_{P\geq 1}$ strategy shows support for a linguist: the number of words with at least one useful suggestion.

4 Evaluation

Evaluation was based on the Polish Wikipedia dumps from the 29th September 2009. Articles were extracted with the help of the *Wikipedia Extractor* tool[4]. A text corpus of the size 172 millions tokens consisting only of articles in textual form (without additional elements like info-boxes, categories etc.) was created. The corpus was pre-processed by TaKIPI [9] (the Polish morpho-syntactic tagger) and next link limits were marked by the "$LB" and "$LE" symbols. All test were performed on the set of noun lemmas (one word and multi-word) extracted from plWordNet version 1.1 (from July 2010). The list consisted of 39 039 noun lemmas including 6 957 multi-word lemmas. As there is no robust shallow parser for Polish, only multi-word lemmas that were covered by the list could be recognised in the corpus.

[4] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor#Related_Work

Two types of evaluation were performed: *direct* during which professional linguists verified manually the accuracy of relation instances extracted by the rules; and *indirect* — instance sets extracted by different rules were used as separate knowledge sources in automatic plWordNet 1.1 reconstruction – improved precision was expected.

During direct evaluation samples of the size 300 instances were randomly drawn from sets extracted by the rules. An instance was placed by linguists in one of the predefined classes:

1. *P*, *M* – proper linguistic hypernymy or meronymy, respectively, as in dictionaries or wordnets;
2. *PT*, *MT* – a form of conceptual hypernymy or meronymy supported by local context;[5]
3. *PG*, *MG* – correct, but given smart linguistic processing tools;[6]
4. *hypo*, *mero* – already added to plWordNet as hyponyms (meronyms);
5. *F* – not slotted into any other class – here treated as errors.

We write *allHypo* to denote the combination PT+P+PG+hypo (and *allMero* for MT+M+MG+mero) and *lingHypo* to denote P+PG+hypo (and *lingMero* for M+MG+mero). Instances extracted by the rules but of the opposite direction relation, i.e. were counted under the label *Hyper*. The results of the manual evaluation are given in Table 1.

Table 1. The accuracy [%] of rules according to the manual evaluation

Rule	No	allHypo	lingHypo	Hyper	allMero	lingMero
R_IInne	11984	**58,00%**	23,00%	0,33%	–	–
R_*To*	10564	33,33%	13,33%	8,33%	–	–
R_*Jest*	15309	39,33%	20,33%	3,33%	–	–
R_*ToRodzaj*_Lnk	678	64,67%	48,67%	2,33%	–	–
R_Dash_Lnk	3220	**65,33%**	33,67%	2,00%	–	–
R_*ToElement*_Lnk	401	11,00%	7,33%	1,67%	57,33%	49,00%
R_Dash_Noun	8617	48,33%	23,33%	0,67%	–	–
R_After_Para	4960	54,67%	17,00%	0,33%	–	–
R_In_Para	924	**59,46%**	48,31%	2,03%	–	–

The number of instances extracted by the heading rules seems to be small, but the rules' applicability was limited to these article terms that were covered by the lemmas extracted from plWordNet 1.1, i.e. less than 10 000 terms. Heading rules explore highly specific patterns and express relatively high accuracy. As construction and testing of the heading rules took only less than two man-work

[5] For instance, examples include a relation linking a named entity with its hypernym signalled by the head noun; a single-word lemma as a remote hypernym in place of the proper multi-word lemma; or hypernymy supported by a role played by some object in the particular local context.

[6] For example: wrong number (Carpathian Mountains versus mountain) or wrong – but semantically related – lemma (tournament versus compete).

days and they are applied to the subsequent versions of Wikipedia, the effort is profitable. The best result among article text rules was achieved by R_*IInne* which is a logical conjunction of a few patterns identifying language constructions representing a kind of enumeration. Among the heading rules the best precision was achieved by R_Dash_Lnk referring to the text under link.

The indirect evaluation was based on the application of the extracted instance sets as additional knowledge sources in automatic AAA-based plWordNet 1.1 reconstruction, cf Sec. 3. Test was performed on lexical units located in the hypernymic structure on the depth equal or greater 4. As a baseline we took the results obtained without the use of the manually written extraction rules. Next AAA results obtained with the data extracted from Wikipedia were compared to the baseline. For the acceptable error range of 2 hypernymic links and the evaluation strategy *All*, the accuracy was increased by 3.7% for the suggestions based on strong fit, 7.7% for weak fit, and 6.2% for the combined result. Concerning the limited size of the instance sets extracted from the Wikipedia, the achieved improvement is valuable. Moreover, we can extract from Wikipedia information about lemmas that are not frequent in the general corpus and for which it is difficult to acquire reliable knowledge sources.

5 Conclusions and Further Research

Our objective is the development of a semi-automatic tool supporting Polish wordnet expansion. As the plWordNet development gradually moves into the domains of infrequent (even in very large corpora) and mostly specific lemmas, the use of publicly available, semi-structured text corpora like Wikipedia is becoming more and more important. We proposed a set of rules extracting instances of the wordnet lexico-semantic relations from the Wikipedia articles. The rules were constructed manually, but for the cost of only less than one man-week of workload. Two groups of rules were developed. Rules of the first group work on articles treated as text documents. They can be applied to a general text corpus too, but they achieve much better results on the Wikipedia due to its informative content. Rules of the second group utilise structural information present in the Wikipedia articles, extract less relation instances, but mostly with better accuracy. The extracted knowledge sources improved the accuracy of the semi-automatic plWordNet expansion.

The most significant problem is the automatic recognition of new multi-word lemmas. Rules are now applied under the assumption that the list of multi-word lemmas is predefined and they all have been syntactically described that facilitates their recognition. We need to develop a method of the automatic acquisition of the linguistically described multi-word lemmas from Wikipedia in combination with a very large corpus. It should increase the coverage of the rules a lot. Moreover, the problem of the Proper Name recognition must be solved.

Acknowledgments. Work co-financed by the European Union within European Economy Programme project POIG.01.01.02-14-013/09.

References

1. Ahn, D., Jijkoun, V., Mishne, G., Müller, K., de Rijke, M., Schlobach, S.: Using Wikipedia at the TREC QA Track. In: Proceedings of TREC (2004)
2. Bunescu, R., Pasca, M.: Using Encyclopedic Knowledge for Named Entity Disambiguation. In: Proc. of the 11th Conf. of the European Chapter of ACL, pp. 9–16. ACL, Trento (2007)
3. Fellbaum, C. (ed.): WordNet – An Electronic Lexical Database. The MIT Press, Cambridge (1998)
4. Gabrilovich, E., Markovitch, S.: Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In: Proc. of the 21st National Conference on AI and the 18th Innovative Applications of AI Conference. AAAI Press, Boston (2006)
5. Gurevych, I., Müller, C., Zesch, T.: What to be? – Electronic Career Guidance Based on Semantic Relatedness. In: Proc. of the 45th Annual Meeting of ACL, Prague, Czech Republic, June 2007, pp. 1032–1039. ACL (2007)
6. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the Conference of the International Committee on Computational Linguistics, pp. 539–545. ACL, Nantes (1992)
7. Nastase, V., Strube, M.: Decoding Wikipedia Categories for Knowledge Acquisition. In: Proc. of the 23rd AAAI Conf., Chicago, pp. 1219–1224 (2008)
8. Nastase, V., Strube, M., Boerschinger, B., Zirn, C., Elghafari, A.: WikiNet: A Very Large Scale Multi-Lingual Concept Network. In: Proc. of LREC 2010, pp. 1015–1022 (2010)
9. Piasecki, M.: Polish tagger TaKIPI: Rule based construction and optimisation. Task Quarterly 11(1-2), 151–167 (2007)
10. Piasecki, M., Broda, B., Głąbska, M., Marcińczuk, M., Szpakowicz, S.: Semi-automatic expansion of polish wordnet based on activation-area attachment. In: Recent Advances in Intelligent Information Systems, pp. 247–260. EXIT (2009)
11. Piasecki, M., Kurc, R., Broda, B.: Heterogeneous knowledge sources in graph-based expansion of the polish wordnet. In: ACIIDS 2011. LNCS (LNAI), vol. 6591, pp. 307–317. Springer, Heidelberg (2011)
12. Piasecki, M., Radziszewski, A.: Morphosyntactic constraints in acquisition of linguistic knowledge for polish. In: Mykowiecka, A., Marciniak, M. (eds.) Aspects of Natural Language Processing (a festschrift for Prof. Leonard Bolc). LNCS, vol. 5070, pp. 163–190. Springer, Heidelberg (2009)
13. Piasecki, M., Szpakowicz, S., Broda, B.: A Wordnet from the Ground Up. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław (2009)
14. Ponzetto, S.P., Strube, M.: Deriving a large scale taxonomy from Wikipedia. In: Proc. of the 22nd Conference of the Advacement of Artificial Intelligence, Vancouver B.C., Canada, July 22-26, pp. 1440–1445 (2007)
15. Ruiz-Casado, M., Alfonseca, E., Castells, P.: Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. In: Szczepaniak, P.S., Kacprzyk, J., Niewiadomski, A. (eds.) AWIC 2005. LNCS (LNAI), vol. 3528, pp. 380–386. Springer, Heidelberg (2005)
16. Zesch, T., Gurevych, I., Mühlhäuser, M.: Comparing Wikipedia and German Wordnet by Evaluating Semantic Relatedness on Multiple Datasets. In: Proc. of NAACL-HLT 2007, pp. 205–208. ACL (2007)

Heterogeneous Knowledge Sources in Graph-Based Expansion of the Polish Wordnet

Maciej Piasecki, Roman Kurc, and Bartosz Broda

Institute of Informatics, Wrocław University of Technology, Poland
{maciej.piasecki,roman.kurc,bartosz.broda}@pwr.wroc.pl

Abstract. The paper presents an algorithm of automatic wordnet expansion on the basis of heterogeneous knowledge sources extracted from a large corpus. The algorithm is the reformulation of the algorithm used in the WordnetWeaver system in terms of the SOM model. Integration of knowledge sources is based on the weighted voting scheme. Several wordnet relations are explored to define attachment points for a new word. Influence of different knowledge sources on the algorithm performance was experimentally investigated. The new version presents better precision than the previous one.

1 Motivation and Related Works

We present a high-accuracy automatic method of identifying semantic similarity of lexical units (LUs)[1] on the basis of knowledge extracted from a corpus. If effective, such a method is directly applicable in semi-automatic wordnet expansion: it makes good suggestions for the linguist. We have tested our method on a new and growing wordnet for Polish – namely plWordNet.

Clustering algorithms applied to words described by semantic relatedness seem to be a natural way in automated wordnet construction. However, clustering usually produces a flat set of clusters. Changing such a set into a hierarchy poses two problems: how to identify the right shape of the tree and how to label higher levels of the cluster tree with the adequately general LUs. In any case, no automatic method can come up with a credible top portion of a wordnet hierarchy due to the highly abstract meanings of LUs occurring on these levels. Thus, we follow a semi-automatic wordnet expansion model: top levels of plWordNet's hypernymy hierarchy have been built manually and automatic methods produce useful suggestions of new LUs for inclusion in plWordNet (henceforth plWN).

Several projects have explored building an extended wordnet over an existing one. The advantage is the possibility of using the wordnet structure already in place, especially the hypernymy structure, as a knowledge source. [1] assigned a meaning representation extracted by Distributional Semantics methods to synsets, and treated the hypernymy structure so labelled as a kind of decision tree. [13] discusses a more radical decision-tree model with recursive upward propagation of meaning descriptions to the root. The description of the upper

[1] We will take a lexical unit, a little informally, to be a lexeme.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 307–316, 2011.

nodes represents the description of descendants. A synset's *semantic description* comprises LUs most similar to its LUs. Distributionally motivated similarity is based on co-occurrences of LUs in corpora. In evaluation on two subtrees from GermaNet, *Moebel* (*furniture*) (144 children) and *Bauwerk* (*building*) (902 children), the best accuracy of exact classification was 14% and 11%, comparable to that in [1].

[12] represents LU meaning by *semantic neighbours* – k most similar LUs. To attach a new LU is to find a site in the hypernymy hierarchy where its semantic neighbours are concentrated. To compute semantic similarity, each LU is first described by the co-occurrence, in a 15-word window, with 1000 most frequent one-word LUs. Evaluation was on the British National Corpus [2] and randomly selected common nouns, 200 each from three frequency ranges: > 1000, $[500, 1000]$ and < 500. Sites identified by the algorithm were compared with their exact hypernyms. The best accuracy of finding the direct hypernym, with no intervening nodes, producing exact reproduction of the wordnet (among 4 highest ranked labels) was 15% for $k = 3$ neighbours taken into account, but the overall classification (considering hypernyms located up to 10 links away from the suggested site in the wordnet structure) gave only 42.63%. The best accuracy of the overall classification was 82.06% for $k = 12$ neighbours considered but the accuracy of the exact placing (finding direct hypernyms) was reduced to 10.15%.

[11] cast the extension of wordnet hypernymy structure in terms of a probabilistic model. Attachment of new elements transforms the former structure **T** into a new structure **T**'. The most appropriate **T**' maximises the probability of the change in relation to the evidence at hand. [11] applied two sources of evidence for extending Princeton WordNet (henceforth PWN) [3]: a classifier-based algorithm of extracting hypernymic LU pairs using lexico-syntactic relations, and a proposed algorithm of extraction of (m, n)-*cousins*. In a "fine-grained" evaluation, [11] manually evaluated randomly selected 100 samples from the first n up to 20000 automatically added hypernymic links. The applied uniform size of the samples equal to 100 for $n > 1000$ was too small to ascribe the results of the evaluation to the whole sets with sufficient statistical confidence. The evaluators were asked: "is X a Y?", where $\langle X, Y \rangle$ is an added link. It is not clear in [11] whether only direct hyponym/hypernyms counted as positive. For each pair of nouns, the algorithm finally selects only the best hit for a new lexeme. The achieved precision of 84% for $n = 10000$ is high, but may be hard to compare with other approaches, including ours (presented in a while), because it is given only for the best hit and the basic criterion cited above is not precise.

2 The Main Ideas

In [7] an algorithm of wordnet expansion was proposed. The algorithm combined several knowledge sources extracted from large corpora. It was successfully applied to plWN expansion, however, the algorithm depended in several steps on heuristic procedures. In this paper we want to reformulate it in terms of unifying basic model and explore its several possible extensions.

We applied heterogeneous knowledge sources for a potential relation of a new LU x with a LU already in the wordnet and thus a relation with some synset. The sources were produced by several methods of extracting lexical-semantic relations (LSR) for Polish. The results of all extraction methods were transformed to sets of LU pairs $\langle x, y\rangle$ such that x and y are semantically related according to the given method and the corpora analysed. Five methods were used:

- a Measure of Semantic Relatedness (MSR) based on the Rank Weight Function (MSR_{RWF}) developed for Polish nouns by [9] (while it extracts closely related LUs with high accuracy, the extracted LU pairs belong to more LSRs than just the typical wordnet relations);
  two sets were produced using MSR_{RWF} – the set $MSRset(y, k)$ of the k units most related to y, where $k = 20$ was used in all experiments, and that set restricted to *bidirectional relations*:
  $MSR_{BiDir}(y, k) = \{y' : y' \in MSRset(y, k) \wedge y \in MSRset(y', k)\}$;
- a C_H classifier [8] which we use for post-filtering LU pairs produced by MSR_{RWF}. C_H is described below in more details. We generated one set by the classifier C_H applied to filtering $MSRset(y, k)$ from LUs not in hypernymy, meronymy or synonymy with y;
- three manually constructed lexico-syntactic patterns in the style of [4]: ⟨NP, NP, ... i inne (*and others*) NP⟩, ⟨NP jest (*is a*) NP⟩ and ⟨NP to (*is a*) NP⟩;
- six manually constructed patterns, similar to the above ones, but focused on mining Wikipedia.
- the *Estratto* method [6] in which lexico-morpho-syntactic patterns automatically extracted from corpus are used to extract LSR instances.

In general C_H was trained on LU pairs extracted from plWN: pairs of LUs associated by the synonymy, hyper/hyponymy (up to the distance of 2 links) and mero/holonymy relation — positive examples, and pairs of LUs which are not associated by any of the three relations — negative examples.

The accuracy of all methods in distinguishing *related LUs* (positive examples in C_H) is around 30%. This is in fact too low to support linguistic work effectively. But positively verified in [7] that by combining the results of different methods we can provide linguist with LU pairs of better accuracy.

Even though the combined method classifies LU pairs as related vs. non interesting well, it still cannot be used to distinguish among different wordnet relations. On the other hand when processing a new LU, all sites of its attachment to the wordnet structure are almost equally important. Thus, we proposed an automatic method of *activation-area attachment*: a new LU is attached to a small area in the hypernymy graph rather than to one synset, cf [10].

The method was inspired by the idea of learning in *Kohonen's networks* [5]. In a Kohonen network, a new learning example is used to modify not only the most similar neuron (the *winner*) but also neurons located close to it. The further the given neuron is from the winner, the smaller is its change caused by this learning example. The distance is measured by the number of links in the graph structure of the network. We aim at finding, for a new LU u, synsets for which we have the strongest evidence of LUs being in the close hypo/hypernym/synonym

relation with u. Ideally that synsets should include near synonyms of the new LU. We assume, that the intrinsic errors in data preclude certainty about the exact attachment point for u. Even if synset t appears to fit, we must consider the evidence for t's close surroundings in the hypernymy structure – all synsets located no further than some upper-bound distance d. We treat the evidence from the surroundings as less reliable than that for LUs in the central synset t. The influence of the context evidence decreases with the distance.

As the knowledge sources overlap only partially, so we will use them all in extending the wordnet. We assume that the subsequent methods explore different pieces of partial information available in corpora. We assume, too, that the application of many different methods allows the use of as much lexical-semantic information as possible. Different sources are not equally reliable; this can be estimated, e.g., by manual evaluation of the accuracy of the extracted pairs. To trust the different sources to a different degree, we introduce weighted voting.

3 Algorithm of Activation-Area Attachment

The algorithm, henceforth AAA, is based on the idea of a *semantic fit*: between two lemmas, as representing two LUs linked by a LSR, and between a lemma and a synset, as defining a LU. The fit is computed from all evidence found in corpora. Next we group synsets that fit the input lemma into the *activation-areas* defining the attachment areas and LUs. AAA was presented in [7]. Here we aim at improving it in three aspects: calculation of lemma-to-lemma and lemma-to-synset *score and fit*, and the *range of relations* utilised. Score and fit functions were made more robust and are expressed now in terms of weighted voting.

In [10] the fit was meant to provide heuristic combination of knowledge sources, some of them sparse but highly accurate. The score was crafted to prioritize fit results and supplement description for pairs with fit equal 0. Score is based on MSR which is defined for every lemma pair, while other sources are partial. We found that approach excessively complicated and we propose an unified measure. Every source is assigned some weight of its reliability, e.g. on the basis of its precision manually evaluated. Next, weighting voting scheme is applied in order to combine support provided by different knowledge sources. Different schemes cab be used to calculate the final value of lemma-to-lemma fit or score. Two functions were used: sigmoid and normalization.

A lemma-to-lemma fit is a function $fit : \mathbf{L} \times \mathbf{L} \rightarrow \langle 0.0, 1.0 \rangle$, where $\mathbf{L}$ is a set of lemmas, which is calculated as following (each $k \in K$ is assigned its weight):
Let $norm = \sum_{k \in K} weight(k)$, then
$fit(x, x') = (\sum_{k \in K \wedge (x,x') \in pairs(k)} weight(k))/norm$
Where $weight$ maps a relation name to a weight (a real number).

AAA consists of two phases: *Phase I*, when we find all synsets that fit a new lemma and *Phase II* when we group the found synsets into connected subgraphs – *activation areas*. **Phase I** is inspired by Self Organizing Map. Input data in SOM is a vector of a finite size. In general SOM requires a similarity function and a distance function. During learning we use similarity function to choose

neuron that is closest to some input. Then the-winner-takes-most strategy and the distance function are used to determine which neurons are adapted to new data. During application the similarity is used to locate data on the map.

In AAA the similarity is identified with the lemma-to-synset fit, which is based on the lemma-to-lemma fit and refers to the synset neighbourhood. Thus the whole surroundings is considered to be a vector describing the synset. The traversing order is bread first search. During **Phase I** we use the winner takes most strategy, as adding a new LU is equal to adapting synsets in close neighbourhood. Attachment of a new LU is similar to the SOM learning phase during which one piece of data is mapped to one SOM area. AAA seems to work in a opposite way: one lemma can be attached to a number of synsets. But, there is no multiple mapping by the same input data as one lemma can have many meanings. Therefore we perform disambiguation. There are several issues to be considered. First of all there is a question whether LUs in synsets should be treated as separate wordnet elements or only as parts of the synset. In the second case the size of the synset is important. The more elements from the synset are related to the new lemma, the better. Then question arises which elements of the neighbourhood should be taken into consideration, i.e., number of links, distance and relation types. Up to now only hypernymy/hyponymy was taken into consideration. However other relations are also valuable for linguists and covered by knowledge sources. Thus we need to take them into account. In **Phase II** the found synsets are grouped into connected subgraphs – activation areas. Let:

- x is a lemma, representing one or more LUs, to be added to the wordnet, S, S' – wordnet synsets, $|S|$ – synset size,
- $hMSR$ (set to 0.4) is the threshold defining highly reliable MSR values,
- $minMSR$ (set to 0.1) is the MSR value below which associations seem to be based on weak, accidental clues,
- $maxSens$ (set to 5) is the maximal number of presented attachment areas.
- $G\{relations\}$ wordnet projection with only links of the $relations$ preserved,
- d is the upper-bound distance defining the hypernymic close surroundings,
- $bfs(S, G)$ returns synset and surrounding synsets in breadth first order.
- $dist(S, S', G)$ is the number of $relations$ links between S and S',
- $dist\_modif(S, S')$ give value to modify fit of supporting synsets.

Phase I. Lemma-to-synset calculation

According to description of **Phase I** presented above functions of distance and similarity must be pre-defined.

1. $similarity(x, S) = \left(fit(x, S) + \sum_{S' \in bfs(S)} fit(x, S') * dist\_modif(S, S')\right)$
2. $strong\_fit(x, S) = \delta(1, similarity(x, S), |S|)$
3. $weak\_fit(x, S) = \delta(hMSR, similarity(x, S), |S|)$

where $\delta : \Re \times \Re \times \Re \rightarrow \{0, 1\}$, such that $\delta(h, n, s) = 1$ if and only if $(n \geq 1.5 * h$ and $s \leq 2)$ or $(n \geq 2 * h$ and $s > 2)$

Phase II. Identify lemma senses: areas and centres

1. $synAtt(x) = \{\mathbf{S} : \mathbf{S} = \{S : strong\_fit(x, S) \vee weak\_fit(x, S)\}$, and **S** is in G$\}$.

2. $maxScore(x, \mathbf{S}) = score\,(x, max_{S \in \mathbf{S}} score(x, S))$
3. Remove from $synAtt(y)$ all $\mathbf{S}$ such that $maxScore(x, \mathbf{S}) < minMSR$
4. Return the top $maxSens$ subgraphs from $synAtt(x)$ according to their $maxScore$ values; in each, mark the synset S with the highest $score(x, S)$.

The upper-bound distance d was set to 2 (**Phase I**), because we observed that no extraction method used can distinguish between direct hypernyms and close hypernyms. The δ function is a means of non-linear quantization from the strength of evidence to the decision. We require more *yes* votes for larger synsets, fewer votes for smaller synsets, but always more than one 'full vote' must be given – more than one synset member voting *yes*. The parameter h of δ relates the function to what is considered to be a 'full vote'. For weak fit, h is set to the value which signals a very high relatedness for the MSR used.

In **Phase II** we identify continuous areas (connected subgraphs) in the hypernymy graph, those which fit the new lemma x. For each area we find the local maximum of the score function for x. We keep all subgraphs with the synset of the maximum score based on the strong fit (the detail omitted above). From those based on the weak fit, we only keep the subgraphs above some heuristic threshold of the reliable MSR result. We also save for the linguists only a limited number of the best-scoring subgraphs ($maxSens = 5$ – it can be a parameter of the application). We do not want to clutter the screen with too many proposals. We do present all subgraphs with the top synset fit based on the strong fit.

4 Evaluation

Evaluation was based on the same automatic method as in [7,10]: AAA is first applied to the reconstruction of a wordnet from which a sample of LUs has been removed; the removed LUs are re-introduced with the help of AAA, and the precision of the process is attributed to AAA as its assessment. Laborious manual evaluation was not applied as its results seem to be correlated with those of the automatic one [7,10]. Two versions of plWN were used during evaluation: the so called plWN Core (plWNCore) version – built manually, the basis of evaluation in [7,10], and a more recent plWN version 1.0, cf [10]. We used the same sample of 1527 noun lemmas from lower parts of the hypernymy structure of plWNCore, as was applied in tests in [7]. In addition the second sample of 1658 noun lemmas was picked from plWN 1.0 in a slightly different way than the firs one: we selected LUs that were located in the lower parts of the hypernymy structure and were weakly connected with the whole structure of plWN 1.0. This sample represents less intensively developed parts of plWN. During one step of evaluation 10 noun lemmas are removed and re-inserted by AAA to the wordnet. If a LU was the only element of a synset, LU was removed, but the empty synset was left in the structure in order not to change the wordnet structure so significantly for the reconstruction step (a missing synset separates hypernymic subgraphs).

Reconstruction results were assessed according to the *single highest-scoring attachment site* strategy (called *One* in [10]) We follow this strategy for the

needs of comparison with other approaches. It is a little artificial in relation to the intended use WNW as tool, but it illustrates AAA behaviour when it works in a fully unsupervised way.

Two AAA variants was tested, as well, as different configurations of AAA parameters. As a baseline we used AAA version from [7], in which lemma-to-lemma fit was calculated according to a heuristic procedure. In all other experiments AAA version described in this paper, i.e., in which lemma-to-lemma fit is based on weighted voting discussed in Sec. 3, henceforth, it is called *weighted AAA*.

In addition to the basic AAA described in Sec. 3 we tested also a version in which not only hypernymic edges, but also edges of other relations were taken into account during calculation of the lemma-to-synset fit. The set of relations along which fit was collected in the synset context included: hypernymy/hyponymy, meronymy and antonymy. Each edge was assigned a weight depending on the lexico-semantic relation it represents. Weights were meant to influence fit values 'transmitted' through them. Hypernymy and hyponymy's factor was set to 1.0 while for the meronymy and holonymy it was experimentally set to 0.7. Factors were later used to modify the value of the score added to synset S by neighbour synset S'. The modifier for fit is calculated as follows: $mod(S, S') = \prod_{e' \in \{e: e \in shortest\_path(S,S')\}} relation\_weights(e')$, where e' is an edge, and $relation\_weights$ maps an edge to the factor.
The fits are then combined [3]: $Sm = \sum_{(s,f)\in(fits,factors)} s * \frac{1}{2^n} * f$.

Experiments were performed on samples of plWN lemmas for AAA variants discussed above. Influence of the values of AAA parameters were tested. Firstly, basic algorithms: the *baseline* and *weighted* AAA were applied on: plWNCore and plWN 1.0. They differ in size and the richness of LU description. We wanted to analyse the impact of the wordnet development on the AAA performance. In addition, we wanted to evaluate how our new weighted model handles test data introduced in [10]. Secondly, we prepared a new test set of different characteristic, i.e., the set consists of weakly connected LUs. This set is meant to show how well the AAA manages deficiencies in the wordnet structure. Thirdly, the AAA variant sensitive to several relations in the context was tested. Next, we analyses how the number of knowledge sources applied impacts the weighted AAA performance especially when they do not overlap.

Finally, the influence of knowledge sources based on the measure of semantic relatedness was investigated. This experiment shows how different types of knowledge sources contribute to the overall AAA performance. Source were divided in two groups: SIM consisting of sources based on the semantic relatedness and PAT – including pattern-based sources.

plWN has been continuously expanded by adding new LUs and links, as well, as correcting errors. Table 1 shows that AAA achieves better precision when applied to a larger and more coherent wordnet. Especially the weighted AAA gains from the richer structure of WN. On one hand hand, weighted AAA is more selective. But on the other hand, bigger network size balances filtering characteristic of the weighted AAA and results in more attachment suggestions. We can observe this on both: old and new test data sets. Interesting results obtained

Table 1. Tab1. core WN vs. WN 1.0: baseline, weighted (one)

	wordnet core - old test sample						wordnet 1.0 - old test sample					
	baseline			weighted			baseline			weighted		
L	S	W	S+W	S	W	S+W	S	W	S+W	S	W	S+W
0&1	40,6%	13,2%	33,3%	13,3%	8,3%	8,6%	46,4%	12,5%	35,4%	64,5%	32,8%	39,1%
2	59,6%	28,4%	51,3%	23,3%	12,9%	13,7%	65,2%	31,5%	54,3%	81,0%	51,7%	57,5%
3	69,3%	35,2%	60,1%	44,4%	24,3%	25,7%	73,0%	39,3%	62,1%	87,9%	60,7%	66,1%
4	76,4%	43,5%	67,6%	55,6%	34,2%	35,6%	79,6%	47,1%	69,1%	92,1%	68,3%	73,0%
5	81,9%	51,6%	73,8%	72,2%	43,0%	45,0%	83,6%	53,9%	74,0%	93,1%	73,6%	77,5%
Hits:	1080	395	1475	90	1206	1296	987	473	1460	290	1171	1461

	wordnet 1.0 - new test sample					
	baseline			weighted		
L	S	W	S+W	S	W	S+W
0&1	26,4%	8,8%	19,0%	37,0%	16,7%	19,2%
2	32,3%	12,1%	23,8%	46,3%	21,3%	24,4%
3	35,7%	14,0%	26,6%	50,9%	23,8%	27,2%
4	36,9%	14,8%	27,6%	51,9%	24,7%	28,1%
5	38,3%	15,4%	28,7%	54,6%	25,9%	29,5%
Hits	504	364	868	108	760	868

Table 2. Tab2 WN1.0(one): baseline, weighted, relations(mero, hipo, anto)

	baseline			weighted			relations		
L	S	W	S+W	S	W	S+W	S	W	S+W
0&1	26,39%	8,79%	19,01%	37,04%	16,71%	19,24%	44,64%	17,02%	20,60%
2	32,34%	12,09%	23,85%	46,30%	21,32%	24,42%	51,79%	21,41%	25,35%
3	35,71%	14,01%	26,61%	50,93%	23,82%	27,19%	55,36%	24,07%	28,13%
4	36,90%	14,84%	27,65%	51,85%	24,74%	28,11%		25,13%	29,05%
5	38,29%	15,38%	28,69%	54,63%	25,92%	29,49%	59,82%	26,06%	30,44%

on the new test data proved that weighted AAA can handle well described areas as well as sparse ones.

For a larger wordnet weighted AAA is better than baseline, cf Table 2. It is, however, restricted by available types of links between synsets. In the second phase (lemma to synset), when neighbourhood is important, additional links improve support for a target synset. This can be observed in the case of the last method presented in Table 2. Reliability of different relation links for AAA must be still investigated. However, as far as hypernymy is very informative, meronymy or antonymy may be used only to the limited extent.

Table 3 shows that increasing overlap between sources of evidence results in better precision without the loss in the number of suggestions. However when overlap is lower, i.e., baseline AAA resorts to the minority voting (weak fit), weighted AAA turns out to be much more selective.

Last set of experiments, cf Table 4, shows how different types of knowledge sources influence results. In general sources based on similarity guarantees large number of suggestions. At the same time other sources helps only in case of the

Table 3. Influence of evidence number[basic, wiki op syntactic, wiki op semantic, head] baseline, weighted

	weighted			w+op_synt			w+syn_sem		
L	S	W	S+W	S	W	S+W	S	W	S+W
0&1	31,4%	16,1%	20,5%	37,0%	16,7%	19,2%	45,5%	24,6%	25,0%
2	40,0%	20,2%	25,8%	46,3%	21,3%	24,4%	54,5%	32,0%	32,4%
3	43,9%	23,0%	29,0%	50,9%	23,8%	27,2%		35,1%	35,5%
4	46,7%	24,3%	30,6%	51,9%	24,7%	28,1%	63,6%	37,5%	38,0%
5	51,0%	26,0%	33,1%	54,6%	25,9%	29,5%	72,7%	40,8%	41,4%
hits total no:	255	639	894	108	760	868	11	578	589

Table 4. Distributional evidence vs. pattern based evidence

	baseline			sim only			no sim		
L	S	W	S+W	S	W	S+W	S	W	S+W
0&1	37,0%	16,7%	19,2%	31,0%	12,1%	19,1%	40,0%	5,7%	7,9%
2	46,3%	21,3%	24,4%	39,0%	15,2%	24,0%	50,0%	7,7%	10,4%
3	50,9%	23,8%	27,2%	42,7%	17,2%	26,6%	56,7%	9,3%	12,3%
4	51,9%	24,7%	28,1%	44,3%	18,5%	28,0%	63,3%	10,4%	13,8%
5	54,6%	25,9%	29,5%	45,8%	19,4%	29,1%	70,0%	12,0%	15,7%
hits total no:	108	760	868	323	552	875	30	441	471

strong match. They cannot improve weak match because of their low quality and weak match represents situations when minority of sources votes.

5 Conclusions and Further Research

We presented a SOM inspired reformulation of the AAA. Secondly, several profitable AAA extensions were discussed, including the treatment of the increasing number of knowledge sources and the utilisation of several lexico-semantic relations present in the wordnet.

In general larger number of knowledge sources improves AAA precision. On the other hand if sources do not overlap then the AAA produces less suggestions due to the lack of synsets supported by the strong fit. Similarly, if we use larger number of relations, i.e., not only hyper/hyponymy links but also meronymy and antonymy ones, AAA comes up with better attachment suggestions for new lemmas. Moreover, we showed that expansion strategies should be adjusted accordingly to the maturity level of the wordnet. Problems of the insufficient description of particular synsets and different importance of relations for AAA must be further investigated. A comparison to other works on automatic wordnet expansion can be misleading. Our primary goal was to construct a tool that facilitates and streamlines the linguists' work. Still, even if we compare our automatic evaluation with the results in [12] during comparable tests on the Princeton WordNet, our results seem to be better. For example, we had 19.2% to 25%(new test data) and 39.1%(old test data) for the highest-scored proposal

(Table 1), while [12] reports a 15% best accuracy for a "correct classifications in the top 4 places" (among the top 4 highest proposals). Our similar result for the top 5 proposals is even higher, 21% up to 41% (new test data) and 77,5% for the old test data. The best results in [1,13], also at the level of about 15%, were achieved in tests on a much smaller scale. [13] also performed tests only in two selected domains. The algorithm of [11], contrary to ours, can only be applied to probabilistic evidence.

Acknowledgements. Work financed by the Polish Ministry of Education and Science, Project N N516 068637.

References

1. Alfonseca, E., Manandhar, S.: Extending a lexical ontology by a combination of distributional semantics signatures. In: Gómez-Pérez, A., Benjamins, V.R. (eds.) EKAW 2002. LNCS (LNAI), vol. 2473, pp. 1–7. Springer, Heidelberg (2002)
2. BNC. The British National Corpus, version 2 (BNC World) distributed by Oxford University Computing Services on behalf of the BNC Consortium (2001), http://www.natcorp.ox.ac.uk/
3. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
4. Hearst, M.A.: Automated Discovery of WordNet Relations. In: Fellbaum [3], pp. 131–153
5. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics 43(1), 59–69 (1982)
6. Kurc, R., Piasecki, M.: Automatic acquisition of wordnet relations by the morpho-syntactic patterns extracted from the corpora in polish. In: 3rd International Symposium Advances in Artificial Intelligence and Applications (2008)
7. Piasecki, M., Broda, B., Głąbska, M., Marcińczuk, M., Szpakowicz, S.: Semi-automatic expansion of polish wordnet based on activation-area attachment. In: Recent Advances in Intelligent Information Systems, pp. 247–260. EXIT (2009)
8. Piasecki, M., Szpakowicz, S., Marcinczuk, M., Broda, B.: Classification-based filtering of semantic relatedness in hypernymy extraction. In: Nordström, B., Ranta, A. (eds.) GoTAL 2008. LNCS (LNAI), vol. 5221, pp. 393–404. Springer, Heidelberg (2008)
9. Piasecki, M., Szpakowicz, S., Broda, B.: Automatic selection of heterogeneous syntactic features in semantic similarity of Polish nouns. In: Matoušek, V., Mautner, P. (eds.) TSD 2007. LNCS (LNAI), vol. 4629, pp. 99–106. Springer, Heidelberg (2007)
10. Piasecki, M., Szpakowicz, S., Broda, B.: A Wordnet from the Ground Up. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław (2009)
11. Snow, R., Jurafsky, D., Ng, A.Y.: Semantic taxonomy induction from heterogenous evidence. In: COLING 2006 (2006)
12. Widdows, D.: Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In: Proc. of NAACL-HLT, pp. 197–204 (2003)
13. Witschel, H.F.: Using decision trees and text mining techniques for extending taxonomies. In: Proc. of Learning and Extending Lexical Ontologies by using Machine Learning Methods, Workshop at ICML 2005 (2005)

Improving Arabic Part-of-Speech Tagging through Morphological Analysis

Mohammed Albared, Nazlia Omar, and Mohd. Juzaiddin Ab Aziz

University Kebangsaan Malaysia, Faculty of Information Science and Technology,
Department of Computer Science
mohammed_albared@yahoo.com, {no,din}@ftsm.ukm.my
http://www.ukm.my

Abstract. This paper describes our newly-developed second order hidden Markov model part-of-speech tagging system specially designed to tag Arabic texts using small training data. The tagger achieves encouraging results. In addition, the paper also presents a hybrid tagging architecture for Arabic, in which our tagger augmented with a weighted morphological analyzer. Finally, we compare the tagger results both standalone and utilizing a highly coverage morphological analyzer. Experimental results are presented and discussed using small training corpus. The experiments show that the best proposed hybrid architecture significantly improves unknown words POS tagging accuracy. 96.6% precision rates are obtained when unknown words occur in the test set.

Keywords: Arabic languages, Hidden Markov model, Morphological analysis, Unknown words.

1 Introduction

Part of speech (POS) tagging systems are programs that assign to each word of an input sentence its most probable POS in context. POS tagging is an essential pre-processing task for most natural langauge processing applications. The task of POS tagging is very difficult due to two main reasons. First, the POS ambiguity, where a word may have a set of possible tags. Second, the existing of unknown words, words that appears in the test data and do not appear in the training data. The problem of unknown words is the main problem in POS tagging [1,2]. Actually, the size of this problem is proportional to many factors such as: size, genre and the quality of the training data.

Most of the current POS taggers require huge amounts of of manually POS annotated data for learning the underling language model. For Arabic,such resources are rare or not freely open for research.The problem of the lack of langauge resources i.e. annotated training corpora is a general problem even for well-studied langauges [3]. The available huge amounts of pre-tagged texts are only suitable for some domains. It well known that taggers trained with the existing hand tagged corpora perform quite poorly in new domains due to unknown words problems [3]. According to Giesbrecht and Evert [4], "the reported

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 317–326, 2011.

state-of-the-art POS taggers tagging accuracies of 97% and 98% have to be understood as optimistic estimates, representing an ideal case for machine learning approaches. In a real-life scenario, accuracies drop below 93%, making the taggers unsuitable for fully automatic processing". As a matter of fact, the situation becomes worse when dealing with resource-scarce languages for which small annotated material is not easily available. All this means that the huge efforts done either for tagging huge amounts of text or for developing complex rule systems are only of limited use[3].

However, our work goes in another direction. The main aim of our work is to develop an accurate POS-tagger for Arabic, a resource-scarce language, using only a small amount of training data. This is can be done by carefully augmenting an Arabic statistical POS tagger, which is tarined using only small manually tagged data, with an external large-lexicon Arabic morphological analyzer. Actually, After obtaining satisfactory results in our preliminary experiments[5] by using our own developed Arabic Bigram HMM tagger,we decided to extend this work to reach a higher level of accuracy. In this work, we focus on two main directions for further enhancements: First, developing a trigram HMM tagger, especially designed and implemented for Arabic in which unknown words are handled using the linear interpolation of both word suffix probability and word prefix probability [5]. Second, augmenting the developed tagger with the external morphological analyzer, to overcome the coverage problem and to better handle unknown words. However, the results show that a high POS tagging precision for Arabic text can be acheived using only small hand annotated training data. So, we argue that our technique successfully overcomes the huge training corpus problem.

The rest of the paper is organized as follows: first, Section 2 discusses related works. Section 3 will give a brief description of the data used in the experiments. Section 4 describes our Arabic Trigram HMM tagger. In section 5,we will show how stochastic tagging can be aided by information from the morphological analyzer. Section 6 gives Experimental results. Finally, conclusions and future work appear in Section 7.

2 Related Work

POS tagging is a well-studied natural language problem. There are several approaches to solve the POS disambiguation problem. These techniques can be classified to machine-learning approaches or rule based approaches. Rule based taggers[6,7] try to assign POS tag to each word based on manually written rules by linguists. On the other hand, machine learning techniques[8,9,10,11] need annotated corpora to pick the most probable POS tag for each word. However, POS tagging is often regarded as a *solved task*. The published overall tagging accuracies of most of these techniques are around 97%. The reported tagging accuracies are achieved when both training data and testing data are from the same genre. However, in the real-life scenario, POS taggers are still unsuitable for fully automatic processing [2],[4]. Giesbrecht and Evert[4]showed that the tagging accuracies of five state-of- the-art POS taggers on German text drop below 93% under real-life conditions.

Previous work on POS tagging has been utilized different kind of features to tackle unknown word POS guessing. These features are mainly based on word substring information, word context information and/or global information. Recently, there has been a lot of interest in Arabic POS tagging. Several tagging methods and tools have been adopted for Arabic language [12,13,14]. For more details about Arabic POS tagging and also about POS tagging techniques, see [15]. Most of the Arabic work has been tested under the closed vocabulary assumption. With this assumption the tagging result is calculated without unknown words. However, the reported tagging accuracies of unknown words on Arabic work are very low compared to what has been achieved in well-studied languages [12]. On the contrary with these works, our aim is to develop an accurate Arabic tagger using only a small training data. Under this condition, the main problem that we focus to solve is the POS guessing of unknown words. To do so, we integrate an external lexicon-based morphological analyzer to improve the performance of our trigram HMM tagger.

3 Data and Tagset

The training data consists of about 22800 manually annotated words. The same training corpus is used for all our tagging schemes. The training data has been annotated using a tag set consisting of 24 grammatical tags, see Table 1. All the models have been tested on a set of randomly drawn 6540 words distinct from the training corpus. It has been noted that 10.7% words in the testing text are unknown with respect to the training set. Both training data and test contain two types of Arabic texts: Traditional Arabic text and Modern Standard Arabic.

Table 1. The Arabic POS Tagset used in annotating our corpus

POS Tag	Label	POS Tag	Label
Conjunction	**CC**	Broken Plural Noun	**BPN**
Number	**CD**	Possessive Pronoun	**POSS_PRON**
Adverb	**ADV**	Imperfective Verb	**VBP**
Particle	**PART**	Non Inflected Verb	**NIV**
Imperative Verb	**IV**	Relative Pronoun	**REL_PRON**
Foreign Word	**FOREIGN**	Interjection	**INTERJ**
Perfect Verb	**PV**	Interrogative Particle	**INTER_PART**
Passive Verb	**PSSV**	Interrogative Adverb	**INTER_ADV**
Preposition	**PREP**	Demonstrative Pronoun	**DEM_ PROP**
Adjective	**ADJ**	Punctuation	**PUNC**
Singular Noun	**SN**	Proper Noun	**NOUN_PROP**
Sound Plural Noun	**SPN**	Personal Pronoun	**PRON**

4 Our Approach: The Trigram HMM Tagger

Hidden Markov Model (HMM) is a well known statistical model in natural langauge processing. HMM POS tagger can predict the tag of the current word

given the tags of one previous word (bi-gram) or two previous words (trigram). HMM POS tagger has been extensively evaluated for English and some other languages. In HMM, the POS problem can be defined as the finding the best tag sequence $t_1^n = t_1 \ldots . t_n$ given the word sequence $w_1^n = w_1 ..... w_n$. The label sequence t_1^n generated by the model is the one which has highest probability among all the possible label sequences for the input word sequence. The trigram HMM can be formally expressed as:

$$t_1^n = \arg\max_{t_1^n} \prod_{i=1}^{n} p\left(t_i | t_{i-1}, t_{i-2}\right) . p\left(w_i | t_i\right) \tag{1}$$

The first parameter $p(w_i|t_i)$ is a known as the lexical probability and second parameter $p\left(t_i|t_{i-1}, t_{i-2}\right)$ is known as the transition probability. These two model parameters are estimated from an annotated corpus by Maximum Likelihood Estimation (MLE), which is derived from the relative frequencies.

Many staistical taggers have been made available to the community for scientific purposes during the last years for example, the TnT tagger [10], HunPos[16] and ACOPOST[17]. However, NLP tools developed for western languages are not easily adaptable to Arabic due to the specific features of the Arabic language[18]. Due to this, we have chosen to develop our own Arabic HMM taggers. This enable us to alter its operation methods whenever required, to integrate it with Arabic natural language processing and also to deal with Arabic specific charctersstics such as the non concatinative nature of arabic word, its writing system and the lack of cpitalization information. Our tagger, named ASPOST(Arabic Statistical POS Tagger) tagger has been implemented using Vis C#. Up to date,we implemented the first and the second order hidden Markov model tagging paradigms (bigram and trigram taggers). The bigram version has been fully described in[5]. In this paper, we are focusing on the trigram version . Our implementation of the basline trigram HMM tagger follows ACOPOST trigram model T3, which in its turn follows the TnT tagger. However for smoothing, ASPOST uses the linear interpolation of unigram, bigram and trigram transition probabilities smoothing method. For unknown word, three methods have been implemented. The first method is TnT suffix guessing algorithm[10]. The second method is the prefix guessing algorithm[5]. The third method is the linear interpolation of both prefix guessing algorithm and suffix guessing algorithm for unknown words[5].

Considering the tagging accuracy as the percentage of correctly assigned tags, we have been evaluated the performance of ASPOST from two different aspects: (1) the overall accuracy (taking into account all tokens in the test corpus) and (2) the accuracy of known words and unknown words. Table 2 summarizes the results of experiments with the prefix guessing algorithem, the suffix guessing algorithm and the linear interpolation guessing algorithm. As in our pervious experiments with the bigram version of HHM,the linear interpolation guessing algorithm, which combine information from both suffix and prefix, gives a considerable rise in accuracy compared to the suffix guessing method, which proved to be a good indicator for unknown word POS guessing in other langauges English

Table 2. The performance of the basic models

Model	% of unknown word	known word	Unknown acc.	The overall acc.
HMM (TnT Suffix guessing algorithm)	10.7	98.15	66.3	94.5
HMM (Prefix guessing algorithm)	10.7	98.14	56.4	93.7
HMM(Prefix+suffix guessing algorithm)	10.7	98.22	71.4	95.3

and German[10]. This is due to the large number of novel phenomena exhibited in the morphology of Arabic language. However, the achieved results for unknown words are still far away from what are achieved in other languages. It is well known that if the tagger encounters known words, error rates for tagging these words decrease substantially compared to tagging words that are unknown. In fact, most of the machines learning algorithms achieve very high accuracy for known words. Improving taggers accuracy therefore often means implementing an advanced method of guessing POS tags of unknown words [19]. In all experiments in this work, we used a small training corpus with only 22800 words. With this small training corpus, the learned lexicon is small and incomplete. The size of this lexicon is 5616 words. This will leads to large number of unknown words when tagging a new text. Under the closed vocabulary assumption (assuming all words are known), the tagger achieves a tagging accuracy over 98.2%. But when, in a more realistic approach, we use an open vocabulary assumption, the accuracy of the tagger is dropped to 93.7%. In the next section, we will discuss our effort to improve the accuracy of the unknown word. We integrate the weighted output of the exteranal morphological analyzer with the linear interpolation guessing algorithm.

5 Morphological Information Integration

In order to improve our tagger accuracy, we integrate the output of a morphological analyzer with our tagger, see Figure 1. As shown in the previous section, the unknown word guessing algorithms which are based on the word internal information do not work well with unknown words POS guessing in Arabic language. In our opinion, this is due to: 1) data sparseness 2) affixes ambiguity 3) the non-Concatenative nature of Arabic word. Arabic words suffixes are ambiguous. Most of the time, Arabic words which are derived from the same root share the same suffix even if they have different POS. For example, Arabic words "كتب","مكتب", "يكتب" and "كتب",which are drived from the same root "كتب", have different POS, PAV,SN,PRV and BP, respectively. In addition, the most frequent suffixes in Arabic are homographs (eg. the ending "ون" can be attached to a verb "يعملون", an adjective "مجتهدون" or a plural noun "مهندسون").

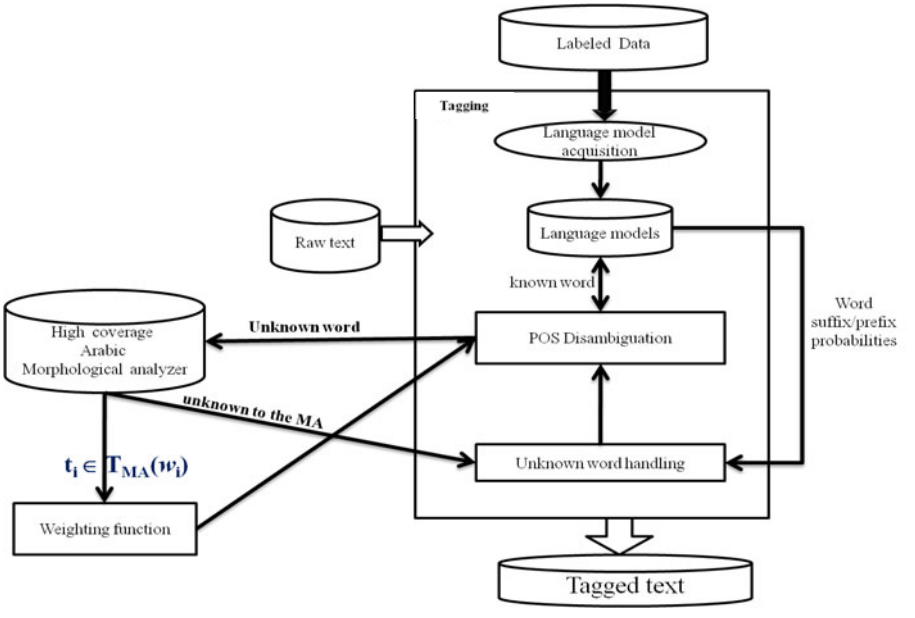


Fig. 1. The Integrated Tagging Architecure

The Morphological Analyzer(MA) is a function that inputs a word w and outputs the set of all its possible POS tags. Note that the size of tags produced by the MA is much smaller than the size of the tagset. Thus, we have a restricted choice of tags as well as tag sequences for a given sentence. Since the correct tag t for w is always in the MA output tags (assuming here that the MA is complete), it is always possible to find out the correct tag sequence for a sentence even after applying the morphological restriction. Since the MA does not assign probabilities to the tags, we address this problem by using a suitable weighting function which give a weight for each tag of the tags proposed by the MA given the word. We have been desigened and implemented a lexicon based Arabic MA. In our system, we utilized several existing Arabic databses [20] and construct other Arabic lexicons for named entities and functin words, to increaze the MA coverage. Due to the differences between used tagsets,a mapping function have been implemented to map each tag in the existing resources to its equivalent in our tagset.

Using the information from the MA, the new lexical models for unknown words are estimated using the following method:

1. If the word w_i is not in the training corpus and it is known to the MA, the MA output the set of possible tags. Then, the lexical probabilities are estimated as follow:
   (a) The tags probability distribution is calculated by using one of the following weighting functions:

i. Assuming a uniform distribution among all the tags proposed by the MA $p(t_j|w_i) = 1\backslash(|\mathrm{T_{MA}}(w_\mathrm{i})|)$, where $\mathrm{T_{MA}}$ is the set of all possible tags of w_i which is proposed by the MA.
ii. For each tag t_j in the $\mathrm{T_{MA}}$, we give it a weight proportional to its occurrence in the training data as in the following equation:

$$p(t_j|w_i) = \frac{p(t_j)}{\sum\limits_{t_m \in \mathrm{T_{MA}}(w_\mathrm{i})} p(t_m)} \tag{2}$$

Where $p(t_i)$ and $p(t_m)$ are calculated from the training data.

(b) Then, the lexical probabilities $P'(w_i|t_j)$ for each tag $t_j$in the $\mathrm{T_{MA}}$ are calculated using the Bayesian inversion.
(c) The final lexical praoabilities are estimated using one of the following steps:
i. The above lexical probabilities $P'(w_i|t_j)$ are used if the word is known to the MA. However, if the word is unknown to the MA, the lexical probabilities of the word $P'(w_i|t_j)$ are calculated using the following formula[5]:

$$P(w_i|t_j) = \lambda P\left(\mathrm{suffix}(w_\mathrm{i})|t_j\right) + (1-\lambda)P(\mathrm{prefix}(w_\mathrm{i})|t_j) \tag{3}$$

where λ is an interpolation factor and $0 \leq \lambda \leq 1$.
ii. The lexical probabilities are estimated using the following formula:

$$p(w_i|t_j) = \alpha P'(w_i|t_j) + (1-\alpha)(\lambda P\left(\mathrm{suffix}(w_\mathrm{i})|t_j\right) + (1-\lambda)P(\mathrm{prefix}(w_\mathrm{i})|t_j)) \tag{4}$$

where α and λ are interpolation factors,$0 \leq \alpha \leq 1$ and $0 \leq \lambda \leq 1$.

6 Experiments and Results

In this section, we study the performance of the Arabic trigram HMM POS tagger aided by the MA. The same training data has been used to estimate the parameters for the HMM tagger. Moreover, the same test set has been used for the evaluation. Several methods to incorporate the weighted output of the MA have been investigated. We have a total of five models. Results are summarized in Table 3. As Table 3 makes clear, when the MA is plugged into the system, a significant increase in the unknown word POS tagging accuracy and consequently in the overall accuracy is clearly noticeable. We find that the use of the MA information with word affixes improves the accuracy with respect to the basic models (see Table 2 and Table 3). It is interesting to see that the lexical models, in which the output(possible tags) of the MA are weighted using uniform distribution, perform better than those, in which the possible tags are weighted proportionally to thier occurrence in the training data. As we have noted already the use of morphological analyzer improves the accuracy of the POS tagger. But what is significant to note is that the percentage of improvement is higher when we linearly interpolate the weighted output of the MA with

Table 3. The performance of the ASPOST aided by the morphological analyzer

Model	% of unknown word	Unknown acc.	The overall acc.
HMM+ (MA Uniform Probability)	10.7	75.6	95.5
HMM+ (MA Proportional Probability)	10.7	73	95.5
HMM+ (MA Uniform Probability)+ (word suffix+ word prefix)	10.7	84	96.6
HMM+ (MA Proportional Probability) +(word suffix+ word prefix)	10.7	80	96.2

the linear interpolation of both word suffix probability and word prefix probability. The results of the experiment, row 3 in Table 3, which linearly interpolate the linear interpolation of both word suffix probability and word prefix probability with the uniformly weighted output of the MA, show a considerable increase over all other approaches. In general, we find that integration of the weighted MA with stochastic tagger improves the tagging accuracy. It can also be seen that the results are comparable and outperformed the best results for Arabic reported in [12] and [21].

7 Conclusion

A Trigram HMM part-of-speech tagger for Arabic language based on small resources has been built. Three lexical models based on internal features of Arabic words to handle unknown words have been proposed and evaluated. In addition, a highly-coverage lexicon-based external morphological analyzer has been integrated with the trigram tagger to overcome the coverage problem and to better handling unknown words. The main advantage of our Arabic tagger is its ability to obtain competitive results using a small training data. So, we argue that our technique successfully overcomes the need for huge training corpus problem. Our future direction is to improve the unknown word predictor. This improvement can be done through several steps. First, we intend to increase the size of training corpus from small sized and to increase the coverage of the MA. Second, we plan to develop more effective weighting functions for the output of the MA. Another future direction is to set up a new test set to re-evaluate the performance of the tagger. The new test set will include annotated data from multiple domains.

References

1. Nakagawa, T.: Multilingual word segmentation and part-of-speech tagging: a machine learning approach incorporating diverse features. PhD Thesis, Nara Institute of Science and Technology, Japan (2006)
2. Fischl, W.: Part of Speech Tagging - A solved problem? Unpublished report, Center for Integrative Bioinformatics Vienna, CIBIV (2009)
3. Marques, N.C., Pereira Lopes, J.G.: Tagging with Small Training Corpora. In: Hoffmann, F., Adams, N., Fisher, D., Guimarães, G., Hand, D.J. (eds.) IDA 2001. LNCS, vol. 2189, pp. 63–72. Springer, Heidelberg (2001)
4. Giesbrecht, E., Stefan, E.: Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus. In: Proceedings of the 5th Web as Corpus Workshop (WAC5), Donostia (2009)
5. Albared, M., Omar, N., Ab Aziz, M.J.: Automatic Part of Speech Tagging for Arabic: An Experiment Using Bigram Hidden Markov Model. In: Yu, J., Greco, S., Lingras, P., Wang, G., Skowron, A. (eds.) RSKT 2010. LNCS, vol. 6401, pp. 361–370. Springer, Heidelberg (2010)
6. Karlsson, F., Voutilainen, A., Heikkila, J., Anttila, A.: Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text. Mouton de Gruyter, Berlin (2010)
7. Samuelsson, C., Voutilainen, A.: Comparing a linguistic and a stochastic tagger. In: Proceedings of the eighth conference on European Chapter of the Association for Computational Linguistics (EACL), Madrid, Spain, pp. 246–253 (1997)
8. Gimenez, J., Marquez, L.: SVM tool: A general POS tagger generator based on support vector machines. In: Proceedings of the Fourth Conference on Language Resources and Evaluation, Lisbon, Portugal (2004)
9. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the International Conference on Machine Learning, MA, USA (2001)
10. Brants, T.: TnT: A statistical part-of-speech tagger. In: Proceedings of the 6th Conference on Applied Natural Language Processing, Seattle, WA, USA (2000)
11. Thede, S., Harper, M.: A second-order Hidden Markov Model for part-of-speech tagging. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics (1999)
12. Emad, M., Sandra, K.: Arabic part of speech tagging. In: Proceedings of LREC, Valetta, Malta (2010)
13. Al Shamsi, F., Guessoum, A.: A hidden Markov model-based POS tagger for Arabic. In: Proceeding of the 8th International Conference on the Statistical Analysis of Textual Data, France, pp. 31–42 (2006)
14. El Hadj, Y., Al-Sughayeir, I., Al-Ansari, A.: Arabic Part-Of-Speech Tagging using the Sentence Structure. In: Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt (2009)
15. Albared, M., Omar, N., Ab Aziz, M.J.: Arabic Part of Speech Disambiguation. International Review on Computers and Software 4(5), 517–532 (2009)
16. Halacsy, P., Kornai, A., Oravecz, C.: HunPos - an open source trigram tagger. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume. Proceedings of the Demo and Poster Sessions, Prague, Czech Republic, pp. 209–212 (2007)

17. Schroeder, I.: A case study in part-of-speech tagging using the ICOPOST toolkit. Technical report, Department of Computer Science, University of Hamburg (2002)
18. Farghaly, A., Shaalan., K.: Arabic Natural Language Processing: Challenges and Solutions, vol. 8(4) (2009), doi:10.1145/1644879.1644881
19. Agi, Ž., Tadi, M., Dovedan, Z.: Improving Part-of-Speech Tagging Accuracy for Croatian by Morphological Analysis. Informatica 32(4), 445–451 (2008)
20. Buckwalter, T.: Buckwalter Arabic morphological analyzer version 2.0 (2004)
21. AlGahtani, S., Black, W., McNaught, J.: Arabic Part-Of-Speech Tagging using Transformation-Based Learning. In: Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt (2009)

Educational Services Recommendation Using Social Network Approach

Krzysztof Juszczyszyn and Agnieszka Prusiewicz

Institute of Computer Science, Wrocław University of Technology
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
{krzysztof.juszczyszyn,agnieszka.prusiewicz}@pwr.wroc.pl

Abstract. In this work we propose a framework for courses recommendation. We use a social network approach to discover relations between users and discover social networks reflecting the patterns according to which the services are used. The topological analysis of this network allows us to detect dense groups of users which tend to use educational. The method for educational services and the results of some experiments are given.

1 Introduction

The concept of recommendation has its roots in a marketing and the problem of recommendation has been studied since eighties. The systems that implement any methods of products recommendation are shortly called recommender systems. A comprehensive survey of recommender systems with a detailed analysis and taxonomy based on many criteria is given [9]. The authors propose such criteria of recommender system classification as: user profile representation, initial profile generation, profile learning technique, relevance feedback, information filtering method, user profile-item matching technique, profile adaptation technique. The recommendation may be realised on the basis of the knowledge of user preferences based on direct questionnaire or analysis of user interactions with the system. In [13] the criterion of recommender systems classification is the user profile and the way of data acquisition, storage and retrieval. The authors distinguished such criteria as: contents of recommendation, explicit entry, user anonymous, data aggregation and use of recommendations. The other way of recommender systems classification is the valuation of recommendation results by means of recommendation accuracy and utility. In [17] the authors proposed such criteria of recommender systems valuation as: satisfaction, efficiency, persuasiveness, effectiveness, trust and transparency.

Some techniques for analyzing large-scale purchase and preference data for the purpose of producing useful recommendations to customers are proposed in [15].

The most popular recommendation method used in a electronic marketing is statistic one, that consists of recommendation the clients the products that are at most evaluated or the most often bought. Also the products bought by similar clients are recommended. In this case the crucial point is to determine the criterion of user similarity obtaining. It is assumed that users are similar if they behave in a similar way. Such an approach is typical for well known collaborative filtering [7].

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 327–336, 2011.

Recommendation techniques may be successfully applied not only to e-commerce but everywhere, where users use widely understood system resources to fulfill their needs. In the scope of our interest is online system for courses enrolment at the Wroclaw University. We propose the method for courses recommendation using social network approach. We assume that the methods for social network discovery can be easy applied to recommender systems. We analyse students' behaviour to detect social networks. The originality of our proposal is that to group students we don't use criteria explicitly stored in the system (concerning courses, marks, demographic data), we apply social network methods for similar students group discovery according to a assumption that the users that belong to the same social network behave similarly i.e. they tend to choose courses on the basis of the social relations, preferring groups their friends already enrolled to., i.e., choose the same or very similar courses. Social network of students is derived from the data from former semesters, then students that belong to the same social group are suggested to enrol to the same classes again. The advantage of our approach is that we use knowledge of user mutual relationships to recommend them services. The work is organised as follows. In Section 2 the problem of Social Network is discussed. In Section 3 the social networks of students are detected and the clusters analysis is conducted. In Section 4 the service recommendation algorithm is introduced, and the evaluation of the results is discussed.

2 Social Networks

Traditionally, the concept of social network is used to describe the relationships between friends, members of the particular society, relatives in the family, etc. It is the set of actors i.e. group of people or organizations, which are the nodes of the network, and ties that link the nodes [1][5][6][18] by one or more relations [5]. Social network indicates the ways in which actors are related [4]. The tie between actors can be maintained according to either one or several relations [5]. Several examples of social networks can be enumerated: a community of scientists in the given discipline who collaborate and prepare common scientific papers, a corporate partnership networks a set of business leaders who cooperate with one other, friendship network of students, company director networks, etc.

The social networks of Internet users differ substantially from the regular ones. Although social networks in the Internet have already been studied in many contexts and many definitions were created, they are not consistent. Also, different researchers name these networks differently. In consequence, these networks are called: computer-supported social networks (CSSN) [19], online social networks [5], web-based social networks, web communities, or virtual communities [1]. Nowadays, based on the kind of service people use, many examples of the social networks in the Internet can be defined. To the most commonly known belong: a set of people who date using an online dating system [2][16], a group of people who are linked to one another by hyperlinks on their homepages [1], customers who buy similar stuffs in the same e-commerce [20], the company staff that communicates with one another via email [21], people who share information by utilizing shared book marking systems such as *del.icio.us* [8].

In this paper we define a social network in the context of particular information system supporting student course enrolment. The relations between users are defined on the basis of their activity and the usage of the educational services available in the system.

3 The Problem Formulation

We consider online education system available at the Wroclaw University of Technology. One of the modules of this system delivers functionalities related to the courses enrolments. The course enrolment is a process that involves at least three main activities: student's authorisation and his rights for a course enrolment verification, finding a course and finally enrolment for a chosen course. A course enrolment is a semantically described service. A service description includes such data as: service name, teacher data, course description, initial requirements for a course enrolment, course duration, time of classes and conditions for a course completion. At the beginning of $t-th$ semester student s_j ($s_j \in S$, S is a set of all students) enrols for the set of courses according to his study profile. A timetable of student s_j in $t-th$ semester is defined as:

$$C_{jt} = \{c_{jt1}, c_{jt2}, \ldots, c_{jti}, \ldots, c_{jti_n}\} \tag{1}$$

where c_{jti} ($c_{jti} \in U_C$, U_C is the universe of all courses) is a course taken by student s_j in $t-th$ semester, described by a set of attributes as follows:

$$c_{jti} = \langle name_i, teacher_i, dur_i, Marks_i \rangle \tag{2}$$

Where $name_i$ denotes the name of $i-th$ course, $teacher_i$ is the name of the teacher of the $i-th$ course, dur_i - the day of the week and tea-time of the $i-th$ course and $Marks_i$ is a set of marks obtained from a course c_i.

Algebra course that runs on each Monday from 9.00 am to 11.00 a.m. by Prof. J. Jones without any preconditions is described as:

$$c_{jti} = \langle Algebra_i, prof.J.Jones, 1, (Monday, 9.00am - 11.00am), \{\} \rangle \tag{3}$$

Aggregated characteristics describing the way of the services usage by one user including demographic user data is represented by a user profile. In case of online education system and courses enrolments user profile up_j ($up_j \in U$, U is a set of all user profiles) is defined as:

$$up_j = \langle dem_j, C_j \rangle \tag{4}$$

where dem_j is a demographic data of user $s_j - th$, C_j ($C_j = \{C_{j1}, C_{j2}, \ldots, C_{jt,\ldots,} C_{jt_j}\}$, C_{jt} denotes a set of *j-th* student courses enrolments in *t-th* semester) is a historical data of system usage.

The functionality of online education system may be extended in order to provide personalised solutions. In case of courses enrolments such solution consists in automatic clusters of students discovery and recommendation of the same courses for the students from the same cluster. In this way first of all the social clusters must be discovered and then the same courses for the members of the same social cluster recommended. In order to discover social clusters the social network approach is applied.

4 Network of Users

Our method of recommendation is collaborative filtering. We analyse students behaviour to detect social networks. The social network is understood as a group of students that behave in a similar way. Assuming that the members of the same social network have similar preferences we can recommend the students that courses that have been just taken by the colleagues form the same social group.

4.1 Definition of the Student Network

In order to create a social network of students, their enrolments will be analyzed with respect to the choices of the same groups. We assume that the link between the students will be created if they enrol to the same group.

Let $S = \{s_1, \ldots, s_m\}$ be the set of students and t - time period (semester) under consideration. We define the strength of the directed relation $RS_t(s_i \rightarrow s_j)$ occuring during t-th semester between students s_i and s_j as the fraction of s_i's groups in which s_j also participated:

$$RS_t(s_i \rightarrow s_j) = \frac{|C_{it} \cap C_{jt}|}{|C_{it}|} \quad \text{if } C_{it} \neq \emptyset, \text{ otherwise} \qquad (5)$$

$$RS_t(s_i \rightarrow s_j) = 0 .$$

Intuitively, $RS_t(s_i \rightarrow s_j)$ is the fraction of the s_i's groups in which also s_j participated. Note also, that $RS_t(s_i \rightarrow s_j) = 1$ means that s_i and s_j show exactly the same behavior (the choise of courses) during the t-th semester.

Having defined the relations we use the network model to capture and process the knowledge about the behavior of the students. Typically, after observing the activity of the students during chosen semester t we are able to detect relations according to the abowe definition, then to build student network with nodes representing students and relations showing their tendency to enrol to the same courses. More formally

Definition 1. A *social network of students* for the t-th semester $SN_t=(S, R_t)$ is a graph consisting of the set of nodes (users) S and a set R_t of weighted edges (relations) between them such that $(s_i,s_j) \in R_t$ if $RS_t(s_i \rightarrow s_j)>0$. The weight of the relation is defined by the value of $RS_t(s_i \rightarrow s_j)$.

Our aim is to use the created network to predict the future (next semester) enrol requests of the students. On the Fig.1 a sample student network derived from the logs of the EdukacjaCL, information system dedicated for the students of the Faculty of

Computer Science and Management, Wroclaw University of Technology is presented. The network consists of 4083 nodes (students) connected by weighted, directed edges created according to (8). For the clarity of the picture it was generated for chosen relationship strenght threshold (with value close to one). Entire network contains over 1.2 million of non-zero relations.

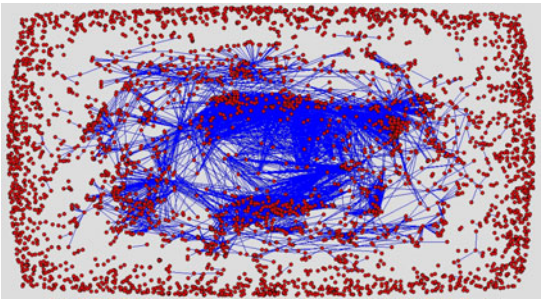


Fig. 1. User network derived from EdukacjaCL data, presented for relation weight threshold $RS_t(s_i \rightarrow s_j)$ equal to 0.9

It should be noted that strong relationships are especially important for our further analysis. Relationship strength $RS_t(s_i \rightarrow s_j)=1$ means that students s_i and s_j behave identically, always enrolling to the same courses. A cohesive cluster detected in the network formed by such relationships (links with strength threshold 1) points to the students who always enrol together, allowing almost certain recommendation.

The discovered network in this section will be used in the next section to discover students' social clusters in order to reason about their behaviour. In the following step the structure of the network will be used to generate recommendations during enrolments.

4.2 Student Cluster Detection

In our case the network communities were extracted using the Clique Percolation Method (CPM) [10][11]. The motivation behind this choice was the structure of the discovered network – it consisted of a large number of overlapping clusters and the CPM is dedicated for this type of network (in contrary to the most of the other existing methods which is their major drawback [10]. It was also successfully applied to the analysis of cluster evolution in large social networks [12] (which can are generally characterized by the massive presense of overlapping and nested communities) [3]. Last but not least, our interpretation of links in social network stressed the importance of dense clusters, especially cliques.

Following [11] we introduce the basic notions of the CPM method. *K-clique*, the key notion of the CPM algorithm, is a complete (fully connected) subgraph of K vertices. The remaining important notions are: (i) *K-clique adjacency*: two K-cliques are adjacent if they share $|K|$-1 vertices, i.e., if they differ only in a single vertex. (ii) *K-clique chain*: a subgraph, which is the union of a sequence of adjacent K-cliques. (iii) *K -clique connectedness*: two K-cliques are K-clique-connected if they

are parts of a *K*-clique chain. (iv) *K-clique percolation cluster (or component)*: it is a maximal *K*-clique-connected subgraph; i.e., it is the union of all *K*-cliques that are *K*-–cliqueconnected to a particular *K*-clique. The CPM algorithm allows to detect clusters composed by adjacent K-cliques and locate the K-clique percolation clusters for any chosen value of K. The interpretation of K is simple: bigger K allows the detection of more dense clusters (which are composed of fully connected subgraphs of size K). Students that behave similarly belong to the l-th social cluster $G_{tl}(k_r)$, where the *r*-th student $k_r \in K$, which is defined as:

$$G_{tl}(k_r) = \left\{ s_j \middle| \forall s_i, s_j \left(s_i, s_j \in G_{tl}(k_r) \right). RS_t(s_i \to s_j) \geq \alpha \right\} \tag{6}$$

It is possible that one student belongs to more than one social cluster. The set of social clusters to which the student s_j belongs in *t-th* semester is defined as:

$$G_{jt}(K) = \left\{ G_{jt1}(k_1), G_{jt2}(k_2), \ldots, G_{jtl}(k_r), \ldots, G_{jtg_l}(k_R) \middle| r,l \in N \right\} \tag{7}$$

$$1) \quad \begin{array}{l} \forall G_{jtl_1}(k_{r_1}), G_{jtl_2}(k_{r_2}), (G_{jtl_1}(k_{r_1}), G_{jtl_2}(k_{r_2}) \in G_{jt}(K)). \\ G_{jtl_1}(k_{r_1}) \cap G_{jtl_2}(k_{r_2}) = \varnothing \; or \; G_{jtl_1}(k_{r_1}) \cap G_{jtl_2}(k_{r_2}) \neq \varnothing \end{array}$$

The set of all social clusters in t-th semester is denoted as:

$$G_t(K) = \left\{ G_{t1}(k_1), G_{t2}(k_2), \ldots, G_{tl}(k_r), \ldots, G_{tg_l}(k_R) \right\} \tag{8}$$

In the case of the investigated network of 4083 students the following clusters were detected for link strength threshold equal 0.9 (Table 1, for chosen values of K):

Table 1. The results returned by CPM on EdukacjaCL network

K	Clusters	Max. cluster size
3	139	201
7	36	21
10	17	18
15	3	16
19 (max.)	1	19

As we may see the maximum value of K in our network was 19 (there was only one fully connected subgraph of the size 19) and the number of clusters varied – dense clusters were found in smaller quantities.

5 The Algorithm for Services Recommendation

Below we present the algorithm for the services recommendation. We assume that in case of empty set of services as a result of *k*-th user request execution the algorithm for services recommendation is applied and the services that have been executed by the users from the same social network to which *k*-th user belongs are recommended.

The idea of the algorithm for services recommendation is as follows. The student s_j enrols for the courses in the $t+1$-th semester. First the social network of students is created and then the CPM algorithm is applied in order to determine social clusters.

Input: The set of student profiles U, student $s_j \in S$.

Output: The set of courses $Rec_{j,t+1} \in U_C$ recommended to student s_j in semester $t+1$.

Begin

Step 1. For given *S* compute the set R_k of relations according to formula (5).

Step 2. Create social network of students SN_t (see Definition 1).

Step 3. Choose the relation strength threshold α (value close or equal to 1 is recommended). Run CPM algorithm on network SN_t in order to create the set of clusters G_t.

Step.4. For student s_j determine the set of social clusters $G_{jt}(K)$.

Step 5. If $G_{jt}(K) = \emptyset$ then $Rec_{j,t+1} = \emptyset$ else go to step 4.

Step 6. Determine the set of recommended courses $Rec_{j,t+1} = \bigcap_{j \in G_{jt\,max}} C_{jt+1}$.

End

Next from the clusters to which student s_j belongs the cluster with maximal value of coefficient K is chosen. Then the courses that have been already taken by the other students from this cluster are recommended.

As we can see, the idea is based on the assumption that the students will bahave similarly during the consecutive semesters and the social relations (which result in the enrolment to the same groups) will be preserved.

The algorithm suggests the actions (course enrolments) which normally are results of communication between students, talks, human-to-human recommendations etc. In this way we collect and process the social interactions which –in this case- are discovered on the basis of the services used by the users.

Note that, however our framework was tested in a specific application scenario, the similar approach may be used in any environment where we have groups of collaborating users which use system services.

In order to evaluate the quality of our method we used data from the two consecutive semesters. This allowed us to compare the recommendations generated on the basis of data from semester *t* (October 2008 – January 2009) with the actual actions of students taken during semester *t+1* (February 2009 – June 2009). The social networks SN_t and SN_{t+1} were created. Then we checked the stability of clusters by using an autocorrelation function *C* to measure the relative overlap between their states in semesters *t* and *t+1*. For given cluster $G_{tl}(k_r)$ it is defined by:

$$C(G_l(k_r)) = \frac{|G_{tl}(k_r) \cap G_{t+1l}(k_r)|}{|G_{tl}(k_r) \cup G_{t+1l}(k_r)|} \tag{9}$$

C is the number of common members of the cluster in *t* and *t+1* divided by the number of the members of union of the cluster states in *t* and *t+1*.

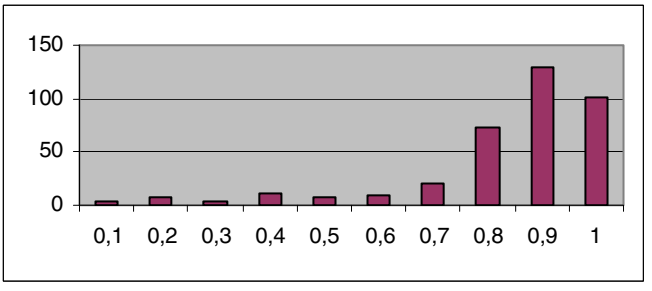


Fig. 2. Distribution of clusters according to their autocorrelation function

As shown on Fig.2 most of the 371 detected clusters have high autocorrelation functions, which suggests good predictive power of our method and confirms that clusters are relatively stable. In consequence – they may be used to predict the future behavior of the students.

On the basis of historical data, gathered from 3 semesters, we conclude that the actual user actions followed recommendations based on our social network analysis in 81% of the cases.

6 Conclusions and Future Work

Our method is based on analysis of the collective actions of users (in our test case: students) in order to determine the similarities in the patterns according to which they use the services of the system. We create the social network and assume that its topology guides the behavior of students. The recommendations are generated according to the gathered knowledge about users' actions.

Future research directions include the analysis of possible synergies between our approach and the classic recommendation methods – we want to check if our method can support them and if they can be used together in order to improve recommendations given to the user. The joint use of classic recommendation techniques could help to generate recommendation for isolated network nodes (users who did not form strong social relationships – visible on Fig.1).

The other issues to be addressed include: 1) The density of the groups, formally expressed by the *K* parameter of the CPM algorithm. In first experiments, for each particular user we used groups detected with the highest possible value of *K*, referring to the dense clusters (for big *K* similar to full graphs). Obviously, any user may belong to several clusters, which may have different density of the connections, and this knowledge may be also used; 2) Generation of recommendations for detected user

clusters – which will imply the processing of the data concerning courses (defined in Section 2) and a strategy for choosing the relationship strength threshold – for first experiments we used values close to one, but the analysis of the networks build from weaker relations could also be useful.

The results will be applied in the recommendation module (currently under development) dedicated for the SOA-based student enrolment system at the Workflow University of Technology.

Acknowledgements

The research presented in this work has been partially supported by the European Union within the European Regional Development Fund program no. POIG.01.03.01-00-008/08.

References

1. Adamic, L.A., Adar, E.: Friends and Neighbors on the Web. Social Networks 25(3), 211–230 (2003)
2. Boyd, D.M.: Friendster and Publicly Articulated Social Networking. In: CHI 2004, pp. 1279–1282. ACM Press, New York (2004)
3. Faust, K.: Models and Methods in Social Network Analysis. In: Carrington, P., Scott, J., Wasserman, S. (eds.) New York (2005)
4. Freeman, L.C.: Centrality in social networks: Conceptual clarification. Social Networks 1(3), 215–239 (1979)
5. Garton, L., Haythorntwaite, C., Wellman, B.: Studying Online Social Networks. Journal of Computer-Mediated Communication 3(1) (1997)
6. Hanneman, R., Riddle, M.: Introduction to social network methods (January 4, 2006); online textbook `http://faculty.ucr.edu/~hanneman/nettext/`
7. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. 22(1), 5–53 (2004)
8. Millen, D., Feinberg, J., Kerr, B.: Social bookmarking in the enterprise. Queue 3(9) (2005)
9. Montaner, M., Lopez, B., De La Rosa, J.L.: Taxonomy of Recommender Agents on the Internet. Artificial Intelligence Review 19, 285–330 (2003)
10. Palla, G., Derenyi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature 435, 814–818 (2005)
11. Palla, G., Derenyi, I., Vicsek, T.: Clique percolation in random networks. Phys. Rev. Lett. 94, 160–202 (2005)
12. Palla, G., Barabasi, A.-L., Vicsek, T.: Quantifying social group evolution. Nature 446, 664–667 (2007)
13. Resnick, P., Varian, H.R.: Recommender Systems. Communications of the ACM 40(3), 56–58 (1997)
14. Rosen, M., Lublinsky, B., Smith, K.T., Balcer, M.J.: Service-Oriented Architecture and Design Strategies. Wiley Publishing, Inc., Chichester (2008)
15. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Analysis of recommendation algorithms for e-commerce. In: Proceedings of the 2nd ACM Conference on Electronic Commerce, pp. 158–167 (2000)

16. Shaw, M.E.: Group structure and the behavior of individuals in small groups. Journal of Psychology 38, 139–149 (1954)
17. Tintarev, N., Masthoff, J.: A Survey of Explonations in Recommender Systems. In: Data Engineering Workshop, pp. 801–810 (2007)
18. Wasserman, S., Faust, K.: Social network analysis: Methods and applications. Cambridge University Press, New York (1994)
19. Wellman, B., Salaff, J., Dimitrova, D., Garton, L., Gulia, M., Haythornthwaite, C.: Computer Networks as Social Networks: Collaborative Work, Telework, and Virtual Community. Annual Review Sociology 22, 213–238 (1996)
20. Yang, W.S., Dia, J.B., Cheng, H.C., Lin, H.T.: Mining Social Networks for Targeted Advertising. In: 39th Hawaii International International Conference on Systems Science (HICSS-39 2006). IEEE Computer Society, Los Alamitos (2006)
21. Zhu, W., Chen, C., Allen, R.B.: Visualizing an enterprise social network from email. In: 6th ACM/IEEE-CS Joint Conference on Digital Libraries. ACM Press, New York (2006)

Working with Users to Ensure Quality of Innovative Software Product despite Uncertainties

Barbara Begier

Institute of Control and Information Engineering, Poznan University of Technology,
pl. M. Sklodowskiej-Curie 5, 60-965 Poznan, Poland
barbara.begier@put.poznan.pl

Abstract. Problems with uncertainty refer to various aspects in knowledge engineering. An open list of uncertain items in expert system development is given and their sources are listed, too. The periodical assessment of a software product by its users is recommended to reduce uncertainties. If symptoms of 'disease' are learnt then the proper diagnosis and therapy may be specified. User involvement is needed to ensure quality of innovative software during its evolutional development. Users provide regular feedback on a product. The experience, related to the expert system assessed by its users, is described.

Keywords: expert system development, user-centeredness, uncertain data, feedback from users, software quality, software product assessment.

1 Introduction

The problem to develop an innovative software product, like an *expert system* or a knowledge-based system, which satisfies its users seems to be more difficult than in the case of a conventional software product. Development of an expert system is burdened by *uncertainty* [4]. This concept applies to various items in knowledge engineering. For example, an expert system may operate uncertain knowledge. Other examples of uncertain items and their sources are given in Section 2.

Probabilistic methods are usually applied to solve problems with uncertainty. But just the humans generate uncertainties. So user involvement and their feedback provided on a product seem to be the promising solution to overcome a problem of uncertainty and to ensure quality of an innovative software product. Many software producers declare their *user-oriented* approach. It means usually working *for* users and respecting their goals when developing a new product. But a point of view of software authors may substantially differ from the real user's point of view. Thus the term of *user-centeredness* remains the term applied previously. Users' involvement may help reducing possible uncertainties which burden innovative software products to build solutions that satisfy their users. Forms of cooperation with users' are presented in Section 3.

The applied life cycle model providing feedback from users on a developed product is presented in Section 4. The recommended here, basic form of users' involvement is a periodical product assessment by questionnaire survey. Real and/or potential users take part in it. The design of a questionnaire and results of the conducted

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 337–346, 2011.

surveys are described in the fourth section. The presented research refers to the real life experience with the unique software product AFIZ (this name is an acronym of the *Analyzer of Facts and Associations*, expressed in Polish). It has been developed for analysts working for the Police and/or the Border Guard. The described user involvement has had a form of regular meetings with the most involved specialists and also the most recommended form of a periodical software product assessment by available user representatives.

2 Uncertainties in an Innovative Product Development

The notion of *uncertainty* refers to many aspects in knowledge engineering. Also the development process of an expert system and information resources required for that purpose are burdened with uncertainty. An open list of uncertain items contains:

- Completeness of requirements.
- Availability of required documents and real life data.
- Definition of a development process – if it is proven in practice to ensure quality.
- Suitable selection of quality criteria and measures of developed software.
- Identification of quality factors and their real impact on an analyzed process.
- Way of assessment of *quality in use* (selected measures, available real users).
- Sense of automating specified activities – how to prove that the introduced solution is correct and welcome in user community in the real life.
- Assumption that product improvements bring the expected results.
- Suitability of the provided infrastructure to the innovative product development.

User involvement is needed to deal with the first eight from the listed above 9 items. There are many sources of an uncertainty (extended version of that in [12]):

- Lack of available experts in the given domain to take part in a collaborative design of ontology or contextual reasoning, for example.
- Domain experts may not be interested to share their knowledge with software developers and to supply them real samples of data – their knowledge has been gathered year by year and becomes the valuable good desired on a market.
- Various experts become the stakeholders in the software process – each expert may represent his/her private and subjective point of view or interest.
- Analyzed data may not be representative for the most of real life cases.
- Some data may be false (at variance with the facts) for unidentified reasons.
- A set of observed and measured items may be insufficient for inference purposes.
- Human cognition is not normalized – this statement refers also to human observations, opinions and considered cases.
- Domain rules and data are noticed in the given moment of time; different values and phenomena may be observed in various periods of time; some rules and data may change before a knowledge-based system is completed.
- Domain experts may differ in specifying relationships between observed symptoms and based on them diagnosis (they may be even wrong).
- There are many possible conclusions and therapies applicable in the given case.

Notion of *causality* is broadly applied in medicine. The diagnosis is based on the observed symptoms which are real life data. Then the therapy is specified. The guiding idea of *uncertain data management* including uncertain reasoning is applied on a high level of abstraction and proven usually in simplified conditions referred to limited and even trivial medical data. Then this idea has been referred to any knowledge-based system [12]. In the author's opinion, it may be also applied to improve software quality by reducing some uncertainties. It requires to:

- Specify a set of symptoms $s_1, \ldots, s_n$ of 'illness' x (inappropriate functioning).
- Identify, in cooperation with users, their sources and specify their importance.
- Gather data coming from observations and tests (expressed by notes and opinions given by software users, for example).
- Propose a set of therapies $t_1, \ldots, t_k$ (specification of software improvements).

Various people, including domain experts, may judge symptoms (measures) differently. But the greater number of participants, the better statistical results. The following steps are recommended to improve quality of an expert system:

- Periodical assessment of product quality by its users is planned and performed according to the plan. Quality criteria and their measures are specified as potential *symptoms* of a software product '*disease*' (poor values given by users) and included in a questionnaire of a survey.
- The *diagnosis*, based on an analysis of obtained values and remarks given by respondents, points out improper solutions and other weak points of a product. In particular, it may find out also a lack of some features and facilities.
- The *therapy* referred to the noticed symptoms is specified in a form as required product improvements.

After therapy t_j the presence of a symptom s_i, (observed by users and expressed by the given values of measures) should decrease. Notes and suggestions given by users are stored and related product improvements are recorded. Thus a history of an evolutional product development is maintained. In addition, problems do not usually appear separately but they interwoven in causal networks [8]. In the development of an expert system most of problems are generated by humans. And in author's opinion, only humans may help to overcome the identified problems.

3 User-Centeredness and a Scope of User Involvement

Designers and programmers used so far to work in isolation from users. They use different language and show other preferences than users. Thus *users' involvement* in the development process is recommended to emphasize social aspects of innovative software product [3] and to ensure a high level of users' satisfaction from it. This idea derives from the Nygaard's *participatory design* initiated in early seventies.

The *user-centeredness* is intended to ensure software quality. However, the meaning of this term may differ representing its various dimensions [6]. The declared *user focus* is usually limited to specifying goals from user's perspective but a fictional user substitutes for a real one. Another dimension refers to *work-centeredness* – there is a focus on effective work of an average but still abstract user whose actions are

supported by the intended system. Next, a concept of *user participation* in the software process means working *with* users instead of working for them. But this idea is often limited to small software project during its design and then testing user interface. Finally, to provide an adaptable *personalization* of a product according to his/her preferences requires involving the diversity of real users in the design process and learning user's behavior when using the system. The author' emphasizes *working with users* to obtain a regular feedback from them.

There are various forms of user involvement in a software process. Most of them refer to requirements specification and software usability testing. In many projects users are stakeholders in software development although their real impact on a process is discussible. Principles expressed in the Agile Manifesto [11] and applied in agile methodologies emphasize *direct communication* between developers and software users. An aim of these activities is to learn who users are, to know their expectations, and their point of view on quality of software under development.

Software developers are often prejudiced and afraid of users' negative impact on project performance which can become more time-consuming and less effective than previously. So there is the question of when user participation is actually helpful. The described survey data from 117 software projects [13] confirm that the highest level of software authors' satisfaction results from a high user involvement in new software projects although at the same time users' expectations are excessively growing. In turn, users were most happy by engaging minimal time in the development process.

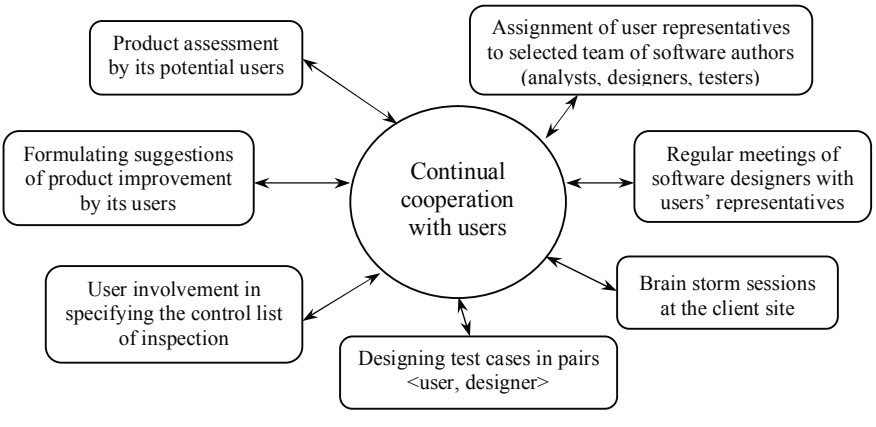


Fig. 1. Some recommended forms of cooperation with users

In some agile approaches users participate in brain storm sessions to build a vision of a future software product and specify its features because software product solution and its applications implicate their daily work. Forms of user participation in a software process, recommended by the author, are shown in the Figure 1. The experiments described in this paper are focused on a *product assessment by its users*.

Moreover the usability evaluation is essential to make sure that the developed software product is easy to learn and to use for real users [5]. There are several methods for usability evaluation: Think Aloud (testing including thinking aloud at

user's workplace instead of at labs), Heuristic Evaluation (usability inspection), and Cognitive Walkthrough (a theory-based method in which every step required to perform a scenario-based task is evaluated to detect potential mismatches).

User's community is not homogeneous – users represent various personality types and differ in age, gender, level of education, profession, and their life experience. Thus the individual representative engaged in the development process of an expert system is not enough to express various expectations and suggestions. The question is how many software quality evaluators need to be involved to present opinion of user community. There is no consensus regarding the optimal sample size in this case – from 4±1 (four or five users are needed to detect 80% of usability problems) [9] up to all population of potential users. In practice, software users working for the same organization share many features, like their educational level, place of living, type of work, acquired skills, trainings, and professional experience. The analysis of data with 102 usability evaluation experiments allows forming a general rule for optimal sample size: it would be 10±2 [5] instead of 4±1. If it is so with usability evaluation then a similar number of users may be sufficient to assess other software quality features.

4 Feedback from Users in a Product Life Cycle. A Case Study

4.1 User Involvement in an Innovative Product Development

Quality problems in an expert system development refer to both its *shell* responsible for communication with users and its *kernel* consisting of various data (knowledge base) and provided mechanisms including an inference engine. Software designers may misinterpret user needs and then the implemented facilities do not meet user expectations. There are also a lot of problems with reliable data for testing purposes.

Any software product is addressed to the specified community of its direct and indirect users – clerks, doctors, engineers, applicants, etc. All of them are experts in the given domain of an application. So the first step is to learn who users are. Users' involvement in software development is recommended here to emphasize social and quality aspects of an innovative product – to identify and overcome problems with uncertainties as much as possible. The direct relationships may cause that users have confidence in software authors and are inclined to entrust them with making use of results of their own professional experience. Otherwise designers cannot expect to learn user expectations and study the real life cases including the required test data.

General aim is to obtain a feedback from users on a developed product to make it acceptable by a wide spectrum of its users. An evolutional product development supported by a continuous feedback from users is recommended here as shown in the Figure 2. Users' representatives take part in software development starting from its early phases. In the described approach the most valuable feedback is obtained in the product periodical assessment by user community. Such approach has been applied by the author successfully to the expert system developed for civil engineers [2].

Users' involvement is not limited to the requirements specification although it's been proved that project is successful when users play an active role in this phase [7]. The general vision of an expert system should be created in cooperation with users. Then potential sources of uncertainty are pointed out during direct meetings of software designers and its potential users.

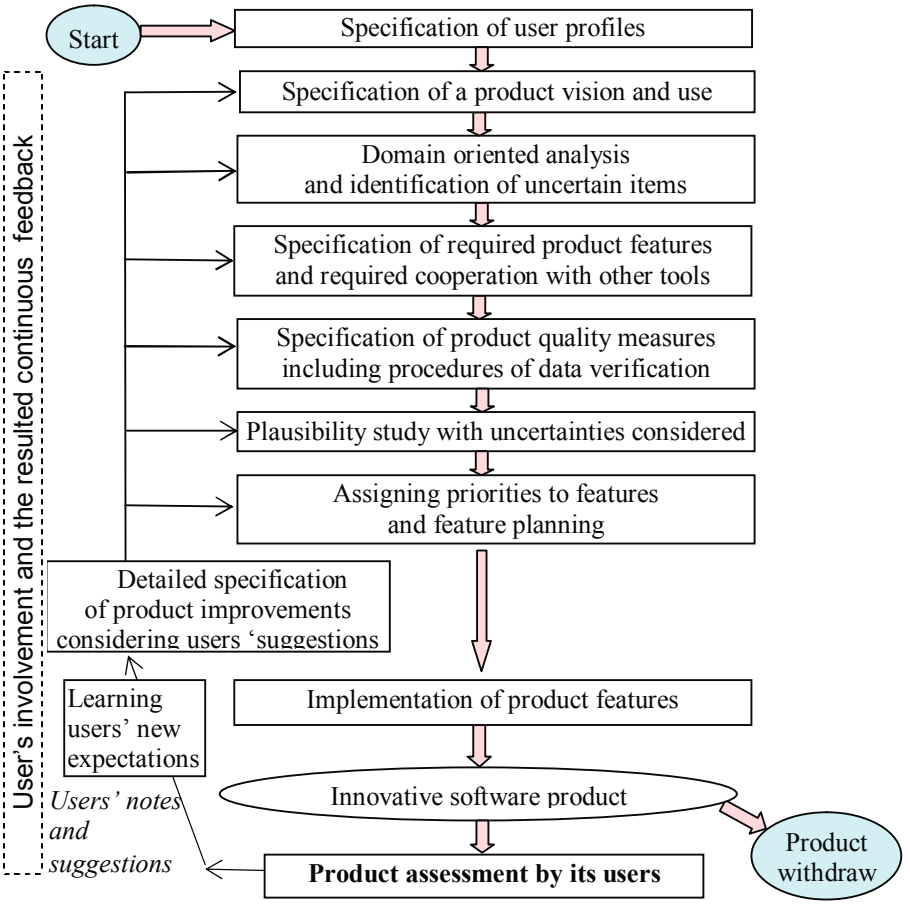


Fig. 2. Evolutional development of an expert system and its improvement based on continuous feedback from users

The described research refers to the product AFIZ (*Analyzer of Facts and Relations* expressed in Polish), which is a part of software developed under the scientific grant entitled *Polish Platform for Homeland Security* [10], the unique scientific initiative beyond European dimension. The undertaken activities are aimed at creating integrated computer tools to support specified efforts to improve public security. The main focus is to support police and other security services, like the border guards, with tools based on modern technologies. Due to the sensitive nature of data and project topics, a main part of research work within the Platform has the classified status. The aim of the considered product is to facilitate analysis of telephone and IMEI (*International Mobile Equipment Identity*) billings. The analysis helps finding a valid telephone number of a wanted person. Location of masts BTS (*Base Transceiver Station*) involved in an analyzed communication may help tracking his/her steps.

Individual experts working for the mentioned above public services, especially investigative ones, participated in all phases of the considered software development

cycle. Applying an iterative-incremental model requires to identify product functional features to be implemented in each iteration. The concept of a *feature* refers to that introduced in the FDD (Feature Driven Development) [1], one of agile methodologies. The FDD and its practices rely on granularity of product functionality. Significance of each specified feature for an entire product is considered. The feature hierarchy is specified according to the assigned priorities and then implemented.

In the described research, the very important form of users' involvement is software product assessment in a survey by questionnaire (Section 4.2). Each working software version is assessed by its users who assign values to the specified quality measures of the considered product. Respondents assess this way to what degree the provided software tool meets their needs. Users add also their suggestions on how to improve the product. The obtained results are not limited to easiness of product use and user interface. They can indicate some missing requirements, point out the proper interpretation of requirements, and implicate the way of their implementation. Results of a product assessment, as described in Section 4.3, are the basis of product further improvements and have real impact on an entire development process.

4.2 Design of a Questionnaire to Assess Software Product by Its Users

Due to the character of an application, senior officers of the respondents have to agree on the design of questionnaire content. The described questionnaire starts with an initial part intended to learn who respondents are. It contains the following phrases:

I. I am a person in the age bracket: A (21 to 35), B (36 to 49), C (over 50)
II. My level of education is: A (master degree), B (bachelor), C (high school)
III. My position at work is: A (managerial), B (ordinary), C (specialist)
IV. Self-assessment of my skills in computing is (<1, 5> using school marks)
V. Frequency of using computer by me in last six months can be estimated as: not at all, once a month, once a week, every 2 days, every day
VI. My self-assessment of using other professional software (<1, 5>)
VII. My place of work is located (<region of the country>).

To assess a product the quality criteria and their measures have to be specified. They should include provided data verification and data exchange mechanisms. Maintained incorrect data are the main source of improper system behavior and its poor usefulness from the users' point of view. Developers are erroneously convinced that these are problems of marginal importance. In the presented case, there are 55 questions in the main part of the designed questionnaire divided into subparts referred to the selected quality attributes of the considered software:

- functionality,
- navigation (as a result of a lucid construction),
- easy way of data input and then results presentation,
- ease of learning to use,
- ease of using,
- data security,
- reliability and efficiency,
- portability,
- general assessment of a product and user satisfaction with it.

The psychometric scale devised by Rensis Likert has been applied to express each respondent's answer. Each item of the questionnaire has a form of a statement, for example: *7. Program maintains and reconstructs steps of data processing making all history of the considered matter available*. Then a respondent is asked to indicate his/her degree of agreement with this statement, using a five-point scale:

5 – *I fully agree*,
4 – *I rather agree*,
3 – *I have doubts*,
2 – *I rather disagree*,
1 – *I completely disagree*.

Such scale is applied at schools in Poland where "5" is the highest mark equivalent to the American "A" and "1" is equivalent to the "E". Free space has been left at the end of the questionnaire for user's suggestions and remarks concerning the expected improvement of a product.

4.3 Results of the Questionnaire Survey

The first assessment of the considered expert system AFIZ took place at the end of training of a large group of its potential users. The paper questionnaires were applied then. The second assessment took place in an electronic form after 2 months. Senior officers gave their agreement on participation of interested officers in the survey but the questionnaires were fulfilled anonymously to tell the truth.

User profile

An average respondent is a relatively young and high educated man who works as a specialist, uses computer daily and knows specific software applied in the profession. As much as 54% of respondents are people in the 21 to 35 age bracket, 42% are 36÷49 years of age and only 4% are above 49. Almost two thirds of respondents have got the master degree. Every fourth respondent occupies a managerial position, 30% is an individual specialist; the others occupy ordinary positions. All respondents use computer every day although only 30% assessed their computing skills as the very good, 61% as good and 9% as sufficient. Most of them (82%) declare their proficiency in using other professional software systems. Only one half fulfilled an item concerning their work location – respondents omitted this item because they did not want to be identified on its basis.

First edition of a product assessment

Respondents had to assess 53 measures from 55 items in the questionnaire; two items concerning product portability were postponed to further assessments. According to directives applied in marketing, the qualified minimum for each item is 70% of obtained answers. In the considered survey more than 30% respondents omitted 6 items ─ many users skipped 2 items concerning ease of learning to use a program, 2 items concerning program security, one question about program installation, and one question about the percentage of daily work supported by the assessed product. Probably more experience with a considered tool is required to answer these items. So 47 items of the questionnaire could be statistically processed and then analyzed.

The author is not entitled to present precisely the statements of the questionnaire and the detailed data concerning assessment of its separate items. But even statistical results may be interesting. As much as 33 product characteristics (on 47) got the mark not less than 4.0, 13 got a mean value from the bracket <3.5, 3.99>, one got a mean value from the bracket <3.00, 3.5>, and only one less than 3.0. The last item is related to the possibly required help of an experienced user of the assessed expert system. Program functionality was highly assessed but with some exceptions. Items concerning ease of learning to use a program were assessed worse.

Four items placed in the last part of the questionnaire (*general assessment of a product and user satisfaction with it*) were assessed quite well. The conformity of program functions with user expectations got the mean value 3.77 of all given marks. Respondents admitted that the considered expert system makes their work easier (mean value 4.0). Using a program is not stressful – the mean value of related marks was 4.33. The level of respondents' satisfaction from a program had a mean value 3.9.

The most important and valuable for the product authors were suggestions of program improvements given by respondents at the end of the questionnaire. Precisely every second respondent gave suggestions and/or remarks. Many opinions were shared by several respondents. After the careful analysis, the list of 31 proposals of program improvements was established and transferred to software authors.

Second edition of a product assessment

During second edition of the survey respondents assessed the unchanged product. They applied the same questionnaire in both surveys of the AFIZ tool in 2010. Again, two items concerning product portability were to be omitted. Questionnaires were fulfilled by almost the same users (only 3 persons were different) so their profile was also almost the same as previously. In 52 cases on all 53 items of the questionnaire the answers were qualified to their statistical processing (minimum 70% answered).

Features of the AFIZ program were assessed highly again – 27 items got the mean value higher than or equal to 4.0, 20 got their mean value from 3.5 to 3.99 bracket and only four had mean values less than 3.5 but more than 3.0. But a lot of measures got a bit worse values than before including 9 measures of functionality (on 12 in total), 7 of interface friendliness (on 10), and only 3 on 7 measures of ease of use. This fact confirms a conjecture that an unchanged product may get with time worse notes than before because user expectations are growing and growing. The percentage of very good and good marks was in 33 cases higher in the first edition, one was equal in both editions, and 19 measures got lower notes at the first time. The essential dispersion of answers related to a given item was observed in many cases.

Almost every respondent (90.5%) gave his/her remarks and suggestions of program improvement. They usually wrote several sentences or a list containing 5 items on average. Again, there were repetitions in the suggested proposals.

5 Conclusions

The problem with uncertainty when referred to the expert systems seems to be more serious than in the case of a conventional software product. User involvement is

recommended and applied to solve problems with uncertainties. Its valuable form, proven in practice is software product assessment by its real and potential users.

The assessment indicates clearly strong and weak points of a product and thus becomes the leading activity in cooperation with users. It breaks off the product authors' isolation from users but it does not interrupt their work decidedly and thus may be accepted by developers' community. Various suggestions on how to improve a product are introduced. This approach does not eliminate other forms of cooperation with users like joint teams, regular meetings, brain storm sessions, and others.

The described experience shows that the same product after a relatively short period of time may be assessed a bit worse by its users than before. So the conclusion is that each software product has to be continuously improved because user needs and expectations are growing after their basic needs are satisfied.

The notion of a software project success from the perspective of developers is at the first sight different than that from software users' point of view. But in a long time perspective the satisfied customer is the primary measure of a project success.

References

1. Agile Software Development using Feature Driven Development (FDD), http://www.nebulon.com/fdd/
2. Begier, B.: Evolutionally Improved Quality of Intelligent Systems Following Their Users' Point of View. In: Nguyen, N.T., Katarzyniak, R., Chen, S.-M. (eds.) Advances in Intelligent Information and Database Systems. SCI, vol. 283, pp. 191–203. Springer, Heidelberg (2010)
3. Begier, B.: Users' involvement help respect social and ethical values and improve software quality. Information Systems Frontiers 12, 389–397 (2010)
4. Hsu, J.S.-C., Chan, C.-L., Liu, J.Y.-C., Chen, H.-G.: The impacts of user review on software responsiveness: Moderating requirements uncertainty. Information & Management 45, 203–210 (2008)
5. Hwang, W., Salvendy, G.: Number of People Required for Usability Evaluation: The 10±2 Rule. Communications of the ACM 53, 130–133 (2010)
6. Iivari, J., Iivari, N.: Varieties of User-Centeredness. In: Proceedings of the 39th Hawaii Conference on System Sciences. IEEE, Los Alamitos (2006)
7. Kujala, S.: Effective user involvement in product development by improving the analysis of user needs. Behaviour & Information Technology 27, 457–473 (2008)
8. Munk-Madsen, A.: Classifying IS Project Problems: An Essay on Meta-Theory, Dept. of Computer Science. Aalborg University, Denmark (2000-2006)
9. Nielsen, J.: Estimating the number of subjects needed for a thinking aloud test. Int. Journal of Human-Computer Studies 41, 395–397 (1994)
10. Polish Platform for Homeland Security, http://www.ppbw.pl/en/p_strona_glowna.html
11. Principles behind the Agile Manifesto (2001), http://agilemanifesto.org/principles.html
12. Puppe, F.: Systematic Introduction to Expert Systems. Springer, Heidelberg (1993)
13. Subramanyam, R., Weisstein, F.L., Krishnan, M.S.: User Participation in Software Development Projects. Communications of the ACM 53, 137–141 (2010)

U2Mind: Visual Semantic Relationships Query for Retrieving Photos in Social Network

Kee-Sung Lee, Jin-Guk Jung, Kyeong-Jin Oh, and Geun-Sik Jo

Department of Computer & Information Engineering, Inha University
{lks,gj4024,okjkillo}@eslab.inha.ac.kr, gsjo@inha.ac.kr

Abstract. This research is to investigate a method that enables social networks to provide a semi-automatic system. The system will allow users to organize their target photos, using the concept of ownership attributes that describe the relationships between objects in the photos. In this paper, we propose formulating a visual semantic relationships query for photo retrieval. A Visual Semantic Relationship Query interface helps users describe their perspectives about the desired photo in a semantic manner. In the ranking process, by interpreting both concepts and relationships, a user's query is transformed into a SPARQL, which is then sent to the JOSEKI server, and the returned photos are evaluated in terms of relevance to each photo. The experimental results demonstrate the effectiveness of the proposed system.

Keywords: Visual semantic relationships query, social network photo ontology, photo retrieval system, semantic photo annotation, U2Mind System.

1 Introduction

In recent years, the amount of personal digital photos from vacations, parties, families, and friends has grown rapidly. All photos are stored on personal storage devices in simple directory structures without meta-information. Moreover, those might be uploaded to a Social Network Service (SNS), such as Flicker, FaceBook, and Twitter to be shared with friends. The traditional common method for SNS is to provide tools for photo browsing, searching, and retrieving from a large database by adding a metadata, keywords or descriptions, to the photo, so that the retrieval process can be performed over annotation words. However, this process becomes more difficult as the number of photos increases.

For example, a user who finds an interesting photo of him and his friend will bookmark it. So, whenever the user wants to search for the desired photo, the searching process would be easier. Over time, a user might forget the site and the path to the desired photo, and they might spend a lot of time visiting more blogs to search it. In particular, the bigger the user's social network, the more relationships between the photos and the blog will have. So, the search task will become more difficult [1].

There are three scenarios to be considered:

- People tend to forget, and it is hard for them to memorize all information about the photos [2].

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 347–356, 2011.

- Secondly, most of the time, the user will only remember the interesting part from photos from the event, which involves attractive objects and interesting places [3, 4].
- Lastly, the existence of a semantic gap between human representation and system resolution, meaning that it will be difficult for the users to represent their memory as a term-based query to be resolved by information retrieval systems [3, 4].

Bookmarking is the best solution for the first scenario, but over time, the number of bookmarks will increase, and it still does not help the user much [5]. Some researchers have proposed solutions that enable users to represent their memory based on term-based queries as well as sketch-based queries [3, 4]. For instance, MindFinder [3] is an image retrieval system that provides user interaction to describe the interesting objects that are located on desired images. An image search by concept map [4] also provides a solution on how the size and location of interesting objects can be drawn using the system. The systems, however, have never provided users with a method to represent relationships between interesting objects. For example, let's consider that a user wants to find a photo showing him and two school friends located on his left and right sides. Then, users could draw three people on the query canvas, but representing the relationships between each other was not considered.

In this paper, we propose a photo retrieval system, called U2Mind. The system will allow users to represent their interesting objects and the relationships between them.The system has been designed for alleviating the semantic gap mentioned in the third scenario. For this purpose, we build ontology, called SNPhoto ontology, to purposely describe objects and the relationships between them. We also design a query interface, called a VSRQ interface, for users to represent objects and relationships. This paper is organized as follows; the next section will describe some of the relevant works to image retrieval systems. In section 3, we will explain the proposed photo retrieval system, including the SNPhoto Ontology, query interface, and evaluation metrics. In Section 4, the implemented system will be illustrated and some results will be shown to emphasize our approach. Finally, the last draws some conclusions of the finding.

2 Related Work

Image search intention concerning spatial layout has been investigated in previous works [3, 4]. In [3], the authors presumed that the Sketch2Photo system is utilized to only stitch the images representing different objects into the resulting image, rather than to find images in the database that meet what exists in the user's memory. The authors of the latter work have mentioned two basic problems for retrieving images: query formulation and query matching. They assumed that query matching highly depends on the query formulation, so they thought that query formulation should be given a primary importance in image retrieval. Based on this idea, they have proposed the MindFinder system, which is a bilateral interactive image search engine using interactive sketching and tagging. Although they have used positional information as well as tagging, they did not convey a clear semantic intention.

The proposed system in [4] focuses on the search intention of the layout of semantic concepts, i.e., what is expected to appear in a specific position on the desired

images is defined by a semantic keyword, but not particularly with a color or sketch. However, they have only considered what was expected to appear in a specific position on the desired images. In contrast to the previous work, we do not only consider what is expected to appear in the specific position, but we also consider the relationship between objects in that position. The relationship between people is animportant factor in social networks. For example, using previous systems that do not consider the relationship between objects, it is a difficult task for users, who want to search photos that include a friend on the left side position, to formulate an implicit query in their thought.

3 Visual Semantic Relationships Query for Photo Retrieval

3.1 System Architecture

The proposed architecture for the photo retrieval system is shown in Fig. 1. We call the proposed system Understanding User Mind (U2Mind). The system consists of three modules; bookmark module, annotation module, and retrieval module. Let us consider a user browsing the Internet using any browser application, like Internet Explorer, Mozilla Firefox, or Google Chrome, and that each browser application contains a bookmark module, known as a plug-in. With this bookmark module, the users can bookmark any interesting photos as they browse their photos on their friend's Web pages or blogs. At this point, the module could gather relevant information including the URL of the bookmarked photo.

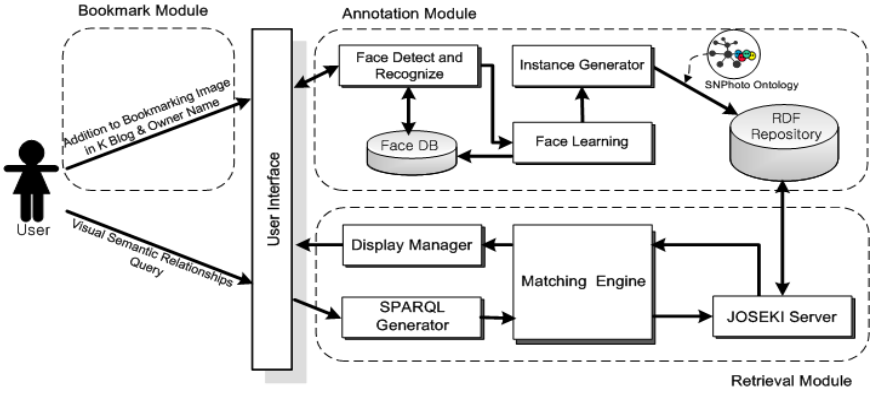


Fig. 1. U2Mind System Architecture

The information and photo are passed into the annotation module. Then, through processing the photo by image processing tools, the annotation module could collect additional information such as the size of the image, names, and locations of people who appear in the photo [7, 9, 10].To retrieve the photos, the user needs to sketch the objects and place the relationships between the objects as lines with tags. With this user convenient technique, a Visual Semantic Relationships Query(VSRQ) is designed.

Based on the user's query, SPARQL Generator in the retrieval module generates a query that is described in SPARQL. This query enters into a JOSEKI[1] server through the Matching Engine located at the center of the retrieval module. The main role of the Matching Engine is to rank photos based on computing the similarity scores between the user's query and photos, which are returned by the JOSEKI server. Finally, the ranked photos are displayed in the GUI.

3.2 Social Network Photo Ontology and Semantic Annotation

Ontology is the formal representation of knowledge. It is a set of concepts within a domain and the relationships between them. Ontology, allow users to organize information on the taxonomies of concepts with their own properties. So, the data that is represented by ontology can be easily understood. To retrieve the desired photos, a user will create a query. The query might have implicit objects and relationships between them. However, the retrieval systems cannot understand the query. The major, purpose of using ontology is to enable the photo retrieval in the proposed system to produce the best fit result corresponding to the user's queries. This is because, when the SPARQL Generator generates the SPARQL[2] query from the user specified query, the generator can use the explicit information and implicit information that are inferred from ontology.

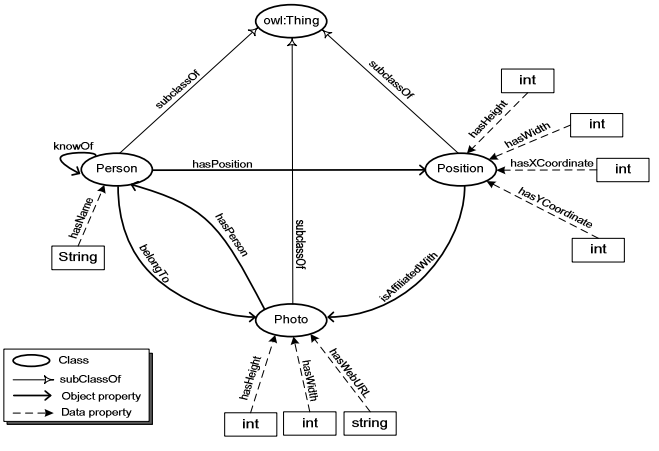


Fig. 2. SNPhoto Ontology

In this context, a small ontology is designed for describing photos that include people. This ontology is called the Social Network Photo ontology (SNPhoto), and it consists of three concepts and some relationships as is shown in Fig. 2. These concepts are the sub-concepts of the "Thing" concept that is the most super-concept in OWL.The Photo concept represents the bookmarked photos. We will assume that the photos must include at least one person. This concept has general metadata such as

[1] http://www.joseki.org/
[2] http://www.w3.org/TR/rdf-sparql-query/

height, width, and URL. For representing the people in photos, we have created the Person concept, and the Position concept is created for the location representation of people in the photos.If the photo is bookmarked, then the next step of the semantic annotation process with SNPhoto is as follows:

> **SYSTEM:** Using image processing tools in OpenCV[3], the face regions are automatically extracted, and the face learner will identify the face. Ontology then will define the relationship between them. Finally, the system will then suggest the semantic annotations to the people and relationships.
>
> **USER:** The current user confirms whether or not all people and relationships are correctly annotated. If not, then the user will modify the mistaken annotations and annotated missed ones.
>
> **SYSTEM:** Instance Generator creates ontology about the photo including people and the position of faces. If the suggested annotations and confirmed annotations are differed, then the face learner learns the mistaken and missed faces using Principal Component Analysis (PCA) [6].

Fig. 3 shows two examples using SNPhoto. Fig. 3(a) shows the relationship between people, and Fig. 3(b) shows an annotated photo that has annotated information of people names, locations, and relationships. In addition, the KnowsOf property has 35 sub-properties such as friendOf, colleagueOf, mentorOf, etc. [8].

```
<Person rdf:about="#SANG-JIN CHA">
    <rdf:type rdf:resource="&owl;Thing"/>
    <closeFriendOf rdf:resource="#H.J.LEE"/>
</Person>
```

(a) RDF representation of relationships

```
<Photo rdf:ID="Photo_27">
      <hasHeight rdf:datatype="&xsd;int">480</hasHeight>
      <hasWebURL rdf:datatype="&xsd;string">
         Http://www.flickr.com/photos/37396664
         @N03/3921780263/</hasWebURL>
      <hasWidth rdf:datatype="&xsd;int">640</hasWidth>
      …....
      <hasPerson rdf:resource="#ME"/>
      <hasPerson rdf:resource="#H.J.LEE"/>
</Photo>
<Position rdf:ID="Position_104">
      <hasHeight rdf:datatype="&xsd;int">71</hasHeight>
      <hasWidth rdf:datatype="&xsd;int">58</hasWidth>
      <hasXCoordinate rdf:datatype="&xsd;int">283</hasXCoordinate>
      <hasYCoordinate rdf:datatype="&xsd;int">137</hasYCoordinate>
      <isAffiliatedWith rdf:resource="#Photo_27"/>
</Position>
<Person rdf:ID="ME">
      <belongTo rdf:resource="#Photo_27"/>
      <hasPosition rdf:resource="#Position_104"/>
</Person>
   ……
```

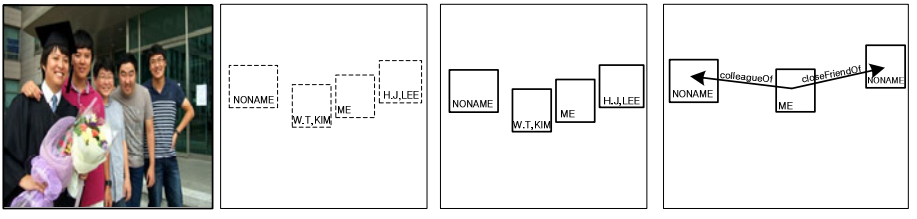
(b) RDF representation of the photo

Fig. 3. Semantic Annotation using SNPhoto Ontology

[3] http://opencv.willowgarage.com/wiki/

3.3 Visual Semantic Relationships Query

In spite of annotated photos with ontology, it is difficult for the retrieval systems to produce the best fit results for a query representing what's in the user's mind with simple keywords. Some researchers have addressed the need of query interfaces for retrieval systems to interpret the user's mind, which users explicitly represent their mind with simple and easy way [3, 4]. To address this need, we have designed VSRQ to allow users to represent their mind with the simple components as shown in Fig. 4.



(a) (b) (c) (d)

Fig. 4. Example of a user's query: (a) Target image (b) Query for considering only names (c) Query for considering both names and positions (d) Query for considering all names, positions and relationships

From Fig. 5, the user's query can be translated using three mechanisms. In the first mechanism, the user only needs to draw a dotted rectangle shape to represent the possible name of the person who may exist in the photos. This mechanism can be used if the user did not know about the position of the person in the picture. In the second

```
PREFIX SNPhoto: <http://eslab.inha.ac.kr/SNPhoto.owl#>
SELECT *
WHERE
{
  ?photo a SNPhoto:Photo.
  ?photo SNPhoto:hasWebURL ?url.

  ?photo SNPhoto:hasPerson ?person1.
  ?photo SNPhoto:hasPerson ?person2.
  ?photo SNPhoto:hasPerson ?person3.

  ?person1 photo:hasName ?name1.
  ?person2 photo:hasName ?name2.
  ?person3 photo:hasName ?name3.

  FILTER (REGEX(?name1, "W.T,KIM") && REGEX(?name2,
"ME") &&  REGEX(?name3, "H.J,LEE"))
}
GROUP BY ?url
```

(a)

```
PREFIX SNPhoto: <http://eslab.inha.ac.kr/SNPhoto.owl#>
SELECT *
WHERE
{
  ?photo a SNPhoto:Photo.
  ?photo SNPhoto:hasWebURL ?url.

  ?photo  SNPhoto:hasPerson ?person1.
  ?photo  SNPhoto:hasPerson ?person2.
  ?photo  SNPhoto:hasPerson ?person3.

  ?person1 SNPhoto:closeFriendOf ?person2.
  ?person1 SNPhoto:colleagueOf ?person3.
}
GROUP BY ?url
```

(b)

Fig. 5. An example of a user's query and an automatically generated SPARQL: (a) Both Fig. 4(b) and Fig. 4(c) are transformed (b) Fig.4(d) is transformed

mechanism, the user needs to draw a rectangle to represent the possible name of the person who may exist along with the possible position of this person in the photos. The third mechanism combines the second mechanism and relationships. Fig. 4(b) depicts that users wanted to find the photos that include at least three people, whose names are W. T. KIM, H. J. Lee, and the current user's name, but the user did not know about the position of the person in the photo. In Fig. 4(c) and Fig. 4(d), the possible positions of

each person are defined onto the specified positions. As shown in Fig. 5, the user queries are transformed into SPARQL by the SPARQL Generator. Both Fig. 4(b) and Fig. 4(c) are transformed into Fig. 5(a) because a location matching process is conducted on the results of the JOSEKI Server by Matching Engine.

3.4 Matching and Ranking

In order to retrieve relevant photos with a user's query, the vector model is utilized since the framework provides partial matching and ranking the photos according to their degree of similarity to the query. In our ranking system, the degree of similarity between the query, Q, and a j-th photo, P_j, in the answer set is calculated by the following equation:

$$Sim(Q, P_j) = w(qa, fa) \times \frac{Q \bullet P_j}{|Q \| P_j|} \tag{1}$$

where the value of weights, *w(fa,qa)*, is computed by using the locations of the user's faces; meanwhile the second component, usually called by cosine similarity in information retrieval literature, is the similarity based on two vectors that are composed of instances and relations of user ontology. This measurement is designed to consider these three important components;

- Users who are annotated or mentioned in a photo and query
- Relations that appear among users
- Positions that contain the faces of users.

The value of elements in vectors for users and relations is in binary form, while the positions are computed by normalization because the sizes of the photos are differs from each others. The weighting scheme is invented to measure how much of the sketched area in the query intersects the face area of the user given in the query. This is based on an assumption that users may draw a shape around the approximate location. In practice, however, this assumption is not concrete so that Euclidean distance is considered for that situation. Let *qa* and *fa* be the photo's rectangle which contains a face and the query's one, respectively, then *p(qa,fa)* is the percent of an intersection area between the two rectangles.

$$p(qa, fa) = \frac{ia}{fa} \times 100 \tag{2}$$

where *ia* is the area which intersects the two.

Euclidean distance between two points in Euclidean 2-space is also used for weighting the similarity. The distance from the point *fc* of the center of the face area in a photo to the point *qc* of one in a query is given by:

$$d(qa, fa) = \sqrt{(fc_x \quad qc_x)^2 + (fc_y \quad qc_y)^2} \tag{3}$$

where the subscript *(x/y)* indicates the co-ordinate. Finally, the value of weight is selected as shown in the following equation.

$$w(qa, fa) = \begin{cases} p(qa, fa) & if\ ia > 0 \\ \dfrac{1}{d(qa, fa)} & otherwise \end{cases} \tag{4}$$

4 Implementation

To demonstrate the effectiveness of the proposed approach, we have implemented the U2Mind system in C#. To extract faces, Haar Cascades[4], which is learned with 7,000

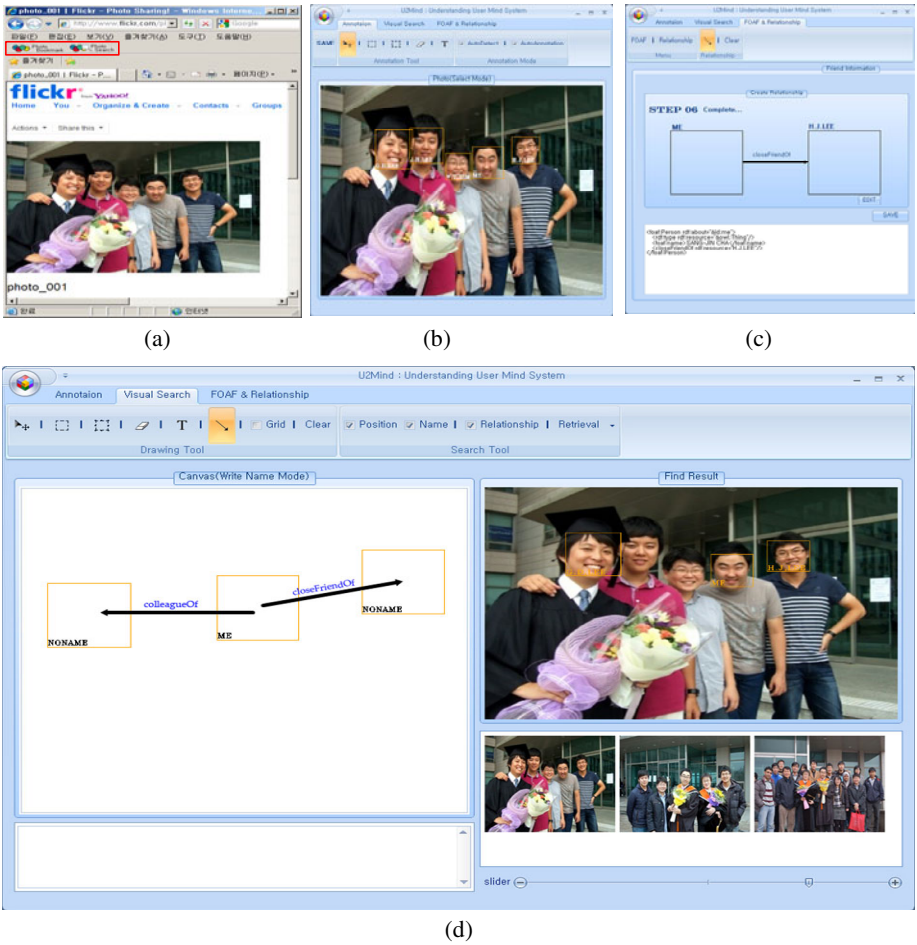


Fig. 6. Screen Shot of U2Mind Interface:(a) Screenshot of Bookmark Module (b) Screenshot of Annotation Module (c) Screenshot of set of relationship(d) Screenshot of Visual Semantic Query module

[4] http://alereimondo.no-ip.org/OpenCV/34

positive samples of face features such as eyes, nose and mouth, is used, and the PCA method in OpenCV is used to recognize the extracted faces. For searching ontology, the JOSEKI Server is used. GUIs for bookmarking photos, annotating ones, querying, and retrieving ones are shown in Fig. 6.

Let's assume that a user wants to find the photo as shown in Fig. 6(a). The user only remembers at least three persons and the relationships between them. The user can then place rectangles for each person on positions in which each person might be located on the desired photo. In addition, the user could put directed lines for relationships on the VSRQ interface. Finally, if the user clicks the Retrieval button, then the system might return the photos as shown in Fig. 6(d).

Fig. 7. Results for each query: (a) The top 12 photos among returned photos to the visual semantic query with an anonymous person and his position. (b) The top 12 photos among returned photos to the visual semantic query with an anonymous person, the user and their positions. (c) The top 12 photos among returned photos to the visual semantic query with two anonymous persons, the user and their positions. (d) Three photos among returned photos to the visual semantic query with two anonymous persons, the user, two relationships, and their positions.

In other words, the user could retrieve the desired photo through interaction with the system as shown in Fig. 7. Fig. 7 illustrates that the top 12 photos among the results, which are sorted in descending order according to the ranking scores, are displayed for each user's query step by step. In other words, if the user should refine his query with additional objects and relationships, then the system might re-rank photos that are not matched with what is on the user's mind.

5 Conclusion and Future Work

For photo retrieval systems in which stored photos might include people and relationships between them, it is a difficult task for systems to interpret the user's mind or intention from the given objects and their positional information. In this paper, we have proposed the novel VSRQ interface to allow users to formulate an implicit query in their mind with objects and relationships between them. In addition, we have designed the SNPhoto ontology to be applied for understanding user's intention in an accurate and explicit manner. Finally, to demonstrate the effectiveness of our approach, the U2Mind system has been proposed, implemented, and evaluated on the personal photos.

One of the limitations of U2Mind is its inability to retrieve photos for a query that has a property such as "anyone who my friend knows" because the proposed system can annotate people with user-centered relationships such as "my friend" and "anyone who I know". In order to overcome this limitation, future work will be dedicated to retrieving photos through inferring relationships between people on the social network in which ontologies, such as FOAF, are being shared.

References

1. Stone, Z., Zickler, T., Darrell, T.: Autotagging Facebook: Social Network Context Improves Photo Annotation. In: IEEE Workshop on Internet Vision, pp. 1–8 (2008)
2. Jung, J.J., Lee, K.S., Park, S.B., Jo, G.S.: Efficient web browsing with semantic annotation: A case study of product images in e-commerce sites. IEICE Transactions on Information and Systems, 843–850 (2005)
3. Wang, C., Li, Z., Zhang, L.: MindFinder: Image search by interactive sketching and tagging. In: 19th International Conference on World Wide Web, pp. 1309–1312 (2010)
4. Xu, H., Wang, J., Hua, X.S., Li, S.: Image search by concept map. In: 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 275–282 (2010)
5. Mathur, P., Karahalios, K.: Using bookmark Visualizations for Self-Reflection and Navigation. In: 27th International Conference Extended Abstracts on Human Factors in Computing Systems, pp. 4657–4662 (2009)
6. Perlibakas, V.: Distance measures for PCA-based face recognition. Pattern Recognit Lett., 711–724 (2004)
7. Kumar, N., Belhumeur, P.N., Nayar, S.K.: FaceTracer: A search engine for large collections of images with faces. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 340–353. Springer, Heidelberg (2008)
8. Davis, I., Jr, E.V.: RELATIONSHIP: A vocabulary for describing relationships between people, http://vocab.org/relationship/
9. Liu, Y., Zhang, D., Lu, G., Ma, W.Y.: A survey of content-based image retrieval with high-level semantics. Pattern Recognition, 262–282 (2007)
10. Kherfi, M.L., Ziou, D.: Image retrieval from the World Wide Web: Issues, techniques and systems. ACM Computing Surveys, 35–67 (2004)

A Personalized Recommendation Method Using a Tagging Ontology for a Social E-Learning System

Hyon Hee Kim

Department of Information and Statistics, Dongduk Women's University
23-1 Hawolkok-Dong, Sungbuk-Gu, Seoul, South Korea
heekim@dongduk.ac.kr

Abstract. As collaborative tagging has become increasingly popular, its role in social e-learning systems has attracted much attention. In this paper, we present a method to provide personalized recommendation services using a tagging ontology for a social e-learning system called TagSES. The TagSES ontology models relationships among students, lecture materials and tags, and the tags are mapped to the domain ontology. Based on the reasoning rules, students with similar interests are clustered and relevant lecture materials are recommended. To validate our method, we have implemented a prototype social e-learning system for a music history domain.

Keywords: tagging ontology, folksonomy, social e-learning.

1 Introduction

In a conventional e-learning system, the learning process is done by downloading and seeing lecture notes or videos, and thus, students are expected to be passively involved in the process. However, in a classroom in real life, they learn from other students' questions and answers, communicate with each other to solve problems, and make a group for a term project. Their social behavior is similar to features of the social Web sites, in that the sites make it possible for users to join a group, to communicate with each other, and to share user-created contents [1]. Furthermore, the social Web sites have functionality of collaborative tagging, which allows users to add keywords to shared contents by their own vocabularies [2].

Collaborative tagging is a powerful tool, which organizes contents for future navigation, filtering, or searching. Ontology-based annotation allows a user to use only pre-defined terms in the ontology [3], but in contrast, in the case of collaborative tagging, a set of tags, also known as folksonomy, has no hierarchies and pre-defined terms. Although such features of folksonomy sometimes lead to the semantic ambiguity such as polysemy, i.e., a single word contains multiple meaning, and synonymy, i.e., different words represent the same meaning, they enable users to find someone who has similar interests by browsing the tags of interests.

Recently, applying features of social Web sites to the traditional e-learning systems has been thus attempted and called a social e-learning system [4]. Actually, some social e-learning sites are available, and provide several social networking facilities.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 357–366, 2011.

In those sites, students freely form a study group, share their contents, suggest their opinion in the course, and collaborate with other students. However, most social e-learning sites do not make full use of folksonomy in the e-learning context. Usually, since folksonomy connects users with resources, it contains valuable information about students such as level of students' understanding, their social connection, and their interaction. Therefore, the use of folksonomy makes a social e-learning system more adaptive and personalized.

In this paper, we present a method to provide a personalized recommendation using a tagging ontology for a social e-learning system called TagSES as a combination of "Tag" and "Social E-learing System". First, we propose the TagSES ontology for folksonomy, which models relationships among students, lecture materials and tags. Second, we show reasoning rules to recommend students with similar interests and relevant lecture materials to a student. The tags in the TagSES ontology are mapped to the domain ontology, and then the reasoning rules are used to categorize concepts of tags and to classify students according the concepts. We have implemented a prototype social e-learning system for the music history domain.

The remainder of this paper is organized as follows. In Section 2, we mention related work. In Sections 3 and 4, we discuss the overview of TagSES and TagSES ontology, respectively. Finally, Section 5 provides concluding remarks.

2 Related Work

Research on ontology-based personalized e-learning has been done [5]. This approach utilized ontologies for three types of resources, i.e., domain, user, and observation, and inferred personalized information using reasoning rules. The main difference between [5] and our approach is the way to generate ontology. While most of existing ontologies are developed by domain experts in advance, TagSES ontology is generated from folksonomy databases on request dynamically. The benefit of TagSES ontology using folksonomy is that a user does not need to know pre-defined words by the domain experts. A user's words reflect on folksonomy, and thus TagSES ontology infers the user's personal preferences based on his tags.

From the point of view of a tag ontology, Gruber formed the basis of a tag ontology, which is represented by an object, a tag, and a tagger [6]. In our approach, TagSES ontology is also represented by the basic concept of [6], but it augments Gruber's model with a concept of cluster, which is grouping of relevant tags, users, and lecture materials. TagSES ontology provides reasoning rules for user information and applies the information to personalized services in a social e-learning environment. To resolve the problem of semantic ambiguity of folksonomy mentioned above, [7] extended the Gruber's model with several classes such as synonym and spellingVariant. The problem of semantic ambiguity in TagSES has not been resolved yet. Similar to [7], augmenting TagSES ontology with classes to represent tags' semantics might be a solution.

Recently, tag-based recommendation systems have been studied [8, 9]. Durao and Dolog [8] provide personalized recommendations by calculation of a basic similarity between tags. Further, they extend the calculus with additional factors such as tag popularity, tag representativeness and affinity between users and tags. In [9], they

proposed a social interest discovery system using user-generated tags, and showed that user-generated tags are effective to represent users' interests and to discover common interests shared by groups of users.

TagSES is very similar to LOCO-Analyst [10] in that both systems apply social Web sites to e-learning systems. LOCO-Analyst is a tool which provides a teacher with students' feedback of his courses by suggesting relevant tags from students' collaborative tagging activities. Relevant tags are chosen using the algorithm of computing context-based relatedness between terms from pre-defined ontologies and from students' folksonomy. LOCO ontology is defined by the domain expert and a teacher maintains LOCO ontology considering students' tags, whereas TagSES ontology is created from users' folksonomy and infers personalized information using reasoning rules.

3 Overview of TagSES

In this Section, we first present the architecture of TagSES, and then show a running scenario. Figure 1 illustrates the basic architecture of TagSES. The system is largely composed of five components: folksonomy databases, ontology repository, a conversion tool, an ontology mapper and a social networking engine. Users can upload lecture materials and add tags to the lecture materials through the web-based user interface. Their tagging activities are stored in folksonomy databases shown in the left side of Figure 1.

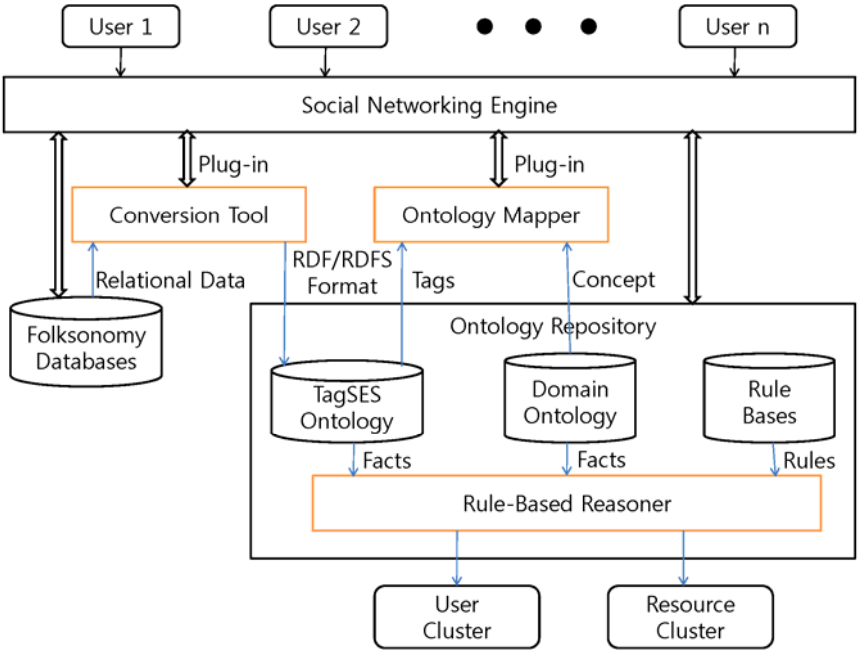


Fig. 1. The TagSES Architecture

The right side of Figure 1 shows the ontology repository which is composed of a domain ontology and a tagging ontology. The domain ontology for music history is an extension of the Music Ontology [11], which describes musical information. The tagging ontology models folksonomy, and folksonomy databases are converted to the tagging ontology represented by RDF/RDFS documents via the conversion tool. Tags in the tagging ontology are mapped to the domain ontology for music history via the ontology mapper. A tag is mapped to the concept of music history, and can be either a class or an instance in the ontology. All of the components are integrated through the social networking engine, shown at the top of Figure 1.

Next, we briefly introduce an application scenario which we will use through the paper. An instructor creates an online class called *music history* and uploads diverse type of lecture materials including lecture notes, audio files representing special genre of music history, video lecture about art history, study guides, etc. Students join the *music history* class, add tags to the lecture materials, and download them. Also, they upload other resources related to the course like essays, relevant images, videos, or audios and add tags to both their resources and others.

Now, the instructor wants students to perform a term project. The term project is studying a particular history of music e.g. baroque, renaissance, and romanticism. When students are asked to form a group for the project, it might be useful to know other students with similar interests. Also, it might be helpful for the system to provide relevant lecture materials about a special topic. TagSES provides these kinds of personalized services using TagSES ontology.

4 The TagSES Ontology and Reasoning

The TagSES ontology is composed of the domain ontology and the tagging ontology. The domain ontology formalizes the concept of music history, whereas the tagging ontology represents users' tagging activities by connecting tags with users and lecture materials. Reasoning rules for personalized services are defined in the rule bases. Let us take a closer look at the music domain ontology, the tagging ontology, and reasoning rules for personalized services.

4.1 The Domain Ontology for Music History

The domain ontology for music history is used to classify students into group of students with similar interests based on their tags which they added. In order to understand usage and pattern of tags in the general music domain, we collected datasets from last.fm [12], one of famous social music websites, using open API. After analyzing a set of tags, we realized that the set of frequently used tags in the music domain can be largely categorized as four major concepts: genre, artist, instrumentation, and work. In addition, the problem of semantic ambiguity mentioned above less exists due to features of specialized domain.

Based on the observation, the domain ontology for the music history is developed as an extension of the Music Ontology [11] which provides a standard base for musical information. The Music Ontology describes music information at three levels of details: Level 1 describes top-level editorial information, level 2 describes the

process behind the production of music, and finally level 3 describes the structure and component events of the music being played. Our domain ontology includes level 1 description of the Music Ontology, and augments it with the concepts of genre according to the musical history. The major classes and their properties of our domain ontology are explained as follows.

tagses:MusicalWork explains a work of music. It describes the title and the subject of a musical work with *hasTitle* and *hasSubject* properties, respectively. Also, the class has *yearofWork* property to explain the year in which a musical work was done. *MusicalWork* class is classified as *MusicalGenre* class via *classifiedAs* property. *MusicalWork* class is performed by artists, and the relationship is represented by *performedBy* property. Musical works, artists, and musical genre influence each other or reflexively, and the relationship is represented by *influencedBy* property.

tagses:MusicalGenre describes a trend of music according to history. It has six subclasses such as *MedievalMusic, RenaissanceMusic, BaroqueMusic, ClassicalMusic, RomanticMusic, and ModernMusic* classes. Examples of instances of *MedievalMusic* class are *chant, mass, motet, chanson,* and *madrigal*. It has *periods* property which describes a time during the life of music history. *MusicalWork* and *MusicArtist* classes are classified as *MusicalGenre* class via *classifiedAs* property.

tagses:MusicArtist describes a person or a group of people who perform music. It has three data properties, i.e., *hasName, periods,* and *region*. *hasName* property explains the name of a music artist and *periods* property describes a time during the life of an artist. Finally, *region* property describes the area of an artist's activity. *MusicArtist* class is also classified as *MusicalGenre* class via *classifiedAs* property. An artist performs a musical work, which is represented by *performs* property.

tagses:Instrumentation describes the particular combination of musical instrument for writing music. Examples of instances of *Instrumentation* class are sonata, concerto, symphony, string quartet, etc. *usedBy* property is used to represent a music artist used an instrumentation, and *usedFor* property is used to represent an instrumentation is used for a musical work.

4.2 The Tagging Ontology

In general, a tagging ontology is modeled as a tripartite relation composed of three common entities: users, tags, and resources [6]. We augment the basic concept of the tagging ontology in [6] with a cluster entity, as shown in Figure 2. The *Cluster* class represents grouping of tags, users, and resources, and therefore has three subclasses, *TagCluster*, *UserCluster*, *ResourceCluster*.

TagCluster class models a set of tags representing a concept of the domain ontology. Since a tag of the tagging ontology is mapped to a class or an instance of the domain ontology, a set of tags representing a musical genre are clustered. We explain the classes and their properties in more detail.

tagses:Users models a user's profile and has two subcalsses, *User* and *GroupOfUsers*. A student or an instructor might be an instance of *User* class. A study group or a social group created by users might be an instance of *GroupOfUsers* class. Each user in the system has a unique identifier and he can be a member of a group. *userID* and *member-of* properties represent the information, respectively. *hasAdded* property represents a user has added a tag to a resource. *Users* class belongs to *Cluster* class, and the relationship is represented by *belongsTo* property.

tagses:Tag models tags which users added, and has two properties, *tagID* and *tagName*. A unique identifier is assigned to a tag with *tagID* property, and a name of a tag is used with *tagName* property. *isTagof* property is used to identify a resource which owns a tag. A tag belongs to a cluster, which is modeled as *belongsTo* property.

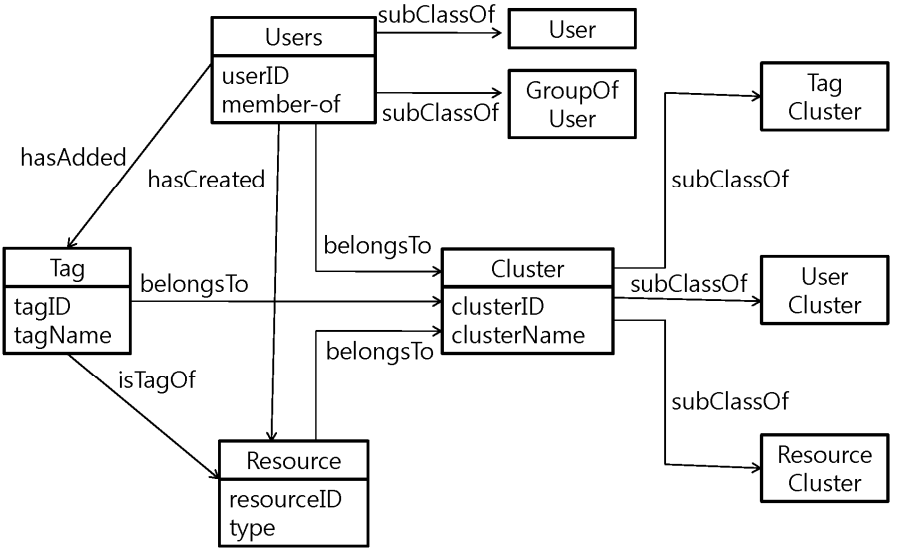


Fig. 2. Tagging Ontology

tagses:Resource models lecture materials such as lecture notes, audio files, video lectures, etc. Each resource has a unique identifier, and for this, *resourceID* property is defined. *type* property is used to identify types of lecture materials. A user can upload lecture materials, and *hasCreated* property is used to define the relationship. *Resource* class also belongs to *Cluster* class with *belongsTo* property.

tagses:Cluster is a core of our tagging ontology model. It has three subclasses, *TagCluster, UserCluster, ResourceCluster* classes. *TagCluster* class is grouping of tags representing a concept of the domain ontology. For example, a set of tags are clustered if those tags notify the concept of romantic music. *UserCluster* class is grouping of users, and users are clustered if their tags belong to the same *TagCluster* class. Finally, *ResourceCluster* class is grouping of resources whose tags represent a

concept of the domain ontology. Each cluster has a unique identifier represented by *clusterID* property, and a name of a cluster is assigned with *clusterName* property.

4.3 Rule-Based Reasoning for Personalized Services

We provide two different types of personalized services for users. One is recommending users with similar interests, and the other is recommending relevant lecture materials based on the cluster of users and resources. The reasoning rules are defined in three steps. First, tags associated with the concepts of the domain ontology are clustered. Second, users are clustered based on the cluster of tags. Finally, resources are clustered based on their tags.

Rules for Clustering Tags: First of all, all of tags are classified according to music history. Users might not know that the tags they used are classified as a specific genre. They just add tags about artist they like, music or instrumentation they are listening. The domain ontology and rule bases help users to know which historical genre they are interested in. There exist three types of rules to classify tags into historical genre: Using *MusicalWork* class, using *MusicArtist* class, and using *Instrumentation* Class. The following Rule 1, Rule 2, Rule 3, and Rule 4 show clustering of tags as romanticism. In the same way, rules for other historical genres are defined.

Rule1. Using MusicalWork class: If a tag is classified as romantic music, then the tag belongs in the cluster of romanticism.

```
(Tag (?X) ∧ hasName (?X, ?Y) ∧ classifiedAs (?Y,
RomanticMusic)) → Romanticism (?X)
```

Rule2. Using MusicArtist class: If a tag is classified as romantic artist, then the tag belongs in the cluster of romanticism.

```
(Tag (?X) ∧ hasName (?X, ?Y) ∧ classifiedAs (?Y,
RomanticArtist)) → Romanticism (?X)
```

Rule3. Using Instrumentation class: If a tag is an instance of instrumentation, and the instrumentation is used for a romantic work, then the tag belongs in the cluster of romanticism.

```
(Tag (?X) ∧ hasName (?X, ?Y) ∧ usedFor (?Y,
RomanticMusic)) → Romanticism (?X)
```

Rule4. Using Instrumentation class: If a tag is an instance of instrumentation, and the instrumentation is used by a romantic artist, then the tag belongs in the cluster of romanticism.

```
(Tag (?X) ∧ hasName (?X, ?Y) ∧ usedBy (?Y,
RomanticArtist)) → Romanticism (?X)
```

Rules for Clustering Users: Let us assume that a user wants to find another user or a group of users with similar interests. Most social Web sites provide categories of subjects, but it is difficult for a user to browse all the categories and to find users with the same interests by visiting each user who belongs in the category. Although some social Web sites suggest users of a specific category, the suggested users are usually famous person or objects. Since most users of an e-learning system are students, suggesting popular users with lots of on-line friends is not useful for students who want to discuss a special topic or to form a group in a class.

In case of last.fm site [12], music recommendations have been done by creating a user profile using the Scrobber, which records a user's history of listening to music. Also, the site suggests friends with similar musical taste based on the music items which users own. However, for using the service, users need to listen to at least 5 music items. In many cases, users cannot listen to all music which they are interested in. In addition, if a user's taste is very specialized, it is difficult to find relevant users.

Tags can give a clue to find relevant users, because tags reflect a user's interest. In order to add tags, users do not need to listen to music. That is, without listening to music, users can represent musical taste with tags. If users' tags are classified as the same cluster of tags, then we can conclude that those users have similar interests. Therefore, they are clustered as relevant users. The following rules show clustering of users who belong to *Romanticism* cluster of tags.

Rule5. If a user added tags which belong to *Romanticism*, then the user belongs to a *SIGRomanticism*, special interest group of romanticism.

```
(User (?X) ∧ hasCreated (?X, ?Y) ∧ hasName (?Y, ?Z) ∧
classifiedAs (?Z, Romanticism))→ SIGRomanticism (?X)
```

In the same way, all of users whose tags are related to the specific historical genre are clustered as clusters of users.

Rules for Clustering Resources: To cluster relevant resources, tags of resources are classified using domain-specific rules. If tags of several resources are classified as the same category, then those resources are clustered as *ResourceCluster*. Let us assume that there are three lecture materials: the video lecture material about impressionism, the audio file "the prelude to the afternoon of a Faun" composed by "Claude Debussy", and the lecture note about "Maurice Ravel" who is one of the impressionist composers. They do not seem to be relevant explicitly. However, all of works might have the same tag "impressionism", and then they can be clustered as *ImpressionismResource*. Rule6 shows how resources are clustered based on their tags.

Rule6. If a resource has a tag which belongs to the impressionism, then the resource is categorized as *ImpressionismResource* .

```
Tag(?X) ∧ isTagOf (?X, ?Y) ∧ hasTagname (?X, ?Z) ∧
classifiedAs (?Z, Impressionism) →
ImpressionismResource (?Y)
```

Now, we explain a running scenario using rules explained above. User "A" added tags like "Bach", "Cantata", and "Alessandro Scarlatti". User "B" added tags like

"Antonio Vivaldi", "Messiah", "Hendel". Based on the rule bases, user A and user B are clustered as a special interest group of baroque music. First, Rule1–Rule4 generates a cluster of tags, *BaroqueMusic*. Next, Rule5 assigns user A and user B to a cluster of users, *SIGBaroqueMusic*. TagSES recommends user A to user B, because they belong to the same interest group. Also, TagSES can recommend lecture materials of user A to user B.

5 Conclusions and Future Work

In this paper, we have presented a method to provide personalized services for a social e-learning system, using collaborative tagging. The TagSES social e-learning system is composed of folksonomy databases, the TagSES ontology, a conversion tool, an ontology mapper, and a social networking engine. TagSES allows students to share lecture materials, to add tags to them, and provides personalized services based on their tags which students added.

TagSES shows that folksonomy can play a crucial role in a personalized social e-learning environment. By integrating the domain ontology and the tagging ontology, the TagSES ontology infers a set of tags which is associated with a concept of the domain ontology, and based on the cluster of tags, infers relevant users and relevant lecture materials, too. Rule bases in TagSES are domain-dependent. We designed the domain ontology for the music history considering folksomony from last.fm [12], and rules to find a set of tags are limited to historical music genre.

In a class in real life, giving a personalized lecture, for example, providing different lecture notes considering students' interest or advising them to suggest other students with the same interest is somewhat limited. From this point of view, we expect that personalized e-learning services using folksonomy can complement the learning process in real life for both students and teachers.

We are implementing the prototype system on top of Elgg [13], an open source social networking engine. MySQL database system is used for the folksonomy databases and Jena [14], a Java-based Semantic Web framework, is used for the ontology repository. For the conversion tool, D2R Server [15] which is a tool for publishing relational databases on the Semantic Web is used. The ontology mapper to relate a tag with a class or an instance of the domain ontology is implemented by PHP. As a future work, we plan to develop a tool for ontology management which reflects changes of folksonomy.

Acknowledgments. This was supported by the Dongduk Women's University grant.

References

1. Kim, W., Jeong, O.R., Lee, S.W.: On Social Web sites. Information Systems 35, 215–236 (2010)
2. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems. J. Information Systems 32(2), 198–208 (2006)
3. Hunter, J., Schroeter, R.: Co-Annotea: A System for Tagging Relationships Between Multiple Mixed Media Objects. IEEE Multimdia 15(3), 42–53 (2008)

4. Kim, W., Jeong, O.R.: On Social e-Learning. In: Spaniol, M., Li, Q., Klamma, R., Lau, R.W.H. (eds.) ICWL 2009. LNCS, vol. 5686, pp. 12–24. Springer, Heidelberg (2009)
5. Henze, N., Dolog, P., Nejdl, W.: Reasoning and Ontologies for Personalized E-Learning in the Semantic Web. Educational Technology & Society 7(4), 82–97 (2004)
6. Gruber, T.: Ontology of Folksonomy: A Mash-up of Apples and Oranges. Int. J. on Semantic Web & Information Systems 3(2), 1–11 (2007)
7. Kim, H., Decker, S., Breslin, J.G.: Representing and sharing folksonomies with semantics. J. of Information Science 36(1), 57–72 (2010)
8. Durao, F., Dolog, P.: Extending a Hybrid Tag-Based Recommender System with Personalization. In: Proc. 2010 ACM Symposium on Applied Computing, SAC 2010, Sierre, Switzerland, pp. 1723–1727 (2010)
9. Li, X., Guo, L., Zhao, Y.: Tag-based Social Interest Discovery. In: Proc. 2008 World Wide Web Conference, WWW 2008, Beijing, China, pp. 675–684 (2008)
10. Jovanovic, J., Gasevic, D., Brooks, C., Devedzic, V., Hatala, M., Eap, T., Richards, G.: LOCO-Analyst: semantic web technologies in learning content usage analysis. Int. J. Cont. Engineering Education and Lifelong Learning 18(1), 54–76 (2008)
11. Raimond, Y., et al.: The Music Ontology. In: Proc. 2007 Int'l Conf. Music Information Retrieval, pp. 417–422 (2007)
12. last.fm, http://www.last.fm/
13. Elgg - Open Source Social Networking Engine, http://elgg.org/
14. Jena Semantic Web Framework, http://jena.sourceforge.net/
15. Bizer, C., Cyganiak, R.: D2R Server: publishing relational databases on the Semantic Web, http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/

Personalization and Content Awareness in Online Lab – Virtual Computational Laboratory

Krzysztof Juszczyszyn[1], Mateusz Paprocki[1], Agnieszka Prusiewicz[1], and Lesław Sieniawski[2]

[1] Institute of Computer Science, [2] Distance Learning Division,
Wroclaw University of Technology,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
{Krzysztof.Juszczyszyn,Agnieszka.Prusiewicz,Mateusz.Paprocki,
Leslaw.Sieniawski}@pwr.wroc.pl

Abstract. The architecture and functionality of virtual laboratory Online Lab, is presented, with respect to its unique features: unified interface for computational tools, scalability, advanced personalization based on social network analysis. Online Lab implements also content- and context-awareness mechanisms and is planned for the future use in content-aware networks' environment. The paper summarizes the first stage of the development of Online Lab working prototype.

1 Introduction - Online Lab

Online Lab (OL) is a generic framework for performing computations from a web browser in education or research. In particular, it allows its users, i.e. students or researchers, to access different kinds of numerical and mathematical software via Python or, in future, other programming languages and perform computations on the provided hardware, without the need for installing and configuring any software on a local computer.

The central concept of Online Lab is a notebook, which is a collection of cells of different kinds, that allows for storing rich contents (text, tables, images, LaTeX rendered mathematics) and source code, and allow for evaluating this source code on remote machines that are equipped with required numerical and mathematical software. The idea is that users are required to have just a modern browser, like Firefox or Chrome, installed on their computers and this is sufficient to be able to perform high-end computing.

Online Lab is a computer software project under development by hp-FEM group at University of Nevada, Reno, as a part of FEMhub project, and at Wrocław University of Technology in Poland, within Operational Programme Innovative Economy "Future Internet Engineering". Note that Online Lab is free software and is issued under a GPL-compatible license, thus given to you free of charge.

In this section the elements of the Online Lab architecture are introduced, the following section discussess the Online Lab approach to personalization and context-aware distributed computing.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 367–376, 2011.

1.1 Architecture

Online Lab (OL) implements three layer architecture consisting of user interface (OL-UI), core server (OL-Core), services and engines (OL-services) – Fig.1. Due to rich and flexible APIs on all levels, this approach allows for replacing parts or entire levels easily, making it possible to tailor Online Lab installation to custom needs. For example, if the default user interface does not fit into existing web sites of some institution, it can be replaced with a new one, tailored to the specific context. This way Online Lab can be embedded into existing web applications as another subcomponent.

Special measures were taken to guarantee that Online Lab can both integrate other web software components and be integrated itself with other web systems. One of the measures to achieve this, was to not use any global identifiers in CSS styles and JavaScript source code, as well as put all functions, symbols and definitions into unique namespace. This way clashes between names defined in Online Lab and other coexisting packages are not possible.

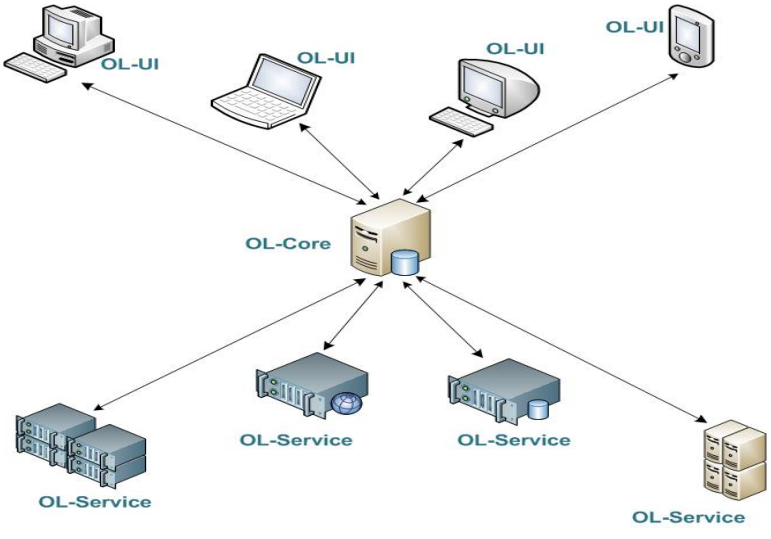


Fig. 1. Online Lab architecture

Communication between different layers in Online Lab is established via JSON-based (JavaScript Object Notation) remote procedure calls (RPC). We chose JSON-RPC over other RPC standards because of its little overhead (compared to e.g. XML-RPC) and natural binding with programming languages we use to develop Online Lab (JavaScrip) and Python). However, communication between services and engines, due to limitations of Python's standard library and testing convenience, is based on XML-RPC. This incoherence is a subject for change in future versions of Online Lab.

There is also RESTful API under development, in parallel with parts of JSON-RPC API, to allow easy programmatic access to Online Lab's installation resources on pure URI level. APIs of this kind are very useful for more advanced users, who would like,

for example, to export their work to some convenient format without the need to load the default user interface of Online Lab or even start a web browser.

1.2 User Interface

Is the most visible part of Online Lab to a typical user. It is a web application emulating a desktop and a window manager. This level is entirely written in JavaScript programming language with extensive use of ExtJS component library (from Sencha Inc.).

Usage of ExtJS library allows for rapid development of UI components due to its rich collection of predefined forms and elements, which can be used directly to created complex specialized components. It also helps with cross-browser compatibility issues, hiding, from Online Lab developers, many subtle differences between web browsers and allows fast development with reduced number of bugs and incompatibilities.

The default desktop-like user interface may seem a little heavy to some institutions, however, due to its rich APIs, Online Lab allows for implementation of other user interfaces tailored for specific needs of a particular users group. This may allow for fitting Online Lab's user interface to a web site design of a school or university that would like to employ this framework in their teaching process, keeping the look and feel of the final service coherent with institution's logotype.

This way, it is also possible to develop, in future, more lightweight user interfaces for mobile devices, like PDAs or modern mobile phones, as well as for touch screens, widening the range of applications of Online Lab. Making a truly and uniformly accessible platform for distributed educational and research purpose.

1.3 Core Server

This is the central point of Online Lab, where communication from different kinds of user interfaces is being handled and messages routed to appropriate services. On this level, there is also a relational database employed for storing user and system data. By user data we mean user's profile, folders, notebooks, comments on other peoples' work etc. System data consists of mainly of session and authentication data and logs.

There are two kinds of messages reaching the core level. Those messages can be either data related, for example when a user created a new notebook or wants to save contents of an existing one, or computations related, when a user wants to evaluate a certain Python expression. The former kind of messages is processed immediately and results are sent back to a client before any other messages are processed. The later kind is dispatched, enveloped and sent via an asynchronous channel to an assigned service.

Other messages can be processed during waiting time, before a service responds, thus during long computations the core server is not blocked and is responsive to other communications, even from the same client (for example it is possible to create evaluation queues).

Service binding is done using a simple least-load algorithm. The core server tries balance the load of every known service by assigning a new resource to a service that has minimal number of resources currently assigned. This may not lead to true load balancing, because system requirements of particular resources are not taken into account (this is a subject for future developments).

Although it may seem that only one core process can be running, it is possible to run several such processes behind a load balancer (e.g. nginx) to provide maximal throughput (there should be a single core process launched per CPU (or CPU core) of core server).

1.4 Services

These are remote, possibly high-end, servers equipped with actual numerical and mathematical software and Online Lab services infrastructure. This infrastructure allows for connecting with a core server and interpreting messages sent from it. Based on the contents of a message, Online Lab service can either start a new engine process, stop an engine process, gather statistics about a running engine process, and start or interrupt a computation. Thus a service may be viewed as a system process manager.

The services are implemented as wirtual machines which are created on demand of the OL-Core server. The virtualization environment for Online Lab is being developed on the basis of Xen, an open source industry standard for virtualization which offers tools for virtualization of x86, x86_64, IA64, ARM, and other CPU architectures [21].

1.5 Engines

This are computer programs running under an OL-Service supervision, that allows for performing the actual computations. Currently, an engine may only be a Python interpreter, or customized Python interpreter. In future other types of engines will be possible, to allow usage of other programming language and computational software directly, without the need for writing Python wrappers, which might be neither easy nor feasible.

2 Content and Context-Aware Based Personalization

Context is any information that can be used to characterize the user. Context may be defined as user activity (motion, speech), location, schedule (time cycles, calendar) and psychology (heat flux, temperature) captured from sensors [2,9,10,14,15,17]. But user context also includes the description of user hardware and software i.e.: operating system, applications that are used, computational and network capabilities. Context awareness means capturing and processing the values of the contextual user information to provide services efficiently [6,11,12]. It is the ability of extracting, interpretation and usage of context information to adapt application functionality to the current context of use [4,7,8]. While content-awareness means that the user data retrieval is performed with respect to user preferences and the specific QoS requirements of the content itself.

2.1 Context and Content-Aware Computing

To provide the services that are fitted to the users, applications must be aware and able to adapt to the changing context [3,5]. By the term of context-aware computing we call the applications that are context-aware. The main goal of the context-aware

systems is to acquire and utilize information on the user context to provide services that are appropriate to the particular people located at the particular time and place [8]. Nowadays the research about context-aware applications is very popular and the problem of the context-awareness is studied by many researchers. According to [7] the context-aware applications can be divided into six types: smart space providing users with smart environment, tour guide guiding the travellers, information system, communication system providing social community, m-commerce and web service.

Knowledge of a user context is used to personalize services. Providing personalized services based only on sensor data is not enough. There are some research about context-aware computing that take into consideration the user preferences. In such cases the user must give his preferences manually to receive personalized services or the system retrieves usage data to discover user preferences [7]. Various recommendation systems have been proposed. The taxonomy of such systems is described in details in [13]. The most popular area of application of recommender systems is electronic market. Several techniques of user data analysis for personalization purposes are given in [16].

We introduce the term of content-aware referring to the content that is personalized. The levels of the context and content awareness are captured on Fig. 2.

In Online Lab *content-awareness* is associated with data exploration and analysis techniques which may be divided into the following groups of functionalities:

- *Content annotation and metadata management.* The notebooks created by the users are annotated with respect to the problems they address and the methods and algorithms they use. The commenting and sharing capabilities will also be available to the users of Online Lab. This implies the use of domain ontologies which will be used for resurce annotation.
- *Recommendation* - finding notebooks and comments concerning the computational problem reported by the user.
- *User activity analysing* - discovering the patterns (temporal, social, geographical) of user activity (the low-level monitoring data used here are gathered on teh context level, discussed below) .
- Personalization - tuning the user interface and content presentation schemes to the role and hardware resources (the OL-UI may be accessed also with smartphones) of the users.

In fact the activity monitoring and personalization are also associated with the *context*, defined as communicational (network and its current state), computational (virtual machines, CPU and GPU power) and memory resources of the system. Thus, on the Fig. 2 the content and context areas mutually overlap. For example - the dimensions of personalization in Online Lab are classified on Fig.3 where personalization takes into account also the context domain.

The context-awarenessof of Online Lab assumes constant r*esource consumption monitoring* - for all notebooks (the actual tasks submitted to the system), OL-Services and their engines, the time of computations, communication delay, memory used by the OL-Services etc are monitored and stored in OL-Core database. In result we are able to directly associate the semantics of the system (i.e. problems being solved with annotated notebooks created by certain users) with access methods and typical resource consumption.

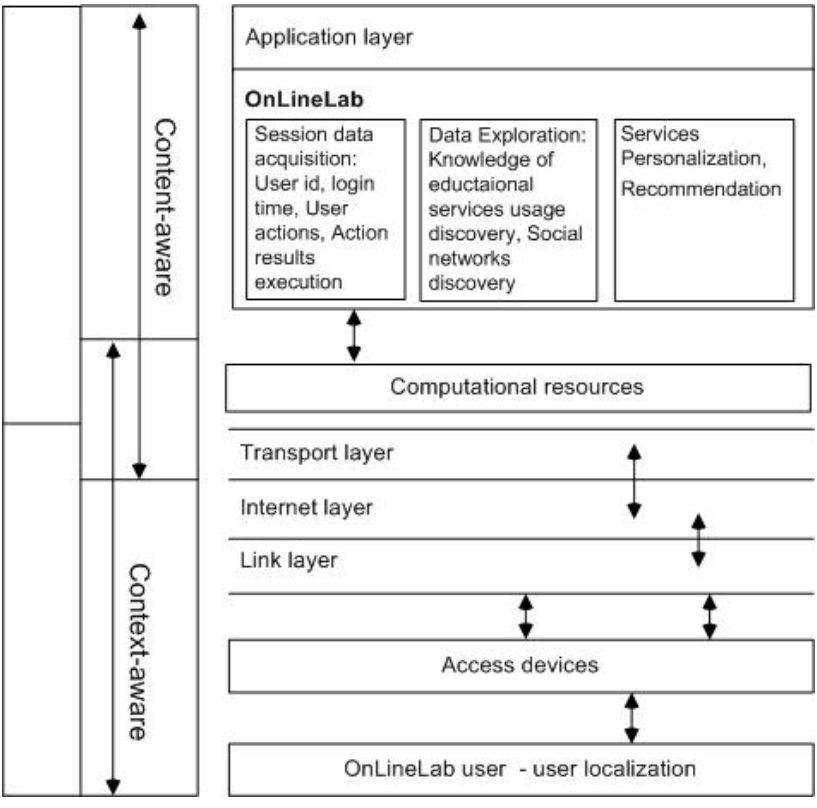


Fig. 2. The levels of the content and context –awareness in Online Lab

During the next stages of development these data will be used to predict the needs of the user community and then use the predictions to resource reservation in order to fulfill the QoS requirements of the users.

In the next section we present in detail an approach to personalisation developed for Online Lab.

2.2 Content-Based Personalization

According to the section 2.1 we can consider services personalization on two levels: context and content obtaining in this way two types of personalization: context- and content-based one. Generally the idea of the content-based personalization applied by the recommender system is given in Fig.4. To realize content-based personalization, first user preferences must be discovered. There are two ways of user preferences acquisition: manually or automatically. In the first case the user must directly give its interests (for example as ratings of a given set of items), in the second one – thepref-erences are discovered based on historical usage data analysis and then stored in user profiles. Several techniques for analyzing purchase for the purpose of producing use-ful recommendations are given in [16]. Among them are association rules and

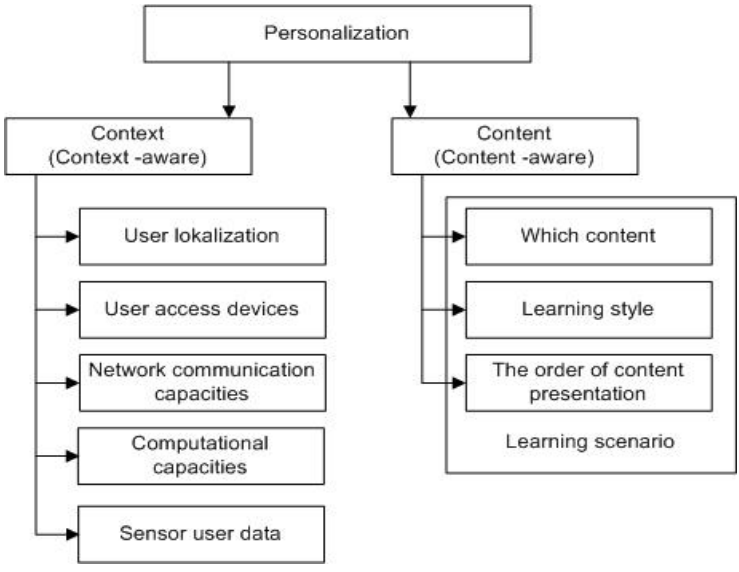


Fig. 3. Dimensions of personalization based on awareness in Online Lab

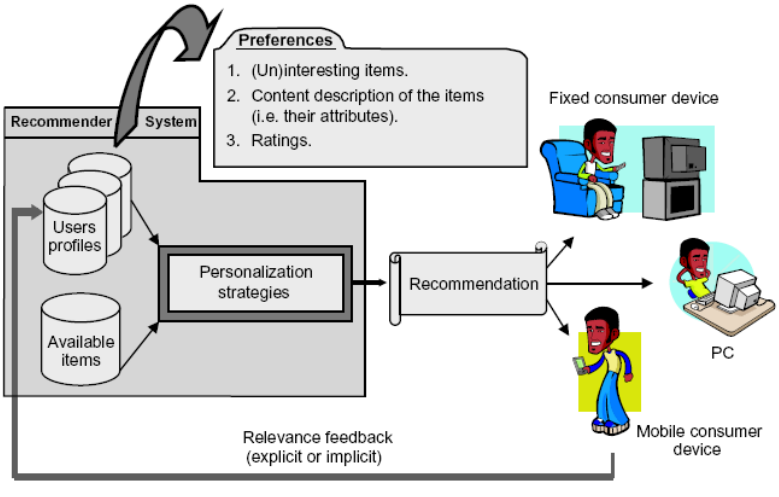


Fig. 4. The general idea of personalized content selection [1]

collaborative filtering as the most accurate recommender system technology. After recommendation is given the user sends explicit (by ratings) or implicit (by the user behaviour analysis in order to decide whether given recommendation is in the scope of user interests) feedback about its accuracy.

For Online Lab user educational services personalization purposes we will apply content-based personalization strategies that will be developed in the future work. To provide services personalization concerning content-aware computing some selected

methods of knowledge retrieval will be applied. The idea of services personalization is as follows: we collect and then retrieve data of Online Lab usage to discover user preferences represented later by their profiles. We will group together the similar users to recommend them the same notebooks. For user grouping and role detection we will apply the mechanisms for social network domain [20].

Social network analysis (SNA) can be defined as "the disciplined inquiry into the patterning of relations among social actors, as well as the patterning of relationships among actors at different levels of analysis (such as persons and groups)". Another definition of SNA was proposed by Valdis Krebs: "Social network analysis (SNA) is the mapping and measuring of relationships and flows between people, groups, organizations, computers, web sites, and other information/knowledge processing entities. The nodes in the network are the people and groups while the links show relationships or flows between the nodes. SNA provides both a visual and a mathematical analysis of human relationships" [18]. Measures (also called metrics) are used in social network analysis to describe the actors' or ties' characteristic features as well as to indicate personal importance of individuals in social network. In this section the methods that are used to evaluate the centrality of a user in the network are presented. Centrality measures called also user position measures can be applied in undirected/directed and unweighted/weighted networks. The high value of different centrality measures indicate various characteristics of the users. Based on these values the following roles (which can be formally defined in terms of measures mentioned above) may be assigned to users:

- Promoter of information is a person who in the most effective way (in comparison with others) spread ideas or information that he has received from other users or from the external environment. This phenomena can be investigated by applying the measures that take into consideration the measures of centrality based on the outdegree of a node.
- Prominent user is a person who is noticed by others as an important member. This phenomena can be investigated by applying the measures that take into consideration the measures of centrality based on the indegree of a node.
- Middleman is a person who acts as go-between in the communication and information exchange. The betweenness centrality is the measure that is utilised in investigating this phenomena.

As concluded in [19], if gathered and used strategically, detailed community behavioral (activity level) information about groups and individual user profiles within a community are more useful than community demographics. The abovelisted data will be gathered and processed in Online Lab environment.

3 Conclusions and Future Work

This paper presents the features of the first prototype implementation of Online Lab along with its specific approach to context- and content-awareness and personalized service delivery. During the next stages of development Online Lab (after introduction to everyday use as a tool for students and academic staff) will be a platform for testing the advanced social network analysis methods, user activity prediction

strategies and personalization schemes. It will be also used by the "Future Internet Engineering" European Regional Development Fund project as a case study for the development of network virtualization techniques and new concepts for content-aware network architectures.

Acknowledgements

The research presented in this work has been partially supported by the European Union within the European Regional Development Fund programme no. POIG.01.01. 02-00-045/09.

References

1. Blanco-Fernández, Y., Pazos-Arias, J.J., Gil-Solla, A., Ramos-Cabrer, E., López-Nores, M., García-Duque, J., Fernández-Vilas, A., Díaz-Redondo, R.P.: Exploiting synergies between semantic reasoning and personalization strategies in intelligent recommender systems: A case study. Journal of Systems and Software 81(12), 2371–2385 (2008)
2. Brunato, M., Battiti, R.: Statistical learning theory for location fingerprinting in wireless LANs. Computer Networks 47(6), 825–845 (2005)
3. Bolchini, C., Schreiber, F.A., Tanca, L.: A methodology for a very small data base design. Information Systems 32(1), 61–82 (2007)
4. Byun, H.E., Cheverst, K.: Utilizing context history to provide dynamic adaptations. Applied Artificial Intelligence 18(6), 533–548 (2004)
5. Dey, A.K.: Understanding and using context. Personal and Ubiquitous Computing 5(1), 4–7 (2001)
6. Figge, S.: Situation-dependent services? A challenge for mobile network operators. Journal of Business Research 57, 1416–1422 (2004)
7. Hong, J.-Y., Suh, E.-H., Kim, J., Kim, S.-Y.: Context-aware system for proactive personalized service based on context history. Expert Systems with Applications 36, 7448–7457 (2009)
8. Hong, J.-Y., Suh, E.-H., Kim, S.-Y.: Context-aware systems: A literature review and classification. Expert Systems with Applications 36, 8509–8522 (2009)
9. Krause, A., Smailagic, A., Siewiorek, D.P.: Context-aware mobile computing: Learning context-dependent personal preferences from a wearable sensor array. IEEE Transactions on Mobile Computing 5(2), 113–127 (2006)
10. Ladd, A.M., Bekris, K.E., Rudys, A., Kavraki, L.E., Wallach, D.S.: Robotics based location sensing using wireless ethernet. Wireless Networks 11(1-2), 189–204 (2005)
11. Lee, I., Kim, J.: Use contexts for the mobile Internet: A longitudinal study monitoring actual use of mobile Internet services. International Journal of Human–Computer Interaction 18(3), 269–292 (2005)
12. Lee, Y.M., Hong, J.Y., Oh, W.I., Kang, H., Suh, E.H.: A review of context aware computing. In: Proceedings of the 11th Annual Conference of Asia Pacific Decision Sciences Institute (2006)
13. Montaner, M., Lopez, B., De La Rosa, J.L.: A Taxonomy of Recommender Agents on the Internet. Artificial Intelligence Review 19, 285–330 (2003)
14. Niemegeers, I.G., Heemstra De Groot, S.M.: FEDNETS: Context-Aware Ad-Hoc Network Federations. Wireless Personal Communications 33(3-4), 305–318 (2005)

15. Samaan, N., Karmouch, A.: A mobility prediction architecture based on contextual knowledge and spatial conceptual maps. IEEE Transactions on Mobile Computing 4(6), 537–551 (2005)
16. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Analysis of recommendation algorithms for e-commerce. In: Proceedings of the 2nd ACM Conference on Electronic Commerce, pp. 158–167 (2000)
17. Satoh, I.: Spatial agents: Integrating user mobility and program mobility in ubiquitous computing environments. Wireless Communications and Mobile Computing 3(4), 411–423 (2003)
18. Krebs, V.: The Social Life of Routers. Internet Protocol Journal 3, 14–25 (2000)
19. Perey, C.: Beyond Eyeballs: Improving Social Networking Metrics, W3C Workshop on the Future of Social Networking Position Paper Christine Perey
20. Wasserman, S., Faust, K.: Social network analysis: Methods and applications. Cambridge University Press, New York (1994)
21. Xen Community Portal, `http://www.xen.org/` (retrieved October 30, 2010)

Workflow Engine Supporting RESTful Web Services*

Jerzy Brzeziński, Arkadiusz Danilecki, Jakub Flotyński,
Anna Kobusińska, and Andrzej Stroiński

Institute of Computing Science
Poznań University of Technology, Poland
{Jerzy.Brzezinski,Arkadiusz.Danilecki,
Anna.Kobusinska,Andrzej.Stroinski}@cs.put.poznan.pl,
Jakub.Flotynski@gmail.com

Abstract. An efficient business process execution and management are crucial for using a Service Oriented Architectures (SOA). Despite there are many applications offering such a functionality for Big Web Services, there is a lack of easy-to-use and well defined tools supporting the alternative approach, called ROA and RESTful Web-Services. In this paper the business process engine implementing a declarative business process language supporting web services compatible with REST paradigm is discussed.

Keywords: SOA, REST, business process engine, workflow, mashup.

1 Introduction

A *Service Oriented Architecture* (SOA) is currently a widely used approach to develop the complex distributed systems. A basic unit of the SOA is *a web service* — an independent software component offering functionality, which can be described, published, and discovered over the network using standard protocols. Web services can be used independently of each other, or they can be composed into a *business processes*, allowing the implementation of the complex system functionality by using already existing services performing simple functions. Besides providing a wide range of functionality, business processes also increase system component's life-time, re-usability and scalability.

The composition of a business process requires the definition of collaboration activities and data-exchange messages among involved web services. In order to enable this, the business processes description languages (workflow languages) are required. However, the description of the business process is not sufficient on its own, it needs some kind of environment that allows to execute it. Such an environment is provided by applications called *workflow engines*.

* Acknowledgment: The research presented in this paper was partially supported by the European Union in the scope of the European Regional Development Fund program no. POIG.01.03.01-00-008/08.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 377–385, 2011.

There are many tools supporting the business processes description and execution, available up to date. Among them are either high level programing languages (e.g. BPEL [13], GWDL [7], BPML [8], AGWL [9], Java Orchestration Language [14], YAWL [12]), or the abstract graphical notations, such as BPMN [16].

Unfortunately, the existing solutions usually fully-support only one of the approaches used to implement the SOA, called in the literature Big Web-Services [10], and based on the SOAP standard. Big Web-Services use a huge stack of defined standards, which are required to be met by developers during the system implementation. An alternative approach to the SOA implementation is the Resource Oriented Architecture [15] and the REST paradigm[10]. RESTful Web Services (web services implemented accordingly to the REST paradigm) have an easy to use and understand interface, based on the HTTP protocol. Moreover, the communication between RESTful Web Services and clients avoids a ballast, associated mainly with a necessity of the XML processing, occurring in the Big Web-Services. The increasing attention the RESTful Web Services gained recently, motivated the need of introducing the workflow language well-suited to this technology [11] and developing the workflow engine for that language. This paper is devoted to the proposed workflow engine, and its architecture.

The paper is structured as follows: Section 2 presents the work related to the existing workflow engines. A Restful Oriented Workflow Language, called ROSWELL is mentioned in the Section 3. Section 4 is devoted to the architecture and implementation of the proposed ROSWELL engine. In Section 5 the results of proposed engine evaluation tests are presented, and finally in Section 6 the conclusions and the future work are described.

2 Related Work

Because of the growing interest in applications compatible with SOA, many tools supporting their creation were proposed. Main disadvantage of existing tools is their focus on Big Web-Services and marginalization of ROA and REST paradigm. Below, the most important currently existing languages and workflow engines are discussed.

Oracle BPEL Process Manager [5] is a commercial engine, which offers many advanced functions, such as management of process versions, tracking of process execution, process dehydration, interaction with a human and the creation of group of servers (*clustering*). Main disadvantages of this product is its high price and only a partial support for REST paradigm. Another BPEL based workflow engine is an open-source Apache ODE. Its interesting feature is the implementation of WSDL 1.1 HTTP binding extensions, which allows the usage of HTTP methods to invoke services. Unfortunately this approach is not sufficient to operate on RESTful Web Services as a set of resources, therefore resource oriented architecture must be hidden behind the standard Web-Service interface. JOpera [4] is a tool distributed as the Eclipse platform plug-in. It introduces the JOpera Visual Composition Language, which is a graphical notation allowing simple and

fast network service programming. Such a description can be transformed to the BPEL code. Other JOpera features are human interactions and Java snippets. Developers preferring Ruby language can choose the free Route-REST [6]. This environment makes the REST API available, extends the Ruby language and allows programming of business processes on the high level of abstraction. Processing — a list of connected activities — is modeled as a finite state machine. Unfortunately, a precise description of activities, business process participants and relations between them is still lacking. The main drawback of a Route-REST is a very inaccurate documentation.

3 ROSWELL — RESTful Oriented Workflow Language

The important issue when choosing a workflow engine, is its support for a programming language. A declarative business process description seems to be a good alternative to the existing imperative languages. It focuses on defining goals to achieve and restrictions to fulfill, rather than listing successive steps in the application. Moreover, because declarative approach is more similar to a natural way of thinking, it represents a higher abstraction level and can speed up a programming process. Unfortunately, the declarative way of describing the business process is still not well supported by the existing workflow engines. There are only a few solutions executing a declarative business process descriptions, their documentation is however usually definitely unsatisfactory.

A business process engine presented in this paper is the implementation of the ROSWELL — a declarative language supporting service oriented architecture and the REST paradigm [11]. The supported language has a syntax similar to a Prolog language, enriched with instructions supporting REST paradigm. Similarly, to the declarative program structure, the ROSWELL is a set of goal declarations. The goal is described by requirements that must be fulfilled (*goal body*) and which are available by a predefined interface (*goal header*). The goal body is a list of *logical conditions,* which can be a *complex logic condition* (*alternative* or *conjunction)* consisting of other *alternatives*, *conjunctions* or *simple logical conditions* (*restrictions*). The *Restriction* can be defined as a goal, math operator or predefined keyword, such as *true*, *false* or a keyword supporting REST paradigm. The last important feature of the ROSWELL is the data representation, which is based on: constants (numbers and strings), variables and structures. A variable type is not declared, but it is set dynamically during a variable substitution. The structure is a hierarchic data type consisting of other structures and variables. Structures can be modified dynamically during the business process execution.

In the presented approach, SOA is implemented as a group of resources available through HTTP protocol. ROSWELL offers a native support for SOA and REST paradigm by special instructions associated with the individual ROA properties [15]: uniform interface, addressablility, statelessness and connectedness. The uniform interface enables access to resources by HTTP protocol methods. The considered language introduces some special keywords — *onGet*, *onPut*,

onDelete, *onPost*, *get*, *put*, *delete*, *post* — allowing defining an uniform interface and calling it from other resources [11]. Addressability is a REST paradigm feature, which means that each resource has its own unique address. This address is specified by resource creation, and during the remote resource calling. Statelessness means that any state can be stored without the specified address between two requests from the same client. ROSWELL, which the presented in this paper workflow engine implements, has a special instruction allowing this, which can be useful especially for applications using HTML forms. Connectedness means that representation of a resource has some links to the associated resources. This is reached by special annotations, which are placed by a programmer in the HTML form and automatically replaced by a business process engine during the business process execution. The accurate description of all mentioned instructions of ROSWELL is contained in [11].

An example of ROSWELL construction is shown below:

```
onPut("http://hospital.pl/patients/{Name}", PData, PResp) :-
    PData->Age>18, addPatientToDB(PData),
    get("http://laboratory.pl/patients/{Name}", Req, Resp),
PResp->Body=Resp.
```

In the considered example, a new resource accessed by HTTP PUT method with the specified address is created, with patient's name, input (`PData`) and output (`PResp`) data. In the first step it is checked whether the patient's age is grater than 18 years. Next, a new patient is added to a database. Finally patient's data from laboratory is got and it is sent to a client as a response.

4 ROSWELL Workflow Engine

Taking pros and cons of the existing workflow engines, described in the Section 2, there is a lack of easy to use, well supported workflow engine fully supporting the REST paradigm. Therefore a new workflow engine implementing ROSWELL language [11], described in the previous Section, has been proposed and implemented. Some of the main features of the proposed solution are discussed below.

The presented ROSWELL workflow engine is compatible with the SOA and the REST paradigm. It is available as the RESTful service and supports the HTTP methods such as GET, PUT, POST and DELETE. The engine enables the interaction with a human, who is either a principal, or the other participant of the business process. A simple web browser can be used as the engine client, due to the fact, that the whole communication between all business process participants goes through the HTTP protocol. It is important to mention that all features of ROA paradigm [15] are supported by the ROSWELL workflow engine.

There are also other features that characterize the presented solution: portability (possibility of installation on many different operating systems), scalability (possibility of multiplication and cooperation between many business process engines), availability (new business process installation doesn't interfere with another users activities) and lightness (listed features are implemented in the way which minimizes system load).

4.1 Architecture Basic Assumptions

The main factor impacting the business process engine architecture is the manner of processing the declarative business process language. The ROSWELL workflow engine was decided to be implemented as a *language translator.* This approach means that ROSWELL language is translated into one of existing programming languages and then application is performed by existing tools and technologies. The main advantage of this solution is the simplicity of translator creation, in comparison with the alternative approach — interpreter implementation. Additionally, the translated and later compiled code usually works much more faster then interpreted code. The portability of this solution depends on the existing tools for the target language. The disadvantage of this approach comes form the fact that all implemented business processes depend on these tools, so any change of a business process implies effects for all existing applications.

4.2 Modular Architecture

The ROSWELL workflow engine has a modular construction. It consists of five modules, which create a chain of processing. The modules are logically separated, they offer the specific functionality and cooperate with other modules in order to reach the listed goals. All modules of the proposed business process engine are presented in Figure 1 and described below.

The *Network interface* is the module responsible for receiving clients requests, unpacking declarative language code and directing it to the next module. An

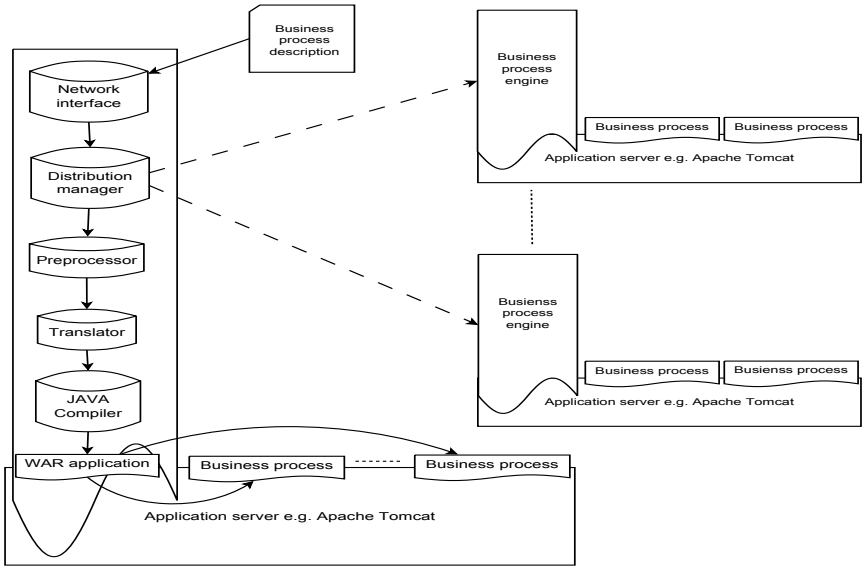


Fig. 1. Architecture and functioning of ROSWELL workflow engine

open-source platform implementing JSR-311 standard, called Jersey, was used to provide compatibility with REST paradigm. Applications supporting this paradigm can be started on any server supporting Java servlet technology. The *Distribution manager* is the module responsible for dividing description of business process into smaller parts, which can be installed and executed by other instances of the proposed business process engine. Such an approach enables distribution and multiplication of the executing business processes, so an efficiency and availability of the whole system is increased. A decision how to divide a process is made by a programmer or a dedicated algorithm. The *Preprocessor* is a module responsible for transformation of the input business process description into a more suitable for the next module, called translator, form. The preprocessor operates only on business process declarative language, so an output description is in the same language (the code is optimized). The first preprocessor task is a deletion of the redundant language constructions i.e. an exchange of facts to rules with an empty body. The second task is an expansion of the special language constructions i.e. generation of the additional code for Java snippets, to make it similar to other declarative languages rules. Preprocessor implementation was developed using Sun JavaCC [3] tool, which is a parser generator for Java language and supports LL grammars.

The *Translator* is the module responsible for transformation of a business process description received from the preprocessor into a description in particular programming language. Currently Java language is supported by presented business process engine. Implementations of Java Virtual Machines are available on the most popular operating systems, such as Windows or Linux, so this choice fulfills the requirement of portability. Similarly to the preprocessor, the translator module is generated by the JavaCC tool. The *JAVA Compiler* is the module responsible for compiling code generated by the translator to an executed code. In the discussed projectth eJava language compiler is used. The *Application server* is the outer tool, allowing the execution of the compiled applications. Such a role can play any Java application server, but the proposed business process engine uses Apache Tomcat [2].

4.3 Performance Tests

Tests comparing implemented ROSWELL engine (working on Apache Tomcat 5.5.28) with the Oracle BPEL Process Manager (Oracle SOA Suite 10g) were performed. The platform on which tests were performed has a subsequent parameters: AMD Sempron 3400+ processor, Gigabyte GA-M55plus-S3G mainboard, 900 MB Kingston CL5 RAM, operating system Windows XP. During the tests the speed of the execution of main language instructions, and the speed of performing main operations the engines offer were evaluated. A special application (a network service) was written for each test. Due to the limited number of article pages, only the supported declarative language application is quoted. The BPEL language listings are omitted.

The first tests type embraces:

- the substitution of the variable for constant string value "`test string`"
  `onGet("Test1/" , Req1 , Res) :- X = "test string".`
- the substitution of the variable for other variable integer value
  `onGet("Test2/" , Req1 , Res) :- INT = 3, X = INT.`
- the switch instruction comparing a variable value with a constant integer value
  `onGet("Test3/" , Req1 , Res) :- X = 5, X1 == 100; X = 30.`

Instructions listed above are simple, and perform very quickly, so they are repeated many times to obtain the reasonable tests results. Thus, instructions performed during tests were executed in a loop. The time was registered always before and after the loop. Each loop provides one average of the instructions execution speed. During tests, speed differences of both engines came to light. Thus, to obtain the credible results, numbers of repetitions should vary for both engines. It was intended to minimize outer short-lived factors, which could had some influence on the tests results. Therefore, loops were also performed many times and after each loop there were some breaks for a short time. A result of the test describes the average value of all executions performed in all loops. The precision of performed tests is 1ms.

The second type of tests contains subsequent comparisons: code compilation speed, the speed of application installations on the engine, utilization of RAM, and the speed of the synchronous requests realization. Each of the above tests were repeated 10 times. First two tests relied on the compilation and installation of the most simple application (doing nothing) on the tested engines. The precision of these tests is determined by Oracle JDeveloper 10g tool and equals 1s. Test of memory utilization was done using "Windows task manager". Memory utilization was measured each time before and after starting of workflow engine. The precision of such a test is 1 KB.

The last of listed above tests — speed of synchronous requests realization — was performed by the application Apache JMeter 2.4 [1]. 11 measurements were obtained for the following numbers of synchronous threads: 1, 100, 200, ..., 1000. Each thread simulated one client connecting to the engine and calling a test service. The time dependencies between starting threads were not set. A number of repetitions was fixed to serve always 10000 requests. The JMeter measured an average time of request realization — time from getting a message to sending a response. As previously, the most simple network service was used. Precision of such a test is 1 ms. Tests results are presented in the table 1.

According to the obtained results, the ROSWELL engine outperforms the Oracle BPM. The ROSWELL workflow engine results turned out to be better in all

Table 1. Tests results

Test	Oracle BPM	Business process engine	Relation OBPM/BPE
Substitution of variable by constant value [ms]	35,2047	0,0019	18895
Substitution of variable by other variable's value [ms]	1,4667	0,0178	82
Switch instruction [ms]	1,4312	0,0167	86
Speed of code compilation [s]	11,60	0,30	38
Speed of installation [s]	28,80	7,30	4
Utilization of RAM [KB]	654046	70863	9

performed tests, especially in the speed of instructions realization. Certainly, it is caused by the compilation of the declarative language to an efficiently executed Java code. The biggest difference was obtained for the test of substitution variable with a constant value. It shows that operations on string values are not a strong point of the Oracle BPM. Execution of the main administration task and utilization of the RAM came out better for ROSWELL engine with the minimal advantage of 4 times. Figure 2 presents times of requests handling in relation to the number of synchronous threads for both engines.

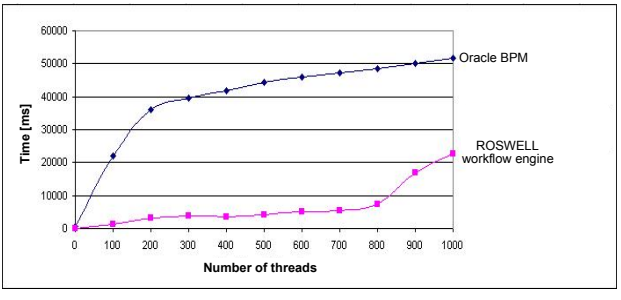


Fig. 2. Times of requests realization in relation to number of synchronous threads

The ROSWELL engine obtained a significant advantage, especially for the number of threads in the range 1..700 — the time of one requests handling is several times lower. There are several reasons of business process engine superiority in all the performed tests. Firstly, the declarative language is translated to the Java code, which is next compiled and executed very fast. The second important issue is the necessity of the XML processing through the Oracle BPM at each time when a new request occurs. This problem doesnot exist in the case of the business process engines with the discretionary form of requests. Moreover, the Oracle BPM is a very complicated system offering many advanced functions, so its memory and computational power demand is higher.

5 Conclusions and Future Work

The main target of the presented work was the creation of an efficient and lightweight business process execution environment. Such an environment should be able to execute a declarative processing and to allow RESTful web-services composition, according to the ROSWELL language assumptions. The presented performance results show a significant advantage of proposed solution — ROSWELL workflow engine — in comparison with the commercial tool, Oracle BPEL Process Manager. The presented solution could be an alternative to existing commercial, complicated, and mainly SOAP oriented systems.

Several extensions of the proposed workflow engine are planned, which can facilitate its usage. The first extension is tracking and modifying business processes

execution. Such a useful solution is currently available in the Oracle BPM. Another proposed improvement is the implementation of the described *Distribution manager* module. It would considerably increase the efficiency and reliability of the discussed environment. The last proposition is an implementation of all extensions, which are expected for the implemented ROSWELL language, such as asynchronous requests or parallel rule execution.

References

1. Home site of Apache JMeter project, `http://jakarta.apache.org/jmeter`
2. Home site of Apache Tomcat project, `http://tomcat.apache.org`
3. Home site of JavaCC project, `https://javacc.dev.java.net`
4. JOpera for Eclipse, `http://www.jopera.org/docs/help/jop.html`
5. Site of Oracle BPM project, `http://www.oracle.com/us/technologies/bpm/index.html`
6. Site of Ruote-Rest project, `http://github.com/jmettraux/ruote-rest`
7. Alt, M., Gorlatch, S., Hoheisel, A., Pohl, H.W.: A grid workflow language using high-level petri nets (2006)
8. Arkin, A.: Business process modeling language (2002)
9. Fahringer, T., Qin, J., Hainzer, S.: Specification of Grid Workflow Applications with AGWL: An Abstract Grid Workflow Language (2005)
10. Fielding, R.T.: Architectural styles and the design of network-based software architectures. Ph.D. thesis, University of California, Irvine (2000), `http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm`
11. Flotynski, J., Stroinski, A.: Declarative business process description supporting the restful paradigm. Tech. rep., Poznan University of Technology (2010)
12. Hofstede, A.H., van der Aalst, W.M.: Yawl: yet another workflow language (2005)
13. Louridas, P.: Orchestrating web services with BPEL (2008)
14. Pautasso, C., Alonso, G.: Jopera: A toolkit for efficient visual composition of web services (2004)
15. Richardson, L., Ruby, S.: RESTful Web Services. O'Reilly, Sebastopol (2007)
16. White, M.A.: Introduction to BPMN (2004)

From Session Guarantees to Contract Guarantees for Consistency of SOA-Compliant Processing

Jerzy Brzeziński, Arkadiusz Danilecki,
Anna Kobusińska, and Michał Szychowiak

Institute of Computing Science
Poznań University of Technology, Poland
{jbrzezinski,adanilecki,akobusinska,mszychowiak}@cs.put.poznan.pl

Abstract. Reliability of service-oriented distributed processing is gaining constantly growing attention. Some attempts to apply well-known fault tolerance techniques have been investigating interaction compensation, service replication or rollback-recovery, among others. For instance, the rollback-recovery approach promises to fully mask the occurrence of faults, allowing the critical or long-running applications (business processes) to automatically restore the consistent processing state. Unfortunately, the notion of consistent state is very ambiguous and has not been formalized in the context of Service-Oriented Architecture (SOA). In this paper we demonstrate how former approaches to specify consistency requirements for distributed shared memory can be adapted to the SOA environment.

Keywords: SOA, fault tolerance, crash-recovery, consistency.

1 Introduction

The *Service-Oriented Architecture* (SOA, [7]) is a paradigm for developing services that may be distributed through a coarse-grained (loosely-coupled) processing environment and shared by numerous applications or business processes. A *service* is an abstract processing entity accessible to consumers (*clients*) through a service invocation *interface* provided along with requirements for interactions. An interaction is typically processed with series of message exchanges (invocation requests and responses) in a particular execution context, i.e. a set of parameters that form a *contract* between a service and a consumer (service *participants*). The sequence of interactions, possibly nested (in which the invoked services become clients for further interactions), that follows from an initial invocation (of a particular business process), will be further regarded as a *session*.

It is typically assumed that services are stateless, in the sense that they do not maintain information about current state of sessions. However, interactions usually cause modifications of resources the services operate on. Thus, the state of processing is directly reflected in the resources state.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 386–395, 2011.

If any system components are prone to failures, the business process may stop at any moment of its execution. We consider in this paper the *crash-recovery* model of failures, i.e. the system components fail by stopping, and are eventually restarted. Since a simple restart of the failed business process from the very beginning will be usually an inefficient waste of time and may lead to unacceptable state inconsistency in distinct resources, a recovery mechanism can be introduced to cope correctly with this problem.

The goal of the recovery of service-oriented processing is to restore a state of the session, which conforms to consistency requirements imposed on the business process, and to resume the processing from the restored state. Unfortunately, there is no common agreement on the notion of consistency for service-oriented processing, and no formal definition of consistency requirements has ever been proposed for the SOA. This gravely prohibits construction of provably correct recovery solutions for the SOA-compliant distributed processing. Although different business processes may specify distinct consistency requirements, some common consistency properties can still be recognized. A subset of such properties could be used then to define a generic consistency model (or different models, according to specific application needs).

In this paper we aim at providing elementary consistency requirements — *contract guarantees* — which can be used to flexibly specify consistency models for recoverable SOA applications. To illustrate the applicability of the proposed contract guarantees, we also present an exemplary consistency model — *atomic consistency*. Our approach is similar to those known from the literature concerning earlier distributed processing paradigms, like the distributed shared memory.

The rest of this paper is structured as follows: the related work on rollback-recovery in distributed systems is outlined in Section 2. Sections 3 and 4 present our proposal and give necessary formal definitions. A sample consistency model for SOA is defined in Section 5. Finally, we briefly conclude the proposal in Section 6.

2 Related Work

2.1 Recovery in Distributed Systems

A wide range of rollback-recovery techniques for general distributed systems and distributed databases have been explored in the literature. In general, such techniques are based on the idea of periodically saving a current system state (a *checkpoint*) during a failure-free execution, to be able to restore an error-free system state from the saved data in case of failures. Depending on when and how checkpoints are taken, different approaches have been investigated [3]. Also for the distributed shared memory (DSM), several approaches have been considered, starting from adaptation of message-passing recovery techniques for memory coherence (consistency management) protocols, ending with sophisticated DSM-specific recovery schemes. Recently, the reliability of the SOA-based systems has been investigated, and some solutions for Web Services have been proposed (e.g. [6,12]). However, most of the existing solutions do not consider

nested invocation dependencies. Thus they are acceptable only for simple application scenarios. If nested interactions are not allowed, then checkpointing local state of some chosen replica of a single service is sufficient for the correct recovery. Such a simplified recovery approach has been commonly adopted in practice by business process execution engines, such as [8]. However, in the complex application scenarios, where nested interactions are used for service composition, distributed checkpointing is necessary to maintain a correct global state useful for recovery. A noticeable fault tolerance framework with distributed checkpointing for Web Services has been proposed in [2]. Unfortunately, this proposal requires complex fault detection and costly global recovery coordination, offering very strict consistency of the recovered processing state. It is clear based on the past experience [4] that many applications could benefit from less restrictive consistency guarantees, allowing the recovery of the processing state in a more efficient way.

2.2 Consistency Guarantees in Distributed Systems

In distributed message-passing systems, the consistency of the processing state is perceived from the inter-process communication point of view. Yet, in the SOA, the processing state is reflected in the state of resources held by invoked services, rather than in the message exchange pattern. Since the state of DSM systems is also maintained in the memory resources (shared objects), the idea of consistency perception of DSM seems to be also convenient in the service-oriented systems. For DSM processing, there exist several notions of consistency, more or less rigid, depending on the required guarantees. Relaxed consistency models offer employing more efficient consistency management techniques, still giving acceptable guarantees about the consistency of access sessions to shared resources.

In particular, *client-centric consistency* models [10] give a set of elementary *session guarantees*, which properly combined together, can lead to intuitive expectations about the system consistency, confirmed by literature ([1,4]) and successful practical projects, both academic (e.g.[9]) and commercial (e.g.[11]). Below, we recall their original definitions introduced for distributed systems with service-replication. In such systems, a client can issue operations on shared and replicated resources towards any available servers holding a replica of the requested object. Typically, only *read* and *write* operations are considered (the resources are considered to be read-write objects, for simplicity).

Read Your Writes (RYW) guarantee assures that, on the next read access (regardless to which one of the available replicas it has been issued), a client will observe results of all the former writes it has performed (even issued to other replicas). *Monotonic Writes* (MW) guarantee ensures global order of all write operations issued by the client (among all replicas). *Monotonic Reads* (MR) guarantee promises the client to always receive a read value no older than from the moment of the previous read issued by that client (whatever replica it has been addressed then). Finally, *Writes Follow Reads* (WFR) guarantee maintains

the causality relation between new writes and former reads issued by the same client.

Distinguishing different session guarantees enables to focus the implementation on delivering each guarantee independently and separately from the others, and individually for each single client (single session). This effectively simplifies system development and verification of desired consistency properties. Such an approach is remarkably attractive also, or especially, for reliability concerns.

In the next section we follow that track, aiming at defining the corresponding consistency guarantees for SOA.

3 Consistency Guarantees for SOA

The idea of correspondence between the consistency guarantees for DSM and those for SOA comes from the observations that a DSM process is considered a sequence of access operations (read/write) performed indirectly (via DSM servers) onto a set of shared data. Also, in the SOA a process consists of a sequence of interactions performed indirectly (via services) onto a set of resources. However, the exact definition of consistency guarantees for SOA needs a careful revision of differences between DSM processing model and SOA processing model.

The essential distinction between both models lies in the style of processing. In DSM, the processing is composed of access operations (on shared variables) issued by clients and performed by replicated DSM servers. The processing session is reflected by the set of operations issued by a given client, and the consistency concerns the order in which different servers perceive those operations. Yet, in the SOA, the processing consists of invocations (possibly nested) of operations provided by a set of loosely-coupled and independent services, and the consistency concerns the resource modifications caused by interrelated invocations.

Consequently, we make the following remarks:

- The notion of consistency applies in the SOA to the state of a set of resources maintained by services; for the sake of simplicity, by a *service state* we mean the state of all mutually related resources maintained by that service;
- A write operation in DSM corresponds in the SOA to such a service invocation which affects the service state ($w(R)$ in figure 1), i.e. its effect remains in the state (of resource R). Obviously, not all the service-oriented interactions change the internal state of services. However, we omitt here those that do not change it, since they do not influence the state consistency;
- A read operation in DSM corresponds in the SOA to presenting the service state to the client ($r(S)$ in figure 1), i.e. the result of the issued invocation is sent back to the client. Event if, typically in SOA, the service state is only partially exposed to the client within the response, that response is intended to update the internal state of the invoking client, and thus, it introduces consistency dependency between the client state and the service state;
- Since we do not consider service replication here (for the purpose of defining the consistency of service-oriented processing we assume the existence of only one instance of each service), therefore, the global order of operations made

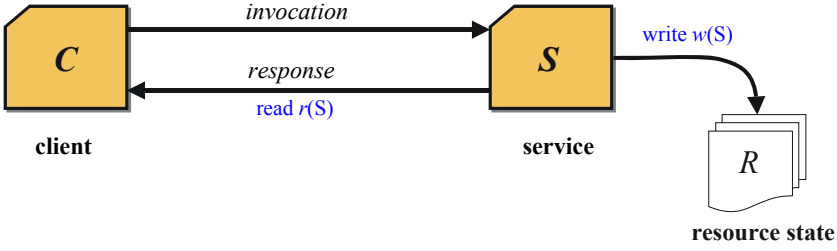


Fig. 1. Analogy between DSM and SOA processing models (blue color denotes DSM-compliant operations)

on a given resource is equivalent to the local order perceived by a single service. As a consequence, Monotonic Writes guarantee is not applicable in the SOA model and will be omitted in further considerations.

Based on the above remarks, we are ready to define the following consistency guarantees for service-oriented processing:

Completeness Guarantee (CG). This guarantee demands that, at the moment of accepting new invocation A issued by client C, the state of invoked service S reflects *effects* of all former (i.e. preceding A in the local order of S) invocations from C. Consequently, CG defines the persistency property of service-oriented interactions from a given client's perspective.

Actually, the business semantics highly influences the importance of particular interactions. Not necessarily all invocations issued by C must indeed be kept persistent. Only client C knows their semantics and can specify the CG requirement for each of them. We allow C to do that by explicitly labeling invocations as "requiring CG". Therefore, we restrict the above Completeness Guarantee only to holding for the explicitly selected invocations. Clearly, the Completeness Guarantee asserts the importance of the selected invocations.

This guarantee corresponds to Read Your Writes session guarantee in DSM.

Service Persistency Guarantee (SPG). Service Persistency Guarantee ensures that every interaction invoked by client C will operate on a service state that is up to date, i.e. comprising the state of the last preceding interaction invoked by C. Consequently, the SPG defines the persistency property of resources shared amongst all the clients. It is up to the service to declare SPG for its resources.

This guarantee corresponds to Monotonic Reads session guarantee in DSM.

Client Persistency Guarantee (CPG). Client Persistency Guarantee protects the business dependency (causality relation) between service invocations made by clients and resource modifications originating from those invocations. That is, the client will never deny the fact it has originated the particular invocation and will permanently accept its consequences (including business ones)

— even when the client shall fall down for a finite period of time (and recover afterwards).

This guarantee corresponds to Writes Follow Reads session guarantee in DSM.

It follows from the above definitions that CG reflects the expectations for the SOA system behavior (also in the presence of failures) that clients require from services. On the other hand, SPG and CPG reflect the resource properties, declared by services for their clients. Each guarantee can be declared for a given single interaction, as a part of the contract between a particular client and a particular service, and it does not necessarily hold for the whole processing session. For that reason, we will refer to them all as *contract guarantees*.

Given the above notions of contract guarantees in the SOA we shall make the following observations:

1. Consistent state of a service-oriented processing may be defined with a conjunction of all or only some of contract guarantees, depending on the desired consistency requirements, namely — *SOA consistency model*.
2. During the failure-free execution of a service-oriented processing, all contract guarantees hold naturally, i.e. the processing state, as we recognize it, is always consistent.
3. The re-execution of a service-oriented processing after a failure may be enabled only from such a restored processing state which preserves all the contract guarantees required by the imposed SOA consistency model.
4. The current state of processing can be identified with the history of operations which have triggered state changes from the beginning of the session (assuming given initial state). In the case of deterministic operations, that history can be represented (and logged for further reference) as an ordered set of invocations. Moreover, responses sent back to the invokers, also follow directly from the history of operations and, consequently, from the ordered sequence of invocations. For the rest of this paper we introduce the following abbreviations: *inv* — for the invocation, *resp* — for the response, and *int* — for the whole interaction.
5. In the case of any client failure and restart (from any state preceding the failure), CG and SPG guarantees for interactions invoked by that client trivially hold after the restart. However, the restored client state may turn out to be "outdated" when compared to the current state of services (remarkably, it is perfectly acceptable by the sole definitions of CG and SPG). In the case of deterministic processing, the client state can be easily "updated", if necessary, without re-execution of the in-between invocations, using the history of operations or — more precisely — the log of interaction responses.
6. If a service fails, the CPG is trivially preserved for all interactions involving that service.

Clearly, in order to propose any solutions and verify their correctness, we need to further formalize the notion of contract guarantees.

4 Formal Definitions of Contract Guarantees

In this subsection, we propose formal definitions of the terms related to the consistency of service-oriented processing. Then, we formally define contract guarantees for the SOA.

4.1 Formal Definitions of Service Interactions and Their Ordering

Interaction $int^x(C_i, S_j)$ is composed of a service invocation issued by client C_i and a response sent back by service S_j. Here, x is a sequence number used to identify the interaction (in our work, without loss of generality, we assume x to be a monotonically increasing value locally assigned to the interaction by client C_i). We distinguish the following events (figure 2):

- Sending invocation request $inv_i^x(C_i, S_j)$ by client C_i;
- Receiving invocation request $inv_j^x(C_i, S_j)$ by service S_j;
- Sending response $resp_j^x(C_i, S_j)$ by service S_j;
- Receiving response $resp_i^x(C_i, S_j)$ by client C_i.

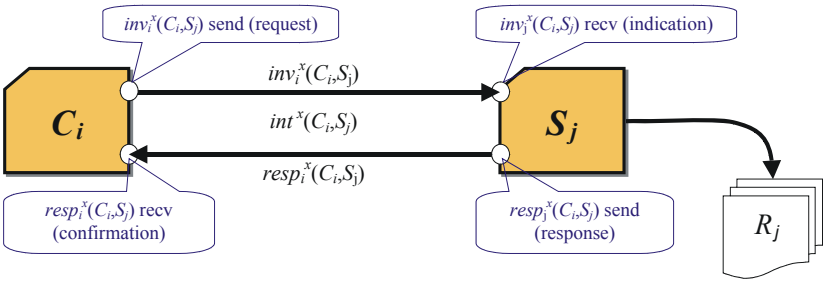


Fig. 2. SOA interaction events

Local precedence $\xrightarrow{C_i}$ *of events* is a total order relation between local events (here, local to client C_i), for which it holds:

$$\forall_{x<y} \; inv_i^x(C_i, S_j) \xrightarrow{C_i} inv_i^y(C_i, S_k) \tag{1}$$

Obviously, it always holds:

$$\forall_x inv_i^x(C_i, S_j) \xrightarrow{C_i} resp_i^x(C_i, S_j) \tag{2}$$

Analogically, the local precedence $\xrightarrow{S_j}$ is a total order of events from service S_j's perspective (formal definition can be omitted here).

Moreover, also *local precedence of interactions* can be distinguished and defined as follows:

$$int^x(C_i, S_j) \xrightarrow{C_i} int^y(C_i, S_k) \Leftrightarrow resp_i^x(C_i, S_j) \xrightarrow{C_i} inv_i^y(C_i, S_k) \tag{3}$$

The local precedence of interaction is only a partial order relation, as — in general — there can be *concurrent* interactions processed in parallel. Even interactions invoked by a single client can be concurrent in the case of *asynchronous* invocations or multithreading service processing.

Causal precedence$\rightarrow$ *of events* is a partial order relation holding the following:

$$resp_i^x(C_i,S_j)\xrightarrow{C_i}inv_i^y(C_i,S_k)\Rightarrow inv_i^x(C_i,S_j)\rightarrow inv_i^y(C_i,S_k) \tag{4}$$

$$resp_j^x(C_i,S_j)\xrightarrow{S_i}inv_j^y(C_k,S_j)\Rightarrow inv_i^x(C_i,S_j)\rightarrow inv_i^y(C_i,S_k) \tag{5}$$

$$inv_i^x(C_i,S_j)\rightarrow inv_j^x(C_i,S_j) \tag{6}$$

$$\left.\begin{array}{l}\wedge\ inv_i^x(C_i,S_j)\rightarrow inv_j^y(S_j,S_k)\\ \quad\ inv_j^y(S_j,S_k)\rightarrow inv_k^z(S_k,S_l)\end{array}\right\}\Rightarrow inv_i^x(C_i,S_j)\rightarrow inv_k^z(S_k,S_l) \tag{7}$$

Causal ordering preserves a local precedence relation perceived by the client (4) and the service (5). It reflects the natural precedence order between invocation send and receive events (6) and follows the nested invocation chain down the session (7).

In the case of synchronous invocations, there always exists a total order of interactions. Thus it holds:

$$inv_i^x(C_i,S_j)\xrightarrow{C_i}inv_i^y(C_i,S_k)\Rightarrow int^x(C_i,S_j)\xrightarrow{C_i}int^y(C_i,S_k) \tag{8}$$

On the contrary, asynchronously invoked interactions can be mutually interlaced. Thus it holds:

$$\neg\left(inv_i^x(C_i,S_j)\xrightarrow{C_i}inv_i^y(C_i,S_j)\Rightarrow resp_i^x(C_i,S_j)\xrightarrow{S_j}resp_i^y(C_i,S_j)\right) \tag{9}$$

As a consequence of the above, the client may receive responses in the order opposite to the invocation order.

4.2 Formal Definitions of Contract Guarantees

Completeness Guarantee. If client C_i requires CG when issuing current invocation $inv_i^{x_{curr}}(C_i,S_j)$, then:

$$\underset{x<x_{curr}}{\forall}\ resp_i^x(C_i,S_j)\xrightarrow{C_i}inv_i^{x_{curr}}(C_i,S_j)\Rightarrow resp_j^x(C_i,S_j)\xrightarrow{S_j}inv_j^{x_{curr}}(C_i,S_j) \tag{10}$$

Service Persistency Guarantee. If service S_j declares SPG, then:

$$\left.\begin{array}{l}\wedge\ resp_j^x(C_i,S_j)\xrightarrow{S_j}inv_j^y(C_k,S_j)\\ \quad\ inv_j^y(C_k,S_j)\xrightarrow{S_j}inv_j^z(C_l,S_j)\end{array}\right\}\Rightarrow resp_j^x(C_i,S_j)\xrightarrow{S_j}inv_j^z(C_l,S_j) \tag{11}$$

One can remark, that if S_j offers the SPG, then for every client C_i the CG guarantee trivially holds.

Client Persistency Guarantee. If service S_j requires CPG, then:

$$inv_j^x(C_i, S_j) \xrightarrow{S_j} resp_j^x(C_i, S_j) \Rightarrow inv_i^x(C_i, S_j) \rightarrow resp_j^x(C_i, S_j) \tag{12}$$

5 Atomic Consistency Model

To illustrate the applicability of contract guarantees for management of consistency, we introduce the atomic consistency model for the SOA.

Definition 1. *The state of interactions between participants of a given SOA processing is* ***atomically consistent*** *if for each service S and its client C, the following holds:*

1. *If client C has invoked service S, then consequences of invoking the service must be eventually reflected in the state of S.*
2. *If the state of C reflects (as a consequence of invocation) reception of interaction response from service S, then the state of S must eventually reflect the execution of this interaction.*
3. *If the state of S reflects the execution of an interaction (in the consequence of invocation from client C), then the state of C must eventually reflect this invocation.*
4. *The invoked interaction must be processed at most once.*

Fulfilling the requirements of atomic consistency can be achieved with contract guarantees. Indeed, atomic consistency involves the fulfillment of the SPG by service S (to accomplish the first condition of the definition) along with the fulfillment of the CG, required by client C for all its invocations (to accomplish the second condition), and fulfillment of the CPG, required from C by service S (the third condition). The important remark is that all these conditions may be accomplished independently and separately from each other. The last condition requires the *at-most-once* delivery semantics of invocations. Given a unique identification of *inv* events, such semantics may be easily provided by the communication sublayer.

6 Conclusions

In this paper, we introduced and formally defined contract guarantees aimed at flexible specification of consistency requirements for SOA-compliant processing. For illustration of their practical applicability, we have used them to define a sample consistency model.

Definition of contract guarantees is only the first step towards the development of correct and efficient recovery protocols for SOA systems. The main advantage of using contract guarantees is the possibility of fulfilling requirements of each guarantee independently of the others. This can highly simplify the consistency management in the case of failures and recovery of interaction participants. Hence, the restoration of a consistent processing state will require

to separately maintain each (and any) of the contract guarantees asserted for the processing (accordingly to the particular consistency model). We believe it will make the recovery protocols remarkably simpler to construct and easier to verify. We have developed some recovery protocols preserving contract guarantees in the scope of IT-SOA project [5]. Evaluation of those protocols and definition of more relaxed consistency models for SOA is a part of the remaining research required in this topic.

Acknowledgment. The research presented in this paper was partially supported by the European Union in the scope of the European Regional Development Fund program no. POIG.01.03.01-00-008/08.

References

1. Chockler, G., Friedman, R., Vitenberg, R.: Consistency conditions for a CORBA caching service. In: Herlihy, M.P. (ed.) DISC 2000. LNCS, vol. 1914, pp. 374–388. Springer, Heidelberg (2000)
2. Dialani, V., Miles, S., Moreau, L., Roure, D.D., Luck, M.: Transparent fault tolerance for web services based architectures. In: Monien, B., Feldmann, R.L. (eds.) Euro-Par 2002. LNCS, vol. 2400, p. 889. Springer, Heidelberg (2002)
3. Elmootazbellah, N., Elnozahy, Lorenzo, A., Wang, Y.M., Johnson, D.: A survey of rollback-recovery protocols in message-passing systems. ACM Computing Surveys 34(3), 375–408 (2002)
4. Friedman, R., Vitenberg, R., Chockler, G.: On the composability of consistency conditions. Information Processing Letters 86(4), 169–176 (2003)
5. IT-SOA project: New information technologies for electronic economy and information society based on soa paradigm (2009), http://www.soa.edu.pl/
6. Chen, J.-Y., Wang, Y.X.: SOA-based service recovery framework. In: Proc. of the 9th International Conference on Web-Age Information Management, pp. 629–635 (July 2008)
7. Laskey, K., McCabe, F., Brown, P., MacKenzie, M., Metz, R.: Reference model for Service Oriented Architecture. OASIS Committee Draft 1.0, OASIS Open (2006)
8. Oracle Corporation. Oracle BPEL Process Manager (2009), http://www.oracle.com/technology/products/ias/bpel/
9. Picconi, F., Busca, J.-M., Sens, P.: Pastis: a highly-scalable multi-user peer-to-peer file system. In: Cunha, J.C., Medeiros, P.D. (eds.) Euro-Par 2005. LNCS, vol. 3648, pp. 1173–1182. Springer, Heidelberg (2005)
10. Terry, D.B., Demers, A.J., Petersen, K., Spreitzer, M., Theimer, M., Welch, B.W.: Session guarantees for weakly consistent replicated data. In: Proc. of the Third Int. Conf. on Parallel and Distributed Information Systems (PDIS 1994), Austin, USA, pp. 140–149. IEEE Computer Society, Los Alamitos (1994)
11. Vogels, W.: Eventually consistent. Commun. ACM 52(1), 40–44 (2009)
12. Zhao, W.: A lightweight fault tolerance framework for web services. In: Proc. IEEE/WIC/ACM International Conference on Web Intelligence, pp. 542–548 (November 2007)

Design of a Power Scheduler Based on the Heuristic for Preemptive Appliances*

Junghoon Lee[1], Gyung-Leen Park[1,**], Min-Jae Kang[2], Ho-Young Kwak[3], and Sang Joon Lee[3]

[1] Dept. of Computer Science and Statistics,
[2] Dept. of Electronic Engineering, [3] Dept. of Computer Engineering
Jeju National University, 690-756, Jeju Do, Republic of Korea
{jhlee,glpark,minjk,kwak,sjlee}@jejunu.ac.kr

Abstract. This paper presents a design and evaluates the performance of a power consumption scheduler in smart grid homes, aiming at reducing the peak load in individual homes or buildings with reasonable computation time. Following the task model consisting of actuation time, operation length, deadline, and a consumption profile, the scheduler first investigates all the allocations for nonpreemptive tasks. Next, for each partial allocation, slots having the smallest power consumption are selected and assigned to the preemptive task, reducing the search space complexity for a preemptive task from $O(M^{\frac{M}{2}})$ to $O(1)$. The performance measurement result, obtained from the implementation of the proposed scheme and comparison with the optimal schedule, shows that the accuracy loss remains below 3.9 % for the number of tasks less than 9 and also below 7.6 % for the space size distribution. Moreover, the proposed scheme can find the optimal schedule more than 80 % for the given parameter sets. After all, our scheme can decide the power consumption schedule promptly with quite a small loss of accuracy.

1 Introduction

The modern power system is becoming more and more intelligent, integrating high-end information technologies in telecommunication, computing theory, and electronic consumer devices [1]. The new trend of power systems, called the smart grid, employs sensors, communication, computing facilities to enhance the overall functionality of the legacy electric power system [2]. From the viewpoint of customers, the smart grid allows them to smartly consume electricity not only by selecting the preferred supplier but also by scheduling the operation of each appliance according to the various conditions such as the price change, current load, and the like. This power network draws more attention to DSM (Demand-Side Management), which encompasses the entire range of management functions

* This research was supported by the MKE (The Ministry of Knowledge Economy), through the project of Region technical renovation, Republic of Korea.
** Corresponding author.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 396–405, 2011.

related to demand-side activities such as program planning, evaluation, implementation, and monitoring [3]. It can be instantiated in residential, commercial, industrial, and wholesale classes.

DSM is aiming at meeting the customer requirement as well as achieving the system goal such as peak load reduction, power cost saving, energy efficiency, and the like [4]. In the smart grid, the electricity rate is decided by a demand response mechanism, which manages the customer-side demand and shapes the load according to the supply conditions and load changes. The demand response pricing policy includes time-of-use rates, critical peak pricing, and real-time pricing. They commonly raise the prices when the demand for a service is at its highest, expecting consumer-side reaction in response to the high price [5]. The reaction is controlled by DSM in homes or buildings in cooperation with smart appliances. Each appliance has its own control strategy by which the customers can program its performance and set optimal performance levels. Anyway, DSM can manage the power consumption more intelligently and autonomously.

Figure 1 illustrates the home electricity management system, or H-EMS in short, in smart grid homes [6]. The DSM functions are performed at the home controller in H-EMS. As a networked appliance, the home controller basically coordinates the interaction among utility companies, home appliances, and monitor devices [7]. First, the controller exchanges the information on price and demand with the utility company via the WAN connection. Next, the controller interacts with home appliances mainly by the home area network, triggering their operation and metering their power consumption [8]. The control target ranges from the basic HVAC (Heating, Ventilating, and Air Conditioning) devices to the electric vehicle that needs to be charged. On the other hand, B-EMS (Building-EMS) performs DSM functions in buildings [9]. After all, we can expect the great increase in the number of devices that are to be controlled and scheduled in the energy management system. This paper will call the home controller and the building controller interchangeably as they share many common functions.

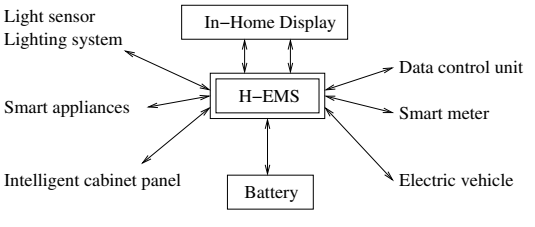


Fig. 1. Home electricity management system

The DSM functions can save money by enabling reduction of energy output during peak-demand hours based on the smart scheduling facility. The scheduling problem is, in some respect, quite similar to process scheduling in the real-time operating system [10]. Electric devices are modeled as real-time tasks having the execution time, the start time, and deadlines. The difference lies in that in energy scheduling, the device can run in parallel as long as the total power

does not exceed the current capacity of the transmission cables. One of the most common scheduling goals is to reduce the peak load as the power provider charges higher rates for the power consumption exceeding the given level. Moreover, if the system-wide peak load exceeds the current power capacity, more power plants must be built. Accordingly, peak load reduction is very important not just in economic but also in environmental aspects. Intelligent scheduling can possibly reduce the peak load. However, as the number of appliances to be managed increases, the scheduling complexity also gets too large. Moreover, some previous researches have revealed that a preemptive task severely increases scheduling complexity. To cope with high complextity, this paper is to design a heuristic-based scheduling algorithm, concentrating on the preemptive task, such as battery charge, which consumes steady amount of power.

This paper is organized as follows: After issuing the problem in Section 1, Section 2 describes the background of this paper and related work. Section 3 explains basic assumptions and designs a power scheduler for the smart grid homes. After performance measurement results are demonstrated and discussed in Section 4, Section 5 summarizes and concludes this paper with a brief introduction of future work.

2 Background and Related Work

As an example of power management, MAHAS (Multi-Agent Home Automation System) adapts power consumption to available power resources according to inhabitant comfort and cost criteria [11]. Based on the multi-agent architecture, the power management problem is divided into subproblems involving different agents, each of which tries to solve its own problem independently to find a solution of the whole problem. Particularly, the control algorithm is decomposed into reaction and anticipation mechanisms. While the first protects constraint violations, the second computes the plan for global consumption considering predicted productions and consumptions. The control function coordinates and negotiates the agent operations, sometimes even eliminating or adding new agents. This scheme seems to be scalable, as it can reduce down the whole search space for the given optimization problem. However, it cannot guarantee obtaining the optimal solution or avoid the complex interaction between the agents.

[12] discusses a scheduling problem for household tasks to help users save money spent on their energy consumption. Assuming the situation the users at home can select an electricity supplier company, its system model relies on electricity price signals, availability of locally generated power, and flexible tasks with deadlines. Particularly, this work adopts a descriptive task model where tasks are either preemptive or nonpreemptive, that is, tasks can be suspended or not, during their operations. A case study shows that cost savings are possible, but fast and efficient solutions are still needed for the scheduler to be deployed in the real-world. This problem stems from the fact that they use relatively fine grained time slots as large as 20 minutes, not employing any heuristic. So, the complexity of search space reaches $O(2^{MN})$, where N is the number of tasks and M is the number of time slots.

Our previous work has designed a power management scheme for smart homes, aiming at reducing the peak power consumption [13]. It finds the optimal schedule for the task set consisting of nonpreemptive and preemptive tasks, each of which has its own consumption profile as in [12]. To compensate for the intolerable scheduling time for the case of large number of tasks and slots, two speed enhancement techniques are employed. First, for a nonpreemptive task, the profile entries are linearly copied into the allocation table without intermittence. Second, for the nonpreemptive task, the feasible combinations are generated in advance of search space expansion based on the length of the task's profile entry and the number of slots between the start time and deadline of a task. Then, the scheduler maps the combination to the allocation table for each search space expansion. This scheme reduces the scheduling time almost to 2 %, compared with [12], that is, time complexity from $O(2^{MN})$ to $O(M^{N_{np}} \cdot (M^{\frac{M}{2}})^{N_p})$, where N_{np}, and N_p are the number preemptive tasks and nonpreemptive tasks, respectively. However, this speed-up is not still enough for practical use, considering the growing number of consumer appliances in homes or building.

3 Scheduling Scheme

To design a scheduling scheme, there are some assumptions. The scheduling function can be performed in the controller, mobile devices, high-capacity computing servers, and the like, while their logic is implemented in the controller. The proposed scheme assumes that the home power management system is able to make local decisions to control residential power consumption as well as to manage the demand through price signals [14]. To this end, the utility company sends a residential load change to the consumer's home, activating load control and consumption reorder. In addition, intelligent appliances and smart chargers can be controlled by the home controller, while the schedule can be generated and modified according to a task set change any time the customer wants.

Each appliance is modeled as a task and each task has its own power consumption characteristics. This paper focuses on two schedulable tasks, namely, nonpreemptive and preemptive tasks, respectively. A schedulable task can be started, suspended, and resumed by the controller. Nonpreemptive tasks can start any time after they get ready as long as they can be completed within their own deadlines. But, their operation cannot be preempted. A dish washer or a laundry machine belongs to this class. On the other hand, the preemptive task can be suspended and resumed within its deadline. The electric car charge is an example of the preemptive task. In addition, the load power profile is practical for characterizing the power consumption behavior of each appliance. Here, as pointed in [12], the load power profile for a washing machine depends on the set program, its duration and the water temperature chosen by the user.

Figure 2 plots a sample power consumption profile for 5 tasks. Task 1 gets ready at time 2 and takes 16 time units and so on. Task 4 has constant power requirement of 2 during the slots of 3 through 7. The length of a time slot can be tuned according to the system requirement on the schedule granularity and the

computing time. In power schedule, the slot length can be tens of minutes, for example, 20 minutes. Tasks 1 through 3 cannot be preempted while Task 4 and Task 5 can. The power consumption pattern for each electric device is aligned to the fixed-size time slot. Actually, each device has its own time scale in its power demand behavior. However, the power regulating functions in homes or devices themselves can keep the power consumption constant during the time slot. Here, we do not explicitly specify the power scale as this term has a relative value. As shown in Figure 2, the peak consumption can reach 11 in slots 6, 7, and 11, without scheduling. However, Figure 3 shows that the peak can be reduced to 7 by an optimal scheduling scheme. The goal of this paper is to reduce the peak power consumption by finding an efficient power schedule within reasonable computation time.

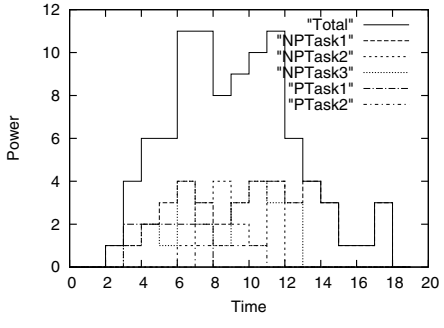


Fig. 2. Task set

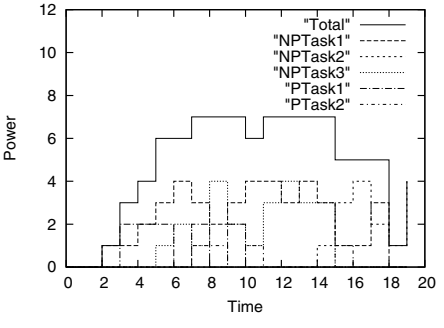


Fig. 3. Optimal schedule

Task T_i can be modeled with the tuple of $< F_i, A_i, D_i, U_i >$. First, F_i indicates whether T_i is preemptive or nonpreemtive. In addition, A_i is the activation time of T_i, D_i is the deadline, and U_i denotes the operation length. Each task is associated with the consumption profile entry which represents the power consumption amount for the U_i time slots. In the optimal scheduling scheme, all feasible allocations are investigated for the given task set. For a nonpreemptive task, for each feasible start time, which can be calculated by subtracting U_i from D_i, the profile entry is just copied to the allocation table one by one, as the task must not be preempted once it has started. The choice option is bounded by M, the number of time slots, hence the time complexity of search space traversal for a nonpreemptive task is $O(M)$. As contrast, for a preemptive task, there are ${}_{(D_i-A_i)}C_{U_i}$ allocations, resulting in the time complexity of $O(M^{\frac{M}{2}})$. However, the number of preemptive tasks is generally small and their power demand remains almost constant during their whole operations. This property allows us to achieve reasonable scheduling time just with small accuarcy loss.

Figure 4 illustrates the detailed scheduling algorithm. The allocation table consists of $M \times N$ fields. The allocation procedure fills the allocation table from the first row, each row being associated with a task. The procedure creates all feasible allocations for nonpreemptive tasks first. That is, if there are only non-preemtive tasks, the accuracy will be guaranteed. For each allocation, instead of

expanding to the preemptive task, say T_i, the scheduler selects U_i slots which have the smallest current power consumption, and assigns them to T_i. This procedure makes the time complexity $O(1)$, as just one allocation is taken. When the allocation procedure reaches a leaf, *EvalAlloc*() is called to check whether an allocation is better than current best in the maximum power consumption. If so, the current best is replaced. In the mean time, *CheckConstraint*() can speed up the search procedure by pruning the unnecessary search tree expansion. If the maximum power requirement of the partial allocation for the tasks from T_0 to T_i already exceeds the current best, it is no use proceeding to T_{i+1}.

```
procedure AllocTab (i)
input : {T_i|(F_i, D_i, A_i, U_i)}
        if i equals to N
            EvalAllloc ()
        end if
        if T_i is nonpreemtive
            for each start time from A_i to D_i − U_i
                copy the profile
                if (! CheckConstraint ()) AllocTab (i+1)
                end if
        else
            for each column from A_i to D_i
                calculate R_i by summing the rows in AllocTab
            select U_i slots having smallest R_i and allocate them to T_i
            map the profile
            if (! CheckConstraint ()) AllocTab (i+1)
            end if
        end if
end procedure
```

Fig. 4. Allocation procedure

4 Performance Measurement

This section implements the proposed allocation method using Visual C++ 6.0, making it run on the platform equipped with Intel Core2 Duo CPU, 3.0 GB memory, and Windows Vista operating system. The experiment sets the schedule length, namely, M, to 20 time units. If one time unit is equal to 20 *min* as in [12], the total schedule length will be 6.6 hours, and it is sufficiently large for the customer appliance schedule. For a task, the start time is selected randomly between 0 and M, while the operation length is also selected randomly, but it will be set to M if the finish time, namely, the sum of start time and the operation length, exceeds M. All tasks have the common deadline, namely, M, considering the situation that all tasks must be done before the office hour begins

or ends. In addition, the power level for each time slot has the value of 1 through 5. As this paper aims at enhancing the computation speed, the performance measurement concentrates on how much accuracy is lost and how much speed-up can be obtained, comparing with them optimal schedule implementation [13]. Accuracy is measured by the difference from the peak consumption obtained from the optimal schedule.

The first experiment measures the peak load reduction according to the number of tasks ranging from 3 to 9, and is carried out for the case of one preemptive task and two preemptive tasks, respectively. For each parameter setting, 50 task sets are generated. Then, the maximum power values, namely, peak load of respective sets are measured and averaged. Figure 5 and Figure 6 plot the accuracy for two cases, respectively. As shown in the figures, no significant difference can be found between two scheduling policies, indicating that our scheme can find the optimal schedule in most cases. The difference from the peak load obtained by the optimal schedule is 2.9 % at maximum when the number of tasks is 3 for one preemtive task case. In addition, for the case of 2 preemptive tasks, the difference increases to 3.9 %, still remaining very small. The more the preemptive tasks, the more search space will be skipped, possibly resulting in the increase of accuracy loss. However, its effect is negligible in our experiment scenario.

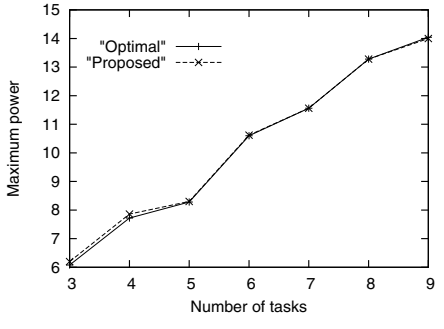


Fig. 5. One preemptive task

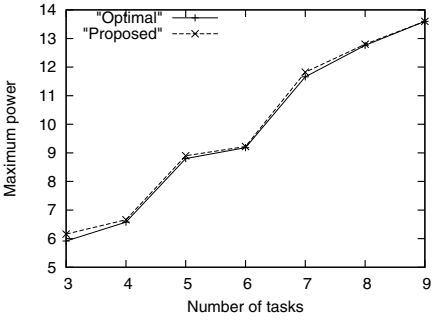


Fig. 6. Two preemptive tasks

The next experiment compares the execution time of optimal and proposed schemes. The execution time is measured using Microsoft Windows *GetTickCount* system call which has the 1 *ms* time granularity. Figure 7 and Figure 8 show the measurement results for the cases of one preemptive task and two preemptive tasks. As shown in figures, two preemptive tasks make the execution time significantly large, compared with the one task case for the optimal scheme, as can be inferred from the much larger values on the y-axis in Figure 8. As contrast, the proposed scheme can maintain the execution time below 1 *sec* until the number of tasks reaches 9 in both cases. Beyond that point, the effect of nonpreemptive tasks dominates the efficiency of the proposed heuristics for the preemptive task. The fluctuation in the execution time of the optimal

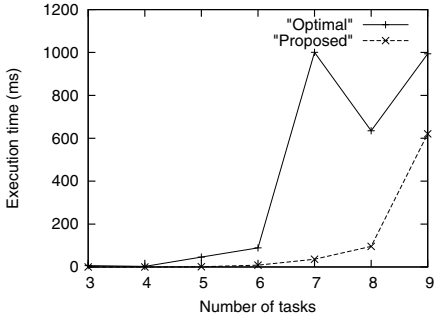


Fig. 7. One preemptive task

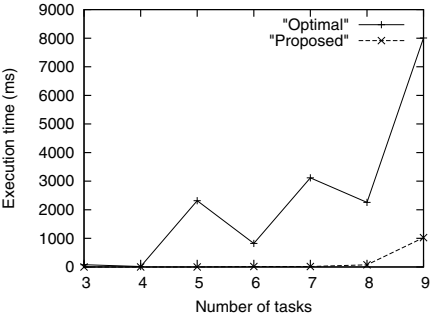


Fig. 8. Two preemptive tasks

schedule stems from the variance in the size of search space and the number of branches pruned by constraint checking.

Next experiment measures the effect of search space size to the execution time and the accuracy of the proposed scheme. In this experiment, the number of total tasks is set to 5, while that of preemptive tasks is set to 2. The value of 50 in the x-axis means each of two preemptive tasks has the search space size from 0 to 50, while the search space size is calculated by ${}_{(D_i - A_i)}C_{U_i}$. Hence, when the space size is 100 in Figure 9, it means that the search space tree could have 100×100 branches for each feasible nonpreemptive task allocation. Figure 9 indicates that execution time of the proposed scheme remains almost at 0, while that of the optimal schedule grows uncontrollably, reaching 180 *sec.* Figure 10 shows the difference in the maximum peak for the cases of optimal and proposed scheduling schemes. The peak powers of both schemes plot the same pattern, due to the dependency on the characteristics of the nonpreemptive task. If the number of suballocations for the nonpreemptive task is small, the number of skipped branches will be also small. Anyway, the maximum gap between two schemes is 7.6 % at maximum when the space size is 250.

We define the optimality as the probability that a scheduling scheme can find the optimal schedule for a given task set, and Figure 11 and Figure 12 plot the

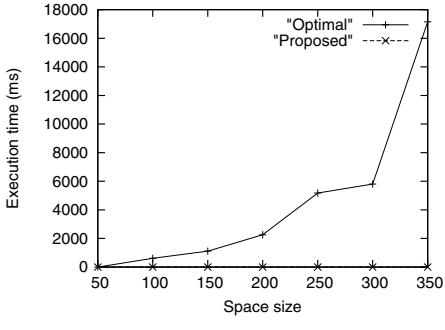


Fig. 9. Execution time

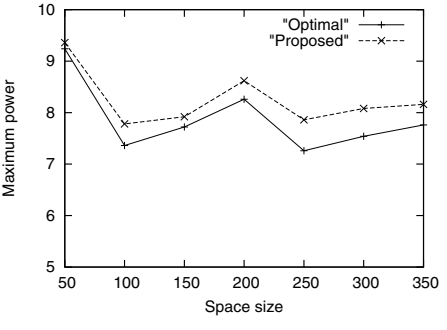


Fig. 10. Power consumption

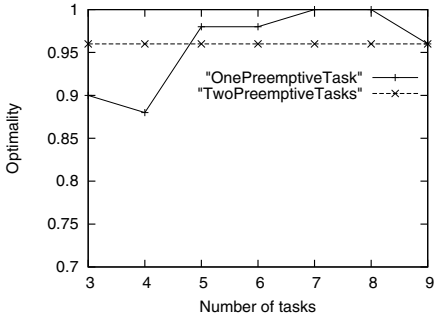


Fig. 11. Optimality vs. # of tasks

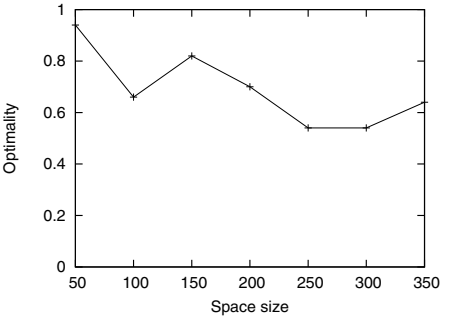


Fig. 12. Optimality vs. space size

optimality of our scheme. In addition to accuracy in the peak power, it is a useful performance metric to evaluate the efficiency of our scheme. In Figure 11, two curves correspond to the case of one preemptive task and two preemptive tasks, respectively. For the first case, when the number of tasks is 7 and 8, our scheme finds the optimal schedule for every 50 task sets for each. In the cases of two preemptive tasks, the optimality constantly remains at 0.96. Figure 12 plots the optimality according to the space size. Here, the number of total tasks is set to 5, while that of preemptive tasks is set to 2 again. The optimality decreases roughly along with the increase of the space size. However, for more than 3 out of 5 task sets, the proposed scheme finds the optimal allocation.

5 Conclusions

This paper has designed a power consumption scheduler capable of minimizing the peak load in individual homes and buildings. To cope with the uncontrollable execution time in finding an optimal schedule, a heuristic is designed for the preemptive task, whose search space complexity is originally estimated to be $O(M^{\frac{M}{2}})$. Following the task model consist of actuation time, operation length, deadline, and a consumption profile, the scheduler first investigates all the feasible allocations for nonpreemptive tasks. Next, for each partial allocation, the scheduler calculates the current load of each slot, select slots having the smallest load, and assigns to the preemptive task one by one. Even if this procedure skips many search space branches, the space complexity is reduced to $O(1)$. The performance has been measured based on the implementation of the proposed scheme and comparison with the optimal schedule already implemented in our previous paper [13]. The analysis shows that the accuracy loss remains below 3.9 % for the number of tasks from 3 to 9 and also below 7.6 % for the space size distribution. Moreover, the proposed scheme can find the optimal schedule more than 80 % for the given parameter sets. As for the execution time, according to the elimination of time complexity stemmed from the preemptive task, the nonpreemptive task becomes the sole complexity factor, allowing to schedule within 1 second.

As future work, we are planning to extend our task model to integrate price change dynamics, power reselling, home-generated electricity, and the like. In

addition, along with the power scheduler, we are also pursuing an efficient method to specify the power trade requirement from both sellers and consumers to design a power trader.

References

1. Gellings, C.: The Smart Grid: Enabling Energy Efficiency and Demand Response. The Fairmont Press (2009)
2. Ipakchi, A., Albuyeh, F.: Grid of the Future. IEEE Power & Energy Magazine, 52–62 (2009)
3. Bonneville, E., Rialhe, A.: Demand side Management for Residential and Commercial End-Users (2006), http://www.leonardo-energy.org/Files/DSM-commerce.pdf
4. Tan, Y., Liu, W., Qiu, Q.: Adaptive Power Management Using Reinforcement Learning. In: IEEE/ACM International Conference on Computer-Aided Design, pp. 461–467 (2009)
5. Spees, K., Lave, L.: Demand Response and Electricity Market Efficiency. The Electricity Journal, 69–85 (2007)
6. Mady, A., Boubekeur, M., Provan, G.: Optimised Embedded Distributed Controller for Automated Lighting Systems. In: First Workshop on Green and Smart Embedded System Technology: Infrastructures, Methods and Tools (2010)
7. Jeong, Y., Han, I., Park, K.: A Network Level Power Management for Home Network Devices. IEEE Transactions on Consumer Electronics 54, 487–493 (2008)
8. Luan, S., Teng, J., Chan, S., Hwang, L.: Development of a Smart Power Meter for AMI Based on ZigBee Communication. Power Electronics and Drive Systems, 661–665 (2009)
9. Lin, S., Guo, X., Chen, W., Zhang, W., Lin, Y.: An Automation Model for the Building Energy Management Systems: A Theoretical Study. In: 10th WSEAS International Conference on Robotics, Control and Manufacturing Technology, pp. 41–46 (2010)
10. Facchinetti, T., Bibi, E., Bertogna, M.: Reducing the Peak Power through Real-Time Scheduling Techniques in Cyber-Physical Energy Systems. In: First International Workshop on Energy Aware Design and Analysis of Cyber Physical Systems (2010)
11. Abras, S., Pesty, S., Ploix, S., Jacomino, M.: An Anticipation Mechanism for Power Management in a Smart Home Using Multi-Agent Systems. In: 3rd International Conference on From Theory to Applications, pp. 1–6 (2008)
12. Derin, O., Ferrante, A.: Scheduling Energy Consumption with Local Renewable Micro-Generation and Dynamic Electricity Prices. In: First Workshop on Green and Smart Embedded System Technology: Infrastructures, Methods and Tools (2010)
13. Lee, J., Park, G., Kim, S., Kim, H., Sung, C.: Power Consumption Scheduling for Peak Load Reduction in Smart Grid Homes (2011) (accepted at ACM Symposium on Applied Computing)
14. Dollen, D.: Intelligrid Consumer Portal Telecommunication Assessment and Specification. EPRI Technical Report 1012826 (2005)

Intelligent Information System for Interpretation of Dynamic Perfusion Brain Maps

Tomasz Hachaj[1] and Marek R. Ogiela[2]

[1] Pedagogical University of Cracow, Institute of Computer Science and Computer Methods, 2 Podchorazych Ave, 30-084 Krakow, Poland
tomekhachaj@o2.pl
[2] AGH University of Science and Technology, Institute of Automatics 30 Mickiewicza Ave, 30-059 Krakow, Poland
mogiela@agh.edu.pl

Abstract. This article presents intelligent diagnostic supporting system for interpretation of dynamic perfusion brain maps. The solution enables the detection, measure and description of abnormalities visualized in CBF and CBV perfusion maps. The algorithm also states prognosis for ischemic lesion evolution in brain tissues. The validation of the algorithm was performed on set of 37 triplets of medical images acquired from 30 different adult patients (man and woman) with suspicious of ischemia / stroke and compared to description done to each case by radiologist. Total error rate of the proposed solution was 23.0%.

Keywords: Dynamic perfusion CT, intelligent information systems, artificial intelligence, diagnostic supporting systems.

1 Introduction

The diagnostic supporting systems (DSS) become well known solutions in contemporary medicine. It is caused by the fact that medical informatics is in the field of interest both physicians and computer scientists. Only few of the DSS that are proposed to medical society becomes the popular and accepted solutions. The popularity and usefulness of DSS is determined by a few important factors. At first the proposed classifier must have proper specificity and sensitivity relation (in case of a binary classifier) [1] or low total error rate (TER) value. The second factor is the performance of the implementation of DSS. In many cases time between patient arrival, diagnosis and medical intervention is crucial for the health or even life of the patient. The last factor is intuitiveness of the mechanism of computer aided decision making process. The set of criteria of decision making should be easy to describe and understand. Also the process of automatic "reasoning" ought to be easy to follow and justify. The algorithms that imitate the way in which physicians state the diagnosis are especially valuable.

A physician has knowledge about important features of the diagnostic data that have to be taken in to account during diagnostic statement. He or she also knows how to interpreted the rise or lowering of analyzed coefficients. It is very difficult (or even impossible) to design valuable DSS without the complete information about medical

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 406–415, 2011.

procedures and domain knowledge about considered diagnostic procedure. Having that knowledge enables us to design the DSS that might facilitate the needs of every day medical practice and be considered as reliable tool.

This article presents a novel automatic method for computer-aided diagnostics of dynamic perfusion computer tomography (dpCT) maps that imitate the procedures performed by radiologists and satisfied the three postulates for DSS system presented above. The system enables the quantitative and quality analysis of visualized symptoms. The quantitative analysis consist detection and measure of potential lesions. The methods for the quantitative analysis was proposed in author's previous works [2], [3]. The quality analysis is based on description and analysis of nature of invalid perfusion regions. The main contribution of this paper is to present new methods that determinate the detected lesion type (hemorrhagic or ischemic) and state prognosis for lesion evolution in brain tissues. The algorithms presented in this article were used to create DSS that can be used when there is a suspicion of pathological states like head injuries, epilepsy, brain vascular disease, ischemic and hemorrhagic stroke that changes blood perfusion. The system is lately called DMD (detection measure and description) DSS system.

The process of an analysis o dpCT proposed by authors is a fusion of image processing, pattern recognition and image analysis procedures that results with accurate computer–aided diagnosis of the visualized symptoms. The output of each step delivers input data to the next procedure (Fig. 1).

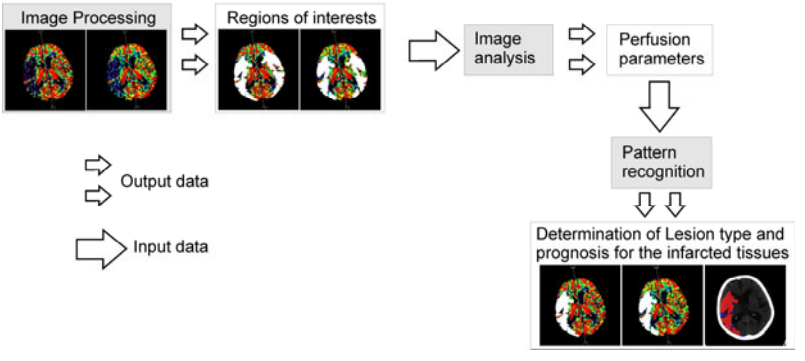


Fig. 1. The process of an analysis of dpCT proposed by authors. The output of each step delivers input data to the next procedure.

For the authors knowledge the system that enables same similarity as ours has not been yet described. The DSS system [4] enables detection of potential stroke regions on 3D MR scans of patients head so it is mainly limited for MR images. This 3D scans are used for atlas–to–image registration (with brain anatomy atlas - AA) in order to perform labeling of brain structures. The Fast Talairach Transform (FTT) for Magnetic Resonance Neuroimages is used. FTT was developed only for 3D Magnetic Resonance (MR) images and it is based on detection some "key points" of brain structure (structural - based registration) that are latterly used for rigid transformation of brain templates. In [5] authors presents a tool especially designed for the PWI-DWI data processing that integrates registration, segmentation, manipulation and visualization techniques. At first,

authors integrate DWI and PWI data in a common coordinate system by applying a registration technique. Second DWI data is processed to identify and measure the infarct area (a semi-automatic segmentation technique based on thresholding). Third, by using the information of the previous steps, PWI data is processed to identify the lesion. Due to the difficulty of processing PWI data segmentation process is supervised by the radiologist.

2 Dynamic Perfusion CT Maps Analysis

Dynamic perfusion computer tomography (dpCT) is a neuroradiology examination that enables to evaluate total and regional blood flows in time unit. In dynamic perfusion treatment the contrasting material is injected into the cardiovascular system relatively quickly (i.e. as an impulse injection). A CT scanner measures the contrast material that remains in the capillary network (on assumption that the blood brain barrier is intact, and if the contrast remains intravascular) creating set of computer tomography images. In some pathologic conditions of the brain that result in a compromise of the normally intact blood–brain barrier (BBB) (such as the presence of highly permeable blood vessels in a tumor), a portion of the tracer could diffuse into the extravascular extracellular space. This proportion of tracer that leaks into the extravascular extracellular space must be estimated with sufficient methods. As a result the time density curve (TDC, also called time intensity curve TIC) for brain arteries and tissues is obtained. Those TDC are the basic set for construction of perfusion maps. The most often used perfusion maps are Time To Peak (TTP), Cerebral Blood Flow (CBF), Cerebral Blood Volume (CBV) and Mean Transit Time (MTT). The value of pixels in TTP perfusion map is measured as maximal value of TDC curves. This map does not require any further calculation. In order to generate the rest of dynamic CT perfusion maps the adaptation of Meier-Zierler model is commonly used. Each pixel of a perfusion map corresponds to the value of perfusion in the given point. The color images help quick diagnosis of an acute stroke in the event of a crisis (Fig. 2).

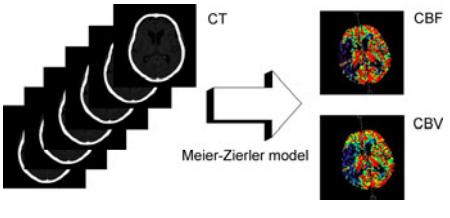


Fig. 2. Generation of perfusion maps. The set of CT images acquired in dpCT treatment is used for direct generation of TTP, and CBF, CBV ant MTT (not shown) with an adaptation of Meier-Zierler.

Brain perfusion imaging is currently used in case of head injuries, epilepsy, and brain vascular disease and especially in stroke diagnosing [6]. The dynamic perfusion imaging enables distinguishing between the ischemic and hemorrhagic stroke which is vital information for the treatment planning.

Despite the fact, that the average values for each of the perfusion parameter has been computed, the diagnosis is based on comparison of relative values of symmetric regions of interest (ROI) of blood perfusion between left and right hemisphere.

Many medical researches were done in order to determine correlation between values of perfusion parameters and long and short-term prognosis for examined brain tissues. It has been shown that CBF and CBV have prognostic values in evaluation of ischemia. In many cases simultaneous analysis of both CBF and CBV perfusion parameters enables accurate analysis of ischemia visualized brain tissues and predict its further changes permitting not only a quality (like CT angiography) but also quantitative evaluation of the degree of severity of the perfusion disturbance which results from the particular type of occlusion and collateral blood.

Despite of some minor differences authors determine average value o CBF for health brain tissues as 55 ml/100 g/min and CBV as 2.5 ml / 100g [7]. The dysfunction of neural cells begins when CBF value drop below 20 ml/100 g/min [8], [9], [6]. Continues drop of perfusion in rage of 10 – 20 ml/100 g/min may affect with cell death in many minutes to hours [6]. CBF of less than 10 mL/100 g of tissue per minute cannot be tolerated beyond a few minutes before infarction occurs causing permanent brain cells damage [8], [9], [6].

Because of some factors [10] the true CBF and CBV values in individual cases may be underestimated in a manner that is difficult to predict. As the result of that some authors prefers to use relative values of CBF and CBF (appropriately rCBF and rCBV). Authors in [10] states that a relative comparison of cerebral blood flow within corresponding areas of both hemispheres of the brain is possible without any limitations because the error of measurement is the same for both high and low CBF values. In [11] and [9] authors describes correlation between relative values of perfusion parameters and prognosis for ischemia evaluation. According to them there are three probable scenarios:

- Tissues that can be salvaged if there is a drop of rCBV and rCBF did not drop beyond 0.48
- Tissues that will eventually become infarcted if there is a drop of rCBV and rCBF did drop beyond 0.48
- Auto regulation mechanism will affects the tissues when CBF is decreased and CBV has correct or increased CBV

The “border values” between those classes was determined with discriminant analysis.

3 The Schema of Diagnostic Algorithm

The process of analysis of dpCT consists of image processing step, pattern recognition step and image analysis procedures. All of these stages will be described in this paragraph.

3.1 Image Processing

Image processing step is consisted of lesion detection algorithm and image registration algorithm. Lesion detection algorithm finds potentially pathologic tissues

(regions of interests - ROI). Image registration algorithm is used for creating a deformable brain atlas in order to make detailed description of visible tissues. Both algorithms are shortly described below.

A Lesion detection

The algorithm used for detection of potential lesions is The Unified Algorithm detailed described in [2] so only the basic concept of it will be presented. In order to perform a perfusion image analysis it is necessary to find lines that separate the left brain hemisphere from the right one (in the following text it is called a "symmetry axis"). In many CT scans the head of the patient is slightly rotated by a small angle towards the bed of tomography (the symmetry axis is not orthogonal towards to OX axis). The symmetry axis of the image is not always the same line that separates the brain hemispheres. This makes the problem more complicated.

The algorithm of symmetry axis derivation proposed by the authors find the symmetry axis by computing centers of masses of image horizontal "slices". The symmetry axis is approximated as the straight line that minimize the least square error between all center of masses coordinates.

After computing the symmetry axis, values of corresponding pixels on the left and right sides of the image are divided, that generates the asymmetry map. The asymmetry map is thresholded (in order to find difference of sufficient value). After that operation image morphological algorithms are used in order to eliminate small asymmetries that cannot be classified as the lesions. The same algorithm for detection of regions of potential lesions can be use for both CBF and CBV maps. After image processing step two symmetric regions are detected in left and right hemisphere. The determination of position of the lesion and its type (ischemic or hemorrhagic) is determined unambiguously in image analysis step.

B Image registration and atlas construction

AA approach for labeling brain images is reliable and broadly used method. The very important issue is to choose proper image registration technique. The goal of registration algorithm is to minimize given error function (maximize similarity function) between to images: the template image (also called static image) and the newly acquired image (moving image). Because of registration of 2D brain images and because of possibility of different gantry tilt between template and moving images it is not possible to use methods that are based on detection some characteristic structural elements in brain images. To overcome this authors used intensity–based approach. The correlation coefficient was used as a function that measures similarity between images before and after registration process.

$$cc(A,B) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(A_{ij}-\overline{A}\right)\left(B_{ij}-\overline{B}\right)}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(A_{ij}-\overline{A}\right)^2\sum_{i=1}^{n}\sum_{j=1}^{m}\left(B_{ij}-\overline{B}\right)^2}}. \quad (1)$$

$\overline{A}$ - average value of pixels colors

For registration purpose authors have chosen free-from deformation algorithm proposed in [12]. One familiar technique to represent a nonrigid transformation is to employ spline functions such as B - splines, because B - splines are locally controlled, they are computationally efficient compared to the other globally controlled splines.

$$ffd(x,y) = \sum_{l=0}^{3}\sum_{m=0}^{3} \beta_l(u)\beta_l(v)\phi_{i+l,j+m} \quad . \tag{2}$$

The parameter $\phi_{i,j}$ is the set of the deformation coefficients, which is defined on a sparse, regular grid of control points placed over the moving image. The functions β_0 through β_3 are the third-order spline polynomials [12]. The minimum of error function (1) was found by using gradient descent method. An example result of image registration algorithm is shown in Fig. 3. After image registration potential lesions are labeled with deformed brain labels.

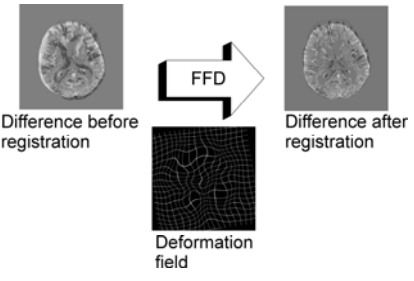


Fig. 3. An example result of image registration algorithm. On the left comparison of two brain CT images (static and moving) before registration process. During FFD registration process the deformation field is computed (in the middle). On the right comparison of the static image after applying deformation field and moving image.

3.2 Image Analysis

Defining the features of entire image after lesion detection step is an easy task. Algorithm measures some important (from medical point of view) features:

- Perfusion in ROI in left and right hemisphere
- Relative perfusion (perfusion in ROI in left hemisphere divided by perfusion in ROI in right hemisphere and perfusion in ROI in right hemisphere divided by perfusion in ROI in left hemisphere)
- Size of ROI

The scaling factors between perfusion map and “real brain” can be derived directly from DICOM files.

3.3 Pattern Recognition

In pattern recognition step algorithm determinate what type of lesion was detected and in which hemisphere. In order to do it is necessary to gather medical knowledge about average perfusion values that was described in first paragraph of this article.

After image processing step two symmetric regions are detected in left and right hemisphere. Author's algorithm compares perfusion in left and right (symmetrical) ROI with average perfusion norms and place potential lesion in hemisphere where modulus of difference between average and ROI value is greater. After this it is an easy task to determinate the type of lesion (hemorrhagic or ischemic) simply by checking if perfusion in ROI is greater or smaller than average. Example results of detected lesions placed in proper brain hemisphere are presented in Fig. 4.

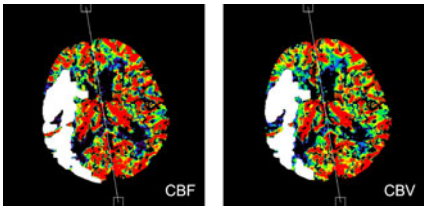


Fig. 4. Example results of detection. Lesions are marked as white region.

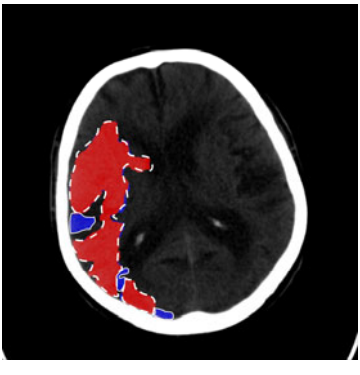


Fig. 5. Lesion description view after detection of potential lesions (lesions are visualized on plain CT axial image). Red region (border with broken line) mark tissues that are prognoses to become infarcted, blue region (border with solid line) - auto regulation mechanism are prognoses to occur in ischemic region.

The last step done by the algorithm is to state prognosis for lesion evolution in brain tissues. As it was written in first paragraph CBF and CBV have prognostic values in evaluation of ischemic evolution. In many cases simultaneous analysis of both CBF and CBV perfusion parameters enables accurate analysis of ischemia visualized brain tissues and predict its further changes permitting not only a quality (like CT angiography) but also quantitative evaluation of the degree of severity of the perfusion disturbance which results from the particular type of occlusion and collateral blood.

Algorithm analyze both perfusion maps simultaneously in order to detect:

- Tissues that can be salvaged (tissues are present on CBF and CBV asymmetry map and values of rCBF did not drop beyond 0.48)

- Tissues that will eventually become infarcted (tissues are present on CBF and CBV asymmetry map and values of rCBF did drop beyond 0.48)
- Tissues with an auto regulation mechanism in ischemic region (decreased CBF with correct or increased CBV)

The example visualized prognoses for brain tissues are presented in Fig. 5. Potential lesions should be compared with corresponding CT / MR image in order to check its presence there. This process will enable proper treatment planning. The detailed (AA based) description of image has been presented elsewhere [3].

4 System Validation

The validation of presented algorithms was performed on set of 37 triplets of medical images acquired from 30 different adult patients (man and woman) with a suspicious of ischemia / stroke. Perfusion maps were generated with Siemens software. Each triplet consisted of perfusion CBF and CBV map and "plain" CT image (one of the image from perfusion treatment acquired before contrast arrival became visible). The algorithm response was compared manually to image description done to each case by radiologist. The results of tests of image registration algorithm was presented elsewhere [3].

The researches showed that when lesion detection algorithm finds the position of perfusion abnormality agreeably with expert description the rest of the DMD algorithm also returns the result agreeably with those description. If the lesion detection algorithm does not find the position of perfusion abnormality agreeably with expert description the DMD system also does not return expected results. In other words the output of the lesion detection process is critical for the whole DSS performance.

The hypothesis to verify was if there is any lesions in perfusion map and if the algorithm found correct position, description and prognosis for them (if the algorithm gave a wrong answer for any of those conditions the case was considered as "error"). Total error rate (the proportion of error instances to all instances) of full automatic detection (without manual correction of position of symmetry axis) was TER=47%.

It can be clearly seen that automatic detection of symmetry axis can be used only to simplify further manual detection of symmetry axis. The huge error factor is mostly caused by overestimation of asymmetry regions (i.e. additional regions between brain hemispheres or in the top / bottom of the perfusion map). Automatic detection of symmetry axis can be very helpful in early estimation of visible lesions because of computation speed and because "real" axis differs only by few degreases, but still further manual correction may be necessary.

TER of semi automatic detection (with correction of position of symmetry axis) was TER=23%.

The errors might be eliminated by more accurate detection algorithm with additional adaptive factors i.e. taking into account the brain volume (for adaptation the minimal lesion size factor), the noise ration in the perfusion map (adaptive median filter size) or average perfusion in whole hemisphere (adaptive threshold for lesion acceptance).

5 Conclusions

In this work intelligent information systems for interpretation of dynamic perfusion brain maps was introduced. The fusion of image processing, pattern recognition, image analysis procedures and medical knowledge enables to perform accurate computer–aided diagnosis of the visualized symptoms. The authors system enables - in a way similar to physician who analyze perfusion data – draw conclusions about the nature of the observed disease process and the way in which this pathology can be overcome using various therapeutics methods.

Authors also plan to improve the system with additional functionalities:

- Include additional stage of reasoning algorithm that determinates relation between lesions visible in CT / MR images and perfusion map giving more accurate prognosis for lesion evolution.
- Include other algorithms for brain atlas construction (Thirion – demon algorithm, fluid registration [13]) that might give more accurate results in the process of image registration.
- Enlarge the knowledge base of the system in order to state more precise diagnoses by including the bloody supply territories atlas (BSTs) [14] that will provide more information for description and reasoning process.
- Improve the recognition rate of the algorithm by adding more adaptive parameters like taking into account the brain volume (for adaptation of the minimal lesion size factor), the noise ration in perfusion map (adaptive median filter size) or average perfusion in whole hemisphere (adaptive threshold for lesion acceptance).

The intelligence DSS systems based on cognitive analysis of medical data has large potential that can be even widened by applying image understanding technology [15] based on applying graph grammar for description of visualized symptoms.

Acknowledgments. This work has been supported by the Ministry of Science and Higher Education, Republic of Poland, under project number N N516 511939.

References

1. van Erkel, A.R., Pattynama, P.M.: Receiver operating characteristic (ROC) analysis: basic principles and applications in radiology. Eur. J. Radiol. 27(2), 88–94 (1998)
2. Hachaj, T., Ogiela, M.R.: Automatic Detection and Lesion Description in Cerebral Blood Flow and Cerebral Blood Volume Perfusion Maps. Journal of Signal Processing Systems 61(3), 317–328 (2010)
3. Hachaj, T.: The registration and atlas construction of noisy brain computer tomography images based on free form deformation technique, Bio-Algorithms and Med-Systems, Collegium Medicum - Jagiellonian University. Bio-Algorithms and Med-Systems 7 (2008)
4. Nowinski, W.L., et al.: Analysis of Ischemic Stroke MR Images by Means of Brain Atlases of Anatomy and Blood Supply Territories. Acad. Radiol. 13, 1025–1034 (2006)
5. Bardera, A., Boada, I., Feixas, M.: A Framework to Assist Acute Stroke Diagnosis, Vision, Modeling, and Visualization (VMV 2005), Erlangen (2005)

6. Latchaw, R.E., Yonas, H., Hunter, G.J.: Guidelines and recommendations for perfusion imaging in cerebral ischemia: a scientific statement for healthcare professionals by the writing group on perfusion imaging, from the Council on Cardiovascular Radiology of the American Heart Association. Stroke 34, 1084–1104 (2003)
7. Eastwood, J.D., et al.: CT perfusion scanning with deconvolution analysis: pilot study in patients with acute middle cerebral artery stroke. Radiology 222(1), 227–236 (2002)
8. Aksoy, F.G., Lev, M.H.: Dynamic Contrast-Enhanced Brain Perfusion Imaging: Technique and Clinical Applications. Semin Ultrasound CT MR 21(6), 462–477 (2000)
9. Hoeffner, E.G., et al.: Cerebral Perfusion CT: Technique and Clinical Applications. Radiology 231(3), 632–644 (2004)
10. Koenig, M., Klotz, E., Heuser, L.: Perfusion CT in acute stroke: characterization of cerebral ischemia using parameter images of cerebral blood flow and their therapeutic relevance. Clinical experiences. Electromedica 66, 61–67 (1998)
11. Sasaki, M., et al.: Procedure Guidelines for CT/MR Perfusion Imaging. Joint Committee for the Procedure Guidelines for CT/MR Perfusion Imaging (2006), http://mr-proj2.umin.jp/data/guidelineCtpMrp2006-e.pdf
12. Parraga, A., et al.: Non-rigid registration methods assessment of 3D CT images for head-neck radiotherapy. In: Proceedings of SPIE Medical Imaging (February 2007)
13. Modersitzki, J.: Numerical Methods for Image Registration. Oxford University Press, Oxford (2004)
14. Duvernoy, H.M., Bourgouin, P., Vannson, J.L.: The Human Brain: Surface. Three-dimensional Sectional Anatomy with MRI and Blood Supply. Springer, Heidelberg (1999)
15. Ogiela, M.R., Tadeusiewicz, R.: Image Understanding Methods in Biomedical Informatics and Digital Imaging. Journal of Biomedical Informatics 34(6), 377–386 (2001)

Development of a Biologically Inspired Real-Time Spatiotemporal Visual Attention System

Byung Geun Choi and Kyung Joo Cheoi*

Chungbuk National University, Dept. of Computer Science
12 Gaeshin-dong, Heungduk-gu, Cheongju, Chungbuk 361-763, Korea
kjcheoi@cbnu.ac.kr

Abstract. In this paper, we present a new spatiotemporal visual attention system. Typical feature integration model is expanded to incorporate motion in our suggested system, and is able to respond to motion stimulus by employing motion fields map as one of temporal features. Proposed system is based on bottom-up approach of human visual attention, but the main difference lies in its temporal feature extraction method, and integration method of multiple spatial and temporal features. Spatial features are integrated into spatial saliency map by weighted combination method. Temporal feature is extracted by SIFT and is analyzed and reorganized into temporal saliency map. Finally, dynamic fusion technique applied to make one spatiotemporal saliency map. To evaluate the performance of the system, we tested with various kinds of real video sequences. We also compared our system with several previous systems to validate the performance of the system.

Keywords: Visual attention, bottom-up, video, spatiotemporal, saliency.

1 Introduction

As there exists a large amount of information in complex scenes, real-time processing is very difficult to many typical computer vision systems[1][3][4][5][6][8]. In order to overcome this problem, a parallel connected processor was suggested, but it is not easy to implement. The one useful method is to reduce the amount and the complexity of information existed in images. Primates have a remarkable ability to interpret complex scenes in real time, despite the limited speed of the neuronal hardware available for such tasks. Intermediate and higher visual processes appear to select a subset of the available sensory information before further processing, most likely to reduce the complexity of scene analysis.

Human visual attention system is studied in two major approaches. One is the bottom-up approach, and the other is the top-down approach. Models of bottom-up attention and saliency for selective visual attention are based on the "a feature-integration theory of attention" proposed by Treisman and Gelade [2]. Visual input is

* Corresponding author.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 416–424, 2011.

first decomposed into a set of topographic feature maps. Different spatial locations then compete for saliency within each map, such that only locations, which locally stand out from their surround can persist. All feature maps feed, in a purely bottom-up manner, into a master "saliency map," which topographically codes for local conspicuity over the entire visual scene. The saliency map is an explicit two-dimensional map that encodes the saliency of objects in the visual environment. Competition among neurons in this map gives rise to a single saliency map. The top-down approach detects visual attention region using pre-knowledge of the target [9][10][11].

Many studies have been progressed based on bottom-up approach, but it has some problems. The system proposed in [3] detected region of visual attention by Winner-Take-All (WTA) method from saliency map. Thus if the same feature existed in the saliency map, it is not easy to detect another same feature because it is treated as a loser feature. The system proposed in [7] used integration method of promoting those maps in which a small number of meaningful high activity areas are present while suppressing the others, using local competition relations and statistical information of pixels in pre-computed maps. However, this method just selects one meaningful major feature map, while neglecting other many feature maps. Some systems are proposed that incorporate temporal information with previous typical spatial attention systems. The system proposed in [1] generated a temporal saliency map using Scale-Invariant-Feature-Transform(SIFT) and RANSAC algorithm. But this system has some problems. If the moving object is very big, the system assigns very high saliency value even if its movement is very little that cannot be perceive. Also, RANSAC algorithm discards some motions. The system proposed in [5] extracted motion feature by Full Search Block Matching (FSBM) algorithm. But it just extracts moved point not object.

In this paper, we present a new spatiotemporal visual attention system. Our system overcomes previous mentioned problems by newly proposed temporal feature extraction method, and integration method of multiple spatial and temporal features. Proposed detailed processing of our system is described in Section 2. In Section 3, experimental results are presented, followed by concluding statements in Section 4.

2 Proposed System

The general operation of the proposed visual attention system is describe+ed in Fig.1.

Fig.1 illustrates proposed model. Starting with current frame (t) RGB image, spatial feature maps for color, intensity, form and orientation are generated. And as a temporal feature map, motion map, is generated by two gray scale images of current frame (t) image and the previous frame (t-1) image using SIFT algorithm. Multiple spatial feature maps are integrated into one spatial saliency map by weighted combination method, and motion map is analyzed and reorganized into temporal saliency map. These two maps are then integrated into one single spatiotemporal saliency map by dynamic fusion technique.

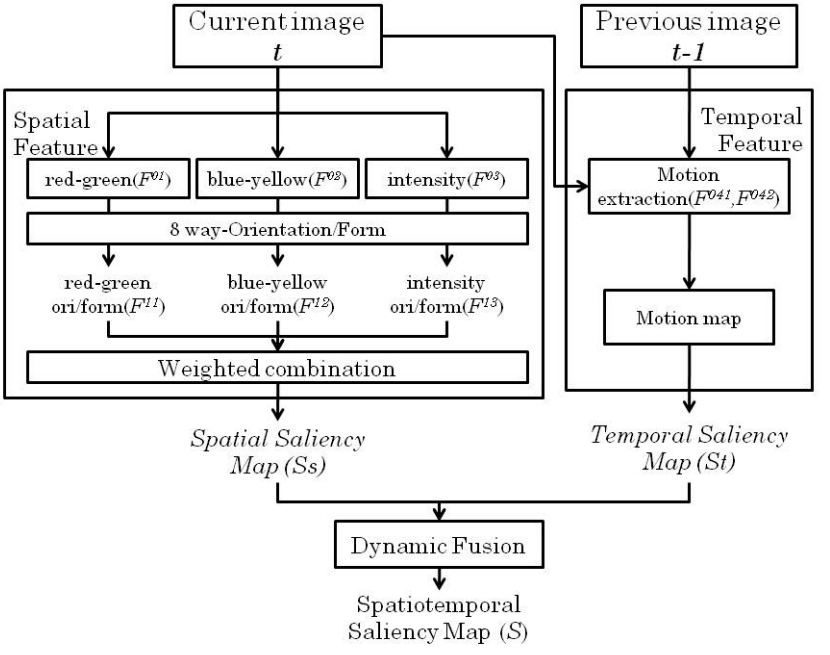


Fig. 1. Schematic diagram of the proposed system

2.1 Spatial Saliency Map

Current (t) RGB image, several spatial features known to influence human visual attention are generated in parallel : two for color contrast(F^{01},F^{02}) intensity contrast(F^{03}).

$F^{01}$and F^{02} are modeled with the two types of color opponency exhibited by the cells with homogeneous type of receptive fields in visual cortex, which respond very strong to color contrast. F^{01} is generated to account for 'red/green' color opponency and F^{02} for 'blue/yellow' color opponency. F^{03} is generated by ON and OFF intensity information of input image. These three features are extracted by similar features extraction method used in [4]. However, unlike the method used in [4], we did not discard minus values in feature maps because we assumed that all extracted features are important. Center-surround computations with 8 orientations ($\theta \in \{0,\pi/8, 2\pi/8,\cdots, 7\pi/8\}$) are performed on F^{01}, F^{02}, F^{03} in order to obtain orientation and form feature. This computation extracts conspicuous local pattern that is different from surround patterns. As oriented ON-center, OFF-surround operator, another bank of multiple-scale, Difference-Of-Oriented Gaussians (DOOrG) filters[4] are used. Through this computation we can have 24 feature maps($F^{0k\theta}$, for k=1,..,3, θ=1,..,8). In order to obtain a unique feature map for each direction(F^{1k}, for k=1,..,3), the maximum response is taken using content-based global nonlinear feature integration

method. Finally, three feature maps are integrated into one spatial saliency map by weighted combination method.

The weight of each feature map is decided by Equation (1).

$$w_n = \frac{\sum_{k=1}^{3} Diff(\mathrm{F^{1k}})}{Diff(\mathrm{F^{1n}})} \tag{1}$$

In Equation (1), n is nth number of the feature map and $Diff(F^{1n})$ means the activity of the feature map F^{1n}. The feature map that has higher activity gets more weight than other feature maps. The spatial saliency map S_s is generated by Equation (2).

$$S_s = \sum_{k=1}^{3} (\mathrm{w_k} * \mathrm{F^{1k}}) \tag{2}$$

2.2 Temporal Saliency Map

Temporal feature, motion, is first extracted by two gray scale images of current frame (t) image and the previous frame (t-1) image using Scale Invariant Feature Transform (SIFT) algorithm[2]. Most of the interesting points extracted by SIFT algorithm is the points in the edge of the objects. In general, interested objects have strong edge. Therefore we can find object-based motion information by matching each interesting points between current image frame (t) image and previous frame (t-1) image. We used the matching method used in [2], and matched interesting point set is generated. From the locations of two matched interesting points, motion vectors are calculated. The motion vectors are classified into 8 orientations. Finally, we can have the motion map.

The motion map is analyzed into temporal saliency map S_t. If one object is moved, all interesting points in that object have the same motion. Based on this information, the motion vectors that have the same direction and the same size are grouped, and grouped motion vector is drawn as a rectangle. The rectangle is decided as maximum x and y and minimum x and y location of interesting points of grouped motion vector. We assigned its saliency value as the motion vector size.

2.3 Spatiotemporal Saliency Map

The spatial saliency map and the temporal saliency map are integrated into one single spatiotemporal saliency map by dynamic fusion technique. Two saliency maps are integrated with the dynamically changing weight. The weight of the each saliency map is decided by Equation (3) and (4).

$$K_t = \frac{count(t-motion)}{count(t)}/3 + avg(avgMotion)/3 + \frac{count(motion-maxMotion)}{count(motion)}/3 \tag{3}$$

$$\mathrm{K_s} = 1 - \mathrm{K_t} \tag{4}$$

In Equation (3), t means the number of blocks in the temporal saliency map. *count(a)* means the number of elements in *a*. *motion* means the number of motion blocks in temporal saliency map. *maxMotion* means the number of motion blocks which has maximum motion value. *avg(a)* means the average value of *a*. *avgMotion* means

'*count(motion) – count(maxMotion)*'. In Equation (4), K_t means the weight of the temporal saliency map and K_s means the weight of the spatial saliency map.

Finally, spatiotemporal saliency map *S* is generated by Equation (5).

$$S = K_s * S_s + K_t * S_t \tag{5}$$

3 Experiments

3.1 Environment

In order to evaluate the performance of the system, we collected various kinds of 80 outdoor real video sequences recorded by camcorder. We first grouped test images into two classes according to the complexity of the spatial feature(the complexity of the spatial feature is high or not). These classified images are then group again into each three classes according to the complexity of the temporal feature(the size of the moving object, the direction of the moving object, the number of the moving object). We also compared our system with several previous systems presented in [3] and [4] to validate the performance of the system.

3.2 Result

Some results of our system is shown in Fig.2. Fig.2(a)~(b) is the two frames of test video sequences. The image contains red tollgate, white car, jade green roof, white sky and traffic safety objects. If we consider only spatial features, interested objects are the traffic safety object, the red tollgate and the jade green roof. Among them, the red tollgate is the most interesting object. However the image contains moving object, the white car which moves left to right side, the moving white car is the most attended object. The Spatial saliency map shown in Fig.2(c) indicates that the most interesting object with only spatial features is the red tollgate, and the next interesting object is the traffic safety object. The moving white car has low spatial saliency. However, the temporal saliency map shown in Fig.2(d) indicates that the moving car is the most interesting object if we use temporal features. The white rectangle in temporal saliency map is the grouped region that has the same motion direction and the same motion size. The spatiotempral saliency map shown in Fig.2(e) indicates that the most salient object is the moving white car, and the next salient object is the red tollgate. Our spatiotemporal saliency map receives the results of both spatial and temporal saliency maps, and does not discard any features. Detected attention regions are ordered by saliency value, and is shown in Fig.2(f). The region printed with lower number means the more attended region than the regions printed with higher number.

We compared our system with the system in [3] and [4]. Some results are shown in Fig. 3. In Fig. 3, the images located in left two columns are the input images of previous frame (t-1) and current frame (t).

Fig.3(a)'s complexity of the spatial feature is simple, and a small object(black car) is moves to one direction. The proposed system detected the moving black car. The system in [3] first detected the tollgate, and second, the yellow traffic safety structure, and third, the red temporal stricter and the green weeds. The system in [4] detected mainly the white sky. Fig.3(b)'s complexity of the spatial feature is simple, and a big

Fig. 2. Some results of the proposed system

object(white car) is moves to one direction. The proposed system detected moving white car. The system in [3] first detected the top region of the red tollgate, and second, the bottom region of the red tollgate, and third, the front of the white car, and the green trees. The system in [4] mainly detected the red tollgate.

Fig.3(c)'s complexity of the spatial feature is simple and some objects are moves to one direction. The proposed system first detected the fast moving silver car, and second, the slowly moving blue car. The system in [3] first detected the front of blue car, and second, the red temporal stricter, and third the red tollgate and the rear of the blue car. The system in [4] detected mainly the red tollgate. Fig.3(d)'s complexity of the spatial feature is simple and some objects are moves to various directions. The proposed system first detected the white car moves to left side, and second, the silver car moves to right side. The system in [3] first detected the white and the silver car, and second, the green roof, red temporal stricter and green temporal stricter. The system in [4] detected mainly the white car and the white sky. Fig.3(e)'s complexity of the spatial feature spatial feature is complicated, and a small object is moves to one direction. The proposed system detected moving object. The system in [3] first detected the green tree, and second, the green tree, the red moving object and the apartment. The system in [4] detected mainly the white sky. Fig.3(f)'s complexity of the spatial feature spatial feature is complicated, and a big object is moves to one direction. The proposed system detected the moving object. The system in [3] first detected the green tree, and second, the green tree, the black tire and the blue mark. The system in [4] detected mainly the white car. Fig.3(g)'s complexity of the spatial feature spatial feature is complicated and some objects are moves to one direction. The proposed system detected the center silver moving cars. The system in [3] first detected the green tree, and second, the green tree, the silver moving car and the green tree. The system in [4] detected mainly the sliver car and the white sky. Fig.3(h)'s complexity of the spatial feature spatial feature is complicated, and some objects are moves to various directions. The proposed system first detected the red moving object, and second, the silver car. The system in [3] first detected the red traffic sign, and second, the green tree, red moving car and silver moving car. The system in [4] detected mainly the white sky.

Fig. 3. Comparison of our system with the system in [3] and [4]

As seen from the above results, the proposed system has better performance than previous system in [3] or [4].

4 Conclusion

In this paper, we presented a new spatiotemporal visual attention system. Our system overcomes previous mentioned problems by newly proposed temporal feature extraction method, and integration method of multiple spatial and temporal features. As a temporal feature, object-based motions are extracted by SIFT algorithm, we can have object-based temporal saliency. All spatial feature maps were taken to integrate spatial saliency map by weighted combination method. From this, we have more spatial information than previous systems. Also, the spatial saliency map and the temporal saliency map are integrated into one single spatiotemporal saliency map by dynamic fusion technique. Two saliency maps are integrated with the dynamically changing weight.

The proposed system has been applied to various kinds of video sequences. Experimental results show that the proposed system can describe the salient regions with higher accuracy than the previous approaches do. The results of our system can be successfully applied to active vision systems. For example, the location of the most interesting region can be used to change the gaze of an image acquisition system. When attention regions are extracted by the attention system, the camera can zoom onto them, and can obtain a more detailed image of the corresponding regions. Moreover, the use of attention system is not restricted to active vision problems, but can be generalized to any system for which efficiency is especially important.

References

1. Zhai, Y., Shah, C.: Visual Attention Detection in Video Sequences Using Spatiotemporal Cues. In: Proceedings of the 14th Annual ACM International Conference on Multimedia, pp. 815–824. ACM, New York (2006)
2. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Key points. International Journal of Computer Vision 60(2), 91–110 (2004)
3. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research 40(10-12), 1489–1506 (2000)
4. Cheoi, K.: Visual attention system based on bottom-up theory of human visual attention. Doctorate Thesis of Yonsei University (2002)
5. Lee, J.: Selective Visual Attention System Based on Motion Information for Active Vision System. Master Thesis of Korea University (2008)
6. Itti, L., Koch, C.: Computational Modeling of visual attention. Nature Reviews Neuroscience 2(3), 194–203 (2001)
7. Treisman, A.M., Gelade, G.: A Feature-Integration Theory of Attention. Cognitive Psychology 12, 97–136 (1980)
8. Itti, L., Koch, C., Niebur, E.: A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. IEEE Transaction on Pattern Analysis and Machine Intelligence 20(11) (1998)

9. Yu, Y., Mann, G., Gosine, R.G.: An Object-based Visual Attention Model for Robots. In: IEEE International Conference on Robotics and Automation Pasadena, CA, USA, May 19-23 (2008)
10. Xiao, J., Cai, C., Ding, M., Zhou, C.: Application of a Novel Target Region Extraction Model Based on Object-accumulated Visual Attention Mechanism. In: Fourth International Conference on Natural Computation, pp. 116–120 (2008)
11. Begum, M., Karray, F., Mann, G.K.I., Gosine, R.G.: A Probabilistic Model of Overt Visual Attention for Cognitive Robots. IEEE Transactions on Systems Man and Cybernetics-Part B: Cybernetics, 1305–1318 (2010)

Knowledge Source Confidence Measure Applied to a Rule-Based Recognition System

Michal Wozniak

Department of Systems and Computer Networks, Wroclaw University of Technology,
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
Michal.Wozniak@pwr.wroc.pl

Abstract. Paper deals with the knowledge acquisition process for the design of the decision support system. Usually in this case the knowledge is given in the form of rules which are formulated by human experts or/and generated on the basis of datasets. Each of experts has different knowledge about the problem under consideration and rules formulated by them have different qualities. The qualities of data stored in the databases are different as well. It might cause differences in quality of generated rules. In the paper we formulate the proposition of a knowledge source confidence measure and we show some of its applications to the decision process e.g., we show how to use it for contradiction elimination in the set of rule. Additionally, we propose how it could be used during decision making.

Keywords: contradiction elimination, knowledge quality, statistical quality measure.

1 Introduction

Machine learning is an attractive approach to build decision support systems (DSS) [17]. For this type of software, the quality of the knowledge plays the key-role. During designing DSS we could gain rules from different sources (i.e., from different experts and on the basis of different databases) and we should take into consideration that their qualities are different. This problem was described for the induction learning [4,7,12] and for the concept description [2,3] partly. We could find several concepts related to the decision making on the basis of the learning set. One of them proposes how produce classifier on the basis of the very small learning set [20]. Those problems might be interpreted as the decision making on the unreliable (non-representative) sources.

The paper concerns on the problem how to assess the quality of rule for the probabilistic reasoning. We propose a kind of measure which could be applied to knowledge acquisition process for diverse forms of rule.

The content of the work is as follow: Section 2 introduces motivation of the work. Then proposition of statistical knowledge quality measure and the areas of its

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 425–434, 2011.

usefulness are presented. In this section contradiction elimination algorithm for the set of the first-order rules is described and interpretation of proposed measure for the typical statistical estimation process is shown also. Section 4 proposes the rule-based decision algorithm based on the Bayes decision rule using proposed quality measure. The last section concludes the paper.

2 Motivation of the Work

During acquisition process we could meet with situation that experts formulate the rules which contradict each other e.g.,

Expert 1 said:	Expert 2 said:	
If A then B	**If A then C**	**B∧C=false**

In this case (for logical interpretation of rule where *if A then B* means $A \Rightarrow B$) two rules given by expert 1 and expert 2 contradict each other. It means that for this case we can not make clear decision. Of course for other interpretation of rule, e.g. in probabilistic reasoning where the mentioned rules might look as follow

If A then B with probability β_1 **If A then C with probability β_2**

$$\beta_1 = P(B|A) \text{ and } \beta_2 = P(C|A). \tag{1}$$

The rules contradict each other only if we can not find any probability distribution for which (1) is true but we allow rules with the same condition and different conclusions.

Let's come back to the logical interpretation of the rules. In this situation we should propose the solution how remove the contradiction. Let's consider two cases:

CASE 1: Contradiction is detected in the rule set given by one expert.
Solution: Probably expert wants to cooperate to eliminate contradictions from his/her own set of rules. Expert modifies the conditions of some of rules or remove some of them.

CASE 2: Rules formulated by different experts contradict each other.
Solution 1: We can ask authors of rules to find the „wrong" rules or part of rules together. It is very expensive method and usually ends in failure.
Solution 2: We decide (arbitrary) which of the rules or their parts should be removed to eliminate contradiction (we are not allowed to modify the rule logic). Probably, we choose rules from the source with the smallest quality (e.g. given by less experienced experts). For computer implementation of the second solution we have to formulate the measure of knowledge source quality.

3 Proposition of Knowledge Confidence Measure

3.1 Definition

Let's notice that during acquisition process rules are obtained under the following assumptions:

1. learning set is noise free (or expert tell us always true),
2. target concept contained in the set of class number $\{1, ..., M\}$.

We consider decision under the first assumption given by the following formulae:

$$P(\text{If A then B}) = 1. \tag{2}$$

During the expert system designing process the rules are obtained from different sources which have different qualities. For the rules given by experts we cannot assume that expert tells us true or/and if the rule set is generated by the machine learning algorithms we cannot assume the learning set is noise free.

Therefore we postulate that we do not have to trust all knowledge we get or we believe it only with the γ factor, proposed as the quality (confidence) measure. It can be formulated as [21]

$$P(\text{If A then B}) = \gamma \le 1, \tag{3}$$

3.2 Contradictions Elimination Algorithm

As we have mentioned above the proposed method of the quality management can be applied to any form knowledge representation as well to the logical one (where "if-then" means logical implication), e.g. for the unordered set [15,17] of logical rules acquisition process we can assign the value of confidence measure to each of rule. It could be used in the case if the contradiction in the set of rules is detected.
Firstly we note the set of rule R consists of the M subsets

$$R = R_1 \cup R_2 \cup ... \cup R_M \tag{4}$$

where R_i denote subset of rules pointed at the i-th class (object can not belong to the two classes in the same time).

For this form of rule the two of them contradict each other if

$$\exists\, x \in X \wedge \exists\, k, l \in \{1, 2, ..., M\}, k \neq l \wedge \exists\, i, j \qquad x \in D_i^{(k)} \wedge x \in D_j^{(l)} \tag{5}$$

Where k, l denote the number of rule. The equation (3) means that we can find observation, which belongs to the decision area $D_i^{(k)}$ of the rule which $r_i^{(k)}$ pointed at class i and decision area $D_j^{(l)}$ of the rule $r_j^{(l)}$ which pointed at different class j. Let $\gamma_i^{(k)}$ and $\gamma_j^{(l)}$ denote the confidence measure of rule $r_i^{(k)}$ and $r_j^{(l)}$ respectively.

The idea of the contradictions elimination algorithm is presented in Fig. 1.

```
1. for i:= 1 to M: //for each class number
2. for k:=1 to |R_i| //for each rule in R_i
3. for j:= i to M: //for each class number bigger than i
4. for l:= 1 to |R_j| ://for each rule in R_j
5. if D_i^(k) ∩ D_j^(l) ≠ ∅ //if r_i^(k) and r_j^(l) contradict each other
6.    then if γ_i^(k) ≥ γ_j^(l)    //if confidence of r_i^(k) is higher
7.                          // than confidence of r_j^(l)
8.     then D_j^(l) := D_j^(l) \ (D_j^(l) ∩ D_i^(k)) //remove the common part
9.                              //from decision area of rule r_j^(l)
10.    else D_i^(k) := D_i^(k) \ (D_j^(l) ∩ D_i^(k))//remove the common part
11.                             //from decision area of rule r_j^(l)
12.   fi
13. fi
14. endfor; endfor; endfor; endfor
```

Fig. 1. Pseudocode of contradiction elimination algorithm

3.3 Using Confidence Measure in Rule-Based System and Minimal Distance Classifier

Linking empirical material is different area, in which the proposed measure might be applied. Firstly we have to answer the question if examples coming from different sources (about the different qualities) should be treated equally. After all they could come from different databases, and quality of stored data (e.g., number of committed typing errors) is different. So let's present how to use the above mentioned quality measure for learning algorithms based on "sequential covering", like e.g., AQ[15, 16] or CN2[5]. The mentioned algorithms protect themselves against overfitting [11] by using heuristic functions which assessing generated rules e.g., in one of the first version of the AQ algorithm, this function takes the cardinalities of three sets into consideration:

1. not covered elements by any early generated rule but covered by the new generated ones with the same label as the assessed rule,
2. not covered elements by any early generated rule and by the new rule with the different label as the new generated rule,
3. covered by the new generated rule with the same label as the rule.

We suggest to take into consideration sum of the quality measure values instead of cardinalities of mentioned subsets.

This quality measure we might use also if we are deciding on using lazy classifiers [1] e.g. *k*-NN algorithm [10, 13]. We suggest to multiple distance between recognized object and learning element by quality measure value of sources which learning

element belongs to. The similar modification was suggested in [6] but authors multiple distance according to the expert assessment.

4 Probabilistic Decision Support Method

Among the different concepts and methods of using "uncertain" information in pattern recognition, an attractive from the theoretical point of view and efficient approach is through the Bayes decision theory. This approach consists of assumption [8] that the feature vector $x = (x^{(1)}, x^{(2)}, ..., x^{(d)})$ (describing the object being under recognition) and number of class $j \in \{1,2...,M\}$ (the object belonged to) are the realization of the pair of the random variables *X, J* . For example in medical diagnosis *X* describes the result of patient examinations and *J* denotes the patient state. Random variable *J* is described by the prior probability p_j , where

$$p_j = P(J = j) \tag{6}$$

X has probability density function

$$f(X = x \mid J = j) = f_j(x) \tag{7}$$

for each *j* which is named conditional density function. These parameters can be used to enumerating *posterior* probability according to Bayes formulae:

$$p(j|x) = p_j f_j(x) \Big/ \sum_{k=1}^{M} p_j f_j(x). \tag{8}$$

The formalisation of the recognition in the case under consideration implies the setting of an optimal Bayes decision algorithm $\Psi(x)$, which minimizes probability of misclassification for 0-1 loss function [9]:

$$\Psi(x) = i \text{ if } p(i|x) = \max_{k \in \{1, ..., M\}} p(k|x). \tag{9}$$

In the real situation *prior* probabilities and the conditional density functions are usually unknown. Furthermore we often have no reason to decide that the *prior* probability is different for each of the decisions. Instead of them we can used the rules and/or the learning set for the constructing decision algorithms [14].

4.1 Rule-Based Decision Algorithm

Rules as the form of learning information is the most popular model for the logical decision support systems. For systems we consider the rules given by experts have rather the statistical interpretation than logical one. The form of rule for the probabilistic decision support system [11] is usually as follows

`if A then B with the probability` β,

where β is interpreted as the estimator of the *posterior* probability, given by the following formulae:

$$\beta = P\left(B|A\right) \tag{10}$$

More precisely, in the case of the human knowledge acquisition process, experts are not disposed to formulate the exact value of the β, but they rather prefer to give the interval for its value

$$\underline{\beta} \le \beta \le \overline{\beta} \tag{11}$$

We propose the following form of rule $r_i^{(k)}$ [22]:

IF $x \in D_i^{(k)}$ **THEN** state of object is i **WITH** the posterior probability $\beta_i^{(k)}$ greater than $\underline{\beta}_i^{(k)}$ and less than $\overline{\beta}_i^{(k)}$

where

$$\beta_j^{(k)} = \int\limits_{D_i^{(k)}} p\left(i|x\right)dx \Big/ \int\limits_{D_i^{(k)}} dx \,. \tag{12}$$

For that form of knowledge we can formulate the decision algorithm $\Psi_R\left(x\right)$

$$\Psi_R\left(x\right) = i \text{ if } \hat{p}\left(i|x\right) = \max_k \hat{p}_k\left(i|x\right), \tag{13}$$

where $\hat{p}\left(i|x\right)$ is the *posterior* probability estimator obtained from the rule set.

The knowledge about probabilities given by expert estimates the average *posterior* probability for the whole decision area. As we see for decision making we are interested in the exact value of the *posterior* probability for given observation.

Lets note the rule estimator is more precise if:

1. rule decision region is smaller,
2. differences between upper and lower bound of the probability is smaller.

For the logical knowledge representation the rule with the small decision area could overfit the training data [11] (especially if this set is small). For our proposition we respect this danger for the rule set obtained from learning set. For the estimation of the *posterior* probability on the basis of rules we assume the constant value of for the rule decision area. Let's define "more specific" relation.

Definition

Rule $r_i^{(k)}$ is "more specific" than rule $r_i^{(l)}$ if

$$\left(\overline{\beta}_i^{(k)} - \underline{\beta}_i^{(k)}\right) \int\limits_{D_i^{(k)}} dx < \left(\overline{\beta}_i^{(l)} - \underline{\beta}_i^{(l)}\right) \int\limits_{D_i^{(l)}} dx \,. \tag{14}$$

Let's define following set of constituents

$$\hat{X} = \left\{\hat{X}_1, \hat{X}_2, ..., \hat{X}_L\right\}, \tag{15}$$

where

$$\forall l \in \{1, ..., L\},\ \ i \in \boldsymbol{M},\ k \in \{1, ..., N_i\}\ \ \left\{\hat{X}_l \cap D_i^{(k)} = \varnothing \text{ lub } \hat{X}_l \subseteq D_i^{(k)}\right\}. \tag{16}$$

Additionally we define the following function:

$$w(l,i,k) = \begin{cases} 1 & \text{if} \quad \hat{X}_l \cap D_i^{(k)} \neq \varnothing \\ 0 & \text{if} \quad \hat{X}_l \cap D_i^{(k)} = \varnothing \end{cases}. \tag{17}$$

Hence definitions (13), (14), (15) and proposed function (16) the proposition of the *posterior* probability estimator $p^{(R)}\left(i_1 | x_1\right)$ is presented in Fig. 2.

```
input: R - set of rules
       x - value of feature vector x
1.  sum := 0
2.  find  X̂_l  where  x ∈ X̂_l .
3.  for each  i ∈ M
4.choose „the most specific" rule  r_i^(k)  from rules for
which  w(l, i, k) = 1
5.  p̂_i(x) := 0.5 * (β̄_i^(m) + β_i^(m))
6.  endfor
7.  if  Σ_{i=1}^{M} p̂_i(x) < 1
8.       then for each class i if  p̂_i(x) = 0
9.            p̂_i(x) := (1/n_0) * (1 − Σ_{k=1}^{M} p̂_k(x)),  //where  n_0  is the number
                        of classes fulfilled this condition
10.          endfor
11. fi
12. for each class  i ∈ M
13.       p_i^(RS)(x) := p̂_i(x) / Σ_{i=1}^{M} p̂_i(x)
14. endfor
```

Fig. 2. Pseudocode of procedure the *posterior* probability estimation on the base of rule set

4.2 Idea of Combining Algorithm

For the considered case, i.e. when we have s different rules bases or learning sets we propose the following decision algorithm:

$$\psi_n^{(C)}(\overline{x}_n,\overline{u}_{n-1}) = i_n \text{ if } \sum_{l=1}^{s} \alpha_l \times p^{(l)}(i_n|\overline{x}_n,\overline{u}_{n-1}) = \max_{k\in \mathrm{M}} \sum_{l=1}^{s} \alpha_l \times p^{(l)}(k|\overline{x}_n,\overline{u}_{n-1}). \quad (18)$$

$p^{(l)}\left(k|\overline{x}_n,\overline{u}_{n-1}\right)$ denotes estimator of the *posterior* probability obtained on base of the l-th learning set or rule base and α_l is the weight of the l-th learning set or rule base which might be interpreted as its quality measure. α_l could be calculated according the procedure presented in Fig.3.

```
1. Let LS be the set of testing examples consists of t
elements.
2. We have s learning sets or rule bases and we
constructed s classifiers
3. Let α_1 = α_2 = ... = α_s = 1.
4. for v:=1 to t do:
5. for l:=1 t do:
6.   if l-th classifier recognized the v-th testing
element correctly then
7.    α_l = α_l + 1
8.   fi
9. endfor
10. endfor
11. Let SUM_s = Σ_{l=1}^{s} α_l
12. for l:=1 to s do:
13.          α_l = α_l / SUM_s
14. endfor.
```

Fig. 3. Pseudocode of quality (confidence) measure values estimation

4.3 Confidence Measure for the Statistical Estimation

The central problem of our proposition is how to calculate the quality measure. For human experts the values for their rules is fixed arbitrary according to the quality of creator. The presented problem we can also find in the typical statistical estimation of unknown parameter β, where we assume the significant level. The significant level can be interpreted as the quality (confidence) measure. While producing the rule set, we have to define somehow the decision areas and then we could use the following statistical model [19]:

- the learning set is selected randomly from a population and there exists two class of points: marked (point at the class $i \in \{1, ..., M\}$) and unmarked (point at the class l, where $l \in \{1, ..., M\}$ and $l \neq i$),
- the expected value for the population is p,
- the best estimator of p is $\hat{p} = m/n$, where n means the sample size and m - the number of the marked elements.

For the fixed significance level α (which we could interpreted as confidence measure) we get

$$P\left(\underline{\beta} < p < \overline{\beta}\right) = 1 - \alpha \,, \tag{19}$$

where

$$\underline{\beta}_i = \frac{n_i}{n_i + (n - n_i + 1)F(\alpha/2,\, 2(n - n_i + 1),\, 2n_i)} \,, \tag{20}$$

$$\overline{\beta}_i = \frac{(n_i + 1)F(\alpha/2,\, 2(n_i + 1),\, 2(n - n_i))}{n - n_i + (n_i + 1)F(\alpha/2,\, 2(n_i + 1),\, 2(n - n_i))} \,. \tag{21}$$

The $F(\alpha, n, m)$ is the value of α quantile of F-Snedecor distribution with (n, m) degrees of freedom. As we mentioned above significant level α could be interpreted as a quality measure of the rule produced using presented statistical model.

5 Conclusions

The paper focused on the proposition of the quality measure of knowledge used by the decision support system. On the basis on aforementioned measure we proposed the idea of the contradiction elimination algorithm for the logical rule base and we discussed its usefulness for *lazy classifiers* and the probabilistic decision support systems..

We hope that the proposed idea of confidence management could be applied to other problems which could be met during the knowledge acquisition process from different sources. Of course presented concept need the further analytical and simulation researches.

Acknowledgments. This research is supported in part by The Polish Ministry of Science and Higher Education under the grant which is realizing in years 2010-13.

References

1. Aha, D.W. (ed.): Lazy Learning. Kluwer Academic Pub., Dordrecht (1997)
2. An, A., Cercone, N.: An Empirical Study on Rule Quality Measures. In: Zhong, N., Skowron, A., Ohsuga, S. (eds.) RSFDGrC 1999. LNCS (LNAI), vol. 1711, pp. 482–491. Springer, Heidelberg (1999)

3. Bergadano, F., et al.: Measuring of Quality Concept Descriptions. In: Proc. of the 3rd European Working Session on Learning, Aberdeen, Scotland, pp. 1–14 (1988)
4. Bruha, I., Kockova, S.: Quality of decision rules: Empirical and statistical approaches. Informatica 17(3) (1993)
5. Clark, P., Niblett, T.: The CN2 Induction Algorithm. Machine Learning 3(4), 261–283 (1989)
6. Cost, S., Salzberg, S.: A Weighted Nearest Neighbour Algorithm for Learning with Symbolic Features. Machine Learning (10), 57–78 (1993)
7. Dean, P., Famili, A.: Comparative Performance of Rule Quality Measures in an Inductive Systems. Applied Intelligence (7), 113–124 (1997)
8. Devijver, P.A., Kittler, J.: Pattern Recognition: A Statistical Approach. Prentice Hall, London (1982)
9. Duda, R.O., et al.: Pattern Classification. John Wiley and Sons, New York (2001)
10. Fukunaga, F., Narendra, P.M.: A branch and bound algorithm for computing k-nearest neighbors. IEEE Transactions on Computers 24, 750–753 (1975)
11. Giakoumakis, E., Papakonstantiou, G., Skordalakis, E.: Rule-based systems and pattern recognition. Pattern Recognition Letters (5) (1987)
12. Gur-Ali, O., Wallance, W.A.: Induction of rules subject to a quality constraint: probabilistic inductive learning. IEEE Trans. on Knowledge and Data Engineering 5(3) (1993)
13. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning; Data Mining, Inference and Prediction. Springer, New York (2001)
14. Kurzynski, M., Wozniak, M.: Rule-based algorithms with learning for sequential recognition problem. In: Proc.of the 3rd Int.Conf. on Information Fusion "Fusion 2000", Paris, pp. TuC5-24–TuC5-29 (2000)
15. Michalski, R.S.: On the quasi-minimal solution of the general covering problem. In: Proc of the 1st Int. Symposium on Information Processing, Bled, Jugosławia, pp. 125–128 (1969)
16. Michalski, R.S., et al.: The multi-propose incremental learning systems AQ15 and its testing application to three medical domains. In: Proc.of the 5th Nat. Conf. on AI, Philadelphia, pp. 1041–1045. Morgan-Kaufmann, San Francisco (1986)
17. Mitchell, T.: Machine Learning. McGraw Hill, New York (1997)
18. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Pub. Inc., San Francisco (1991)
19. Sachs, L.: Applied Statistic. A Handbook of Techniques. Springer, New York (1984)
20. Schapire, R.E.: The boosting approach to machine learning: An overview. In: Proc. of MSRI Workshop on Nonlinear Estimation and Classification, Berkeley, CA (2001)
21. Wozniak, M.: Concept of the Knowledge Quality Management for Rule-Based Decision System. In: Klopotek, M.A., et al. (eds.) Intelligent Information Processing and Web Mining, pp. 575–579. Springer, Heidelberg (2003)
22. Wozniak, M.: Case and Rule-based Algorithms for the Contextual Pattern Recognition Problem. In: Kumar, V., Gavrilova, M.L., Tan, C.J.K., L'Ecuyer, P. (eds.) ICCSA 2003. LNCS, vol. 2667, pp. 89–98. Springer, Heidelberg (2003)

A New Frontier in Novelty Detection: Pattern Recognition of Stochastically Episodic Events

Colin Bellinger and B. John Oommen*

School of Computer Science, Carleton University, Ottawa, Canada
{cbelling,oommen}@scs.carleton.ca

Abstract. A particularly challenging class of PR problems in which the, generally required, representative set of data drawn from the second class is unavailable, has recently received much consideration under the guise of *One-Class* (OC) classification. In this paper, we extend the frontiers of OC classification by the introduction of a new field of problems open for analysis. In particular, we note that this new realm deviates from the standard set of OC problems based on the following characteristics: The data contains a *temporal* nature, the instances of the classes are "interwoven", and the labelling procedure is not merely impractical - it is almost, by definition, impossible, which results in a poorly defined training set. As a first attempt to tackle these problems, we present two specialized classification strategies denoted by Scenarios $S1$ and $S2$ respectively. In Scenarios $S1$, the data is such that standard binary and one-class classifiers can be applied. Alternatively, in Scenarios $S2$, the labelling challenge prevents the application of binary classifiers, and instead, dictates a novel application of OC classifiers. The validity of these scenarios has been demonstrated for the exemplary domain involving the Comprehensive Nuclear Test-Ban-Treaty (CTBT), for which our research endeavour has also developed a simulation model. As far as we know, our research in this field is of a pioneering sort, and the results presented here are novel.

Keywords: Pattern Recognition, Novelty Detection, Noisy Data, Stochastically Episodic Events.

1 Introduction

A common assumption within supervised learning is that the distributions of the target classes can be learned, either parametrically or non-parametrically. Moreover, it is assumed that a representative set of data from these classes is available for the training of supervised learning algorithms; indeed, the latter implies the former.

Beyond this, there exists a special form of Pattern Recognition (PR), which is regularly denoted *One-Class* (OC) classification [4,5,6,7,10,11]. This

* *Chancellor's Professor; Fellow: IEEE and Fellow: IAPR.* The Author also holds an Adjunct Professorship with the Dept. of ICT, University of Agder, Norway.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 435–444, 2011.

"exceptional" category of binary classification is noteworthy in lieu of the fact that drawing a representative set of data to compose the second class (ω_2), is abnormally arduous, if not altogether impossible.

PR tasks of this nature have previously been constituted as involving outlier (or novelty) detection as the vast majority of the data takes a well-defined form that can be learned, and that samples from the ω_2 class will appear anomalously – outside the learned distribution. Although such problems represent a significant challenge, the results reported in the literature demonstrate that satisfactory results can often be obtained (see [4,5,6,7,10,11], for example).

In the subsequent section, Section 2, we introduce an advanced category of OC learning. Section 3, proceeds to draw a conceptual distinction between the target domain and those to which OC classifiers have traditionally been applied. The set of OC learners applied in this research are considered in Section 4. Section 5 describes an experiment based on the exemplary task of verifying the Comprehensive Nuclear Test-Ban-Treaty (CTBT). The results of the experiments, and a subsequent discussion, are contained in Section 6 and Section 7 respectively. Finally, Section 8 consists of our concluding remarks.

2 SE Event Recognition

To expand the horizon of the field, we observe that there exists a further, and yet more challenging subset of the OC classification domain of problems, which remains unexplored. We have denoted this class of problem as Stochastically Episodic (SE) event[1] recognition.

The problem of SE event recognition can be viewed in a manner that distinguishes it from the larger set of OC classification tasks. In particular, this category of problems has a set of characteristics that collectively distinguish it from its more general counterparts. The characteristics of this category can be best summarized as follows:

- The data presents itself as a time sequence;
- The minority class is challenging to identify, thus, adding noise to the OC training set.
- The state-of-nature is dominated by a single class;
- The minority class occurs both rarely and randomly within the data sequence.

Typically, in OC classification, the accessible class, and in particular, the data on which the OC classifier is trained, is considered to be well-defined. Thus, it is presumed that this data will enable the classifier to generalize an adequate function to discriminate between the two conceptual classes. This, for example,

[1] Events of this nature are denoted stochastic because their appearances in the time-series are the results of both deterministic and non-deterministic processes. The non-deterministic triggering event could, for example, be the occurrence of an earthquake, while the transmission of the resulting p- and s-waves, which are recorded in the time-serise, are deterministic.

was demonstrated in [5], where a representative set of the target computer user's typing patterns, which are both easily accessible and verifiable, were utilized in the training processes.

The classification of SE events is considerably more difficult because deriving a strong estimate of the target class's distribution is unfeasible due to the prospect of invalid instances (specifically members of the ω_2 class erroneously labelled ω_1) in the training set.

Under these circumstances, we envision two possible techniques for discriminating between the target class and the SE events of interest. The first scenario, denoted S1, involves application of standard clustering/PR algorithms to label both the classes appropriately. Alternatively, there are no instances of the ω_2 class available in S2, and the ω_1 class is poorly defined. Thus, novel applications of traditional OC classifier are applied.

3 Characteristics of the Domain of Problems

To accentuate the difference between the problems that have been studied and the type of problems investigated in this research, we refer the reader to Table 1. This table displays an assessment of six classification problems that have previously appeared in the literature on OC classification. In addition, we include the CTBT verification problem, which we present as a model SE event recognition problem. The first column indicates whether the problem has traditionally been viewed as possessing an important *temporal* aspect. The three entries with an asterisk require special consideration. In particular, we note that while traditionally these domains have not been studied with a temporal orientation, they do, indeed, contain a temporal aspect. The subsequent column signals whether the manual labelling of data drawn from the application domain is a significant challenge. This is, for example, considered to be a very difficult task within the field of computer intrusion detection, where attacks are well disguised in order to avoid detection.

Table 1. A comparison of well-known OC classification problems. The explanation about the entries is found in the text.

Dataset	Temporal	ID Challenge	Imbalance *Type I*	Imbalance *Type II*	Interwoven
Mammogram	No	Low	Yes	Medium	No
Continuous typist recognition	No	Low	Yes	Medium	No
Password hardening	No	Low	Yes	Medium	No
Mechanical fault detection	No*	Low	Yes	Medium	No
Intrusion detection	No*	High	Yes	High	No
Oil spill	No*	High	Yes	Medium	No*
CTBT verification	Yes	High	Yes	High	Yes

The following two columns quantify the presence of class imbalance. In the first of these, we apply a standard assessment of class imbalance, one which relies on the determination of the *a priori* class probabilities. Our subsequent judgement departs slightly from the standard view, and considers class imbalance that arises from the difficulty of acquiring measurements (due to cost, privacy, *etc.*). The final column specifies if the minority class occurs rarely, and randomly (in *time* and magnitude), and if it occurs within a time sequence dominated by the majority class.

To summarize, in this section we have both demonstrated the novelty of this newly introduced sub-category of PR problems, and positioned the CTBT verification task within it. We additionally note that the fault detection, intrusion detection, and oil spill problems could be reformulated to meet the requirements of our proposed category. This, indeed, suggests a new angle from which these problems can be approached.

4 Classification

In all brevity, we mention that the binary classifiers used in this study were the Multi-layer Perceptron (MLP), the Support Vector Machine (SVM), the Nearest Neighbour (NN), the Naïve Bayes (NB) and the Decision Tree (J48), all of which are fairly well known, and so their descriptions are omitted here.

Alternatively, this work employed the following OC classification techniques: *a)* autoassociator (AA) [6], the Combined Probability and Density Estimator (PDEN) [5], one-class Nearest Neighbour (ocNN) algorithm [3], and the the scaled ocNN (socNN) [2].

Each of the applied classifiers has been implemented in the Weka machine learning software suite.

4.1 Classification Scenarios

Two possible SE event recognition problem exist. The S1 scenario assumes that through PR, or Clustering, means, all the instances of the minority class can be separated from the majority class for training purposes. Furthermore, the ω_2 class is large enough to explore binary classification.

In S2, however, the primary class will not be well-defined, as it is likely to contain erroneously labelled instances of the outlier class. This is a result of the impracticality of manually identifying and labelling them. In addition, the hidden minority class is extremely small.

For S2, we propose the application of OC classifiers in an unsupervised manner. In particular, they are trained on datasets in which the vast majority of instances have correctly been extracted from the background class. However, the impracticality of identifying the rare SE events implies the probable presence of some erroneous training instances.

We submit that by utilizing estimates of the state-of-nature, the problems associated with the noisy training set can be overcome through the appropriate parametrization of an internal *rejection rate* parameter.

The general performance of the classifiers are examined across all of the simulated detonation ranges. In addition, the performance as a function of distance is examined.

5 Experimental Setup

In this section, we present a series of experiments based on the verification of the CTBT. These experiments are designed to both illustrate the domain of SE events, and to exhibit a first attempt at SE events recognition.

5.1 Application Domain

The CTBT aims to prevent nuclear proliferation through the banning of all nuclear detonations in the environment. As a result, a number of verification strategies are currently under study, aimed at ensuring the integrity of the CTBT. The primary verification technique being explored relies on the quantity of radioxenon measured continuously at individual receptor sites, distributed throughout the globe. Radionuclide monitoring, in general, has been identified as the sole technique capable of unambiguously discriminating low yield nuclear detonations from the background emissions. More specifically, verification of the treaty based on the four radioxenon isotopes, ^{131}Xe, ^{133}Xe, ^{133m}Xe and ^{135}Xe, has been promoted due to the relatively low background levels, their ideal rates of decay and inert properties [8,9].

In general, the measured radioxenon levels are expected to have resulted from the industrial activities, such as nuclear power generation and medical isotope production. However, they are also the byproducts of low yield clandestine nuclear weapons tests, which are the subject of the CTBT.

5.2 Procuring Data: Aspects of Simulation

As a means of acquiring experimental datasets for this research, we utilized the simulation framework presented by Bellinger and Oommen in [1]. Their simulation framework models SE events, such as earthquakes, nuclear explosions, etc., as they propagate through the background noise, in this case representing radioxenon emitted from industry into the earth's atmosphere.

While it is generally beneficial to develop and study classifiers on "real" data, this is, indeed, impossible within the CTBT verification problem due to the absence of measured detonations, and the limited availability of background instances.

6 Results

In this section, we present the results that were obtained according to the four assessment criteria motivated in the previous sections, on the first classification scenario. We commence our exploration of PR performance by examining the AUC scores produced by each classifier over the 23 detonation ranges.

6.1 Results: Scenario 1

We first present the experimental results obtained for Scenario S1.

General Performance: With regard to the results, we include a general overview of the performance levels of each of the considered classifiers on the simulated CTBT domain. More specifically, we present an assessment of the five binary classifiers and the four OC classifiers, in terms of their AUC scores averaged over the 230 datasets that spanned the 23 detonation ranges.

In light of the fact that the SE events, which are to be identified, will, in practice, occur at random and unpredictable distances, these results yield a particularly insightful overview of the general performance levels.

The SVM classifier is, surprisingly, by far the worst performing classifier on this data, and in spite of its bias, it is, on average, worse than the OC classifiers, AA and socNN. This is demonstrates in Table 2, which contrasts the mean AUC scores of AA and socNN as 0.656 and 0.603, respectively, with the mean value for the SVM classifier of 0.528. Moreover, all four OC classifiers appear superior to the SVM when considered in terms of their maximum AUC scores.

The binary classifier, the MLP, stands out as the superior classifier, with J48, NN, and NB contending for the intermediate positions. The results posted in Table 2 confirm that the MLP is the strongest of the classifiers considered here. Furthermore, it indicates that the J48 and NB are very similar, and that the NN is the fourth ranking binary classifier according to the mean and maximum scores. However, the NN is second when ranked according to the minimum AUC scores.

Table 2. This table displays the general classification results, in terms of AUC

	Mean	Max	Min	STDV
NB	0.772	0.939	0.504	0.074
MLP	0.869	0.976	0.674	0.067
NN	0.741	0.913	0.584	0.071
J48	0.774	0.98	0.500	0.148
SVM	0.528	0.813	0.500	0.065
ocNN	0.540	0.875	0.496	0.087
PDEN	0.487	0.943	0.182	0.156
socNN	0.603	0.842	0.405	0.094
AA	0.656	0.970	0.251	0.140

Notably, of the set of OC classifiers, PDEN produced the most variable range of the AUC scores. It is our suspicion that this variability resulted from the PDEN's generation of an artificial second class in its training process. However, further exploration of this matter is required. In general, the AA classifier is identified as the strongest OC classifier, both with respect to its mean and median values. While the socNN classifier achieved the second highest mean, it is more stable than the AA, with a lower standard deviation.

Performance as a Function of Distance: These results are particularly interesting, as they provide greater insight into performance trends, and suggest a performance scale for successively sparser receptor networks.

The performance plots depicted in Figure 1 reflect the ensemble mean of each classifier's performance at the 23 detonation ranges.

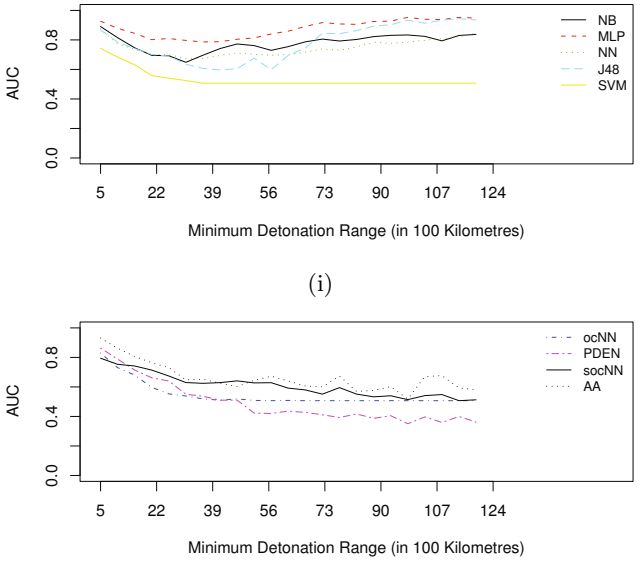


Fig. 1. In this figure, plot (i) displays the performance of the five binary classifiers, in terms of their AUC scores, as a function of distance. Similarly, plot (ii) displays the performances of the four OC classifiers as a function of distance, according to their AUC scores.

Within Figure 1 (see Plot (i)), the MLP classifier is identifiably the superior classifier to the remaining four binary learners in terms of the AUC, across the range of detonation distances. In addition, it is not subject to the abrupt fluctuations that J48, and to a lesser extent NB, incur.

Plot (ii) in Figure 1 presents the performance of the OC learners. All of the OC classifiers follow a similar downward trend, which occurred between 0.8 and 0.9, towards, or beyond in the case of the PDEN, an AUC of 0.5. The AA and the socNN degrade in a slower, and in a more linear fashion than PDEN and ocNN.

6.2 Results: Scenario 2

In this section, we explore the very intriguing classification scenario S2. More specifically, we present an assessment of the four OC classifiers, in terms of their AUC scores on the 230 datasets that covered the 23 detonation ranges.

General Performance: We first present a general overview of the performance of the set of OC classifiers on the simulated CTBT domain.

In light of the fact that the SE event will, in practice, occur at random and unpredictable distances, these results are particularly insightful.

Table 3 contains a compilation of the mean, maximum, minimum and standard deviation of the each classifier's overall results.

Table 3. This table displays the general classification results, in terms of AUC

	Mean	Max	Min	STDV
ocNN	0.505	1	0.496	0.042
PDEN	0.507	1	0.075	0.185
socNN	0.587	1	0.292	0.171
AA	0.621	1	0.024	0.225

Our assessments of Table 3 reveal that, similar to our findings on the S1 scenario, the AA classifier is superior, in terms of its mean, and median scores, to the other OC classifiers. Indeed, on this, which is a more challenging task, its mean and median values are only slightly lower than in the previous task. However, within this second scenario, it has the lowest minimum AUC scores. It is also extremely unstable, with results ranging from perfect to near zero.

The classifier, socNN, ranks second after the AA according to its mean, and was considerably more stable, while the ocNN and PDEN classifiers produced values that were near or below 0.5.

Performance as a Function of Distance: As in the case of Scenario S1, we have also studied the performance of the classifiers as a function of distance, where the latter is assessed according to the AUC.

The AA and socNN are, once again, roughly identifiable as the best of the four classifiers in Figure 2. However, all of the classifiers, with the exception of ocNN, which rapidly converges to 0.5, suffer from significant and essentially random fluctuations. These fluctuations in performance suggest that the classifiers' results were as dependent on the nature of the SE events in the 230 datasets, as on the distance at which the events originally occurred.

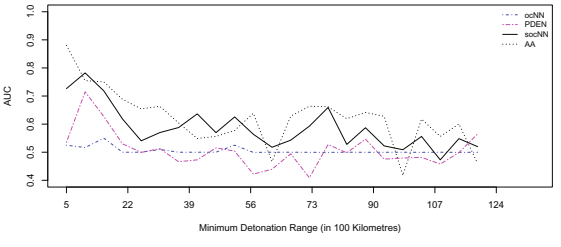


Fig. 2. This figure displays the performance of the four OC classifiers as a function of distance, according to their AUC scores

7 Discussion

The relatively low mean AUC scores produced by the OC classifiers, and variability in their results on the CTBT feature-space, clearly illustrate the many challenges in the application of OC learners. However, we suspect that the results are not so imbalanced, as Hempstalk *et al.*, in [5], noted that such comparisons are generally biased towards binary learners.

The initial performances of the OC classifiers suggests that they are very capable of associating anomalously high levels of radioxenon with the SE event class. However, the binary learners are not only well adapted to classifying anomalously highly levels as members of the SE event class, they are also capable of classifying anomalously low levels, which commonly result from detonations that occurred well beyond the radial distance to the background source and advected from a different direction.

The instability in performance that is depicted with respect to distance that appears in S2, results both from the erroneous instances in the training sets, and the variability in the classification challenges presented by the few members of the SE event class in the test sets. Indeed, the generation of random SE events over a domain as vast as the simulated CTBT domain, will inevitably produce both very easy, and nearly impossible classification tasks. Thus, when randomly including only a minute number of these events in the test sets, it is probable that performance on the SE event class will fluctuate significantly. This is, of course, why a large number of receptors are required in the global receptor network.

However, while the ensemble mean performance fluctuates considerably over the successive experiments, when considered in terms of the overall means, the performance of the OC classifiers on the S2 task is only slightly lower than on the S1 task. This is, indeed, a promising result.

8 Conclusion

In this research, we have extended the frontiers of novelty detection through the introduction of a new field of problems open for analysis. In particular, we noted that this new realm deviates from the standard set of OC problems based on the presence of three characteristics, which ultimately amplify the classification challenge. They involve the *temporal* nature of the appearance of the data, the fact that the data from the classes are "interwoven", and that a labelling procedure is almost, by definition, impossible.

As a first attempt to tackle these problems, we presented two specialized classification strategies as demonstrated within the exemplary scenario intended for the verification of the Comprehensive Nuclear Test-Ban-Treaty (CTBT). More specifically, we applied the simulation framework presented by Bellinger and Oommen, in [1], to generate CTBT inspired datasets, and demonstrated these classification strategies within the most challenging classification domain. More specifically, we have shown that OC classifiers can successfully be applied to classify Stochastically Episodic (SE) events, which are unknown, although present, at the time of training.

The problem of including the temporal aspects of SE events in a PR methodology (for example, by invoking a time series analysis) remain open.

References

1. Bellinger, C., Oommen, B.J.: On simulating episodic events against a background of noise-like non-episodic events. In: Proceedings 42nd Summer Computer Simulation Conference, SCSC 2010, Ottawa, Canada, July 11-14 (2010)
2. Bellinger, C., Oommen, B.J.: Unabridged version of this paper (2010)
3. Datta, P.: Characteristic concept representations. PhD thesis, Irvine, CA, USA (1997)
4. Ghosh, A.K., Schwartzbard, A., Schatz, M.: Learning program behavior profiles for intrusion detection. In: Proceedings of the Workshop on Intrusion Detection and Network Monitoring, vol. 1, pp. 51–62 (1999)
5. Hempstalk, K., Frank, E., Witten, I.H.: One-class classification by combining density and class probability estimation. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part I. LNCS (LNAI), vol. 5211, pp. 505–519. Springer, Heidelberg (2008)
6. Japkowicz, N.: Concept-Learning in the Absence of Counter-Examples: An Autoassociation-Based Approach to Classication. PhD thesis, Rutgers University (1999)
7. Kubat, M., Holte, R.C., Matwin, S.: Machine learning for the detection of oil spills in satellite radarimages. Machine Learning 30(2), 195–215 (1998)
8. Saey, P.R.J., Bowyer, T.W., Ringbom, A.: Isotopic noble gas signatures released from medical isotope production facilities – Simulation and measurements. Applied Radiation and Isotpes (2010)
9. Stocki, T.J., Japkowicz, N., Li, G., Ungar, R.K., Hoffman, I., Yi, J.: Summary of the Data Mining Contest for the IEEE International Conference on Data Mining, Pisa, Italy (2008)
10. Tarassenko, L., Hayton, P., Cerneaz, N., Brady, M.: Novelty detection for the identification of masses in mammograms. In: IEE Conference Publications 1995(CP409), pp. 442–447 (1995)
11. Tax, D.M.J.: One-class classification; Concept-learning in the absence of counter-examples. PhD thesis, Technische Universiteit Delft, Netherlands (2001)

Iterative Translation by Monolinguists Implementation and Tests of the New Approach

Anna Potępa, Piotr Płonka, Mateusz Pytel, and Dominik Radziszowski

Department of Computer Science,
AGH University of Science and Technology,
Al. Mickiewicza 30, 30-059 Kraków, Poland
dr@agh.edu.pl
http://itm-services.eu

Abstract. The paper states Iterative Translation by Monolinguists (ITM), the new approach to the translation process in which a translation is carried out by persons without a common language who work in a cooperation, e.g. by two monolinguist instead of a translator. To prove the effectiveness of the approach, the pilot ITM platform as well as its dedicated extensions for both text translation and instant messaging communication have been designed, implemented and tested. Results of the tests confirm ITM's usability both in terms of quality and efficiency. Measured translation quality is acceptable and translation provisioning time is significantly lower. The concept is likely to bring a reduction of the multilingual information flow barrier and to be a transitory solution during the process of creation of an infallible machine translation engine.

Keywords: iterative translation, monolingual human translation, web services, machine translation, translation service platform, translation quality testing.

1 Introduction

Pierre from France and Lisa from the USA want to communicate with each other, but they know only their native languages. They don't want to use professional translator's help because it would be too expensive, uncomfortable or embarrassing. The only solution they have is to use one of Machine Translation (MT) engines, which provides a fast and free way of having a text translated. Although widely used and continuously enhanced, MTs still introduce many errors to produced translations. What should Lisa do with the machine translated phrases she does not understand? In our solution Lisa highlights incomprehensible fragments of the translated text, and then Pierre, seeing the problem, tries to say the same thing using simpler words. Hopefully, this time the previously paraphrased text is translated by the MT engine correctly and Lisa is able to understand the sense although it was impossible. If the MT engine fails again, the process may be repeated. As a result, after one or more iterations Lisa should receive correct and comprehensible text.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 445–454, 2011.

The above real live example is based on almost any (especially highly motivated) person's ability to recognize and to correct errors in texts written in his or her native language and to express ideas in different ways (provide paraphrase). An exploitation of this facts in combination with machine translation at its current stage of maturity, gives an opportunity to prepare acceptable quality translations. This approach has been called **ITM – Iterative Translation by Monolinguists**. With ITM, translation is carried out by two or more people without a common language who work in a cooperation, e.g. by monolingual users instead of a translator. Such a translation is a solution which aims to be more reliable than automated algorithms, but far cheaper and faster than a traditional approach involving a professional translator.

To prove the effectiveness of such an approach a pilot implementation of a ITM system has been successfully developed and tested. In this paper the results of the research on the ITM platform and its dedicated extensions for both text translation and instant messaging communication are presented. Section 2 focuses on a classification of machine translation output, which is crucial for the arrangement of ITM's decision flow and a translation model proposed in Section 3. Section 4 gives details on ITM platform implementation and is followed by results of effectiveness tests collected in Section 5. Section 6 and 7 present conclusions and benefits that usage of the ITM system can bring to different translation parties.

2 Quality Classification of MT Results

As the automatically translated segments are far from being perfect, and happen to be completely incomprehensible, the MT produced output has to be divided into quality related categories that determinate indispensable corrective actions. From the point of view of target language monolinguist, the following categories were introduced:

1. VALID - comprehensible and correct output.
2. CORRECTABLE - comprehensible output which requires some post edition changes in style, punctuation, vocabulary or grammar. Example:
   MT(EN): *This solution dedicate to companies who reduce capital expenditure.*
   Correct(EN): *This solution is dedicated to companies which...*
3. INCOMPREHENSIBLE such translation may be nonsensical, there is no chance a monolingual target language speaker will know how to post-edit it without additional piece of information. Example:
   Source segment (FR): *A' l'impossible nul n'est tenu.*
   MT(EN): *In the impossible no one is required.*
   Correct translation (EN): *No one is bound to do the impossible.*
4. DUBIOUS - the output is comprehensible (correct or may need some post-edition changes), but it is possible that it contorts the original meaning. The chance of this incompatibility is determined primarily on the basis of the text's context. Example:
   MT: *I went to bed, I could only think of a sleeve.*
   Possible correction: 'sleeve' to 'sleep' indicated by the word 'bed', but one cannot be certain, that the the original sentence did not mention 'sleeve'.

5. FALSE - comprehensible and correct target language output which carries improper meaning. Such output introduces false information into translated text. Example:
   Source(FR): *C'est un homme qui a annonce qui avait gagne une enchere.*
   Translation(EN): *This is the man who announced that he had won an auction.*
   Correct(EN): *This is the man who announced who had won an auction.*

There are existing systems which help human translators to manage and to correct an output of translation memory [1,7] or MT engines [8] by displaying both original and translated segments. A bilingual person can then compare them and fix possible mistakes. ITM approach addresses the problems in groups 2, 3, 4 and 5 so that they can be resolved by monolingual users.

A MT output can be easily enhanced by a target language speaker. Someone who is familiar with a target language can recognize accurate sentences from the first category as well as correct mistakes of the sentences from the second group and therefore introduce a more appropriate style to the final text.

When another user who knows the source language is added to the process new possibilities emerge. Supposing target language speaker cannot understand a segment or finds it dubious, a paraphrase of its original form can be requested. The same phrase written in other words and carrying the same meaning can be translated to the target language by a MT again, with a bit of luck this time the translation is understandable, or gives a hint on the meaning of the original one. If the translation of the paraphrase is not sufficient there are other ways to clarify the meaning, such as images and comments which are discussed later.

The main advantage of this method is that it is easier to find two monolingual people, than one who speaks both necessary languages. This benefit is even more significant, when two languages are not likely to be known by the same person. In that case the standard approach is to have the text translated to English and then to target language, which doubles the work and may introduce more errors. What is more, translation in cooperation enables more than two users to concurrently work on the same text.

3 ITM Translation Model

A number of features need to be provided by the ITM system which enables users to translate texts or messages in a cooperation to ensure comprehension is achieved. At the beginning of the process, the text is split to segments [9], and than each segment is machine translated. At this point, the output returned by the MT is presented to the interlocutor who speaks the target language. For each segment one decides which category it belongs to and undertakes an appropriate action.

User actions - Fig. 1 - contains target language monolinguist decision flow. Diagram includes three basic ITM actions (in black) and optional actions (in grey), which define additional means of communication able to significantly speed up the entire process.

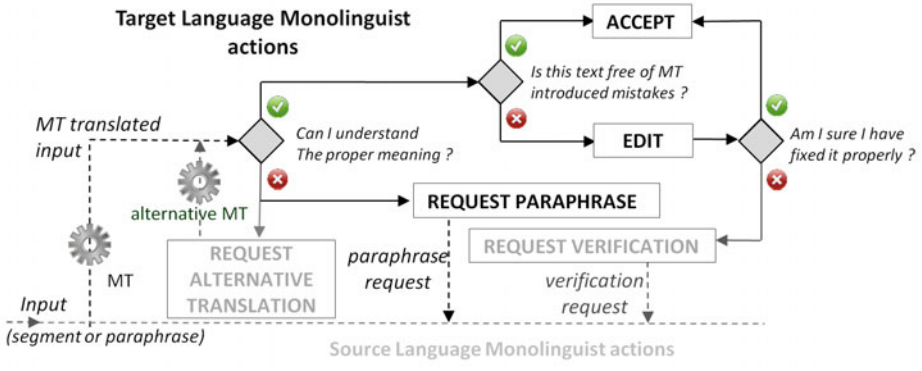


Fig. 1. Target language monolinguist actions and decision flow

Basic actions cover:

- **Acceptance** - Segment is found comprehensible and correct.
- **Edition** - Segment is grammatically wrong or any other kind of mistake introduced by the MT can be fixed. After edition a user may accept the segment. In the background a better translation suggestion is sent back to the MT engine or stored in a database.
- **Paraphrase Request** - segment or its part is incomprehensible, a target language user requests a paraphrase and put further editing of the text segment on hold until a response is delivered. Once the translated paraphrase is available the user should try to fix the segment translation using the provided information.

Optional actions supported by ITM system:

- **Alternative Translation** - System may use two or more MT engines. Provided with multiple translations of a given segment, target language monolinguist can pay special attention and request a paraphrase of differently MT translated text, reducing an amount of false facts introduced to the translation - errors in category 4 or 5 (comprehensible, but the sense is contorted).
- **Verification Request** - Sometimes the editing user cannot decide whether the fixed sentence reveals the real meaning of the original one in a source language. The edited sentence should then be reverse-translated and confirmed or rejected (and parphrased) by a source language user.

Source language monolinguist's actions are performed on demand, according to the flow in Fig. 2. There is only one basic action:

- **Providing Paraphrase** - Whenever a segment is incomprehensible for a target user, a source language speaker should provide a redundant source phrase which means the same as an original segment for which help has been requested. If necessary the sentence should be analyzed in context of surrounding segments.

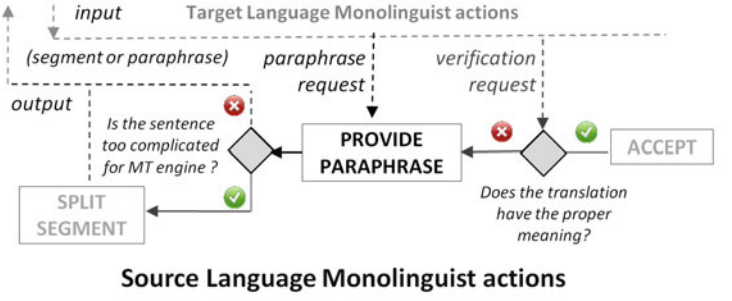


Fig. 2. Source language monolinguist actions and decision flow

Optional actions include:

- **Acceptance** of reverse-translated segments, which a target user post-edited and was unsure if the edition was correct. This process can introduce further errors, but in our tests it proved to be a valuable addition to the system.
- **Segment splitting** segments and replacing long sentences with two or more shorter ones, which possibly can be easier to translate for a machine translation engine.

To sum up, in a presented method of translation the target language user is responsible for the final appearance of a translated text. He or she alone decides whether help is needed and when a sentence can be accepted, while source language user's help is also provided on demand. The presented translation model was used in an implementation of the experimental ITM platform as well as its graphical interface which enables users to perform described actions and therefore utilize the ITM process to provide a translation of the text.

4 Pilot Implementation

During the research, the pilot implementations of the ITM platform was developed. The main aim of the platform is to provide basic functionality and programmer interfaces, to enable its usage in a number of services designed for different translation purposes. To fulfill this task the ITM Platform is based on number of machine translation engines (including Google Translate[4] and Babel Fish) as well as using data storage and access possibilities. Overall system architecture is presented in Fig. 3. It is implemented in Java as a 3-tier application. The bottom, data persistence layer currently supports two different storage possibilities - relational database and Java Content Repository[2] which was chosen because of scalability reasons. Middle-level layer gathers platform business logic and exposes it as a package of services and also SOAP[3] web services for remote usage. The platform itself provides a common WebService interface which enables other applications to provide ITM process driven translation.

The core system functionality can be extended with a number of plug-ins that enables ITM usage within different context and technologies. Our research efforts focused mostly on:

– Translation of a text – with the main aim to produce coherent, accurate translation understandable for a target language user and possibly free of grammar and style mistakes.
– Real time chat – aimed to provide means of communication to monolingual users. The messages exchanged during the chat should be understandable for all participants. There is no need to enhance sentences which contain even serious grammar or style errors but are comprehensible. In this case collaborative translation service should provide means of indicating problems in comprehension and resources indispensable to clarify expressed thoughts.

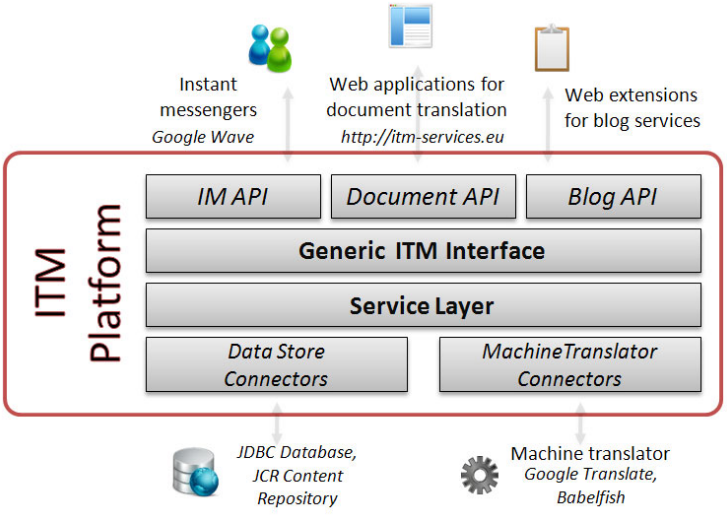


Fig. 3. ITM - General Platform Architecture

There are other possible areas of the practical application of the ITM technology including social network services and an exchange of internal documents in international enterprises, they will be under research in the near future. Currently implemented extensions are: an online web application providing users with tools they need to successfully translate a text, and a Google Wave technology integrated solution, which allows different monolinguists for a multilingual instant communication.

5 Effectiveness Tests

The main goals of effectiveness tests were to verify the quality of an ITM translation in comparison to a traditional process and to estimate the time and effort needed to perform the translation. Texts of three different specialisations were translated in three language combinations - Tab. 1. Professional translations of these texts were the patterns used in the effectiveness evaluation. ITM translation process was performed by three native speakers of English, Polish and

Table 1. Tested texts - characteristics

Property \Test	I	II	III	IV	V	VI
Language Combination	EN-PL	EN-PL	EN-PL	PL-EN	PL-EN	FR-EN
Type of text	Adventure	Technical	Article	Technical	Article	Article
Number of segments	33	45	28	20	28	10
Number of words	633	804	610	321	533	192

Table 2. BLEU results

Results \Test	I	II	III	IV	V	VI
BLEU (5-grams) for MT	0.152	0.207	0.101	0.192	0.172	0.243
BLEU (5-grams) for ITM	0.301	0.356	0.221	0.341	0.316	0.381
Improvement in BLEU	98%	71%	118%	77%	83%	56%

Table 3. Adequacy and fluency results

Results \Test	I	II	III	IV	V	VI
Rating(1-5) - MT	2.2 / 2.0	3.1 / 2.1	2.9 / 2.0	3.2 / 2.9	3.0 / 1.9	3.3 / 3.1
Rating(1-5) - ITM	3.0 / 3.1	4.2 / 4.2	3.9 / 3.5	4.3 / 4.0	3.8 / 3.2	4.1 / 4.1
Improvement	36%/55%	28%/52%	34%/75%	43%/38%	26%/68%	24%/32%

French, who could communicate using only their own mother language making use of the ITM system.

To measure effectiveness metrics proposed by Callison-Burch et al. in [10] were used. A quality of translations was evaluated using different metrics:

- Automated evaluation - BLEU
- Human evaluation - scores based on 5-point adequacy and fluency scale
- Human evaluation - ranking translated sentences relative to each other.

Bleu [11] is the standard in a machine translation quality assessment. It calculates n-gram precision and a brevity penalty and can also use multiple reference translations. In human evaluation the five point scale for adequacy indicates how much of the meaning expressed in the reference translation is also expressed in a hypothesis translation (1-None, 2-Little, 3-Much, 4-Most, 5-All). The second five point scale indicates how fluent the translation is (1-Incomprehensible, 2-Disfluent, 3-Non-native, 4-Good, 5-Flawless).

The results achieved - Tab. 2 - showed considerable improvement in terms of BLEU (56% to 118%). In human assessment by bilinguals - Tab. 3 - scores showed that a produced text appears correct and most of the meaning was preserved. The most significant improvement was noted on fluency (even 75% in comparison to 'bare' MT) and an average score can be described as good.

Because fluency and adequacy ratings are not a standard and depend only on judges' feelings, [12] suggests trying a separate evaluation where people are asked to rank translations in relation to another. During this relative assessment

Table 4. Segment classification

Category \Test	I	II	III	IV	V	VI
VALID	6%	5%	4%	20%	7%	20%
CORRECTABLE	22%	40%	28%	25%	43%	50%
INCOMPREHENSIBLE	36%	20%	46%	15%	32%	20%
DUBIOUS	36%	35%	22%	40%	18%	10%
FALSE	3%	0%	3%	0%	7%	0%

Table 5. Translation time comparison

Category \Test	I	II	III	IV	V	VI
Working time - target language editor	50	90	45	25	50	12
Working time - source editor	18	25	20	5	15	3
Working time - professional translation	200	260	150	110	165	60
Working time - professional translator with CAT*	170	210	130	85	145	50

only 5-15% of MT output segments were perceived as correct and fluent as professional translation (mostly very short sentences) while 26% up to 44% of ITM produced segments were considered of as good quality as the reference professional translation.

Tab. 4 presents a classification of the segments translated by the MT engine described in Section 2. A number of segments in each category varies considerably for different types of text. For all texts, less than half of the segments were accepted by the target language editor immediately or after post-edition. For texts with long, complex and difficult sentences (e.g. Text 1) MT engine returned correct or nearly correct translation of only 28% of segments. Effectiveness of the MT engine is not a subject of this research, but the conclusion is that ITM must be prepared to handle any type of text and make no assumptions on how many paraphrases/verifications will be needed.

FALSE (contorting the general meaning) segments were counted by bilingual judges after translation had been done. The fact that only a few segments were categorized as FALSE is noticeable. Most of the MT introduced false facts were eliminated thanks to available ITM mechanisms.

The results achieved by ITM, clearly show that average quality of performed translations is good (4.1 in 1-5 scale) or at least acceptable (3.1 - non native) for the worst test case. The main advantage of ITM usage is of course time (and money). Tab. 5 gathers time statistics of three translation methods ITM, traditional translation and modern CAT tool (XTRF-TM). Overall working time of both persons involved in ITM process is at least half shorter than time needed by a professional bilingual translator (even working with CAT tools).

* Translation time with CAT is shorter mostly not because of segment repetitions but unique XTRF-TM[14] tool ability to operate concurrently and perform a proofreading just after a translation of a segment.

6 Benefits for MT Engines

Thanks to the data obtained during work on translations, the ITM system may improve the effectiveness of MT engines. For example, the web version of Google Translate Service[4] provides the interface to contribute a better translation for the given segment. Each human revised and accepted phrase in the ITM system can be submitted via mentioned interface or made available via web services. This way the most significant improvement can be achieved in style and reader friendliness. A human proofreader can express the meaning in more accurate words or remove ambiguity introduced by the MT engine in a way that no algorithm can achieve.

Not only does the ITM system collect improved translations, but it also has the ability to deliver a set of paraphrases which are supplied by users of the source language whenever the target language user does not understand a machine-translated sentence. As it has been proved that a paraphrase corpus can be crucial for training MT engines [12]. There are existing methods to extract paraphrases automatically from texts [13], but what ITM technology offers is a way to provide human created, fluent paraphrases for specific sentences, which are currently translated incorrectly.

7 Conclusions

Iterative Translation done by Monolinguists (ITM) is a very promising way of international rendition. Created platform proves not only implementability of the concept, but has also been tested in a real runtime environment. Results of the tests confirms ITM's usability both in term of quality and efficiency. The measured translation quality is acceptable, parties involved in the translation process are required to know only one language and a translation provisioning time is radically lower. The concept is likely to bring a significant reduction of the multilingual information flow barrier and change current translation market. The most important promises of ITM are:

- an increase of the number of participants in multilingual communication, as well as the number of people who may be engaged in the translation process - shift from number of persons with good knowledge of at least two languages to all Internet users,
- a chance to expand the translation market by the spheres that until now, have been treated as unprofitable due to the cost and time of the translation,
- a significant improvement of machine translation engines effectiveness, by providing feedback information about segments' translation quality, paraphrases and correct translations.

In conclusion it could be seen as a transitory solution, during a process which aims to produce an infallible fully automated translation engine.

References

1. Casacuberta, F., Civera, J., Cubel, L., Lagarda, A.L., Lapalme, G., Macklovitch, E., Vidal, E.: Human Interaction For High-Quality Machine Translation. Communications of the ACM (2009)
2. Content Repository for Java Technology, http://www.day.com/specs/jcr/
3. Simple Object Access Protocol W3C, http://www.w3.org/TR/soap/
4. Google Translate Homepage, http://translate.google.com/
5. Google Wave, http://wave.google.com/
6. BabelFish Homepage, http://babelfish.yahoo.com/
7. Barrachina, S., Casacuberta, F., Cubel, E., Lagarda, A., Tomas, J., Vilar, J.-M., Bender, O., Civera, J., Khadivi, S., Ney, H., Vidal, E.: Statistical Approaches to Computer-Assisted Translation, Computational Linguistics (2007)
8. Barrachina, S., Casacuberta, S.F., Cubel, E., Lagarda, A., Tomas, J., Vilar, J.-M., Bender, O., Civera, J., Khadivi, S., Ney, H., Vidal, E.: Correcting Automatic Translations through Collaborations between MT and Monolingual Target-Language Users (2009)
9. Milkowski, M., Lipsk, J.: Using SRX standard for sentence segmentation in LanguageTool (2009)
10. Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., Schroeder, J.: Evaluation of Machine Translation
11. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL (2002)
12. Callison-Burch, C., Koehn, P., Osborne, M.: Improved Statistical Machine Translation Using Paraphrases
13. Shinyama, Y., Sekine, S., Sudo, K.: Automatic Paraphrase Acquisition from News Journals. In: Proceedings of Human Language Technology Conference (2002)
14. XTRF-TM Translation Management System, http://www.xtrf.eu/

Attribute Mapping as a Foundation of Ontology Alignment

Marcin Pietranik and Ngoc Thanh Nguyen

Institute of Computer Science, Wroclaw University of Technology,
Wybrzeze Wyspianskiego 27, 50-370, Wroclaw, Poland
{marcin.pietranik,ngoc-thanh.nguyen}@pwr.wroc.pl

Abstract. In this paper we present preliminary results of our work on creating a mapping function between two ontologies. Careful investigation of many different approaches to this task, has brought us to the point where we were able to distinguish independent ontological levels, differing on their granularity- starting from the structural level, through concept level and attribute level. This distinction drew our attention to the necessity of precise definition of attribute's structure. In this paper we propose our ideas about developing consistent methodology of expressing attributes' complexity and in further parts possible approaches and requirements about the distance function that is able to compare these structures. A comparison with other related works and brief description of potential development directions is given as a summary.

1 Introduction

In distributed, web-based environments that are expected to handle knowledge management task, ontologies play important role. By formal definitions they can be treated as a representation of part of reality, with all of it's behavior and specific features. Recent approaches to defining ontologies has brought the whole topic closer to practical applications, basing on simplification of ontological structure to straightforward labeled graphs that are able to describe basic relations between real world objects, represented as nodes of these graphs.

Such simplification concentrates only on the surface of the task, that is easy to implement and visualize, but omits the grounding level of the issue. This has brought us to the point where we asked ourselves about the foundations of ontologies. What are their building blocks? Can they be divided into some smaller, atomic elements? How can we define them? How can we retain expressiveness? Eventually, we have noticed the necessity of coming up with the answer to these questions.

Our contribution to the topic is expanding ontology definition with it's basic building blocks. Furthermore, we provide reliable way of calculating distances between them, along with formal definitions of requirements that any future modifications or other approaches should meet. We claim that it should be preliminary step in finding ontology alignments. Effective mapping ontologies requires the analysis of their content, on every level of granularity.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 455–465, 2011.

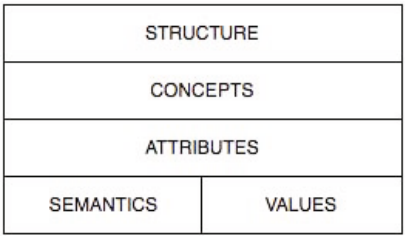


Fig. 1. Ontological Stack- Levels of Abstraction within ontologies

Starting with the foundations, we claim that the good beginning is to process the structure of attributes (which are building blocks of concepts), by spreading them into two levels: semantics and valuation. What is more, we provide our attempt to find and define their characteristic features and how these features relate to other feature that can be expressed in ontologies.

After careful analysis of the problem we were able to distinguish four levels of abstraction that occur within ontology. On Fig.1 we present the ontological stack, which is the way we see ontologies as a sum of it's components.

Beginning with former approaches that treat ontologies as a labeled graphs, we have put on top *the structure level* which holds the knowledge about binary relationships between concepts. It can refer to hierarchical connections or any other links between them. Below we have put *the concept level* that contain the way of describing real world objects, their names and structures. Next, we need to define mentioned structure of concepts, which can be achieved in *the attribute level*. In this level the necessity of expressing every aspect of attributes appears. We present this on the lowest level of ontological stack, which contains two elements. By *semantics* we call the description of basic, atomic features of attributes- the way they relate to real world features spreaded across many attributes. On the other hand, *valuations* of attributes describe possible states they can accept.

We claim that distinguishing explicit attributes' semantics from their valid valuations is consistent with intuitive way we see the real world. For example, consider an attribute *weight*. Despite the same semantics (which is describing actual weight of some object) within different ontologies, their valuation may differ. In one it can be expressed with pounds (*lb*) and in the other with kilograms (*kg*). Similar issue occurs when two attributes share the same valuation (for example finite subset of natural numbers), but differ in semantics (for example one attributes refers to *age* and some other to *length*). Third issue that naturally appears is naming conventions used to label attributes. Obviously names, used just as they are, can cause uncertainties in identifying common entities from ontologies that refer to the same objects from the real world.

Described problems play crucial role in finding correlations and similarities between ontologies. Approaching this task (also called *ontology alignment*)

require taking into consideration these aspects of modeling real world. We see former ontology alignment methods (described briefly in section 3), that heavily rely on labels, as vague and insufficient.

In this paper we provide formal definitions of every level of ontological stack from Fig.1 along with our ideas about calculating distances between elementary objects within ontologies. Further sections of this article are organized as follows. Part 2 contains detailed formal definitions of elements of every ontological level illustrated on Fig.1. Part 3 contains overview of work that has been done and described in literature. Section 4 contains description of our work on defining attribute distance function. The last part gives brief overview of possible future works and short summary.

2 Basic Notions

Recently developed and described in literature ([4] and [15]) definitions of ontologies have become too complex and narrowed only to implementation aspects. Despite this fact, most of these approaches didn't specify the basic and atomic building blocks of ontologies (for example concepts or attributes). Therefore, they lack the expressing power to provide unequivocal method of describing semantics of the smallest entities that can be found in such structures.

In the following paper we would like to revise approaches to ontologies' definitions. Treating [16] as a starting point, we propose a simplification of considered topic, but also precise definition of ontologies on every level of granularity, which is a level of the whole ontology as a structure of describing knowledge, a concept level which carries the meaning of basic objects within ontologies and attribute level, which contains the basic semantical features used further in defining any other entity within ontology.

Following the definition taken from [16] we define ontology as a triple:

$$O = (C, R, I) \tag{1}$$

where C is a set of concepts, R is a set of relationships between concepts (defined as $R \subseteq C \times C$) and I is a set of instances. Next, we define every concept c from set C also as a triple:

$$c = (Id^c, A^c, V^c) \tag{2}$$

in which Id^c denotes a concept label (that is an informal name of a concept), A^c is a set of attributes belonging to the particular concept and V^c is a set of domains of attributes from A^c. The triple c is called *concept's structure*.

Let $A = \bigcup_{c \in C} A^c$ and $V = \bigcup_{c \in C} V^c$ Henceforth, we will call Real World or Universe of Discourse the tuple (A, V). Such structure contains all of the possible attributes along with all of their possible valuations. Therefore, it contains possible states that can be expressed in ontologies build on top of such framework.

In this point we have noticed the need of defining inner semantics of attributes, that would describe their basic properties. We define semantics of attributes as follows.

We assume that there is a defined and finite set X of atomic descriptions for attributes semantics. An element of set X is an elementary description given in natural language. For example:

- "The Day of Birth" for attribute *Birth_Day*
- "The Tax Identification Number" for attribute *TIN*
- "The Length of Life" for attribute *Age*

We now assume the set S in the set of symbols identifying the semantics' elements (elementary description) used to build complex descriptions of semantics of attributes. Let L_x be the formal language, in which we use symbols from S and logic operators $\neg, \vee, \wedge$ for forming formulas. L_s is a sublanguage of the sentence calculus language. For example, for $S = \{s_1, s_2, s_3, s_4, s_5\}$ we can give set $L_s = \{s_1 \wedge s_2, s_3 \wedge \neg s_4, s_5\}$.

Definition 1. By semantics of attributes we call a function:

$$S_o : A \rightarrow L_s \tag{3}$$

Thus, the semantics of an attribute $a \in A$ is $S_o(a)$, being some formula from language L_x.

Definition 2. By valuation of an attribute $a \in A$ we call a partial function

$$F_a : I \rightarrow V_a \tag{4}$$

In order to be able to easily compare semantics of whole concepts we need to be able to recognize different correspondences between concepts' building blocks, which are attributes. Owing definitions 1 and 2 we can now define the following relations between attributes.

Definition 3. Two attributes $a, b \in A$ are equivalent referring to their semantics (*semantical equivalence*) if the formula $S_o(a) \Leftrightarrow S_o(b)$ is a tautology.

Definition 4. Two attributes $a, b \in A$ are equivalent referring to their values (*value equivalence*) if $F_a = F_b$. Two attributes $a, b \in A$ are equivalent if they are in semantical equivalence and value equivalence. Developing this approach we can define further relations between attributes, which are *generalization* and *contradiction*.

Definition 5. Attribute $a \in A$ is more general then attribute $b \in A$ referring to semantics (*semantical generalization*) if formula $S_o(a) \Rightarrow S_o(b)$ is a tautology.

Definition 6. Attribute $a \in A$ is more general then attribute $b \in A$ referring to values (*value generalization*) if there exists a function from $F_a(I)$ to $F_b(I)$.

Definition 7. Attributes $a, b \in A$ are in contradiction if formula $\neg(S_o(a) \wedge S_o(b))$ is a tautology.

3 Related Works

3.1 Ontology Mapping

Ontology mapping task (also called ontology alignment) is a problem widely discussed in literature. [4] contains the description of the state of the art of this topic along with careful investigation and detailed descriptions of both contemporary and former literature. Basic approach (that is common to all approach featured in [4]) treats ontology alignment task as a task of finding a set *Align* containing tuples of the structure: (id, e, e', r, n), where id is a unique identification of particular tuple from *Align*, e and e' are entities belonging to two different ontologies that are about to be mapped, r is a relationship holding between these entities (it can be *equivalance*, *disjointness* and *generalization*) and n is a confidence degree which is a normalized similarity value (from set $[0, 1]$) between two entities.

The main problem of finding such mappings is finding efficient and reliable similarity function which will return confidence values for any two objects from ontologies, that can be further confronted with fixed threshold. Such an approach allows to easily filter proper pairs of entities from ontologies and return a set *Align* that satisfy given requirements.

In [11] authors propose a semantic distributed system that makes existing mapping information exchangeable and convenient to share, by collecting existing mappings (aquired from other systems) and aggregating them. Therefore, they provide an indirect approach to solving ontology alignment, in order to create a solution to reliable query transformations.

3.2 Propositional Distances

In literature there can be found many different approaches to this problem, but all of them can be divided into two groups: purely syntactic and model-based.

We assume that logical clauses are composed with propositional letters taken from the finite set P, also called alphabet, containing all of the propositional letters (symbols) used to describe basic atomic features, that can be utilized to express more complex properties. Furthermore, we will incorporate the notation taken from [14] that denotes two logical expressions as d and q (which are abbreviations for *document* and *query*). These expressions will be built with letters from P and standard logical connectors $\neg, \vee, \wedge$.

The task of comparing two clauses (or set of clauses) comes down to finding a distance function D between two elementary entities taken from set X, which is a set of all of possible clauses. Therefore desired function is defined on the Carthesian product $X \times X$ and operates with non-negative real values. As described further in [13], it is also required that it meets conditions of standard metric space.

The first and the most intuitive approach to comparison of two logic statements is incorporation of one of the many different methods of calculating distances between finite sets, such as Jaccard's distance [10] or Hamming's distance [9]. In [7] and [6] the generic approach has been presented. It assumes that both d

and q are single clauses and from there authors introduce number of parameters used for calculating the overall distance: n, the number of propositional letters existing only in d, m, the number of propositional letters existing only in q, l, the number of propositional letters existing both in d and q.

Subsequently, the Jaccard's distance function is defined as follows:

$$D(d,q) = \frac{n+m}{n+m+l} \tag{5}$$

Such an approach can be easily generalized into the method of calculating distances between sets of clauses- for example complex statements expressed in disjunctive normal form. Assuming that now d and q are in the respective forms $a_1 \wedge a_2 ... \wedge a_s$ and $b_1 \wedge b_2 ... \wedge b_t$, equation 5 can be rewritten in two ways as follows:

$$D(d,q) = \min_{i \in [1..s], j \in [1..t]} D(a_i, b_j) \tag{6}$$

$$D(d,q) = \frac{1}{s+t} \sum_{i=1}^{s} \sum_{j=1}^{t} D(a_i, b_j) \tag{7}$$

Equation 6 simply returns minimal distance between elements from two sets of clauses. Therefore, it meets all of the requirements listed in the beginning of this section. Nevertheless, it also has some major disadvantages. First of all suppose that we want to compare two large sets containing completely different clauses, but also containing two clauses that are identical. This method will eventually return distance equal to 0, suggesting that given sets are also identical. Intuitively we can say that such approach to comparing logical expressions won't return plausible results.

On the other hand equation 7 returns average value of distances between elements of compared sets. On first sight, such approach fixes the problem appearing in equation 6, but suppose we are given two sets containing the same clauses a and b. Assuming such input data, considered approach won't return 0. Despite the identity of compared sets, it will return 0,5 instead (by comparing all of the elements pairwise from given sets).

Both of the described methods are not reliable because of few other disadvantages. First of all it is purely syntactic approach, therefore they omit semantical analyze of compared input. What's more, they do not include the uncertainty of information expressed in compared statements- they see no difference between the situation in which some clause contains negation of some propositional letter and the situation in which the same clause does not contain this letter at all. In other words- uncertainty caused by the lack of some feature (we cannot say if modeled entity have or not) is not captured.

In [14] an interesting approach to calculating similarities between prepositional clauses has been introduced. Authors apply techniques taken from the field of belief revision to get a measure of similarity between documents stored in some knowledge base and queries fired against such base. Having the raw set of propositional letters P and the set of standard logical connectors authors in

[14] enhance them with semantics, explicitly given by interpretations, which are functions from P to set $\{true, false\}$. In other words- particular interpretations is a standard valuation assigned to every propositional letter. A model of some formula α is simply an interpretation mapping this formula into *true*. For the simplicity, authors treats interpretations (and therefore models) as sets of letters mapped into *true*. Furthermore, authors denote set of all models of a formula with *Mod* operator (e.g. set of all models for α is denoted as $Mod(\alpha)$). We can say that formula α entails β if and only if $Mod(\alpha) \in Mod(\beta)$ (this will be further denoted as $\alpha \models \beta$). Next, authors introduce "partial vectors" to conveniently represent formulas as sets. For example formula $\alpha = a \wedge b \vee c$ will be represented as $\{\{a, b\}, \{c\}\}$.

Note that interpretations (and models) are also represented as sets of letters from formulas mapped into *true* and all considered formulas are expressed with disjunctive normal form. Therefore authors define set of models of a formula as a set of partial vectors containing truth values of elementary conjunctions combined with atoms from alphabet P not appearing in particular conjunction. For example, assume the propositional alphabet $P = \{a, b, c, d\}$ and a formula $\alpha = a \wedge \neg b$. The model of this formula can be represented as $Mod(\alpha) = \{\{a\}, \{a, c\}, \{a, d\}, \{a, c, d\}\}$. Having such structures for representing any formula and it's models authors incorporate Dalal's measure (described in details in [2]), which calculates the distance between two formulas as the number of propositional letters on which interpretations of these formulas differ. This can be calculated as a symmetrical distance- having two sets A and B it is given as:

$$Dist(A, B) = |(A \cup B) \setminus (A \cap B)|\,. \tag{8}$$

Henceforth, we will return to the situation in which we want to calculate the distance between two formulas representing some document d from knowledge base and some query q. Generally speaking a measure of distance between formulas is given as a distance between their models. Authors divide the task of comparing these two formulas into two cases: the one in which a document has only one model and the second, in which document can have several different models. In the first case this measure is defined as the distance between the model of d and the closest model of q.

$$D(d, q) = \min_{m \in Mod(q)} Dist(m, q) \tag{9}$$

The intuition behind this approach is very straightforward- the distance tells us about the minimal number of letters the should be changed in the document to satisfy the query. In other words- how much the particular document should be modified in order to ensure that the entailment $d \models q$ begins to hold.

Authors of [14] expand this simple method to the point where it can deal with documents that have several models. They propose to calculate average distance from models of the document to the set of models of the query, as follows:

$$D(d, q) = \frac{\sum_{m \in Mod(d)} Dist(Mod(q), m)}{|Mod(d)|} \tag{10}$$

Of course, in literature can be found many other approach to described task- for example [3] takes into account the probability of validity of matching two formulas. This solution, due to the assumption that the distribution to mentioned probability is explicitly given, does not correspond to our needs. Therefore we see model-based approaches as efficient methods of computing distances between documents and queries represented as disjunctive normal form logical formulas. Flexible handling of uncertainties of formulas, inclusion of the whole alphabet in comparison process and adding semantics (given by interpretations) seems to be fitted enough to meet our needs of comparing semantics of concepts' attributes within ontologies.

4 Attribute Mapping Function Definition

In order to find possible mapping between two ontologies we need at first find a mappings between elementary objects from them, which are concepts. Moving this idea forward, we need to begin with comparing structures of these concepts. So the task of comparing concepts comes down to the task of comparing their sets of attributes and attributes itself.

We claim that finding alignments between ontologies can occur only within the same universe of discourse (A, V). According to intuition, it is impossible to find correspondences between two unrelated topics- for example *computer science* and *chemistry*. Omitting these assumption can cause unreliable results, that can easily interfere with expected behavior of the application in which ontology alignment is incorporated.

For mapping an attribute $a \in A$ to attribute $b \in A$ we need to define a distance function between them. This function $d(a, b)$ has two components: semantical and value distance.

Definition 8. By the semantical distance between attributes we call a following function:

$$d_s : A \times A \rightarrow [0, 1] \tag{11}$$

which satisfy the following conditions:

1. $d_s(a, a) = 0$ for all $a \in A$
2. $d_s(a, b) = 0$ if a and b are in semantical equivalence
3. $d_s(a, b) = 1$ if a and b are in semantical contradiction
4. $d_s(a, b) = d_s(b, a)$ for all $a, b \in A$

Since the semantics for particular attributes from A are formulas from language A_x (according to equation 3 in definition 4) we can assume that these formulas are expressed in disjunctive normal form. Having this assumption, we can apply the modification of model based approach (described in section 3.2). Loosening the requirement of triangle inequality, but keeping our conditions from current section we propose following method of calculating distance between two attributes' semantics.

Definition 9. Having two attributes $a \in A$ and $b \in A$, by the semantical distance between them we call a following function:

$$d_s(a,b) = Dist(S_o(a), S_o(b)) = \frac{\sum_{m \in Mod(S_o(a))} min_{m' \in Mod(S_o(b))} Dist(m,m')}{|Mod(S_o(a))|} \tag{12}$$

Where function *Mod* return models of formula in accordance with the definition from section 3.2 differing to the fact that it does not use the whole propositional alphabet S but only the sum of all atomic descriptions used in formulas $S_o(a)$ and $S_o(b)$. The *Dist* function in a standard symmetrical distance from equation 8.

Considering two attributes with complex semantics we want to gain the knowledge about features they both refer to. Due to intuition we believe that returning average of minimal distances between components of two semantics is the most appropriate to achieve this requirement.

Definition 10. By the value distance function between attributes we call a function:

$$d_v : A \times A \to [0,1] \tag{13}$$

which satisfy the following conditions:

1. $d_v(a,a) = 0$ for all $a \in A$
2. $d_v(a,b) = 0$ if a and b are in value equivalence
3. $d_v(a,b) < 1$ if a is more general then b
4. $d_v(a,b) = d_v(b,a)$ for all $a,b \in A$

Since automatically we can only determine if two valuations from definition 2 are equivalent or not, it is very difficult to define a value distance function between to attributes. We think that declaring whether or not the function $F_a(I)$ to $F_b(I)$ exists heavily relies on expert's knowledge. Therefore, we cannot provide automatic method of calculating considered distance.

Our initial idea is nevertheless including the level of coverage of a domain and codomain of function transforming $F_a(I)$ to $F_b(I)$ in comparison to the cardinality of initial sets $F_a(I)$ and $F_b(I)$.

5 Future Works and Summary

In this article we have made our first attempt to expand ontology definitions taken from our previous works. We have added explicit semantics to the smallest building blocks of ontologies, which are concepts' attributes. In further parts we have investigated possibilities of calculating distances between these atomic entities. We have taken into consideration all of their features, both semantic content and valid valuations. Also, we have formulated formal requirements that any future modifications and developments must meet in order to be treated as reliable.

In the future we would like to concentrate on expanding the structure of semantical descriptions of attributes, enriching them with relationships between atomic descriptions of semantics. We claim that more general attribute should

be closer to more specific attributes, than two specific attributes to each other. We want to create a solution that will be able to include formal reasoning in distance calculation.

Second direction of our work will concentrate on including attribute distance function in defining distance between two concepts from different ontologies. Owing precise definition of their structure, we believe that it will be possible to generalize common ontology alignment methods, which return simple one-to-one mappings between concepts into standardized approach of denoting complex many-to-many mappings.

What is more, we think that our work can be treated as a starting point in developing a standardized methodology of creating ontologies itself, simplifying their analysis and application.

Acknowledgment. This research was partially supported by Polish Ministry of Science and Higher Education under grant no. N N519 407437.

References

1. Bernstein, P., Halevy, A., Pottinger, R.: A vision of management of complex models
2. Dalal, M.: Investigations Into a Theory of Knowledge Base Revision: Preliminary Report. In: Proceedings of the 7th National Conference on Artificial Intelligence, pp. 475–479 (1988)
3. Esposito, F., Malerba, D., Semeraro, G.: Classification in noisy environments using a distance measure between structural symbolic descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 390–402 (1992)
4. Euzenat, J., Shvaiko, P.: Ontology Matching, 1st edn. Springer, Heidelberg (2007)
5. Fensel, D.: Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce. Springer, Berlin
6. Ferilli, S., Basile, T.M.A., Biba, M., Di Mauro, N., Esposito, F.: A general similarity framework for horn clause logic. Fundamenta Informaticae 90(1), 43–66 (2009)
7. Ferilli, S., Biba, M., Basile, T., Di Mauro, N., Esposito, F.: k-Nearest Neighbor Classification on First-Order Logic Descriptions. In: IEEE International Conference on Data Mining Workshops, ICDMW 2008, pp. 202–210 (2008)
8. Flouris, G., Plexousakis, D.: Belief Revision in Propositional Knowledge Bases. In: Proceedings of the 8th Panhellenic Conference on Informatics, pp. 412–421 (2001)
9. Hamming, R.W.: Error detecting and error correcting codes. Bell System Technical Journal 29(2), 147–160 (1950)
10. Jaccard, P.: Etude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Socit Vaudoise des Sciences Naturelles 37, 547–579 (1901)
11. Jung, J.J.: Ontology Mapping Composition for Query Transformation on Distributed Environments. Expert Systems with Applications 37(12), 8401–8405 (2010)
12. Konieczny, S., Prez, R.: Merging with integrity constraints. In: Symbolic and Quantitative Approaches to Reasoning and Uncertainty, pp. 233–244 (1999)
13. Li, M., Chen, X., Li, X., Ma, B., Vitnyi, P.M.B.: The similarity metric. IEEE Transactions on Information Theory 50(12), 3250–3264 (2004)

14. Losada, D.E., Barreiro, A.: A logical model for information retrieval based on propositional logic and belief revision. The Computer Journal 44(5), 410 (2001)
15. Maedche, A., Staab, S.: Ontology learning for the semantic web. IEEE Intelligent Systems and Their Applications 16(2), 72–79 (2001)
16. Nguyen, N.T.: Advanced Methods for Inconsistent Knowledge Management (Advanced Information and Knowledge Processing). Springer, Heidelberg (2007)

Multiagent-Based Dendritic Cell Algorithm with Applications in Computer Security

Chung-Ming Ou[1], Yao-Tien Wang[2], and C.R. Ou[3]

[1] Department of Information Management, Kainan University, Luchu 338, Taiwan
cou077@mail.knu.edu.tw
[2] Department of Computer Science and Information Engineering, Hungkuang University, Taichung 433, Taiwan
ytwang@sunrise.hk.edu.tw
[3] Department of Electrical Engineering, Hsiuping Institute of Technology, Taichung 412, Taiwan
crou@mail.hit.edu.tw

Abstract. Agent-based artificial immune system (ABAIS) is applied to intrusion detection systems (IDS). A multiagent-based IDS (ABIDS) inspired by the danger theory of human immune system is proposed. The intelligence behind ABIDS is based on the functionality of dendritic cells in human immune systems and the danger theory, while dentritic cells agents (DC agent) are emulated for innate immune subsystem and artificial T-cell agents (TC agent) are for adaptive immune subsystem. This ABIDS is based on the dual detections of DC agent for signals and TC agent for antigen, where each agent coordinates with other to calculate danger value (DV). Agents coordinate one another to calculate mature context antigen value (MCAV) and update activation threshold for security responses.

1 Introduction

Most solutions for network security are static, which collect, analyze and extract evidences after attacks. These methods have a common disadvantage, namely, they lack the abilities of self-learning and self-adapting. Therefore these solutions cannot prevent unknown attacks. Artificial immune system (AIS) is realized for computer security as a new research focus of biologically inspired computational approach.

Most AIS researches focus on the development of specialized AIS algorithms inspired by theories such as the negative selection theory or the danger theory. Forrest et al. [1] proposed an instance of computer immunology to protect the computer systems using the principles of human immune systems. Gu et al. [2] proposed an architecture of combining multiagent systems and dendritic cells. Applying AIS algorithm to intrusion detection systems (IDSs) can be traced back to [3].

For information security, one important aspect learned from immunology is the following: a computer security system should protect a host or network of

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 466–475, 2011.

hosts from unauthorized intruders, which is analogous in functionality to the immune system protecting the body from invasions by foreign pathogens. For example, anti-virus software has recently adopted some features analogous to the innate immune system, which can detect malicious patterns. However, most commercial products do not yet have the adaptive immune system's ability to address novel threats. Multiagent systems provide the ability of distributed information security management [4]. Each agent has its specific security task and goal, while coordination and cooperation can be achieved by MAS. The motivation of this research is the following. Can agent-based information security systems learn itself to effectively determine whether the abnormality is "actually" incurred by some malicious attacks. This paradigm is very similar to the danger theory proposed in the immunology [5]. In this way, host-based IDS obtained tunable and adaptable threshold values determining danger signals.

Greensmith et al. [6,7] proposed the Dendritic Cell Algorithm (DCA) whose purpose is to correlate data in the form of antigens and signals, then to identify groups of antigens as normal or anomalous. It is believed that a DC is better performed by agent technology while considering its adoption to network environment. DC agents in our agent-based intrusion detection system (ABIDS) will evaluate antigens and corresponding signals according to DCA to determine whether antigens are malicious nor not.

The arrangement of this paper is as follows. In section 2, preliminary knowledge such as AIS, danger theory, intelligent agents system and computer security are introduced. For section 3, IDS based on ABAIS model inspired by DT is discussed Simulations of two agent-based algorithms for ABIDS based on different scenarios will be given in section 4.

2 Background

2.1 Intrusion Detection Systems (IDS)

Intrusion detection systems (IDS) focus on exploiting attacks, or attempted attacks, on networks and systems, in order to take effective measures based on the system security policies, if abnormal patterns or unauthorized access is being suspected. However, there are two potential mistakes by IDS, namely, false positive error (FPE) and false negative error (FNR). For FPE (FNE), a pattern is mistakenly determined as abnormal (normal).

A host-based intrusion detection system (HIDS) consists of an agent on a host that identifies intrusions by analyzing system calls, application logs, file-system modifications and other host activities and state. In a HIDS such as application-based IDSs, sensors usually consist of a software agent.

2.2 Artificial Immune System (AIS)

The human immune system (HIS) consists of the antibodies and lymphocytes, which include T-cells and B-cells. It can be categorized as innate and adaptive immune system. The innate immune system is characterized by three roles,

namely, host defense in the early stages of infection, induction of the adaptive immune response and determination of the type of adaptive response. The main characteristics of the adaptive immune system are recognitions of pathogens.

Artificial Immune Systems (AIS), which is based on HIS, have been applied to anomaly detections [2,3,8,9,10]. AISs have been developed according to negative selection algorithm and clonal selection algorithm which are based on the classical self-nonself theory. This so-called self-nonself classification theory had been challenged while failing to explain several immunological phenomena. Some alternative theories have been proposed, for example, the danger theory (DT).

Dendritic Cells. The dendritic cell (DC) is a vital link between the innate and adaptive immune system and provides the initial detections of pathogenic invaders. Once activated, they migrate to the lymphoid tissues where they interact with T-cells and B-cells to initiate the adaptive immune response. DCs are the first defense line for HISs which will arrive at the locations where antigens intrude. DCs can combine the danger and safe signal information to decide if the tissue environment is in distress or is functioning normally.

T-Cell. DCs will influence the differentiation of T-cells by releasing particular cytokines. They drive the T-cell to react to the antigen in an appropriate manner. Naive T-cells are those who have survived the negative and positive selection processes within the thymus, and have migrated to circulation system between the blood and lymphoid organs. Naive T-cells reach an activated state when the T-cell receptor (TCR) on their surfaces binds to the antigen-peptide-MHC molecules on the surfaces of DC's, and co-stimulatory molecules are sufficiently upregulated on the surface of the DCs to show the degree of danger signals.

2.3 Danger Theory (DT)

Matzinger [5] proposed the Danger Theory, which has become more popular among immunologists in recent years for the development of peripheral tolerance (tolerance to agents outside of the host). DT states that the immune system will only respond when damage is indicated and is actively suppressed otherwise. DT proposes that DCs have danger signal receptors (DSR) which recognize signals sent out by distressed or damaged cells. DCs are activated via the danger signals then are able to provide the necessary signals to the T-cells (more precisely, T-helper cells) which control the adaptive immune response.

2.4 Agent-Based Artificial Immune Systems

Agent is an entity that has the ability of consciousness, solving problem, self-learning and adapting to the environment, which is very similar to immune systems in functionalities. AIS can be contributed to the learning mechanism of agents. It may be used to optimize the agent's responses to intruders. Functionalities of the biological immune system, such as content addressable memory, adaptation, etc., can be utilized in intelligent agents. Three major stages for ABAIS

inspired by the clonal selection theory are diversity generation, self-maintenance and memory of nonself. The last two properties define the adaptiveness of the IBMAS. These steps are carried out by agents distributed over the MAS.

3 Intrusion Detection Mechanism Based on ABAIS

We will propose an improved IDS based on ABAIS which will first detect danger signals emitted by computer hosts. These signals are based on some security threat profile, which defines by system calls. Several agents are generated which can communicate each other to emulate functionalities extracted from DT-inspired AIS (Fig 1).

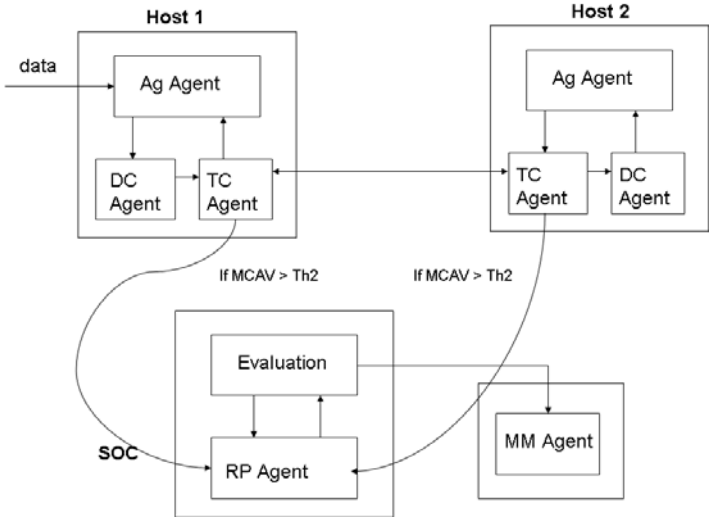


Fig. 1. Architecture of ABIDS

3.1 Antigen

An antigen is defined as an information vector extracted from network packet. For example, antigens are binary strings extracted from the IP packets, which include IP address, port number, protocol types, etc. However, the antigen defined at our ABIDS is related to system calls rather than network packets. Each system call is captured and converted into an antigen [11].

3.2 Intelligent Agents in AIS

We design a MAS with antigen agents, DC agents, T-Cell agents, Responding agents and Memory agents to perform functionalities of ABIDS such as DCAs.

Antigen Agent (Ag Agent). Antigen agents, which are installed at computer hosts, represent data item from the nonself dataset. They extract and record selected attributes from these data items. According to the agent-based DCA (ABDCA) which will be described later, each Ag agent sends corresponding DC agent a picked message when a nonself antigen appears.

Dendritic Cell agent (DC agent). DC agents are the kernel of the ABAIS which are installed and distributed at each computer hosts. When an Ag agent issues a picked signal, corresponding DC agent will evaluate the risk state facing by the host by calculating the danger value (DV). Similar to nature DC, each DC agent has three stages, namely immature, semimature and mature. DC agents are started from immature stages. When a picked signal issued from the Ag agent, DC agent executes data processing function such as the ABDCA. When a DC agent is at either semimature state or nature state, it returns the mature context to corresponding TC agent.

T-Cell agent (TC agent). TC agents are also installed at each computer hosts. They are activated by the signals from DC agents when the DVs exceed thresholds. Each TC agent have three numerical values associated with it; these represent the accumulated certainties and severities of attack: *T-cell activation threshold, Th1 activation threshold, Th2 activation threshold*. These TC agents will communicate each other for exchanging these numerical values. For TC agent issuing malicious act warning for corresponding antigen, it informs TC agents "nearby" by exchanging its numerical values. For TC agent not issuing malicious act warning for this antigen, it simply update its numerical values according "nearby" TC agents issuing malicious act warning.

Responding Agent (RP Agent). Ag agents, DC agents and TC agents are coordinating one another to perform malicious act responses. After DV exceeds threshold value, TC agents will inform RP agent, which is installed at SOC and each computer host. RP agents will activate some control measure to such malicious act.

3.3 Algorithm: Agent-Based Dendritic Cell Algorithm (ABDCA)

If an antigen is collected in an environment of danger, the context of this antigen is marked as anomalous and such antigen collected by the DC agent is potentially an intruder. We observe that multiple DCs present multiple copies of the same antigen type for invoking an immune response; it leads to an error-tolerance immune system as a single DC is far from stimulating a "false positive" error. The ABDCA is listed as follows.

Three temporary output signals are PAMP signal, danger signal and safe signal. The definitions of PAMP, danger and safe signals, by way of parameters of computer hosts such as CPU usage, memory load, network connection (number) and bandwidth saturation, are as follows.

Algorithm 1. Agent-based DCA (ABDCA)

```
input : Antigens and signals
output: Antigen context (0 for safe/1 for danger)
Initialize DC agents (Immature state);
for each DC_agent_i do
   get antigen (Ag);
   store antigen;
   get input_signal;
   calculate output_signal_i;
   if output_signal_i > Activation_Threshold_i then
      Ag_context is assigned as 0;
      State of DC_agent_i="semi-mature";
   else
      Ag_context is assigned as 1;
      state of DC_agent_i="mature";
   end if
end for
update cumulative output_signals;
```

1. PAMP Signal: Network Connection > th_netconnection AND bandwidth Saturation > th_bandsaturation.
2. Danger Signal: CPU Usage > th_cpuload OR Memory Load > th_memoryload.
3. Safe Signal: All parameters < th_parameter.

Where "th_parameter" represents the threshold value of the parameter. Now the computation of output signals by three temporary output values is the following.

$$output\_signal = \frac{W_P \cdot C_P + W_D \cdot C_D + W_S \cdot C_S}{W_P + W_D + W_S} \tag{1}$$

According to empirical experiments [6], the weights for this output signal is $W_P = 2$, $W_D = 1$, $W_S = 1$. C_P, C_D, C_S represent PAMP signal, danger signal and safe signal. The principle for this weights is the following: PAMP signal will decide whether the danger signal will be a "really" harmful one. Whether output signal is harmful or nor is defined by the Table 1.

3.4 Algorithm: Agent-Based IDS

Now according to ABDCA, we propose an algorithm describing agent-based IDS. ABDCA can provide information not only a network packet but also a group of

Table 1. Definitions of Output Signal

Output Signal	C_P	C_D	C_S
Normality	0	0	1
Harmless Abnormality	0	1	0
Harm Abnormality	1	1	0

Algorithm 2. Agent-based IDS (ABIDS)

Input: Antigen and Signals
Output: Antigen Types
1. Data Processing:
1.1 Ag agent extracts antigen from network traffic signal according to ag agent attribute.

2. Agent response:
2.1 DC agent returns context value according to threat profile; also according to the policy of the SOC if necessary.
2.2 DC agents will response to such antigen by updating its state according to DCA.

3. Danger Signal processing:
3.1
if this DC agent returns mature values to Ag agent **then**
Ag agent transfers it to the TC agent.
end if
3.2 TC agent categorizes such antigen according to the(updated)threat profile.

4. MCAV generation:
4.1 TC agent generates MCAV, which is sent to the RP agent.
4.2 RP agent determines if the corresponding antigen is malicious or not.
5. Antigen Memory:
5.1 Memory Agent stores necessary MCAV information

network packets is anomalous or not. This is achieved by the generation of an anomaly coefficient value, namely MCAV.

For the step 4 of the Algorithm 2, each nonself antigen gets a binary string of mature contexts from every selected DC agent installed at corresponding computer host. MCAV can be calculated through the number of context "1" divided by the number of all contexts. According to [2], it is similar to a voting system, where the antigen is the candidate and the DC agents are the voters. If the context is "1" ("0"), it means the DC agent determines this antigen is malicious (benign). The MCAV is actually the probability of that this antigen is being malicious.

One advantage of ABIDS is the communications and coordination between DC agents and TC agents. Useful knowledge obtained by DC agents and TC agents should be stored in some databases which could be shared by each DC and TC agent to improve their detection efficiencies. If a TC agent determines its host is under malicious attack by analyzing the corresponding MCAV, it will inform other TC agent "nearby". Such communication is direct and does not pass through proxy server or SOC. The activation threshold of this TC will be updated according to the following equation.

$$T(t) = T(t-1) + \alpha E(t-1)(E(t-1) - T(t-1)), t = 1, 2, \ldots, t_b \quad (2)$$

where E_i is the excitation level of TC agent, and t_b is the attacking time by this antigen.

4 Simulation

4.1 Malicious Behaviors Determined by ABDCA

One advantage of ABIDS is the following: it can determine some nonself-antigenic behavior which is at the verge of normality and abnormality. For example, for those malicious behaviors with short attacking time. The threshold of S, C and T are the following $S_{th} = 0.50$, $C_{th} = 0.50$ and $T_{th} = 0.5$. The number of computer hosts within this network is 2000. Each host has its fixed vector of weights.

Nonself Antigen with High Severity, Low Certainty and Short Attacking Time. While an antigen intrudes a computer host with behaviors of high severity but low certainty and short attacking time, it is difficult sometimes to determine if this is a malicious attack. In this case, we consider the threat profile TP=$[0.9, 0.01, 0.01]$ for the ABDCA algorithm.

We expect the simulation should be "stable" for reasonable many hosts as Fig. 2. MCAVs are around 0.03 and 0.05, which shows that this nonself antigen is normal.

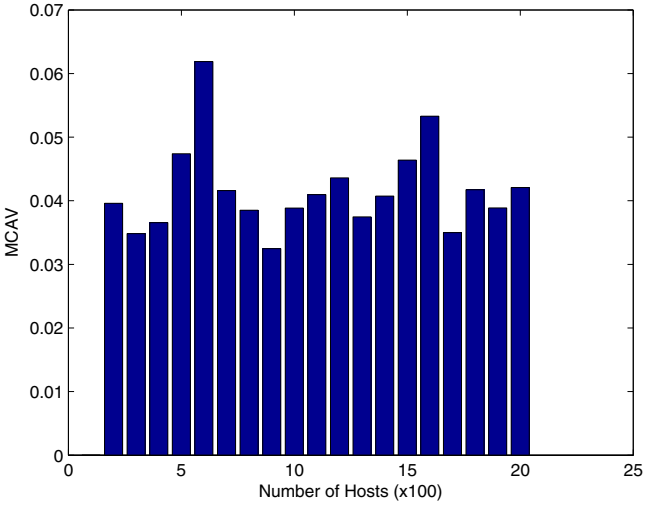


Fig. 2. Average MCAV for an Antigen with High Severity solely

4.2 Simulation of TC Agents' Coordinations

Interactions among Three TC Agents. We consider three computer hosts with TC agents interactions. TC1, TC2 and TC3 are informing each other for confirmed attack from the corresponding antigen. After TC agent communicate each other, each agent will update its activation threshold value according to equation (2). For simplicity, we assume that E is some "finitemixture" function. Initial activation value for each TC agent are all set to 0.01; parameter α for each TC agent are 0.90, 0.75 and 0.5, respectively. SImulation result is shown as Fig. 3. WHen the attacking behavior is at its peak around time 80, three TC agents adjust their activation threshold values simultaneously. When the attack is ceased, such threshold values are reduced and stable afterwards.

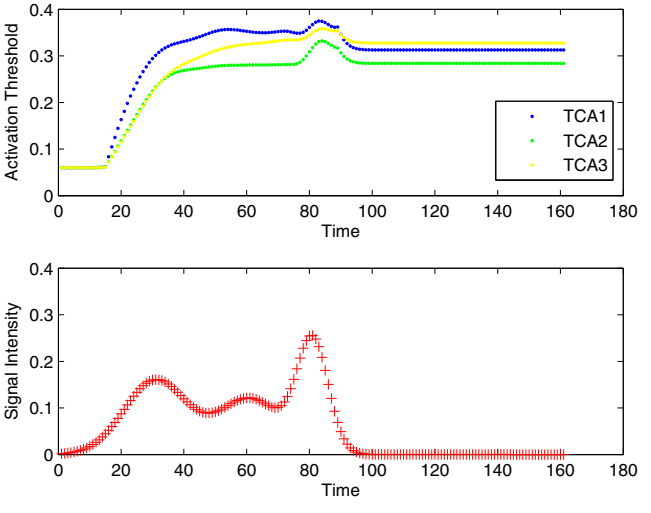


Fig. 3. Activation Thresholds of Three TC agents within the attacking cycle

5 Conclusions

We propose a multiagent-based intrusion detection system. The intelligence behind such system is based on the danger theory of human immune systems. In particular computations of danger values with dynamic thresholds will reduce the false positive rate of danger signals issued by computer hosts. Three agents, namely, Ag agent, DC agent and TC agents are coordinated to exchange information of intrusion detections. The evaluations of three factors S, C, and T are pragmatic issues. On the other hand, threshold values have to be defined by SOCs according to security profiles.

References

1. Forrest, S., Hofmeyr, A., Somayaji, T.: Longstaff: A sense of self for unix processes. In: Proceedinges of the 1996 IEEE Symposium on Research in Security and Privacy, pp. 120–128 (1996)
2. Gu, F., Aickelin, U., Greensmith, J.: An agent-based classification model. In: 9th European Agent Systems Summer School, EASSS 2007 (2007)
3. Aickelin, U., Bentley, P., Cayzer, S., Kim, J.: Danger theory: The link between AIS and IDS. In: Timmis, J., Bentley, P.J., Hart, E. (eds.) ICARIS 2003. LNCS, vol. 2787, pp. 144–165. Springer, Heidelberg (2003)
4. Yang, J., Liu, X., Li, T., Liang, G., Liu, S.: Distributed agents model for intrusion detection based on ais. Knowledge-Based Systems 22, 115–119 (2009)
5. Matzinger, P.: Tolarance, danger and the extended family. Annual Review in Immunology 12, 991–1045 (1994)
6. Greensmith, J., Feyereisl, J., Aickelin, U.: The dca: Some comparison. Evolutionary Intelligence 1(2), 85–112 (2008)
7. Greensmith, J., Aickelin, U., Cayzer, S.: Detecting danger: The dendritic cell algorithm. Robust Intelligent Systems 12, 89–112 (2008)
8. Zhang, J., Liang, Y.K.: Integrating innate and adaptive immunity for worm detection. In: Proceedings of the Second International Workshop on Knowledge Discovery and Data Mining, pp. 693–696 (2009)
9. Fu, H., Yuan, X., Wang, N.: Multi-agents artificial immune system (maais) inspired by danger theory for anomaly detection. In: 2007 International Conference on Computational Intelligence and Security Workshops, pp. 570–573 (2007)
10. Harmer, P., Williams, P., Gunsch, G., Lamont, G.: An artificial immune system architecture for computer security applications. IEEE Transactions on Evolutionary Computations 65(3), 252–280 (2002)
11. Greensmith, J., Aickelin, U., Tedesco, G.: Information fusion for anomaly detection with the dendritic cell algorithm. Information fusion 11(1), 21–34 (2010)

Secured Agent Platform for Wireless Sensor Networks

Jan Horacek and Frantisek Zboril jr.

Brno University of Technology, Faculty of Information Technology
Bozetechova 2, 612 66 Brno, Czech Republic
{ihoracek,zborilf}@fit.vutbr.cz
http://www.fit.vutbr.cz

Abstract. Wireless sensor networks (WSN) are designed for usage in a wide range of applications where large amount of sensor nodes do some observation for a quite long time period. Although today applications of WSN are based rather on active messages then on active mobile codes we study security issues when mobile codes (agents) are used there. Securing such codes then means to assure their integrity, confidentiality and also their proper interpretation. On the other hands the hosting devices must be also protected from misbehaviour of malicious agents. In this text we present how we are making agents for the WSN and also how we protect these agents against some possible risks.

Keywords: mobile agents, security, distributed networking, wireless sensor network, WSN, ALLL, microcontroller.

1 Introduction

Todays approaches of making implementations for WSN are mostly based on library or if you like operation system TinyOS (v.2). Program is then written in C's dialect named NesC that allows making modular event/command driven design of application. Behaviour of the whole distributed system is driven by some messages that triggers on some specific behaviour of the nodes. These messages are called "active messages" and de facto works as a facility for remote events invocation. Such approach is fine for WSN because of its energy efficiency. On the other hand, mobile codes or agents mean quite big communication overhead. We advocate our effort of making agents for systems like wireless sensor networks that there may be situation when temporary extension of mote's computational abilities may bring more value than the costs of energy spent. Also there are still some efforts in connecting agents with WSN in the research community. We mention systems like SensorWare [4], Agilla [6] or ActorNet [7] that tries to make agents as a tiny and easily interpretable codes, or MAPS [1] and AFME [9] where are proposed relatively small, but Java-based agents.

We discuss this more in [14]. Anyway we try to make the mobile agent code as tinny as possible. On the other hand we address some issues that would guarantee

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 476–485, 2011.

some important security objectives like integrity and confidentiality. These issues are charge of the section 2. After a brief introduction of our *WSageNt* system in section 3 we follow with description of devices that support security built-in systems in section 4. Our approach to protect agents and data in WSN is charge of section 5 and an outline how to locate possible attacker is discussed in section 6. Finally we conclude with some remarks about our future work.

2 Security Issues in WSN with Mobile Agents

Talking about security means mostly talking about CIA (confidentiality, integrity and accessibility) aspects of security. These three aspects are also the ones that should be guaranteed in the WSN systems. We follow known problems for mobile agents [13], which rise from possibilities that:

- A proper agent is endangered during transport from one agent platform to another.
- A proper agent is intentionally misinterpreted or corrupted by a malicious platform.
- The agent program is an attacker itself.

From this point of view in a secure WSN system we should guarantee CIA for communication and transport channels, agent platforms and also we should guarantee that no invalid agent will be accepted for any secure mote within the system.

Let us have an agent program P that is in its open form. So P is a string from an agent language. We see the start-point of agent's life in the system is when an agent program P is sent from a base-station to a first WSN node (mote) or a set of nodes. So now there must be a mechanism, which ensures that a reliable agent is sent in its correct form to a reliable and addressed node. When an agent appears at the destination node, then it must be ensured that the program P will be correctly interpreted, will not be improperly modified and also in some cases the agent program including agent's beliefs, goals and intention must be confidential. That means that two general threats must be eliminated. The first one is that an invalid agent platform is pretending correct interpretation of the code but it does not behave in correspondence to the program or the platform provides bad information from the mote's inputs (sensor, radio, memories, etc.) to the agent. The second threat is that an attacker loads the program from a stolen (potentially correct and reliable) mote and substitutes it with another program P'. This situation also causes that an attacker could obtain some data from the program that should be confidential.

Above mentioned are the risks, which we try to reduce when our agent is working inside the WSN. This text goes on with introduction of our *ALLL* agent language and the platform/interpreter system for small devices called *WSageNt*.

3 ALLL Agents and WSAgeNt System

3.1 ALLL Language

We start with brief introduction of *ALLL* language as it is currently implemented for the WSN. Closer look at the *ALLL* interpreter and the language itself can be found in [14]. For this paper we just briefly describe programming language and the concept of our interpretation and management agent platform.

Idea of the *ALLL* agents is that the control code or program is intended to be a sequence of actions structured in some hierarchy. Each agent also includes its belief and plan base, input buffer for agent's incoming messages, set of registers and a stack of plans that represents its current intention. All these parts are written in the *ALLL* code.

Particular actions should be of several kinds. An action could represent some work with agent's belief base (addition, deletion, testing of beliefs), it can represent extension of actual intention with a new plan, there can be an action that accepts a message from agent's input buffer or an action that sends a message to another agent. Finally there can be an action that executes some of platform services.

There is no specification of arithmetical or logical expressions, no iterations or complex logical operations for logical reasoning etc., because such computations are a part of platform services. For this reason the agent program can be tiny but expressive enough to determine agent's behaviour within the system.

3.2 WSageNt Platform

WSageNt platform is an agent platform that operates on WSN nodes. Currently supported network nodes are MicaZ and Iris motes and we are also able to run such platform in simulators like T-Mass [15]. The platform contains an interpreter of the *ALLL* language. Our *WSageNt* system is programmed in NesC programming language. We have tried to choose components of new secured system to be as similar to the original motes as possible to let us run there just slightly modified version of our original *WSageNt* platform. Following paragraphs describe basic operations and services that our platform provides:

- Message sending – basic TinyOS interface provides sending of messages, which can be 28 bytes long at most. We often need to send messages longer than this size is. Ideally we need to send messages of variable length. Message is then divided into packets that are sent to the destination node where the whole message is reassembled.
- Agent mobility – this service includes copying or moving agent from one node to another. It is based on message sending action, because agent code is usually longer than 28 bytes.
- Set of operations – this is a wide range set of actions. Our platform provides a set of operations for working with lists, some arithmetic, logical and relational operations and so on.

- Interface for reception – this mean sensing data from sensors. We have also implemented circular log area for storing such data so we can see history of measured data. We can also start process that periodically senses such data in some intervals and this is done assynchronously of interpretation of an agent. There are also some statistic operations on such sensing process like: "Give me the maximum value from a temperature sensor, which had been obtained during last 10 periods".
- Other services – as an example we can mention a neighbour nodes discovery service.
- Interpreter synchronisation – calling interpreter to do one more step when lower layers has finished previous operation. This section contains waiting for incoming messages and suspending interpreter for given amount of time.

For improvement of system security, especially confidentiality and integrity of agents and messages we had to modify message sending and agent mobility services. Section 5.3 discusses this in more details. We must also consider that we can trust only to our secured microcontrollers. Due to this fact we had to also modify interface for reception and sensed data history a little bit. That will be briefly discussed in section 5.2.

4 Secured Microcontrollers

We have actually designed two versions of secured WSN nodes. We are using AT90SC144144CT or AT90SO128 secured 8/16 bits microcontrollers for both solutions. Both of them are similar to AVR microcontrollers as is ATmega128L, which is used in modified form at MicaZ motes. They are secured against side channels such as light or temperature emission and they have also certification of EAL4+. This certification guarantees level of security, which is sufficient for industrial and banking area.

One of the main reasons why we have decided to use these microcontrollers is size of their RAM memory. There is also for example AT90SO4 that provides only 2kB of RAM that is not enough to run any regular agent program. MicaZ nodes provides 4kB of RAM memory and because the *WSageNt* system has been originally designed for such platform we have set 4kB as the minimum for decision what microcontroller we will use.

Other parameters remain the same. We will focus on AT90SO128 microcontroller in the following text. It can be powered with 2 AA batteries (as well as the MICAz or IRIS motes), it provides low power idle modes and there are two 16-bits timers too. Secured microcontroller has built-in random number generator and also DES, 3DES and AES hardware modules. Finally this microcontroller provides also a checksum accelerator.

On the other hand it provides less input/output ports then usual ATmega128L. Another disadvantage is higher energy consumption, which is important factor in the WSN systems. In full workload AT90SO128 consumes around 20mA while usual ATmega128L consumes just 6mA. This is significant increase of energy

consumption, but it is still acceptable for the WSN area. We notice that these values are achieved on full workload only and that AT90SO128 runs at higher frequency (26MHz against 8MHz) [3].

5 Secured Nodes in WSN

5.1 Architectures of Secure Nodes

Figure 1 demonstrates our first solution. AT90SO128 is connected to the MicaZ node through SPI or I2C bus. MicaZ AT microcontroller just redirect communication and provides interface to the sensors. We have identified some security threads of such solution. ATmega128L can be reprogrammed and used for an attack called man in the middle. For this reason every communication going through this MCU must be ciphered.

We must take into account that an attacker has also possibility to put his microcontroller between radio module and AT90SO128. All agent platforms and their data must be then stored in secured AT90SO128. Also usage of flash memory is disputable. We cannot trust data, which are stored in flash anymore. One of the main disadvantage of our first solution is its' high energy consumption, but we can use present sensor boards.

Our second solution is introduced on Figure 2. We have decided to develop our own WSN nodes. Such node consists of secured AT90SO128 MCU and Zigbee 802.15.4 radio module. We use I2C bus to connect such radio module. Particular sensors will be connected through this bus also.

Advantage of this solution is obvious. We have to power just the radio module, MCU and sensors. There is no additional flash module, no led diodes and so on. On the other hand, we are able to connect sensors only via I2C bus and also we will need to create new sensor boards.

As we said alreadywe will not study validity of sensed data in this paper, but we will talk about security risks of storing such data. We note that an agent system can profit from trustworth/reputation systems usually used in such area.

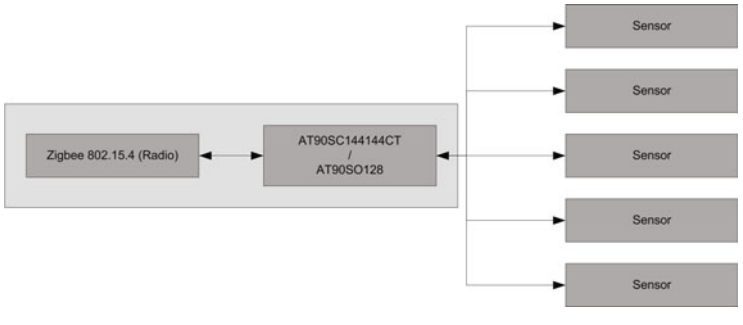


Fig. 1. First solution with usual MicaZ/Iris mote

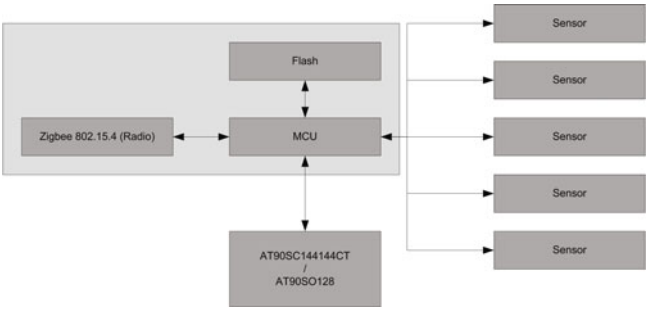


Fig. 2. Second solution with our own WSN node

5.2 Securing History of Sensed Data

As we wrote in section 3, our platform also provides an interface for messages reception. Basically history of sensed data is stored in flash memory of MicaZ node. Usage of additional flash memory raises new security risks. To reduce such risks we may take one of the following approaches.

First approach is to cipher all data stored in the flash memory. We will also need signatures of all data (or their parts) to be stored in our secured MCU.

Second stance is to think of data from history as a regular data from sensors. If an attacker can modify data from sensors, we have no need to try to keep them safe in the flash. We have to use additional control mechanisms for sensed data as well as history of data taken from the flash.

This paper does not contain description of such mechanisms, but it will be one of main efforts in our future work.

5.3 Securing an Agent Code vs. Securing Message

At the beginning we must ask a question if there are any differences between sending regular message and sending agent code from one node to another one. We can regard the agent code as a message, but we can also say that when we compare an agent code size with size of common message then we find that we must set two different ways of transferring such information.

We suppose that there must exists some input buffer for regular messages and it is designed with respect to the fact that its' content may be dropped. For example in a situation when CRC of a message does not match then we need to resend the whole message again. So we can say that dropping of content is usual operation. An agent can continue after such situation even when something wrong happens.

On the other hand agent code has much larger size and due to this fact only one agent may exist in a platform at one moment. We must be sure that we communicate with trustworthy node at the beginning of communication. There is no step back in replacing an agent code.

Figure 3 shows such pretest phase before we start sending encrypted data packets. Sender (B) starts transfer of an agent with passing header packet.

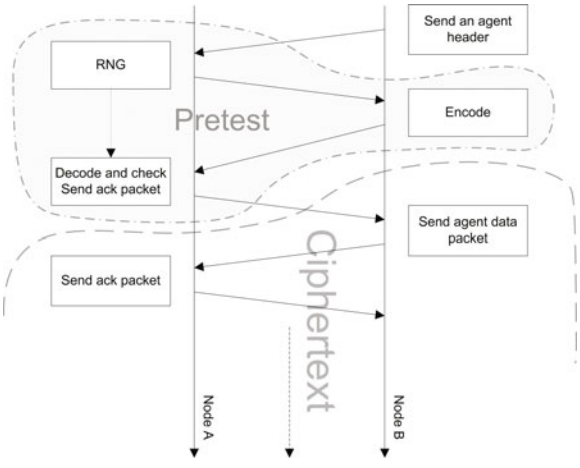


Fig. 3. Pretest phase for sending agent code

Receiving node (A) generates random number sequence and sends it to B. Side B encodes this sequence with a shared key and sends it back to side A. Encoded sequence is then decoded with the same shared key on side A and then its' content is compared with the original sequence. If both sequences match then an ack packet is send and communication can continue.

Communication then continues and all data packets are encrypted and decrypted with another shared key. Usage of two shared keys improves resistance against attack as we could use known pairs - open form and cipher form of random number sequence. Attacker could possibly break this part of communication more easily, but it will give him or her just an ability to start sending/receiving of ciphered data packets. In the worst case, he will be able to delete agents from nodes, but he will be not able to send (or receive and decode) any working agent. Node A now believes that it is communicating with friendly mote and permits incoming agent. Interpretation of possible existing agent is terminated and the agent is deleted. New agent is being interpreted after successful transfer.

As we have already said, sending data message may cause droping of the whole message. At this case we do not need such pretest phase. After sending a message header, node A just sends ack signal and node B can start sending ciphered data packets.

6 Finding an Attacker in the Network

One of the important issues is to find possible attacker in the network. Let us consider situation shown on Figure 4.

An attacker in the network wants to compromise our WSN. Main limitation of the attacker in the WSN is the fact, that he has (as any other WSN node)

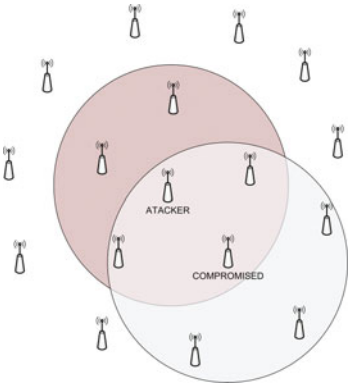


Fig. 4. An attacker is trying to interfere into communication

limited action radius. In this area he can be passive (sniff a communication, etc. . .) or active. Especially he can do these two main things:

- Jam a communication
- Modify a packet

We will consider the jamming problem firstly. When an attacker is jamming a communication, it will cause a problem for nodes to send a message to motes in attacker's action radius. We can detect it when sending mote does not receive ack signal from the receiver. Sending side starts counting these problems and when problems occur too many times per some interval, sending node evaluate such situation in the way that attacker is close to the target node. Sending node starts sending an alert to the whole network.

Let us describe the second situation. As we have mentioned already, the whole message is sent, decrypted and then its' CRC is checked. When CRC doesn't match it indicates that there is a problem in communication. Also when this problem occurs too many times, receiving node may suspect an attacker. But in this case node starts broadcast a warning alert to close-by nodes. This message says that node is compromised. This message will be dropped in action radius of attacker node. But action radius of attacker node does not cover whole radius of the compromised node. This message is then received with close-by node and this node finally sends an alert to the whole network.

Attacker can also send an alert to the network, but such behaviour goes against attackers' needs, because an attacker wants to stay undiscovered in the network. So we do not consider such problem.

We can continue on Figure 5. Atacker attempts of interfering with our network create a hole in the network. We can see that there are some nodes around such hole, which sends alert to the network.

We can also assume that we know a topology of our network. A lot of work about this topic has been already written. We can refer for example to [5]. This system operates with *WSageNt* platform and can create an agent that walks

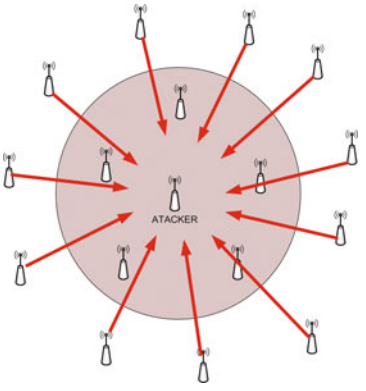


Fig. 5. Hole in the network that locate an attacker

through the network and returns back with database of neighbour nodes. When agents walk through whole network we can establish a network topology.

Now we have two main possibilities how to locate the attacker. We know that there are some attack alerts in network and we also know a location of hole that was created. At this moment an operating staff may send someone to check what has really happened or he or she may disconnect such part of network.

7 Conclusion

We introduced basic principles of securing of WSN system when mobile agents are used there. Main problems of these systems are integrity and confidentiality of agent's and messages when they are sent from one mote to another. Such secure architecture that we presented enables to hide such data and protect them against possible attacker. In our future work we intend to analyze energy consumption when unsecure and secure motes are used.

Acknowledgement. This work was partially supported by grant BUT FIT-10-S-1 and the research plan MSM0021630528.

References

1. Aiello, F., Fortino, G., Guerrieri, A., Gravina, R.: MAPS: A Mobile Agent Platform for WSNs based on Java Sun Spots. In: Proceedings of ATSM (2009)
2. Atmel. Embedded Security, http://www.atmel.com/dyn/products/devices.asp?family_id=700
3. Atmel. AT90SO128 Summary, http://www.atmel.com/dyn/resources/prod_documents/AT90SO128_Summary.pdf
4. Boulis, A., Han, C., Srivastava, M.B.: Design and Implementation of Framework for Efficient and Programmable Sensor Networks. In: Proceedings of the 1st International Conference of Mobile Systems, Applications and Services, pp. 187–200 (2003)

5. Gabor, M.: Web Interface For Wireless Network Monitoring. Master thesis, Brno, FIT VUT v Brně (2010)
6. Georgoulas, D., Blow, K.: In-Motes: Intelligent Agent Based Middleware for Wireless Sensor Networks. WSEAS on Communications Journal, 515–522 (2006)
7. Kwon, Y., Sundersh, S., Mechitov, K., Agha, G.: ActorNet: An Actor Platform for Wireless Sensor Networks. In: Proceedings of 5th AAMAS 2006, pp. 1297–1300 (2006)
8. Mitchell, C.: Trusted computing, Institution of Electrical Engineers, 313 (2005); ISBN 0863415253
9. Muldoon, C., O'Hare, G.M.P., Collier, R., O'Grady, M.J.: Agent Factory Micro Edition: A Framework for Ambient Applications. In: Alexandrov, V.N., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2006. LNCS, vol. 3993, pp. 727–734. Springer, Heidelberg (2006)
10. Pecho, P., Zboril, F., Drahansky, M., Hanacek, P.: Agent Platform for Wireless Sensor Network with Support for Cryptographic Protocols. Journal of Universal Computer Science 15 (2009)
11. Tarig, M.: Using Secure-Image Mechanism to Protect Mobile Agent against malicious Hosts, World Academy of Science Engineering and Technology (2009)
12. Wilhelm, U.G., Staamann, S.M., Buttyán, L.: A Pessimistic Approach to Trust in Mobile Agent Platforms, Ecole Polytechnique Fédérale de Lausanne (2000)
13. Yee, B.S.: A Sanctuary for Mobile Agents, Internet Programming. Springer, New York (1997)
14. Zboril, F., Horacek, J., Spacil, P.: Intelligent Agent Platform and Control Language for Wireless Sensor Networks. In: Proceedings of 3rd EMS, Atény, GR, p. 6. IEEE CS, Los Alamitos (2009); ISBN 978-0-7695-3886-0
15. Zboril, F., Zboril, F.V.: Simulation of Wireless Sensor Networks with Intelligent Nodes. In: 10th International Conference on Computer Modelling and Simulation, Cambridge, GB, p. 6. IEEE CS, Los Alamitos (2008), ISBN 0-7695-3114-8

Multiagent-System Oriented Models for Efficient Power System Topology Verification

Kazimierz Wilkosz and Zofia Kruczkiewicz

Wrocław University of Technology Wybrzeże Wyspiańskiego 27,
50-370 Wrocław, Poland
{Kazimierz.Wilkosz,Zofia.Kruczkiewicz}@pwr.wroc.pl

Abstract. The paper deals with the multiagent systems for the power system topology verification. Two of such multiagent systems are considered. The idea of topology verification utilized in each of them allows considering the topology verification of a whole power system as the set of many local verification processes. Each of the described multiagent systems performs the topology verification over a different protocol. Types of agent defined for these systems are different. For each of the considered multiagent systems, analysis and design are outlined using the Multiagent Systems Engineering. For their design models, the performance evaluation has been carried out. Then, a comparison of the investigated multiagent systems is made, paying special attention to their performance effectiveness.

Keywords: modeling, multiagent system, interaction protocol, simulation, power system topology verification, performance evaluation.

1 Introduction

The modern monitoring a Power System (PS) assumes automatic realization of real-time modeling of PS and first of all building a PS topology model (i.e. the PS connectivity model) and verification of this model. In the paper, the verification of the PS topology model with the use of the method described in [1] is considered. In [1], the idea is adopted, that Topology Verification (TV) for the whole PS is decomposed into many Local TVs (LTVs). LTV is performed for the nearest neighborhood of a substation. It was found that for the realization of the mentioned distributed PS TV the Agent Technology (AT) is very useful [2]. Three Multi-Agent Systems (MASs) for PS TV were considered in [3]. For one of them, that have been denoted in [3] as MAS-3, the maximal number of created messages is the lowest. MAS-3 is also considered in this paper as a reference system. Now, it is called as MAS-1. In the paper, we consider also other MAS called as MAS-2. In MAS-2, there are three types of agents, i.e. the agent *Dispatcher*, nodal agents and substation agents, whereas in MAS-1, we have the two first ones. The substation agent is associated with a substation with more than one electrical node. The nodal agent is associated exactly with one electrical node. The described difference between MAS-2 and MAS-1 allows obtaining more advantageous features of MAS-2 comparing with MAS-1.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 486–495, 2011.

In the paper, during the process of modeling and designing, we pay special attention to evaluation of performance of the analyzed MASs. The main goal of the paper is comparison of the mentioned systems from the view point of system performance. The performance evaluation of MAS-1 and MAS-2 is done using the performance experiments based on simulation of the test system. In the paper, as it was in [3], the maximal numbers of created messages, the mean times of LTV processes and the duration times of TV are considered.

It should be emphasized, that according to performance engineering of software systems [4], [5] when a new MAS is designed, performance requirements have to be considered for each phase of a life cycle. There are many papers, in which performance evaluation of different MASs is presented. This problem is considered in [6] for the systems: ZEUS, JADE, Skeleton Agents, in [7] for Aglets IBM, Concordia, Voyager.

Effectiveness of development of MAS can be enhanced using modern methodologies [8] and platforms. In this paper, the MaSE (MultiAgent Systems Engineering) methodology [9] are taken into account.

2 The Idea of the Power System Topology Verification with Use of Unbalance Indices and Artificial Neural Networks

The base of PS TV in the paper is the idea, which is described in [1]. This idea of TV assumes use of so-called unbalance indices, which are inputs for Artificial Neural Networks (ANNs) with radial basis functions. The unbalance indices are defined on the basis of relationships among measured quantities in PS. The mentioned measured quantities in PS are active and reactive power flows at ends of branches (power lines, transformers) and voltage magnitudes at (electrical) nodes in PS. Values of the unbalance indices create sets which are characteristic for different topology errors. The decisions on existence of topology errors are taken using outputs of the utilized ANNs.

One ANN is associated with one node in PS. The ANN allows taking a decision on correctness of modeling each of the branches, which are connected with the mentioned node. The measurement data, that are required in the decision process, are from an area, which contains as a central point the node with which the ANN is associated, all branches connected with this ANN and all adjacent nodes. Decisions regarding to correctness of modeling a particular branch are taken by two ANNs, which are associated with the terminal nodes of that branch. The final decisions is taken analyzing the mentioned partial decisions.

TV of a whole PS consists of many verifications of modeling particular branches. In this situation, it is possible to develop MAS for PS TV, as it was described in [2], [3]. In this system, there are agents associated with nodes (the node agents) and branches (the branch agents) of a power network. The goal of the node agent is performing LTV with the use of the unbalance indices and ANN. The goal of the branch agent is taking a final decision on correctness of modeling the branch on the base of two decisions taken by the node agents associated with terminal nodes of the branch.

3 General Description of the Considered Multiagent Systems

The IEEE 14-bus test system, which is shown in Fig. 1, is utilized in the process of performance evaluation of the considered MASs.

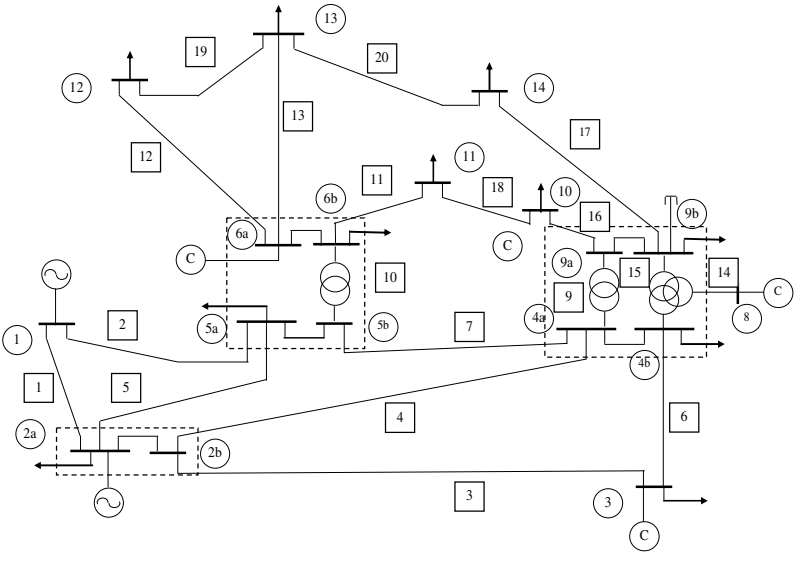


Fig. 1. The IEEE 14-bus test system

MAS-1, constructed for this system, is shown in Fig. 2a. In MAS-1 there are 19 nodal agents, labeled with A1-A19. Fig. 2b shows MAS-2 with the agents A1-A11. The test system includes three multi-node substations. These substations are served by the following substation agents: A2, A4 and A5. Other agents are the nodal agents.

The TV process begins when any symptom of occurrence of a Topology Error (TE) appears. As a symptom of occurrence of TE is considered at least one of the following events:

$$W_{Pi} \notin \left[-\delta_{WPi}, \delta_{WPi}\right], \quad W_{Qi} \notin \left[-\delta_{WQi}, \delta_{WQi}\right] \tag{1}$$

where: W_{Pi}, W_{Qi} - unbalance indices for the i-th node for active and reactive power, respectively [1]; δ_{WPi}, δ_{WQi} - positive constants.

At this moment, the nodal and substation agents perform LTV. They take decisions on correctness of modeling of the branches connected with their nodes. The nodal agents send the results of their LTVs to the *Dispatcher* agent. These are partial decisions on correctness of modeling branches which are connected with nodes in one-node substations when MAS-2 is considered. Such partial decisions are also decisions on correctness of modeling internal branches in multi-node substations in the case of MAS-1.

The substation agents, which are in MAS-2, take the partial TV decisions for all branches connected with the nodes served by these agents. The substation agent takes also the final TV decisions for branches connecting internal nodes of the substation. These agents do not take TV decisions for branches, of which one terminal node is in other substation. The substation agents send the results of their TVs to the agent *Dispatcher*. It collects received final decisions, and those which are base for taking final decisions on its level.

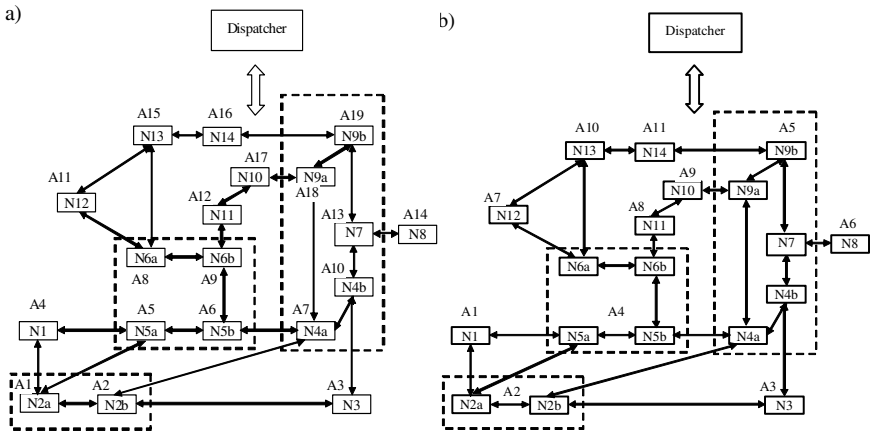


Fig. 2. The considered MASs for TV of the IEEE 14-bus test system: a) MAS-1 with nodal agents, b) MAS-2 with nodal- and substation agents

For both MASs, the agent *Dispatcher* takes the final decisions for branches which are among different substations. The mentioned substations can be one-node substations or multi-node substations. Additionally, for MAS-1, the agent *Dispatcher* takes the final decisions for branches which are internal branches in multi-node substations.

4 The Analysis Model of the Multiagent System for PS TV

The analysis model of the considered MASs is built by using the AgentTool_1.8.3 tool of the MaSE technology [9]. In this model, one distinguishes goals, roles and tasks.

The main goal of the analyzed MASs is PS TV (Fig. 3). The subgoals of MASs are: management of TV and agents associated with nodes and substations of PS (the goal 1.1), executing the process of TV (the goal 1.2) and executing the process of measuring distinguished quantities in PS (the goal 1.3).

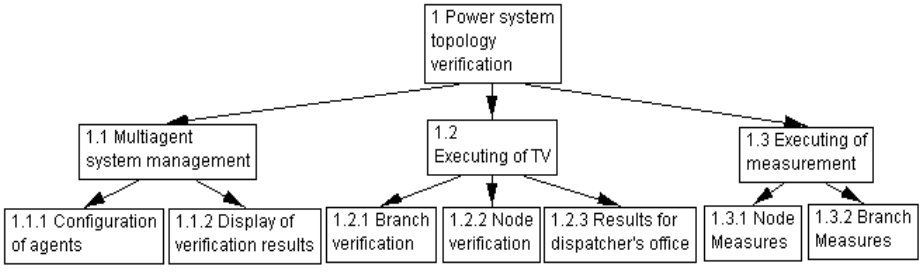


Fig. 3. Goals diagram of MAS for PS TV

The considered systems include one agent which plays the *Dispatcher* role (a rectangle). Each of other agents plays one instance of the roles *Node* or together *Node* and *Substation* (Fig. 4). The *Node* role (Fig. 4) fulfils the following subgoals: 1.2.2, 1.2.3, 1.3.1, 1.3.2, 1.3 (Fig. 3). For the *SubStation* role there is only one subgoal 1.2.1 (Fig. 3). Other goals (Fig. 3) are secured by the *Dispatcher* role. Each role performs a few tasks (ellipses). Each task is modeled as the statecharts diagram. Using tasks, the agents of roles exchange messages with each other according to the suitable external protocols (solid lines). Internal protocols (dashed lines) are used when tasks of the same role exchange messages. MAS-1 includes two types of agents: the first one playing the *Dispatcher* role and the second one such as the nodal agent playing the *Node* role for each node of PS (Fig. 6). Additionally, MAS-2 contains the substation agents, playing the *Substation* and the *Node* roles (Fig. 6). In the last case, the *Node* roles represent particular nodes of the multi-node substations of PS.

The sequences of messages exchanged by tasks are used to fulfill goals (Fig. 3) of roles (Fig. 4). Fig. 5 presents such sequences of the one instance of the *Dispatcher* role and the four instances of the *Node* role for MAS-1 and additionally the one instance of the *Substation* role for MAS-2. The labels of arrows represent the messages (Fig. 5) exchanged between the tasks of the roles with the use of the external protocols (Fig. 4).

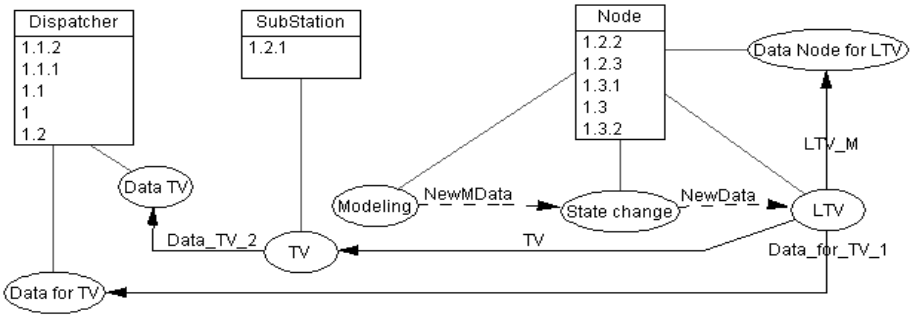


Fig. 4. The role diagram of the experimental MAS-1 and MAS-2 for realization of the TV process

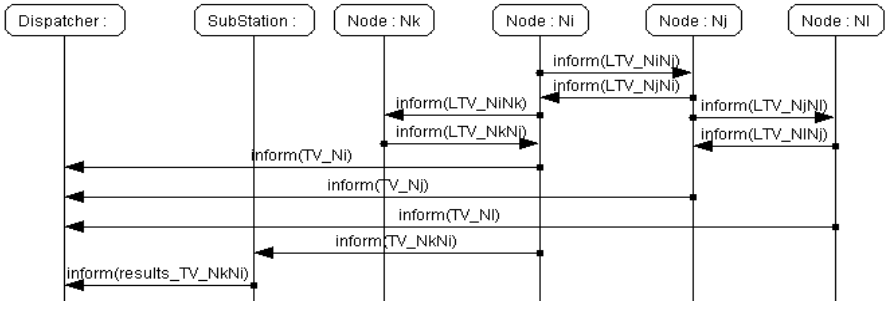


Fig. 5. The sequences of the messages exchanged by the tasks of MAS-1 and MAS-2

Further, it is assumed that the considered power network has n nodes, m branches, and the number of branches, which are connected with the considered node, is k. Additionally, from the view point of MAS-2 (with the nodal and substation agents) the m branches divide into branches of two kinds. The branches of the first one connect nodes belonging to different agents playing the *Substation* or *Node* roles. Let's assume that the number of such branches is equal to a. The branches of the second one connect nodes belonging to the agents of the *Substation* role and they connect nodes inside the multi-node substation. The number of the branches, which are considered now, is equal to b. The number of agents is equal to d. In particular, in the MAS-1 system where there are only agents of the *Node* roles, d is equal to n.

In MAS-1, independent decisions regarding correctness of modeling of branches are taken by the agents of the *Node* role when any symptom of occurrence of TE is detected (Section 3) in the *LTV* task. Additionally, in MAS-2 the *Node* and *Substation* roles are fulfilled by the substation agent. The *Substation* role executes TV for branches, which are inside one substation.

The detection of symptoms of topology errors is based on the testing the nodal unbalance indices, which is done by all the *State change* tasks. The nodal unbalance indices are calculated using measurement data of active and reactive power flows at the ends of branches. All the *Modeling* tasks build the model of a PS topology. The *Modeling* tasks internally send their data to the *State change* tasks using the internal *NewMData* protocol. After these actions the *State change* tasks internally send their data by the internal *NewData* protocol to the *LTV* tasks for LTVs.

LTV uses the external *LTV_M* protocol (Fig. 4). During the LTV process the total number of messages exchanged among each *LTV* and the *Data Node for LTV* tasks of neighboring nodes belonging to the different agents (the nodal agents or also the substation agents) is equal to $2a$ or $2(m - b)$ (each one of the $2r$-units size). In other words, this number is equal to the sum of the numbers of: (i) all the *inform(LTV_NjNl)* and *inform(LTV_NlNj)* and similarly *inform(LTV_NiNj)* and *inform(LTV_NjNi), inform(LTV_NiNk)* and *inform(LTV_NkNi)* messages (Fig. 5) exchanged among different agents of the *Node* role in the case of MAS-1; (ii) all the *inform(LTV_NjNl)* and *inform(LTV_NlNj)* and *inform(LTV_NiNj)* and *inform(LTV_NjNi)* messages exchanged among agents of the *Node* and *Substation* roles in the case of MAS-2. It is assumed that instances of roles *Nj* and *Nl* represent the nodal agents, whereas instances of roles *Ni* and *Nk* are fulfilled by substation agents. The total size of sent messages is as follows: $FM_{LTV\_MAS\text{-}1} = 4mr$, in the case of MAS-1 and $FM_{LTV\_MAS\text{-}2} = 4ar$ or $FM_{LTV\_MAS\text{-}2} = 4(m\text{-}b)r$ in the case of MAS-2. It should be underlined that in MAS-1 there are only nodal agents and in MAS-2 we have the nodal and substation agents.

The TV process is carried out after the appropriate LTV processes finish. In MAS-1, the *LTV* tasks (Fig. 4) send at most n messages (of the kr-units size), having the total size $FM_{TV\_MAS\text{-}1} = 2m\,r$. The considered messages are the *inform(TV_Ni), inform(TV_Nj), inform(TV_Nl)* messages (Fig. 5) for the *Data for TV* task of the agent of the *Dispatcher* role (Fig. 4). If only some of agents of the *Node* roles detect symptoms of occurrence of TEs, they send the *inform(TV_Nj)* messages to the agent of the *Dispatcher* role. The size of each such a message is equal to the number kr. The number of the *inform(TV_Nl)* and *inform(TV_Ni)* messages is equal to the number ofbranches connected with nodes with symptoms of occurrence of TEs. For each of such branches at least for one terminal node the mentioned symptom is detected. The size of each of these messages is equal to r.

In the case of MAS-2 as it is in the case of MAS-1, the nodal agents send messages with LTV decisions to the agent of the *Dispatcher* role. More complex situation is with respect to the substation agents. The complete TV process for the set of b branches, which are inside substations, is carried out by the *Substation* role of the substation agents. The final TV decisions for the mentioned branches are taken by these agents instead of the agent of the *Dispatcher* role. During performing the decision process by a substation agent, the *inform(TV_NkNi)* message containing the LTV information from the *LTV* task is sent to the *TV* task. Then, the *TV* task sends the *inform(results_TV_NkNi)* message with a final TV decision to the *Data TV* task of the *Dispatcher* agent. The substation agent does not take final decisions for the branches which go out from its substation. For such branches only LTV decisions are taken and messages with these decisions are sent to the *Data for TV* task of the *Dispatcher* agent. The total size of messages sent by the nodal and substation agents to the *Dispatcher* agent is $FM_{TV\_MAS\text{-}2} = (2a+b)\ r$ or $FM_{TV\_MAS\text{-}2} = (2m\text{-}b)\ r$.

Summarizing, when all nodes identify symptoms of occurrence of TEs, the total size of messages of the complete TV process is following: $FM_{LTV+TV\_MAS\text{-}1} = 6m\ r$ for MAS-1 and $FM_{LTV+TV\_MAS\text{-}2} = (6a+b)\ r$ or $FM_{LTV+TV\_MAS\text{-}2} = (6m\text{-}5b)\ r$ for MAS-2.

The result from the presented analysis is that MAS-2 with substation agents generates smaller network traffic than MAS-1 with only the nodal agents. Measure of the improvement is the number $5b$, where b is the internal branches of multi-node substation.

5 Design of Multiagent Systems for PS TV

After the analysis model is worked out, the design model of MAS-1 and MAS-2 are created [9]. The design models are made as the *Agent Template diagram* (Fig. 6). The *Agent Template diagram* is mapped from the analysis model. This diagram shows the *Agent Communication Language* (ACL) messages [10] exchanged between agents. The *Agent Template Diagram* for MAS-1 and MAS-2 is depicted in Fig. 6. In Fig. 6, each instance of the *Node_Agent1* and *Node_Agen2* agents represents the nodes of PS which are connected with each other. The instances of the *SubS_Agent1* and *SubS_Agent2* agents represent substations of PS. The *Dispatcher* agent manages the whole PS TV.

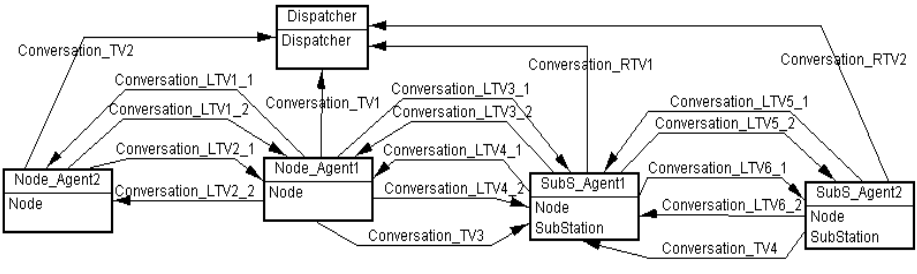


Fig. 6. The *Agent Template Diagram* for MAS-1 and MAS-2. MAS-1 contains the *Dispatcher*, the *Node_Agent1* and *Node_Agent2* agents. MAS-2 contains agents of MAS-1, and *SubS_Agent1* and *SubS_Agent2* agents.

In MAS-1 and MAS-2, LTV is performed, whenever symptoms of occurrence of TEs are detected. Independent decisions of the *Node_Agent1* and *Node_Agent2* agents initiate LTVs and they send the results of LTVs to the *Dispatcher* agent which executes TV for appropriate branches. Additionally, in MAS-2 the *SubS_Agent1* and *SubS_Agent2* agents execute TV (Section 4) for branches, which are internal for substations represented by the considered agents, and the final TV results are sent to the *Dispatcher* agent.

6 Idea of Performance Experiments

The aim of the performance experiments was to evaluate maximal total size and also maximal total time of transfer for all messages connected with LTV processes and maximal total size and maximal total time of transfer for all messages connected with taking final TV decisions when decisions from the LTV processes are known. The mentioned parameters were evaluated for both the considered MASs.

The presented performance experiments have been done, using the IEEE 14-bus test system (as in [1]), shown in Fig. 1. The parameters of the test system are as follows: $n = 19$, $m = 25$, $a = 16$, $b = 9$ The start time of TV is a time instant, when a symptom of occurrence of TEs is detected, and the finish time of TV is a time instant of taking the last TV decision on the dispatcher level, being an effect of the mentioned symptom.

During the investigation it was assumed that: a) all branches are actually in operation, b) all measurement data are burdened with Gaussian noise [1]; c) the number of changes of measurement data utilized by MAS-1 and MAS-2 is such that all agents are activated, i.e. TV is performed for the whole system, d) a number of considered TV cases is equal to 1000. The computer used in experiments has the following parameters: Intel(R)Core(TM)2 Duo CPU T7500 @2.20 GHz, RAM 3072 MB, 32-bits Windows Vista Business. Table 1 shows the maximal total size of messages in a whole TV process and in its stages for MAS-1 and MAS-2 (Fig. 2). In Table 1 (also in Table 2), *X* means MAS-1 or MAS-2. Table 2 presents maximal total time required for transfer of all messages in a whole TV process and in its particular stages for both the considered MASs. The results of calculations given in Table 2 correspond to the numbers of messages exchanged among agents in the considered MASs.

Table 1. The maximal total size of all messages in a whole TV process and in its particular stages for the considered MASs

X	FM_{LTV_X}	$\frac{FM_{LTV\_X}}{FM_{LTV\_MAS-1}}$ %	FM_{TV_X}	$\frac{FM_{TV\_X}}{FM_{TV\_MAS-1}}$ %	FM_{LTV+TV_X}	$\frac{FM_{LTV+TV\_X}}{FM_{LTV+TV\_MAS-1}}$ %
MAS-1	100 *r*	100.0	50 *r*	100.0	150 *r*	100.0
MAS-2	64 *r*	64.0	41 *r*	82.0	105 *r*	70.0

Table 2. Maximal total time of transfer of all messages in a whole TV process and in its particular stages for the considered MASs

X	MT_{LTV_X} ms	$\frac{MT_{LTV\_X}}{MT_{LTV\_MAS-1}}$ %	MT_{TV_X} ms	$\frac{MT_{TV\_X}}{MT_{TV\_MAS-1}}$ %	MT_{LTV+TV_X} ms	$\frac{MT_{LTV+TV\_X}}{MT_{LTV+TV\_MAS-1}}$ %
MAS-1	562	100.0	229	100.0	791	100.0
MAS-2	325	57.8	201	87.8	526	66.5
MT_{LTV_X}, MT_{TV_X}, MT_{LTV+TV_X} - maximal total times of transfer of all messages in LTV processes, of all messages with LTV decisions sent to the *Dispatcher* agent and of all messages in a whole TV process, respectively.						

The savings in the network traffic in MAS-2 are achieved by the use of substation agents apart from the nodal agents. Such a solution can reduce the number of exchanged messages with the unbalance indices needed for LTVs and also the number of messages sent to the *Dispatcher* agent which, in the considered case, gathers: (i) final TV decisions, regarding inner branches of substations, (ii) LTV decisions being base for taking final TV decisions for branches among different substations. The mentioned fact means decreasing of network traffic. A further consequence of the considered fact is decreasing the duration time of TV.

7 Conclusion

In the paper, two practical MASs, which are based on the idea of TV from [1], are presented. Differences between these MASs are a result of different types of utilized agents as well as different organization of realization of TV for a whole PS (a system architecture). In MAS-1, there are the *Dispatcher* agent and the nodal agents. In MAS-2, additionally, there are the substation agents. In MAS-1, all the final decisions are taken by the *Dispatcher* agent whereas in MAS-2, the part of the final decisions is taken by the substation agents. The consequence of the introducing the substation agents in MAS-2 is decreasing a number of messages exchanged among agents. Therefore for MAS-2, the time of performance of TV is shorter.

The considerations in the paper show that appropriate definition of agents and collaboration among them can significantly reduce communication complexity and have positive impact on time of performance of assumed tasks.

References

1. Lukomski, R., Wilkosz, K.: Method for Power System Topology Verification with Use of Radial Basis Function Networks. In: Sandoval, F., Prieto, A., Cabestany, J., Graña, M. (eds.) IWANN 2007. LNCS, vol. 4507, pp. 862–869. Springer, Heidelberg (2007)
2. Wilkosz, K.: A Multi-Agent System Approach to Power System Topology Verification. In: Yin, H., Tino, P., Corchado, E., Byrne, W., Yao, X. (eds.) IDEAL 2007. LNCS, vol. 4881, pp. 970–979. Springer, Heidelberg (2007)

3. Wilkosz, K., Kruczkiewicz, Z., Rojek, T.: Multiagent Systems for Power System Topology Verification. In: Corchado, E., Yin, H. (eds.) IDEAL 2009. LNCS, vol. 5788, pp. 815–822. Springer, Heidelberg (2009)
4. Smith, C.U., Lloyd, G.W.: Performance Solutions, A Practical Guide to Creating Responsive, Scalable Software. Addison - Wesley, Canada (2002)
5. Babczyński, T., Kruczkiewicz, Z., Magott, J.: Performance Analysis of Multiagent Industrial System. In: Klusch, M., Ossowski, S., Kashyap, V., Unland, R. (eds.) CIA 2004. LNCS (LNAI), vol. 3191, pp. 242–256. Springer, Heidelberg (2004)
6. Camacho, D., Aler, R., Castro, C., Molina, J.M.: Performance Evaluation of ZEUS, JADE, and SkeletonAgent Frameworks. In: 2002 IEEE International Conference on Systems, Man, and Cybernetics, vol. 4, p. 6 (2002)
7. Dikaiakos, M., Kyriakou, M., Samaras, G.: Performance Evaluation of Mobile-Agent Middleware: A Hierachical Approach. In: Picco, G.P. (ed.) MA 2001. LNCS, vol. 2240, pp. 244–259. Springer, Heidelberg (2001)
8. Wooldridge, M., Rao, A. (eds.): Foundations of Rational Agency. Kluwer Academic Publishers, The Netherlands (1999)
9. Deloach, S.A.: The MaSE Methodology. In: Bergenti, F., Gleizes, M.-P., Zambonelli, F. (eds.) Methodologies and Software Engineering for Agent Systems. The Agent-Oriented Software Engineering Handbook Series: Multiagent Systems, Artificial Societes and Simulated Organizations, vol. 11. Kluwer Academic Publishing, Dordrecht (2004)
10. Specification of FIPA, `http://www.fipa.org/specs/`

Intelligent Safety Verification for Multi-car Elevator System Based on EVALPSN

Kazumi Nakamatsu[1], Toshiaki Imai[2], and Haruhiko Nishimura[2]

[1] School of Human Science and Environment, University of Hyogo, Himeji, Japan
nakamatu@shse.u-hyogo.ac.jp
[2] Graduate School of Applied Informatics, University of Hyogo, Kobe, Japan
{aa09t404,haru}@ai.u-hyogo.ac.jp

Abstract. In this paper, we introduce the basic idea of a multi-car elevator safety verification based on an paraconsistent annotated logic program called EVALPSN with taking a three-car/shaft elevator system as an example for the safety verification. The EVALPSN based multi-car elevator safety verification system is constructed by formalizing the system safety properties to be secured in EVALPSN.

Keywords: multi-car elevator, defeasible deontic reasoning, logical verification, annotated logic program, EVALPSN(Extended Vector Annotated Logic Program with Strong Negation).

1 Introduction

We have already developed a paraconsistent annotated logic program called Extended Vector Annotated Logic Program with Strong Negation(abbr. EVALPSN)[3,9] which can be applied to various real-time intelligent control and safety verification systems such as pipeline process safety verification [6,7]. The features and advantages of EVALPSN based safety verification are: (1) since EVALPSN can deal with deontic notions such as forbiddance, safety properties in deontic expression can be easily translated into EVALPSN; (2) logical safety verification for various kinds of control can be easily carried out as logic programming; (3) since some restricted fragment of EVALPSN can be implemented on microchips as electronic circuits, the EVALPSN safety verification is suitable for real-time control.

Recently, multi-car elevator systems in which more than one car are operated in the same shaft have been practically implemented and started their services in some countries. However in order to avoid car crash in the same shaft assuring the safety for multi-car elevator control is strongly required. Our work is aiming to provide multi-car elevator control systems based on EVALPSN safety verification, which includes two phases, one is verifying the safety for car operation and another one is selecting the most appropriate car operation among more than one secured car operations. As the first step to our goal, we have already introduce the idea of single car elevator control based on EVALPSN safety verification [10]. In this paper, we extend the idea of EVALPSN based single-car

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 496–505, 2011.

elevator safety verification to three-car elevator safety verification as the second step to our goal.

This paper is organized in the following mannar: first, EVALPSN is reviewed briefly, and the outline of the multi-car (three cars/shaft) elevator system and its safety properties are introduced; next, the safety properties are translated into an EVALPSN and simple examples of the EVALPSN based safety verification are provided; last, the future development for EVALPSN based multi-car elevator control is introduced in the conclusion.

2 EVALPSN

We review EVALPSN briefly[3]. Generally, a truth value called an *annotation* is explicitly attached to each literal in annotated logic programs[1]. For example, let p be a literal, μ an annotation, then $p:\mu$ is called an *annotated literal*. The set of annotations constitutes a complete lattice. An annotation in EVALPSN has a form of $[(i,j),\mu]$ called an *extended vector annotation*. The first component (i,j) is called a *vector annotation* and the set of vector annotations constitutes the complete lattice,

$$\mathcal{T}_v(n) = \{\ (x,y) | 0 \leq x \leq n, 0 \leq y \leq n, x, y \text{ and } n \text{ are integers } \}$$

in Figure 1. The ordering($\preceq_v$) of $\mathcal{T}_v(n)$ is defined as : let (x_1,y_1), $(x_2,y_2) \in \mathcal{T}_v(n)$,

$$(x_1,y_1) \preceq_v (x_2,y_2) \ \text{ iff } \ x_1 \leq x_2 \ \text{ and } \ y_1 \leq y_2.$$

For each extended vector annotated literal $p:[(i,j),\mu]$, the integer i denotes the amount of positive information to support the literal p and the integer j denotes that of negative one. The second component μ is an index of fact and deontic notions such as obligation, and the set of the second components constitutes the complete lattice,

$$\mathcal{T}_d = \{\bot, \alpha, \beta, \gamma, *_1, *_2, *_3, \top\}.$$

The ordering($\preceq_d$) of $\mathcal{T}_d$ is described by the Hasse's diagram in Figure 1. The intuitive meaning of each member of $\mathcal{T}_d$ is $\bot$(unknown), α(fact), β(obligation), γ(non-obligation), $*_1$(fact and obligation), $*_2$(obligation and non-obligation),

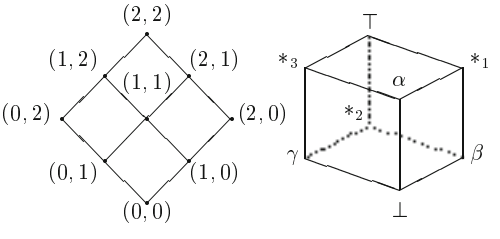


Fig. 1. Lattice $\mathcal{T}_v(2)$ and Lattice $\mathcal{T}_d$

$*_3$(fact and non-obligation), and $\top$(inconsistency). Then the complete lattice $\mathcal{T}_e(n)$ of extended vector annotations is defined as the product $\mathcal{T}_v(n) \times \mathcal{T}_d$. The ordering($\preceq_e$) of $\mathcal{T}_e(n)$ is defined as : let $[(i_1, j_1), \mu_1]$ and $[(i_2, j_2), \mu_2] \in \mathcal{T}_e$,

$$[(i_1, j_1), \mu_1] \preceq_e [(i_2, j_2), \mu_2] \quad \text{iff} \quad (i_1, j_1) \preceq_v (i_2, j_2) \ \text{ and } \ \mu_1 \preceq_d \mu_2.$$

There are two kinds of *epistemic negation* ($\neg_1$ and $\neg_2$) in EVALPSN, both of which are defined as mappings over $\mathcal{T}_v(n)$ and $\mathcal{T}_d$, respectively.

Definition 1. (epistemic negations $\neg_1$ and $\neg_2$ in EVALPSN)

$$\begin{aligned}
&\neg_1([(i,j),\mu]) = [(j,i),\mu], \quad \forall \mu \in \mathcal{T}_d,\\
&\neg_2([(i,j),\bot]) = [(i,j),\bot], \quad \neg_2([(i,j),\alpha]) = [(i,j),\alpha],\\
&\neg_2([(i,j),\beta]) = [(i,j),\gamma], \quad \neg_2([(i,j),\gamma]) = [(i,j),\beta],\\
&\neg_2([(i,j),*_1]) = [(i,j),*_3], \quad \neg_2([(i,j),*_2]) = [(i,j),*_2],\\
&\neg_2([(i,j),*_3]) = [(i,j),*_1], \quad \neg_2([(i,j),\top]) = [(i,j),\top].
\end{aligned}$$

If we regard the epistemic negations as syntactical operations, the epistemic negations followed by literals can be eliminated by the syntactical operations. For example, $\neg_1(p\!:\![(2,0),\alpha]) = p\!:\![(0,2),\alpha]$ and $\neg_2(q\!:\![(1,0),\beta]) = p\!:\![(1,0),\gamma]$.

There is another negation called *strong negation* ($\sim$) in EVALPSN, and it is treated as well as classical negation.

Definition 2. (strong negation $\sim$) [2]
Let F be any formula and $\neg$ be $\neg_1$ or $\neg_2$.

$$\sim F =_{def} F \rightarrow ((F \rightarrow F) \wedge \neg(F \rightarrow F)).$$

Definition 3. (well extended vector annotated literal)

Let p be a literal.

$$p\!:\![(i,0),\mu] \quad \text{and} \quad p\!:\![(0,j),\mu]$$

are called *weva(well extended vector annotated)-literals*, where $i, j \in \{1, 2, \cdots, n\}$, and $\mu \in \{\ \alpha,\ \beta,\ \gamma\ \}$.

Defintion 4. (EVALPSN)
If $L_0, \cdots, L_n$ are weva-literals,

$$L_1 \wedge \cdots \wedge L_i \wedge \sim L_{i+1} \wedge \cdots \wedge \sim L_n \rightarrow L_0$$

is called an *EVALPSN clause*. An *EVALPSN* is a finite set of EVALPSN clauses.

We note that if the annotations α and β represent fact and obligation, notions "fact", "obligation", "forbiddance" and "permission" can be represented by extended vector annotations, $[(m,0),\alpha]$, $[(m,0),\beta]$, $[(0,m),\beta]$, and $[(0,m),\gamma]$, respectively in EVALPSN, where m is a positive integer.

3 Multi-car Elevator Safety Verification

In order to verify the safety of the multi-car elevator system we introduce the same logical safety verification method for railway interlocking and pipeline valve control introduced in [4,6].

3.1 Multi-car Elevator System

For simplicity, we suppose that our multi-car elevator system physically consists of three cars c_i, c_j and c_k in the same shaft with thirty passenger floors(1F - 30F) fl_1, fl_2, $\cdots$, fl_{30} and two underground garage floors(B1F and B2F) fl_{B1} and fl_{B2}. We introduce a logical entity called a *route* from one floor to another floor, which can be represented by just its origin and destination floors, for example, "the route from floor fl_2 to floor fl_5". Routes are designated by arrows in Figure 2. In our multi-car elevator system a passenger is supposed to push the destination floor button, which is called a *passenger request* and is logically translated into a route called a *floor call*. We use some special names to identify floors. A floor where a passenger request is issued is called the *call floor* of the passenger request(floor call), and the destination of the passenger request is called its *destination floor*, which are usually represented by symbols fl_c and fl_d. Moreover a floor occupied by car c_x $(x = i, j, k)$ is called the *current floor* of the car, which is designated by fl_x corresponding to car c_x $(x = i, j, k)$. A route from the current floor of a car to the call floor of the floor call is called

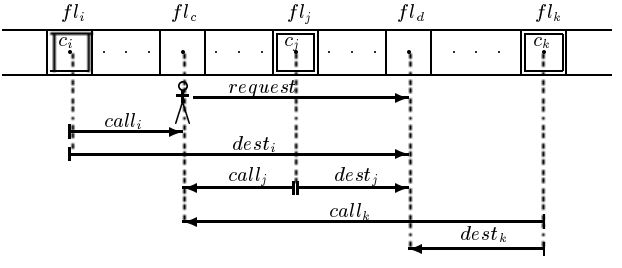


Fig. 2. Multi-car Elevator System

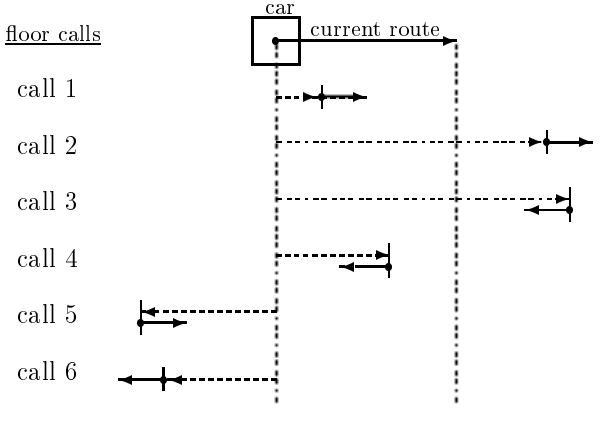


Fig. 3. Six Kinds of Floor Calls

the *call floor route* of the car, and a route from the current floor of a car to the destination floor of the floor call(passenger request) is called the *destination floor route* of the car. Car c_x call and destination floor routes are designated by $call_x$ and $dest_x$ $(x = i, j, k)$ in Figure 2, respectively.

Here we introduce the basic elevator service policy of our multi-car elevator system. Floor calls can be classified into six kinds, **call 1** $\cdots$ **call 6**, in accordance with the relations of floor calls(solid arrows), car current floors and call floor routes(dotted arrows), which are shown in Figure 3. Now we consider which kinds of floor calls should be processed. In order to process the call floor route of floor calls **call 5** and **call 6** the car should quit the current service and turn, and similarly in order to process the floor call **call 4** (passenger request route) the car also should quit the current service and turn after picking up the passenger. Since reacting to those kinds of floor calls seems to cause serious inconvenience for other passengers, we assume that: three floor calls **call 4**, **call 5** and **call 6** are forbidden from being processed in our elevator system.

3.2 Safety Verification Based on EVALPSN

Safety verification is one of the most crucial issues in various control systems such as railway interlocking, pipeline valve control, process control and so on. Some EVALPSN based safety verification systems have been introduced in [4,10,5].

In our multi-car elevator safety verification system, if a passenger request is issued, in order to avoid car crash if both the call floor route and the destination floor route can be secured(set safely) or not is verified by EVALPSN programming. We suppose the following **Safety Properties** to be verified in our multi-car elevator system.

SP-1. it must not be a case that there exist a route set by another car in the route to be set;

SP-2. it must not be a case that there exists a car whose evacuation (avoidance) route cannot be set on the route to be set, where evacuation routes are for avoiding car crash.

We will translate the safety properties into EVALPSN in the following section. In the EVALPSN based safety verification system, the safety properties in EVALPSN clauses are stored as the safety verification EVALPSN. The EVALPSN based safety verification is carried out as follows: the current state of the multi-car elevator system is translated into EVALPSN clauses and added to the safety verification EVALPSN; then it is verified if the object route can be safely set or not in EVALPSN programming.

4 EVALPSN Multi-car Elevator Safety Verification

First of all, we define two EVALPSN literals used for the EVALPSN based multi-car elevator safety verification.

Route (set/unset) from floor fl_m to floor fl_n by car c_x at time t is formalized in the EVALPSN clause,

$$ro(fl_m, fl_n, c_x, t) : [\mu_1, \mu],$$

where $\mu_1 \in \{\bot(0,0), \mathtt{s}(1,0), \mathtt{x}(0,1), \top(1,1)\}, \mu \in \mathcal{T}_e$, and annotations $\mathtt{s}(\mathtt{x})$ show that the route is set(unset), respectively. For example, EVALPSN clause, $ro(fl_1, fl_{10}, c_j, t) : [\mathtt{x}, \beta]$ can be intuitively interpreted that the route from floor fl_1 to floor fl_{10} must not be set($\mathtt{s}$) by car c_j at time t.

Direction (upward/downward) from floor fl_m to floor fl_n is formalized in the EVALPSN clause,

$$dir(fl_m, fl_n) : [\mu_2, \mu],$$

where $\mu_2 \in \{\bot(0,0), \mathtt{up}(2,0), \mathtt{no}(1,1), \mathtt{dw}(0,2), \top(2,2)\}, \mu \in \mathcal{T}_e$, and annotations $\mathtt{up}/\mathtt{dw}/\mathtt{no}$ show that the directions of the route, upward/downward/no-direction, respectively, where no-direction represents that floor fl_m is identical to floor fl_n. For example, EVALPSN clause, $dir(fl_2, fl_5) : [\mathtt{up}, \alpha]$ can be intuitively interpreted that the direction of the route from floor fl_2 to floor fl_5 is upward.

4.1 Floor Call Constraint in EVALPSN

The three kinds of floor calls **call 4**, **call 5** and **call 6** are excluded from processing as we mentioned in the previous section, therefore we need to translate the constraint as the forbiddance from processing those three types of floor calls into EVALPSN before translating the safety properties. We show how to formalize the forbiddance in EVALPSN with Figure 4[1].

In the case of floor call **call 4**, if the call floor is on the car c_x current route and its floor call has the reverse direction to the car c_x current route, it is forbidden from setting the call floor route, which is translated into the EVALPSN clauses,

$$\begin{aligned} & ro(fl_x, fl_o, c_x, t) : [\mathtt{s}, \alpha] \wedge \\ & dir(fl_x, fl_o) : [\mathtt{up}(\mathtt{dw}), \alpha] \wedge dir(fl_c, fl_d) : [\mathtt{dw}(\mathtt{up}), \alpha] \wedge \\ & dir(fl_o, fl_c) : [\mathtt{dw}(\mathtt{up}), \alpha] \\ & \rightarrow ro(fl_x, fl_c, c_x, t) : [\mathtt{x}, \beta], \qquad x \in \{i, j, k\}. \end{aligned} \tag{1}$$

In both the cases of floor calls **call 5** and **call 6**, if the car c_x call floor route has the reverse direction to the car c_x current route, it is forbidden from setting the call floor route, which is translated into the EVALPSN clause,

$$\begin{aligned} & ro(fl_x, fl_o, c_x, t) : [\mathtt{s}, \alpha] \wedge \\ & dir(fl_x, fl_o) : [\mathtt{up}(\mathtt{dw}), \alpha] \wedge dir(fl_x, fl_c) : [\mathtt{dw}(\mathtt{up}), \alpha] \\ & \rightarrow ro(fl_x, fl_c, c_x, t) : [\mathtt{x}, \beta], \qquad x \in \{i, j, k\}. \end{aligned} \tag{2}$$

[1] The call floor route is designated by cf in Figure 4.

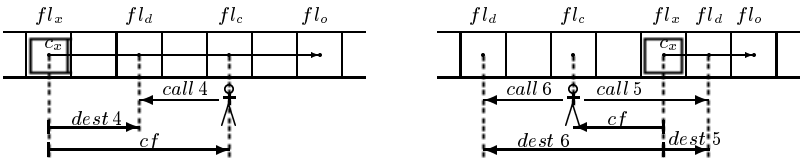


Fig. 4. Forbiddance from Call-4, Call-5 and Call-6

4.2 Safety Properties in EVALPSN

We suppose there are two adjoining cars c_m and c_n in Figure 5; the object route to be verified for car c_m is from floor fl_m to floor fl_d; and the route from floor fl_n to floor fl_o has been set by car c_n. Then considering safety properties **SP-1** and **SP-2**, the object route is forbidden from being set in the following two cases:

1 the direction of the route set by car c_n is the reverse to that of the object route and the destination floor fl_o of the route set by car c_n is in the object route;

2 the direction of the route set by car c_n is not the reverse to that of the object route,[2] car c_n is on the object route and the evacuation route of car c_n is forbidden from being set.

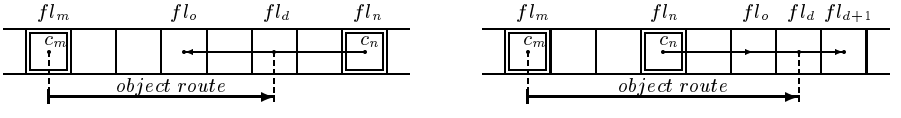


Fig. 5. Safety Verification with SP-1/2

One instance of case **1** and one counter instance of case **2** are shown by the left and right figures in Figure 5, respectively. The above forbiddance derivation cases can be formalized in the EVALPSN clauses,

$$\begin{aligned}
&ro(fl_n, fl_o, c_n, t) : [\mathtt{s}, \alpha] \wedge \\
&dir(fl_m, fl_d) : [\mathtt{up(dw)}, \alpha] \wedge dir(fl_n, fl_o) : [\mathtt{dw(up)}, \alpha] \wedge \\
&\sim dir(fl_d, fl_o) : [\mathtt{up(dw)}, \alpha] \\
&\rightarrow ro(fl_m, fl_d, c_m, t) : [\mathtt{x}, \beta], \\
&\text{where } (m, n) = \{(i, j), (j, k), (k, j), (j, i)\},
\end{aligned} \tag{3}$$

$$\begin{aligned}
&ro(fl_n, fl_o, c_n, t) : [\mathtt{s}, \alpha] \wedge ro(fl_n, fl_{d+1(d-1)}, c_n, t) : [\mathtt{x}, \beta] \wedge \\
&dir(fl_m, fl_d) : [\mathtt{up(dw)}, \alpha] \wedge \sim dir(fl_n, fl_o) : [\mathtt{up(dw)}, \alpha]
\end{aligned}$$

[2] If car c_n does not have a set route, the direction of the route set by c_n can be treated as no-direction.

$$\rightarrow ro(fl_m, fl_d, c_m, t) : [\mathtt{x}, \beta], \\ \text{where } (m, n) = \{(i, j), (j, k), (k, j), (j, i)\}. \tag{4}$$

We also need to translate another kind of physical constraint against the three cars, c_x $(x = i, j, k)$ into EVALPSN, that is to say, car c_i cannot access to floors fl_{29} and fl_{30}, car c_j cannot do to floors fl_{30} and fl_{B2}, and car c_k cannot do to floors fl_{B1} and fl_{B2}, which are formalized in the following EVALPSN clauses,

$$ro(fl_i, fl_{29}, c_i, t) : [\mathtt{x}, \beta], \quad ro(fl_i, fl_{30}, c_i, t) : [\mathtt{x}, \beta], \\ ro(fl_j, fl_{30}, c_j, t) : [\mathtt{x}, \beta], \quad ro(fl_j, fl_{B2}, c_j, t) : [\mathtt{x}, \beta], \\ ro(fl_k, fl_{B1}, c_k, t) : [\mathtt{x}, \beta], \quad ro(fl_k, fl_{B2}, c_k, t) : [\mathtt{x}, \beta]. \tag{5}$$

Lastly, we need the EVALPSN clause to derive permission for setting routes,

$$\sim ro(fl_m, fl_n, c_z, t) : [\mathtt{x}, \beta] \rightarrow ro(fl_m, fl_n, c_z, t) : [\mathtt{x}, \gamma]. \tag{6}$$

4.3 Examples for Safety Verification

We show simple examples of the EVALPSN safety verification with Figure 6.

If a passenger request issued, the floor call is divided into two routes, the call and destination floor routes, and both the routes should be secured in the safety verification system. We suppose two cases shown in Figure 6. In the two cases, the floor call from floor fl_c to floor fl_d (designated by *request* in Figure 6) has been posed to car fl_i, there are car c_j that is free and car c_k that has the current route from floor fl_k to floor fl_o on the car c_i destination floor route (designated by *destination* in Figure 6). The difference of the two cases is just the direction of the car c_k current route. Then, the safety for the car c_i call floor route (designated by *call* in Figure 6) is easily verified because there are neither cars nor a route set by another car in both the cases. Therefore, we show how the safety for the car c_i destination floor route is verified by the EVALPSN safety verification system.

– First of all, it is checked if the forbiddance from setting the car c_i destination floor route,

$$ro(fl_i, fl_d, c_i, t) : [\mathtt{x}, \beta] \tag{7}$$

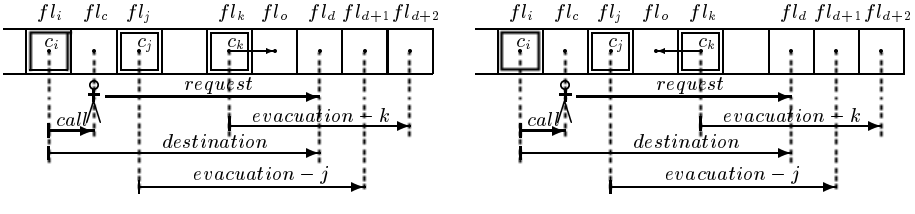


Fig. 6. Safety Verification Examples

can be derived or not by EVALPSN clauses (1), (2), (3) and (4)[3];
- then, since car c_j on the route has no current route, the forbiddance (7) cannot be derived by EVALPSN clauses (1), (2) and (3);
- however, if setting the car c_j evacuation route is forbidden, the forbiddance (7) is derived by EVALPSN clause (4);
- next, it is checked if the forbiddance from setting the car c_j evacuation route (designated by $evacuation - j$ in Figure 6),

$$ro(fl_j, fl_{d+1}, c_j, t) : [\mathtt{x}, \beta] \tag{8}$$

can be derived by EVALPSN clauses (1), (2), (3) and (4);
- then, since there is car c_k on the car c_j evacuation route, it is checked if the forbiddance from setting the car c_k evacuation route (designated by $evacuation - k$ in Figure 6),

$$ro(fl_k, fl_{d+2}, c_k, t) : [\mathtt{x}, \beta] \tag{9}$$

can be derived or not;
- since the forbiddance (9) can be derived in the left shaft in Figure 6 by none of EVALPSN clauses (1), (2), (3) and (4), the car c_j evacuation route and the car c_i destination floor route are secured and we have the permission for setting those routes by EVALPSN clause (6);
- on the other hand, since the forbiddance (9) can be derived in the right shaft in Figure 6 by EVALPSN clause (2), neither the car c_j evacuation route nor the car c_i destination floor route are secured.

5 Conclusion and Future Work

In this paper, we have introduced the basic ides of the EVALPSN safety verification for multi-car elevator system with three cars/one shaft though, one final goal is to construct an EVALPSN safety verification based multi-car/multi-shaft elevator control system, which includes automatic car assignment, that is to say, automatically selecting the most appropriate car among cars whose safety has already been secured. In order to construct the car selection system, the idea of plausible reasoning used for judgment may be applicable to selecting the most appropriate car for passenger service. For example, if we take passenger waiting time and energy saving as judge standards, we could select the most appropriate car for service. Since it has been shown that EVALPSN can deal with plausible reasoning [8], we will construct the car assignment system in the EVALPSN safety verification based multi-car elevator control system.

Moreover, we are implementing a PC simulation system for the EVALPSN multi-car elevator safety verification aiming at practical use.

We can conclude that our EVALPSN based multi-car elevator safety verification system has high extendability such as three-car systems to four-car ones

[3] We ignore EVALPSN clauses (5) for deriving forbiddance in this example.

or single shaft systems to multiple ones and it is quite easy to be implemented as both software and hardware, although such implementation issues have not been addressed in this paper.[4]

References

1. Blair, H.A., Subrahmanian, V.S.: Paraconsistent Logic Programming. Theoretical Computer Science 68, 135–154 (1989)
2. da Costa, N.C.A., Subrahmanian, V.S., Vago, C.: The Paraconsistent Logics P$\mathcal{T}$. Zeitschrift für Mathematische Logic und Grundlagen der Mathematik 37, 139–148 (1989)
3. Nakamatsu, K., Abe, J.M., Suzuki, A.: Annotated Semantics for Defeasible Deontic Reasoning. In: Ziarko, W.P., Yao, Y. (eds.) RSCTC 2000. LNCS (LNAI), vol. 2005, pp. 432–440. Springer, Heidelberg (2001)
4. Nakamatsu, K., Abe, J.M., Suzuki, A.: A Railway Interlocking Safety Verification System Based on Abductive Paraconsistent Logic Programming. In: Soft Computing Systems (HIS 2002), Frontiers in Artificial Intelligence and Applications, vol. 87, pp. 775–784. IOS Press, Amsterdam (2002)
5. Nakamatsu, K., Seno, T., Abe, J.M., Suzuki, A.: Intelligent Real-time Traffic Signal Control Based on a Paraconsistent Logic Program EVALPSN. In: Wang, G., Liu, Q., Yao, Y., Skowron, A. (eds.) RSFDGrC 2003. LNCS (LNAI), vol. 2639, pp. 719–723. Springer, Heidelberg (2003)
6. Nakamatsu, K.: Pipeline Valve Control Based on EVALPSN Safety Verification. J. Advanced Computational Intelligence and Intelligent Informatics 10, 647–656 (2006)
7. Nakamatsu, K., Mita, Y., Shibata, T.: An Intelligent Action Control System Based on Extended Vector Annotated Logic Program and its Hardware Implementation. J. Intelligent Automation and Soft Computing 13, 289–304 (2007)
8. Nakamatsu, K., Imai, T., Abe, J.M., Akama, S.: An Introduction to Plausible Reasoning Based on EVALPSN. In: Nakamatsu, K., Phillips-Wren, G., Jain, L.C., Howlett, R.J. (eds.) New Advances in Intelligent Decision Technologies. Studies in Computational Intelligence, vol. 199, pp. 363–372. Springer, Heidelberg (2009)
9. Nakamatsu, K., Abe, J.M.: The development of Paraconsistent Annotated Logic Program. Int'l J. Reasoning-based Intelligent Systems 1, 92–112 (2009)
10. Nakamatsu, K., Abe, J.M., Akama, S., Kountchev, R.: Introduction to Intelligent Elevator Control Based on EVALPSN. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010. LNCS (LNAI), vol. 6278, pp. 133–142. Springer, Heidelberg (2010)

[4] This work is financially supported by Japanese Scientific Research Grant (C) No. 20500074.

Multi Robot Exploration Using a Modified A* Algorithm

Anshika Pal, Ritu Tiwari, and Anupam Shukla

Soft Computing and Expert System Laboratory
ABV-Indian Institute of Information Technology and Management, Gwalior, India
anshikapal@yahoo.com, tiwariritu2@gmail.com,
dranupamshukla@gmail.com

Abstract. Exploration of an unknown environment is one of the major applications of multi robot systems. A popular concept for the exploration problem is based on the notion of frontiers: the boundaries of the current map from where target points are allocated to multiple robots. Exploring an environment is then about entering into the unexplored area by moving towards the targets. To do so they must have an optimal path planning algorithm that chooses the shortest route with minimum energy consumption. Aiming at the problem, we discuss a modification to the well known A* algorithm that satisfies these requirements. Furthermore, we discuss improvements to the target allocation strategy, by pruning the frontier cells, because the computation burden for optimal allocation is increases with the number of frontier cells. The proposed approach has been tested with a set of environments with different levels of complexity depending on the density of the obstacles. All exploration paths generated were optimal in terms of smoothness and crossovers.

Keywords: Multi Robot System, Area Exploration, A* Algorithm, Frontiers.

1 Introduction

Area exploration has been investigated by the robotics research community through the years. Indeed, this problem lends itself to several applications in different fields, from industrial, such as lawn-mowing or vacuum-cleaning, to military such as de-mining , or humanitarian, such as search and rescue operations.

The robot area exploration is defined as the process of effectively covering an environment by a single or a team of robots in order to gain as much information about the world as possible in a reasonable amount of time [1]. Several approaches, ranging from grid decompositions of the environment to the development of heuristics, have been proposed. More recently, formulations which extend the problem to the multi-robot context have been introduced [4, 5, 6, 29]. The idea is to take advantage of the cooperation among the robots to provide higher robustness as well as to lower the time required to complete the task [2].

Indeed, three major issues define the area exploration problem, according to [3]: coordination of robots, integration of information collected by different robots into a

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 506–516, 2011.

consistent map, and dealing with limited communication. Apart from these issues, the process also requires a path planning algorithm for navigating to target locations. Several methods have been used in the past. We use an A* algorithm [7] with modification in heuristic function to repeatedly calculate paths towards unknown space. To verify the effectiveness of the path planning algorithm in the exploration task, we use the method presented by Agusti Solanas [8] as a base algorithm for exploring the unknown environment. The next contribution of the paper is the pruning of the frontier cells. Our target selection method reduces the computation burden through pruning the frontiers, and thus accomplishes the exploration task quickly.

2 Related Work

The algorithm proposed by Yamauchi [9] is the simplest exploration strategy. Each robot simply looks for a frontier cell that can be reached with the lowest cost. It is a greedy strategy with no coordination among robots. The algorithm by Burgard et al. [10], [11] aims at reducing the exploration completion time and thus coordinates the robots to explore as much area as possible. In [12], the robots methodically sweep the unknown space by advancing in close line formation. The formation is broken and restored as necessary whenever new obstacles are encountered. This strategy requires a very tight coordination among all robots.

In [13] a collaborative exploration problem is studied, for a team of mobile robots. An algorithm applied to multi-robot teams for the complete coverage of a connected area with unknown obstacles is proposed in [14]. Following a different scheme based on potential fields, [15] proposes a deploying strategy in which a cluster of robots is evenly distributed over the environment by considering repulsive forces: robots repel each other and obstacles repel robots. The robots keep moving until repulsive forces cancel each other.

The exploration algorithm based on k-means clustering proposed in [8][16] has overcome the greedy characteristic of the previous strategies by sending robots to different regions of the workspace. Each robot is then assigned to its closest region according to the Euclidean distance from the robot to the region's centroid [8]. The assignments of robot to regions are improved by an optimization strategy instead of by assigning robots to regions on a sequential basis [16]. In [20] the robots work in pairs and coordinate their actions. This technique tries to decrease the overlap between the robots as much as possible. If we come to communication constraints, many algorithms have been implemented and results verified [25, 26, 27, 28].

In the problem of exploring an unknown environment with multiple robots, they incrementally decide where to go in order to perform some predetermined tasks. If the robots' traveling paths through the unknown space are shortened, the completion time of the exploration task is reduced. The approach proposed in [17] finds the nearly most optimal path of the robot using Genetic, ANN and A* algorithms. In [18] MNHS based Robot Path Planning method is proposed. The motivation is to make the problem robust against the uncertainties that might arise like the sudden discovery that the path being followed does not lead to the goal. Paper [19] solves the problem of robotic path planning using a combination of A* algorithm and Fuzzy Inference.

3 Environment and Robot Model

The environment to be explored is modeled as a 2D occupancy grid. There are stationary obstacles of arbitrary shape in the area. During the exploration, each cell of the grid has one of three statuses: occupied, free or unknown. Here "occupied" means that the cell is occupied by obstacles; "free" means that no obstacle exists in this cell; and "unknown" means that the occupancy of the cell is not known yet. The physical shape of a robot can be modeled as a square. For simplicity, its dimension is set as the size of a grid cell [21]. It is assumed that each robot has mapping, localization and communication capability. The robot uses its range sensor to detect the status of the cells (i.e. free or occupied) within its sensor region. A sensor region of a robot is a region around its immediate surroundings, in a domain limited by a square inscribed within the circle defining its effective sensor range S_r (Figure 1).

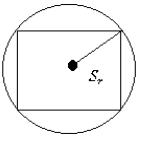


Fig. 1. Sensor region of a robot

4 Path Planning Algorithm

Path planning is planning a path from the current location to the next target. In exploration, both the current and the target locations belong to the explored region, there must exist a viable path inside the explored region. We solve the problem using A* algorithm with additionally including an energy factor, and call it Energy-efficient A* (EA*) algorithm. In A* algorithm the path scoring function is given by

$$F=G+H \tag{1}$$

Where G is cost of getting from source to current node; and H is estimated cost from the current node to target. Here cost is the Euclidean distance of two 2D points. It hence tries to optimize the total path length to be travelled by the robot. In the real world, however only shortest route might not give a very realistic picture of the problem, energy conservation also play a vital role. Stops and turnings cause acceleration and deceleration that consumes significant energy.

Our EA* algorithm is slightly different from A*, where selection of next node is not only based on the distance, it also includes robot movement direction. Robot's state is represented as its location (x,y) and direction θ; < x, y, θ >. We assume the robot uses 45^0 as the unit for turning, since we only allow the robot to move from one cell to one of its eight neighbors. We consider the energy for stops and turns if the two states have different directions. Table 1 shows the energy consumption rate for different turns and stops. In EA* algorithm the path scoring function is given by

$$F=G+E+H \tag{2}$$

Where parameter G and H are same as A* algorithm. E is the energy consumption from source to current node. Figure 2 shows an example with a path from A* and EA* algorithm. The path generated by EA* algorithm is more smooth compared to A* with less number of turns.

Table 1. Energy consumption for stops and turns

Turns/Stops	Penalization Rates
Stop	0.5
45^0	0.4
90^0	0.6
135^0	0.8
180^0	1.0

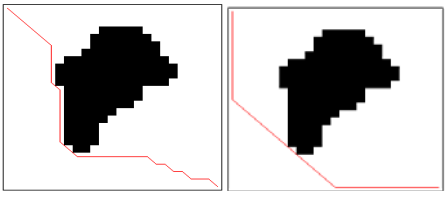


Fig. 2. Path from A*(left) and EA*(right),upper left corner is the source and lower right corner is the target, black regions represents the obstacles

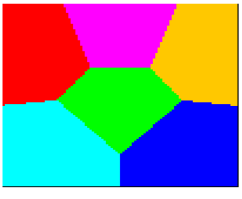


Fig. 3. Example of environment partitioning, colored regions are the clusters produced by k-means

5 Exploration Algorithm

The operational overview of the system from a mission perspective, illustrated in Figure 4. We use the method proposed by [8] as a base for the work presented here.

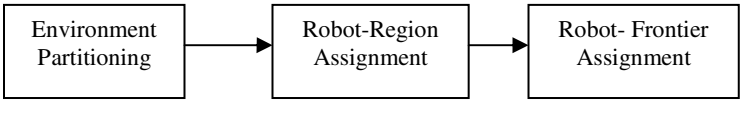


Fig. 4. Operational flow of the algorithm

5.1 Environment Partitioning

Unexplored area is partitioned into as many areas as available robots by applying k-means [22]. In this case, *K* coincides with the number of available robots and the objects are unknown map cells represented by their 2D coordinates. The centroids of the k clusters are initially set to random basis. The algorithm then iteratively modifies those centroids until convergence. The result is a set of k disjoint regions, whose centers of mass are the resulting centroids. For a detailed explanation consult [8]. Figure 3 shows the clustering example.

5.2 Robot-Region Assignment

After the unknown space is divided into K regions, the algorithm decides which region is assigned to each robot. The goal is to assign a robot to every region in such a way to minimize the total cost. Therefore, it is necessary to design an optimal solution to assign robots to regions fairly. In this proposal, the assignment between robots and regions is formulated as a assignment problem via Hungarian Method [23]. Consider a 'n x n' cost matrix which represents the cost of all individual assignments of robots to regions. Here, each entry in the cost matrix is the length of the path the corresponding robot has to travel to reach the designated target points. The distance from a robot to region is defined according to EA* path planning algorithm, presented in section 4, where robot locations are the source and region's centroid are the targets.

$$\begin{bmatrix} C_{11} & C_{12} & C_{13} & C_{1n} \\ C_{21} & C_{22} & C_{23} & C_{2n} \\ C_{31} & C_{32} & C_{33} & C_{3n} \\ C_{n1} & C_{n2} & C_{n3} & C_{nn} \end{bmatrix} \tag{3}$$

The Hungarian method, which is able to find the optimal solution with the minimal cost given this matrix, can be summarized by the following steps [24]:

1. Compute a reduced cost matrix by subtracting from each element the minimal element in its row. Afterwards, do the same with the minimal element in each column.
2. Find the minimal number of horizontal and vertical lines required to cover all zeros in the matrix. In case exactly n lines are required, the optimal assignment is given by the zeros covered by the n lines. Otherwise, continue with Step 3.
3. Find the smallest nonzero element in the reduced cost matrix that is not covered by a horizontal or vertical line. Subtract this value from each uncovered element in the matrix. Furthermore, add this value to each element in the reduced cost matrix that is covered by a horizontal and a vertical line. Continue with Step 2.

This optimal assignment ensures that the total sum of path lengths of robot moving to their assigned regions is minimized.

5.3 Robot-Frontier Assignment

Having decided the assignment of a robot to every region, this section describes to which frontier cell each robot should travel. We choose as a destination the frontier cell that leads to minimum cost. The cost of a frontier cell F_j for a robot R_i, is defined as the traveling cost C_{ij}, which are calculated by applying path planning algorithm. The cost function is defined as

$$cost_{ij} = \begin{cases} C_{ij} + O_j & F_j \in r_i \\ \Delta + d(F_j, c_i) + O_j & F_j \notin r_i \end{cases} \tag{4}$$

Where r_i is the optimal region assigned to robot R_i and Δ is a constant that represents the maximum distance measurable on the map (e.g., the length of the map's diagonal).

Hence, frontier cells that do not belong to the region assigned to a robot receive a significantly high penalization Δ. c_i is the centroid of the region r_i, and d represents the Euclidean distance between two locations. O_j is the frontier penalty, which is initially set to zero. In order to avoid overlap of the sensor's footprint, once a frontier F_j is assigned to robot R_i, then the other frontiers that lie within the sensor range (S_r) of the robots' target location (i.e. F_j), is penalized with a constant αS_r with α≥2. Since all regions are disjoint by construction, this rule ensures that robots separate from each other. For every robot, the frontier with minimum cost according to (4) is chosen as its destination frontier. When a robot reaches its destination frontier, it scans the environment and updates the global map. Then a new destination frontier is determined by applying the same procedure. Once a predefined number of unknown cells (P_n) are discovered by the team the remaining cells are redistributed by reapplying the clustering phase. The process stops when no unknown cells are left.

5.4 Frontier Pruning

The above mentioned frontier exploration method is heavily influenced by the computational time for calculating the real path traveling cost. So, an additional consideration is that it is undesirable to compute cost for each frontier, because some of the frontiers are less favorable, which are very far from the current position, it requires lot of computation for calculating the real path travelling cost for that cells. One of the solution is that to prune the frontier list on the basis of their estimated traveling cost d_{ij}. An admissible heuristic is used for cost estimation.

$$d_{ij} = \left|F_j - R_i\right| \tag{5}$$

R_i is the robot's current location, | | is the Euclidean distance of two points. According to admissible fact, the estimated cost must always be lower than or equal to the actual real path cost. So by considering this fact we can prune the cells which has higher estimated cost without computing their actual cost, which is computationally time consuming. The proposed pruned frontier based exploration approach is described below:

1. Sort the frontier cells with increasing order of their estimated cost.
2. Make pruned frontier list by selecting top K % frontier cells.
3. Each robot than choose a target from that pruned list using equation (4).

While this method is not necessarily optimal, it is fast, and in our experience entirely sufficient for exploration mission.

6 Experiments

To evaluate the proposed method, a Java based self-developed simulator is built. The workspace was taken as an input in form of an image. The image depicted the obstacles as black regions and the accessible area as the white region. Figure 5 shows two different workspace on which the experiments have been carried out.

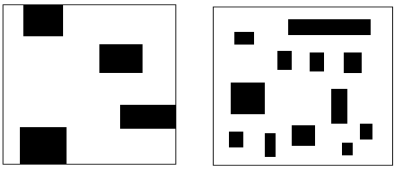


Fig. 5. Two Simulation environments, Map m1 (left), and Map m2 (right)

The three approaches have been compared under the same configuration: (i) method proposed by Augusti [8] with A* path planning algorithm, (ii) EA* based exploration, and (iii) Pruned frontier based exploration.

Table 2. Parameters for experiments

Parameter	**Value**
Map Size	100 X 100
Team Size	4, 6, and 8 robots
Sensor Range (S_r)	8 Cell units
α	2
K	50
P_n	(Map size * S_r) / 100
Robots' initial location	Randomly selected

Figure 6 shows the final trajectories followed by six robots after fully exploring the map m1 and m2. The traversed path is shown by lines with different colors. From the figure, we can see that, the paths generated by EA* algorithm are more smooth with minimum turns and have less crossovers, crossovers again minimized by pruning the frontiers. There are proper separations in paths in order to minimize the overlap. Efficiency is measured in a variety of ways, (i) total exploration time (in min.) needed for exploring the entire environment, (ii) average energy consumed by each robot, and (iii) average distance traveled by each robot during exploring the environment.

As can be seen from the table 3 the team using pruned frontier based approach significantly outperforms the other two approaches, especially in terms of time. This is mainly due to the fact that the size of the frontier list is reduced by 50%, and all the computations are performed on that reduced list not on the complete list. It is very fast in making decision like where to move next. We can also see that as the number of robots increases the values of all the efficiency parameters are decreases, which shows the utilization of multiple robots. Results also shows that the behavior of pruned frontier and EA* based exploration strategy is quite similar in terms of traveled distance and energy expanded. The reason is that, in almost cases, the target locations selected in each step is same in both approaches. However, time is much reduced by pruning the frontier list (figure 7). This proves that by pruning the frontier list we are getting the most favorable candidate locations.

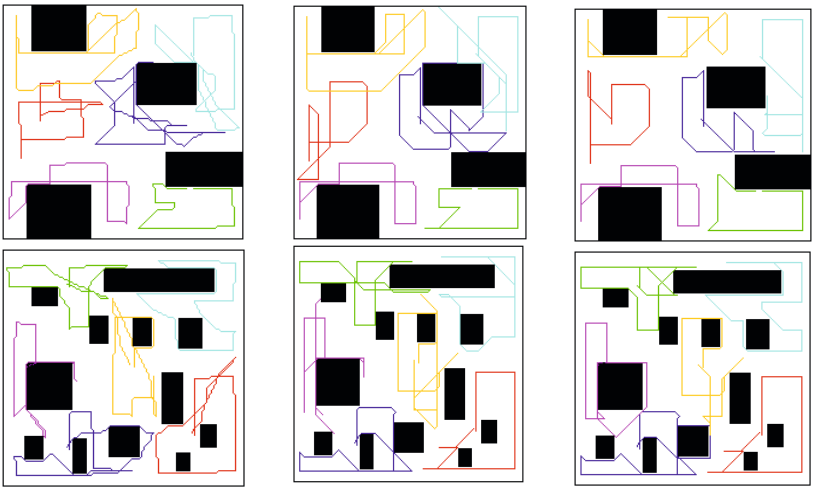


Fig. 6. Trajectories for map m1(top) and map m2(bottom) with three approaches Augusti(left), EA*(middle), and Pruned Frontier(right) in 6 robot configuration

Table 3. Experimental results for map m1 and m2

#Robots	Approach	Map m1			Map m2		
		Time	Energy	Dist.	Time	Energy	Dist.
4	Augusti	32	36	264	18	49	300
	EA*	15	18	260	15	23	272
	Pruned Frontier	11	18	262	4	22	270
6	Augusti	6	34	216	10	54	236
	EA*	5	13	215	8	19	220
	Pruned Frontier	1	11	170	2	17	205
8	Augusti	4	26	138	4	28	173
	EA*	3	12	145	4	15	169
	Pruned Frontier	1	12	142	1	11	156

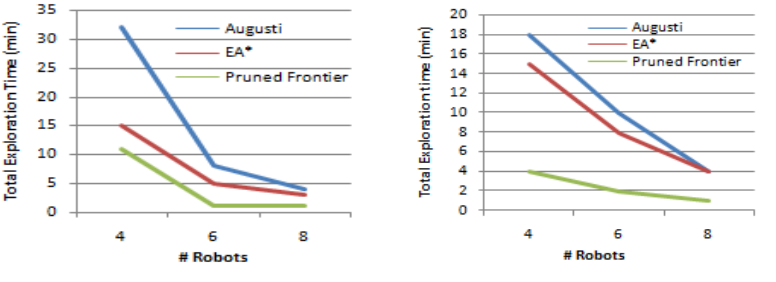


Fig. 7. Number of robots vs. time, for map m1 (left), and map m2 (right)

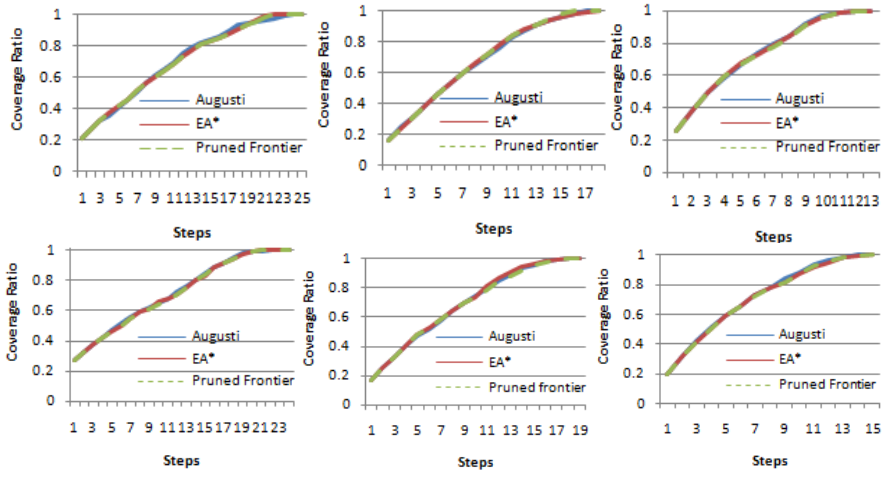


Fig. 8. Number of steps vs. coverage ratio for map m1(top) and map m2 (bottom) with 4(left), 6(middle), and 8(right) robots

Finally, we investigate the percentage of area covered by the robots in each step. This is formally defined as:

$$Coverage\ Ratio = \frac{number\ of\ explored\ free\ cells}{total\ number\ of\ accessible\ free\ cells} \tag{6}$$

Figure 8 shows the coverage ratios for different configurations, the behavior of all three approaches is almost same, however we can notice that the number of steps are reduced as team size increases.

7 Conclusion and Future Work

We have presented a multi robot exploration method that takes into account the path planning problem. The method is an enhancement of the exploration algorithm based on unsupervised clustering. However our main concern is not only to produce an efficient path but also to complete the exploration mission as soon as possible. To do so, an improved pruned frontier based exploration strategy is presented to reduce the computational time. The exploration efficiency is investigated through multiple tests by taking various performance parameters. Experiments presented in this paper illustrate that a team of robots using our approach can complete their exploration mission in a significantly shorter period of time with minimum energy consumption. This method does not require any explicit coordination or synchronization between robots. Future work will consist of dealing with environments of varying size, and including the conditions where only limited range communication is possible.

References

1. Yamauchi, B.: A Frontier-Based Approach for Autonomous Exploration. In: IEEE International Symposium on Computational Intelligence in Robotics and Automation, pp. 146–151 (1997)
2. Guzzoni, D., Cheyer, A., Julia, L., Konolige, K.: Many robots make short work. AI Magazine 18(1), 55–64 (1997)
3. Fox, D., Ko, J., Konolige, K., Limketkai, B., Stewart, B.: Distributed multi-robot exploration and mapping. In: Proc. of the IEEE. Special Issue on Multi-Robot Systems (2006)
4. Cyrill, S., et al.: Efficient exploration of unknown indoor environments using a team of mobile robots. Annals of Mathematics and Artificial Intelligence 52, 205–227 (2008)
5. Noa, A., et al.: The giving tree: constructing trees for efficient offline and online multi robot coverage. Annals of Mathematics and Artificial Intelligence 52, 143–168 (2008)
6. Visser, A., Slamet, B.A.: Balancing the information gain against the movement cost for multi robot frontier exploration. In: European Robotics Symposium (2008)
7. Rich, E., Knight, K.: Artificial Intelligence, pp. 29–98. McGraw- Hill, New York (1991)
8. Solanas, A., Garcia, M.A.: Coordinated multi-robot exploration through unsupervised clustering of unknown space. In: Proc. Int. Conf. on Intelligent Robots and Systems, vol. 1, pp. 717–721 (2004)
9. Yamauchi, B.: Frontier-based exploration using multiple robots. In: Proc. of Int. Conf. on Autonomous Agents, pp. 47–53 (1998)
10. Burgard, W., et al.: Collaborative multi-robot exploration. In: Proc. Int.Conf. on Robotics and Automat., vol. 1, pp. 476–481 (2000)
11. Burgard, W., Moors, M., Stachniss, C., Schneider, F.: Coordinated multi robot exploration. IEEE Transactions on Robotics 21(3), 376–386 (2005)
12. Latimer, D., et al.: Towards sensor based coverage with robot teams. In: IEEE Int. Conf. on Robotics and Automation, pp. 961–967 (2002)
13. Yamauchi, B.: Decentralized coordination for multirobot exploration. Robotics and Autonomous Systems 29, 111–118 (1999)
14. Ge, S.S., Fua, C.H.: Complete Multi-Robot Coverage of Unknown Environments with Minimum Repeated Coverage. In: IEEE Intl. Conf. on Robotics and Automation, pp. 727–732 (2005)
15. Howard, A., Mataric, M.J., Skhatme, G.S.: Mobile sensor network deployment using potential fields: a distributed, scalable solution to the area coverage problem. In: 6th Int. Symp. on Distributed Autonomous Robotics Systems, pp. 299–308 (2002)
16. Wu, L., et al.: Balanced Multi-Robot Exploration through a Global Optimization Strategy. Journal of Physical Agents 4(1) (2010)
17. Kala, R., et al.: Mobile Robot Navigation Control in Moving Obstacle Environment using Genetic Algorithm, Artificial Neural Networks and A* Algorithm. In: Proceedings of the IEEE World Congress on Computer Science and Information Engineering, pp. 705–713 (2009)
18. Kala, R., Shukla, A., Tiwari, R.: Robotic Path Planning using Multi Neuron Heuristic Search. In: Proceedings of the ACM 2009 International Conference on Computer Sciences and Convergence Information Technology, pp. 1318–1323 (2009)
19. Kala, R., Shukla, A., Tiwari, R.: Fusion of probabilistic A* algorithm and fuzzy inference system for robotic path planning. Artificial Intelligence Review 33(4), 275–306 (2010)

20. Al-Khawaldah, M., Livatino, S., Lee, D.: Frontier based exploration with two Cooperative Mobile Robots. International Journal of Circuit, Systems and Signal Processing 4(2) (2010)
21. Sheng, W., Yang, Q., Tan, J., Xi, N.: Distributed multirobot coordination in area exploration. Robotics and Autonomous Systems 54, 945–955 (2006)
22. Hartigan, J.A., Wong, M.A.: A k-means clustering algorithm. Journal of Applied Statistics 28, 100–108 (1979)
23. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Research Logistics Quarterly 2(1), 83–97 (1955)
24. Wurm, K.M., et al.: Coordinated Multi-robot Exploration using a Segmentation of the Environment. In: International Conference on Intelligent Robots and Systems (2008)
25. Rooker, M.N., Birk, A.: Multi-robot exploration under the constraints of wireless networking. Control Engineering Practice 15(4), 435–445 (2007)
26. Doniec, A., et al.: Distributed constraint reasoning applied to multi robot exploration. In: 21st IEEE International conference on Tools with Artificial Intelligence (2009)
27. Ferranti, E., et al.: Rapid exploration of unknown areas through dynamic deployment of mobile and stationary sensor nodes. Autonomous Agents and Multi Agent systems 19(2), 210–243 (2009)
28. Pei, Y., et al.: Coordinated Multi-Robot Real-Time Exploration with Connectivity and Bandwidth awareness. IEEE International conference on Robotics and Automation (2010)
29. Vallejo, D., et al.: A Multi-agent Architecture for Multi-robot Surveillance. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) ICCCI 2009. LNCS (LNAI), vol. 5796, pp. 266–278. Springer, Heidelberg (2009)

Fuzzy Ontology Building and Integration for Fuzzy Inference Systems in Weather Forecast Domain

Hai Bang Truong[1], Ngoc Thanh Nguyen[2], and Phi Khu Nguyen[3]

[1] University of Information Technology,
Ho Chi Minh City, Vietnam
bangth@uit.edu.vn
[2] Institute of Informatics, Wroclaw University of Technology, Poland
thanh@pwr.wroc.pl
[3] University of Information Technology, Ho Chi Minh City, Vietnam
khunp@uit.edu.vn

Abstract. Weather forecast is an environment where there are inevitable uncertainties associated with weather phenomena. In such a fuzzy environment, inference systems for weather forecast and its evaluation have been explored to tackle difficult major in formulating forecast policy, and to cope up with vague and/or abnormal (chaotic) meteorological information. In this paper, a framework of building a fuzzy ontology for representing the meteorological knowledge is proposed. The weather fuzzy inference system has been suggested, which takes the fuzzy ontology and the corresponding instances as its knowledge base. A method for fuzzy ontology integration is introduced for solving inconsistency among weather services' knowledge.

Keywords: Weather, Weather ontology, Fuzzy ontology, Ontology alignment, and Fuzzy Inference System.

1 Introduction

In the recent years, fuzzy technique has drawn considerable attention towards handling this kind of complex and non-linear problems. The technique has been widely applied to many meteorological problems such as prediction of ceiling and visibility using fuzzy logic [10]. This work combines fuzzy logic and case-based reasoning to produce forecasts of airport cloud ceiling and visibility. Firstly, identify the attributes to be used to indicate similarity between cases and, secondly, describe degrees of similarity between such attributes to locate k-nearest neighbors from airports' historical databases. These nearest neighbors are adapted to produce values of forecast parameters. Mitra [9] presents the concept of fuzzy inference system (FIS) in the prediction of fog. This approach addresses the issue of linguistic fuzzy rule-based modeling from available daily current weather observations in winter season over New Delhi. The parameters that are depicted, as fuzzy sets in this paper are dew point, dew point spread, wind speed, sky condition (Skc), and rate of change of the dew point spread (Rcdps). This goal is achieved by modifying the rule antecedents to produce a flexible and interpretable output space. One of the first successful applications of ANNs in forecasting is reported

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 517–527, 2011.

by Lapedes and Farber [7]. Their results show that ANNs can be used for modeling and forecasting nonlinear time series with very high accuracy. Following Lapedes and Farber, a number of papers were devoted to using ANNs for weather forecast because of their ability to model an unspecified nonlinear relationship between load and weather variables. [12] Taylor and et al presents a ANN with a single hidden layer and input variables from analysis of daily load data for England and Wales weather. And explain how ensemble predictions can be used to improve the accuracy of the ANN load forecasts. The results show that the average of the load scenarios is a more accurate load forecast than that produced using traditional weather forecasts. Hung and et al [6] use an Artificial Neural Network technique to improve rainfall forecast performance. The training data is generated from years of hourly data of 75 rain gauge stations in Bangkok, Thailand. The parameters is a combination of relative humidity, air pressure, wet bulb temperature and cloudiness. Aimed at providing forecasts in a near real time schedule, different network types were tested with different kinds of input information. Results show that ANN forecasts have superiority over the ones obtained by the persistent model.

In this paper, a framework of building a fuzzy ontology for representing the meteorological knowledge is proposed. The weather fuzzy inference system has been suggested, which takes the fuzzy ontology and the corresponding instances as its knowledge base. A method for fuzzy ontology integration is introduced for solving inconsistency among weather services' knowledge.

2 Integration Fuzzy Logic on Ontology

Traditional ontologies have been developed by using crisp logic such as first-order logic and description logic [1], so it can not afford to provide well-defined means. Conceptual formalism of the ontologies can not be fully representative for imprecise and vague information (e.g. "rainfall is very heavy") in many application domains. Therefore, in recent years, several studies on the methods of integrating fuzzy logic into the ontology to extend the traditional ontology more suitable for resolving problems of uncertain inferences. For example, [3] first steps toward the realization of a theoretical model and a complete structure-based ontologies can consider the nuances of natural language by integrating fuzzy logic to the ontology. By integrating ε-connection to describe fuzzy logic, Lu and co-authors [13] propose a new approach to combine both features fuzzy logic and distribution within the description. Their main contribution is to propose an algorithm to achieve discrete tableau inference fuzzy logical ontology system.

At first, let us remind the definition of a fuzzy set. Let U be a classical set of objects, called the universe, for every generic elements $x \in U$. Membership in a classical subset A of U is often viewed as a characteristic function , μ from U to $\{0, 1\}$ such that $\mu_A(x)$= 1 if $x \in A$, and $\mu_A(x)$= 0 if $x \notin A$. $\{0, 1\}$ is called a valuation set. If the valuation set is allowed to be the real interval [0, 1], A is called a fuzzy set [2]. $\mu_A(x)$ is the membership value of $x \in A$. The closer the value of $\mu_A(x)$ is to 1, the more x belongs to A.

Here, we explain on how to integrate fuzzy values on different elements of an ontology. Most often an (non-fuzzy) ontology is defined by the following elements:

- C a set of concepts (classes);
- R set of binary relations defined on C;
- Z set of axioms, which formulas of the first order logic and can be interpreted as integrity constraints or relationships between instances and concepts, and which can not be expressed by the relations in set R.

In general, on the basis of the aforementioned literature, we can state that an ontology is fuzzy if one of the above mentioned elements is fuzzy. In our study however, Z is out of focus. We agree with [3] that there exists a generic membership function f : $(concepts \cap instances) \times properties \rightarrow property\_value \times (0, 1]$ with the meaning that $f(o; p)$ is the value that a concept or an instance o assumes for property p with associated degree. For example, in a weather ontology, $f(Rain, depth)$ = (150,0.8) means that for the property *depth*, the concept *Rain*, has value 150 with degree 0.8. Here, the fuzzy of the concept is displayed via its instances. Thus it does not support fuzzy in abstract level. In our study, we address three fuzzy levels in a fuzzy ontology including fuzzy concept, fuzzy instance and fuzzy relation. That can be expressed as follows:

- Fuzzy concept: f : $(concepts \times properties) \rightarrow (0, 1]$ that means each property belongs to a fuzzy concept with a degree. For example, we consider the concept *Weather* with its structure $Weather(hasRain : 0.7, hasWind : 0.6, hasVisibility : 0.3, hasHumidity : 0.4)$. There exists a membership function $f(Weather \times hasRain) = 0.7$, in which, *Rain* belongs to *Weather* with a degree 0.7.
- Fuzzy instance: similar to [3] we agree that f : $instances \times$ properties $\rightarrow$ property_value $\times(0, 1]$. For example, given a concept $Car(hasColor, hasPrice)$ with f : $ShelbyCobra \times hasColor$= (Yellow, 0.5).
- Fuzzy relation: f : $concepts \times concepts \rightarrow$ (0, 1].

We assume a real world $(\boldsymbol{A}, \boldsymbol{V})$ where $\boldsymbol{A}$ is a finite set of attributes and $\boldsymbol{V}$ is the domain of $\boldsymbol{A}$. Also, $\boldsymbol{V}$ can be expressed as a set of attribute values, and $\boldsymbol{V} = \bigcup_{a \in \boldsymbol{A}} V_a$ where V_a is the domain of attribute a. We consider domain ontologies referring to the real world (A, V), such ontologies are called (A, V)-based. As a fuzzy (A,V)-based ontology we understand a triple:

$$O = (C, R, Z) \tag{1}$$

where:

- $\boldsymbol{C}$ is a set of concepts. A fuzzy concept is defined as a triple:

$$concept = (c, A^c, V^c, f_c) \tag{2}$$

  where c is the unique name of the concept, A^c is a set of attributes describing the concept and $V^c \subset V$ is the attributes domain: $V^c = \bigcup_{a \in A^c} V_a$ where V_a is the domain of the the attribute a and f_c is a fuzzy function:

$$f_c : A^c \rightarrow [0, 1] \tag{3}$$

  representing the degrees to which concept c is described by attributes. Triple (A^c, V^c, f_c) is called the fuzzy structure of concept c.

– ***R*** is a set of fuzzy relations between concepts, $R = \{R_1, R_2, \ldots, R_m\}$ where

$$R_i \subset C \times C \times (0, 1] \tag{4}$$

for i = 1, 2,. . . ,m. A relation is then a set of pairs of concepts with a weight representing the degree to which the relationship should be;
– ***Z*** is a set of axioms, which can be interpreted as integrity constraints or relationships between instances and concepts. This means that Z is a set of restrictions or conditions (necessary & sufficient) to define the concepts in ***C***.

Notice that, a fuzzy concept is called a concrete fuzzy concept if it represents a fuzzy set without any attribute. The definition of concrete fuzzy concept in [4] as follows:

Definition 1 (Concrete Fuzzy Concept). *The concrete fuzzy concept* $cf \subseteq$ C *is defined as a following 4-tuple:*

$$cf = (V_{cf}, V'_{cf}, L_{cf}, f_{cf}) \tag{5}$$

where, cf *is the unique identifier for the concept. The* $V_{cf} \subseteq \boldsymbol{V}$ *is a concrete set, which called domain of the concept.* $V'_{cf} \subseteq (0, 1]$ *presents fuzzy values of the concrete set* V_{cf}*. The* $f_{cf} \in F$ *is a concrete fuzzy predicate on* V_{cf} *(considered as a membership function),* $\forall\, v \in V_{cf}$*,* $f_{cf}(v) \in V'_{cf}$*.* $L_{cf} \subseteq V$ *models linguistic qualifiers, is determined by the strength of the property value in* V_{cf}*.*

3 Fuzzy Ontology-Based Weather Inference System

In this section, we present the main ideas underlying our proposed Fuzzy Ontology-based Weather Inference System (F-WIS). There are following goals of F-WIS:

– The system solves inconsistencies among weather services' fuzzy ontologies representing its knowledge base to generate the best representation that is considered as F-WIS's knowledge base.
– The system provides an inference mechanism that tackles difficult major in formulating forecast policy, and to cope up with vague and/or abnormal meteorological information.

To highlight the involved system, *Fig* 1 shows the basic components of an F-WIS. It includes four principle components such as an *Ontology Integration*, a *Fuzzification*, a *Weather Knowledge Base*, *Decision-Making Supprot*, and a *Defuzzification*.

The *Ontology Integration* makes consistency among weather services' knowledge. We assume that each weather service owns an ontology as its knowledge base. The F-WIS's knowledge base is acquired as the best representation of a number of the weather services. The word best representation mentioned above means the following criteria:

(1) The result of integration should be guaranteed the completeness, that is all meteorological information included in the component elements will appear in the integration result.

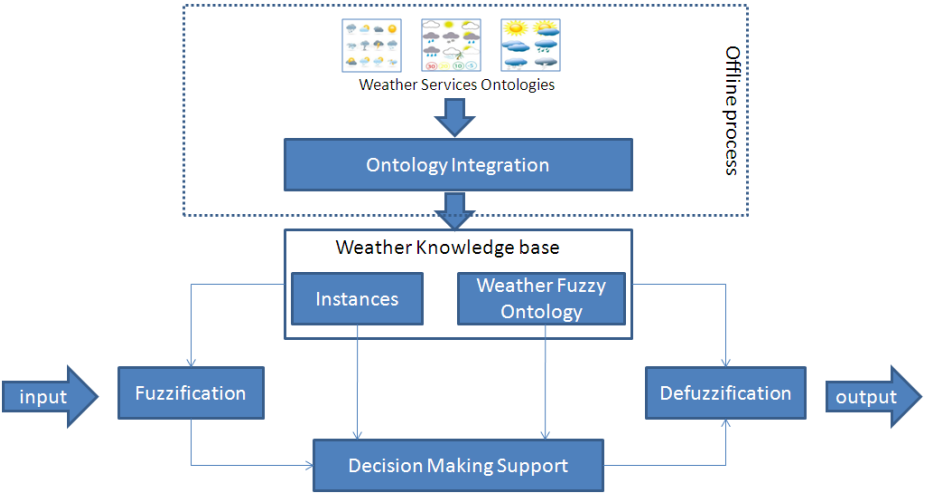


Fig. 1. Weather Inference System

(2) All conflicts appearing among elements to be integrated should be solved. It often happens referring to the same subject different elements contain inconsistent information. Such situation is called a conflict. The integration result should not contain inconsistency, so the conflicts should be solved.

The *Fuzzification* which transforms the input variables into degrees of match with linguistic values in fuzzy ontology, including three functions as follows:

- Normalizing the values of input variables
- Perform a scale mapping that transfers the range of values of input variables into corresponding fuzzy ontology.
- Converting values of input variables into corresponding linguistic values.

The *Weather Knowledge Base* includes weather fuzzy ontology and its instances. Fuzzy ontology characterizes the control goals and control policy of the weather by means of a set of linguistic control rules.

The *Decision Making Support* is the kernel of the F-WIS, which performs the inference operations on the fuzzy weather ontology. This includes an adaptive neural network for fuzzy inference.

The *Defuzzification* is opposite with the Fuzzification, which performs the fuzzy results of inference into a crisp output.

4 Reusing Ontologies for Weather Fuzzy Ontology Building

Here, we describe a top-down approach for reusing existing ontologies to building a weather ontology. In the top-down approach, first, we use Wordnet to generate a hierarchy of abstract concepts (called weather upper ontology) in weather domain. These

concepts are maped to concepts belonging to utility ontologies on Sweet and weather ontologies that is provided by other weather services. Domain experts can compose the weather ontology by inheriting the concepts and inferencing capabilities in the mapping. Second, the concepts in weather upper ontology matches to parameters belonging to JMBL[1] Based on these matches, we collect names that are used to develop the weather ontology. Third, domain experts use an ontwiki to modify and supplement fuzzy properties (attributes and relations) into the target ontology. To discover the membership values of the fuzzy ontology, a neural network is used to learn these membership values from real data. During developing process, the ontology and fuzzy membership values are evaluated by meteorological experts. The ontology building process is shown as *Figure* 1.

Wordnet is an upper ontology, is provided under the direction of George A. Miller, is a high-level, domain-independent ontology, providing a framework by which disparate systems may utilize a common knowledge base and from which more domain-specific ontologies may be derived. Its lexical database organizes nouns and verbs into hierarchies of *is-a* relations. In version 2.0, there are nine noun hierarchies comprising 80,000 concepts, and 554 verb hierarchies comprising 13,500 concepts [8]. The concepts expressed in such an ontology are intended to be basic and universal concepts to ensure generality and expressivity for a wide area of domains. We use the Wordnet as the foundation for deriving concepts in the weather domain. In this way, the weather ontology designer takes advantage of the knowledge and experience already built into the Wordnet.

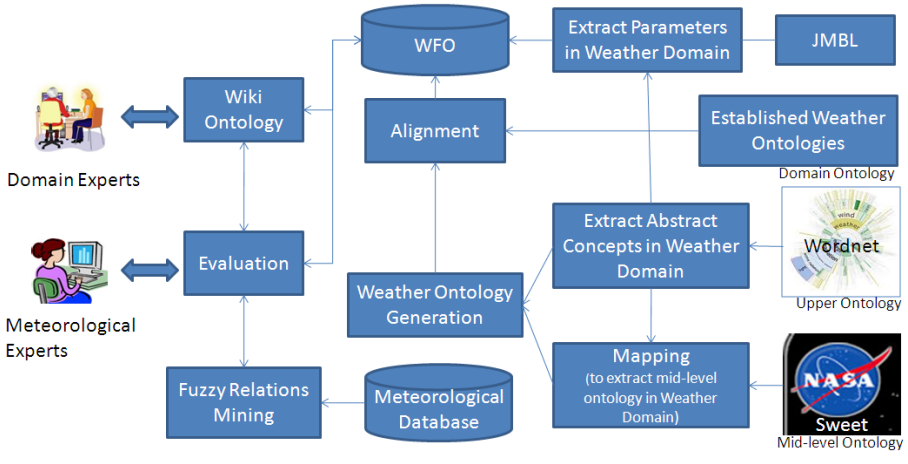


Fig. 2. Ontology Bulding Process

SWEET is a mid-level ontology, is provided by, which serves as a bridge between abstract concepts defined in the Wordnet and low-level domain specific concepts

[1] JMBL is a specification for a standard language that will broker the exchange of information between meteorological and oceanographic (METOC) data providers and user applications.

specified in a weather ontology. Ontologies on SWEET may provide more concrete representations of abstract concepts found in Wordnet. This ontology category also encompasses the set of ontologies that represent commonly used concepts, such as Time and Location. These commonly used ontologies are sometimes referred to as utility ontologies. With the existence of such ontologies, weather ontology designers can compose their weather ontology using these utility ontologies and inherit the concepts and inferencing capabilities provided by them. Further, concepts in the utility ontology could be mapped to concepts in the upper ontology.

The weather ontology is composed by importing mid-level ontologies from SWEET. Domain experts also extend concepts defined in the mid-level ontologies and Wordnet. Additionally, reusing well established weather ontologies such as ... in the development of the target weather ontology allows one to take advantage of the not only semantic richness of the relevant concepts and logic, but also difficult majors already built into the target ontology. The intended use of Wordnet is for key concepts expressed in a weather ontology to be derived from, or mapped to, concepts in Wordnet. Using common the mid-level ontologies and Wordnet is intended to ease the process of integrating or mapping weather ontologies in applications of the target weather ontology.

Fuzzy Relations Mining includes neural networks to discover fuzzy relationships among weather phenomena and other phenomena from meteorological data. The fuzzy relationships are evaluated by meteorological experts. In this paper, it is not intended to deal with them yet and it will be a subject of another work in the future.

5 Fuzzy Weather Ontology Demonstration

The Glossary of Meteorology [6] defines weather as the state of the atmosphere, mainly with respect to its effects upon life and human activities. As distinguished from climate, weather consists of the short-term (minutes to months) variations of the atmosphere. Popularly, weather is thought of in terms of temperature, humidity, precipitation, cloudiness, brightness, visibility, and wind.

Here, the weather ontology takes account aspects of this definition by assigning the attributes aforementioned to the concept *Weather*, and the concepts *Weather Phenomenon*, *Atmospheric Phenomenon*, *Oceanic Phenomenon*, *Storm*, ... are parts of *Weather* (see *Fig*. 3.). Each attribute belongs to the concept *Weather* associated with an important degree (called membership value). The membership values are discovered from meteorological data via a perceptron.

Storm is defined [5] as any disturbed state of the atmosphere, especially as affecting the earth's surface, and strongly implying destructive or unpleasant weather.

We consider the majors of storm are *Tropical cyclones*, *T-storms*, and *Tornados*. *Lightning* as well as the *Thunder* are essential to *T-storms*. Both *Tornadoes* and *Tropical Cyclones* are atmospheric vortices, but are otherwise very different from each other, i.e., the life time of the former is mostly longer than a week, but the latter is mostly less than 10 minutes.

There are several kinds of Tropical cyclones such as *Tropical Depressions*, *Tropical Storms, Hurricanes* and *Typhoons*. These Tropical cyclones is distinguished according to intensity and region. In particular, *Tropical Depressions* cause wind speed less than

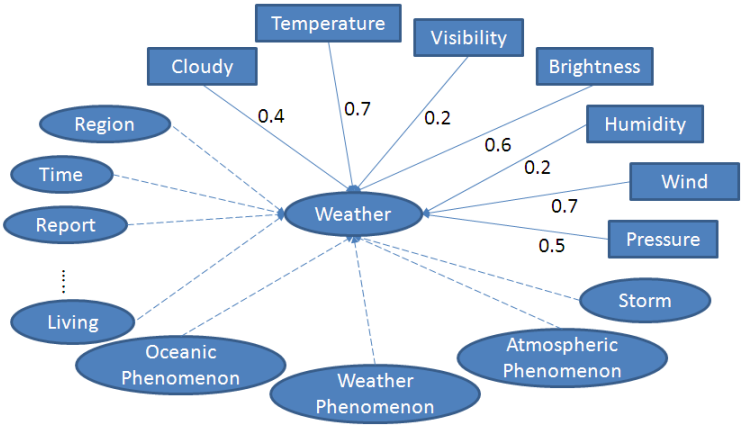


Fig. 3. First Level of Weather Ontoloy

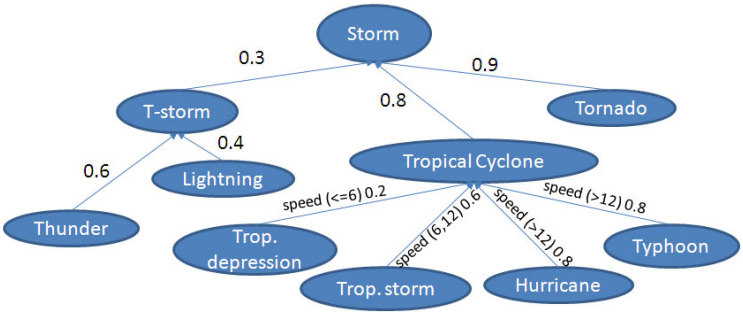


Fig. 4. A part of hierarchy of the concept Storm

or equal to Beaufort force 6, Tropical Storms' wind speed are stronger than 6 and less than 12, and hurricanes and typhoons are the same with even stronger wind.

According to intensity of kinds of storm, domain experts assign fuzzy membership values to subsumption relationships among them (see *Fig*. 4.).

Here, we present some other relations among phenomena in weather domain. In the *Fig*. 5., We consider several phenomena causing *Drought*.

Drought is a climatic anomaly, characterized by deficient supply of moisture resulting either from sub-normal rainfall, erratic rainfall distribution, higher water need or a combination of all the factors. Droughts are the resultant of acute water shortage due to lack of rains over extended periods of time affecting various human activities and lead to problems like widespread crop failure, unreplenished ground water resources, depletion in lakes/reservoirs, shortage of drinking water and, reduced fodder availability etc.

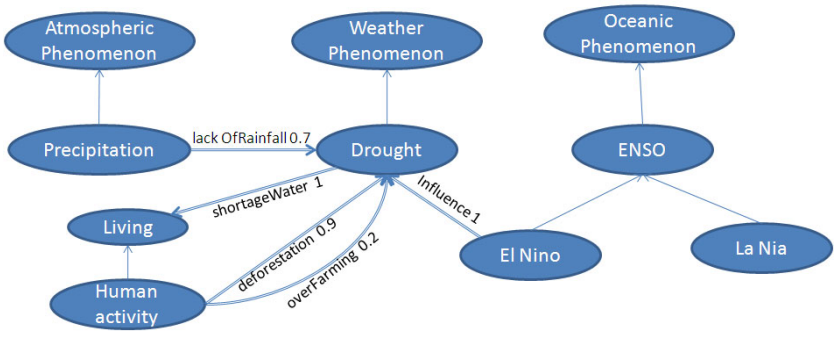


Fig. 5. Fuzzy relations causing Drought

The most concepts of leaf level in Weather fuzzy Ontology are concrete fuzzy concepts. Let's see the *Fig*. 7., the concrete fuzzy concept *Rain* associated with models linguistic qualifiers such as *Nil, Trace, Light, Moderate,* and *Heavy* to talk about the depth of rainfall.

6 Fuzzy Ontology Integration

Ontology Integration is very often realized for knowledge integration. Ontologies have well-defined structure and the result of ontology integration is also an ontology that should be satisfied two aforementioned criteria (1) and (2). With first criterion, It seems simple since one can make the sum of all sets of concepts, relations and axioms from component ontologies in the integration process. However, this may contain inconsistency in the sense that some of the component ontologies may be in conflict. The conflicts in the result of ontology integration will be solved by satisfying the second criterion [11,4].

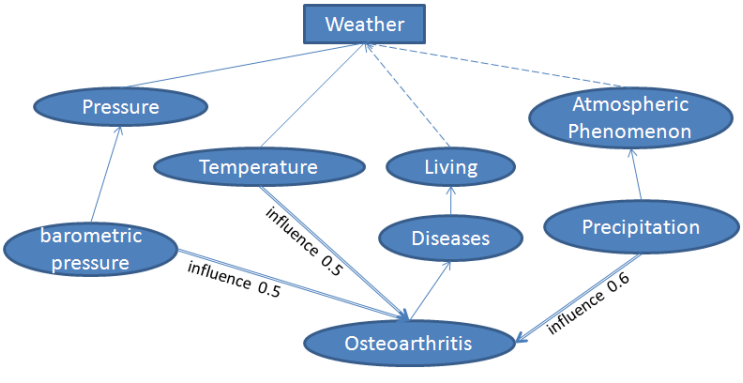


Fig. 6. Fuzzy relations *Influence* with different membership Values

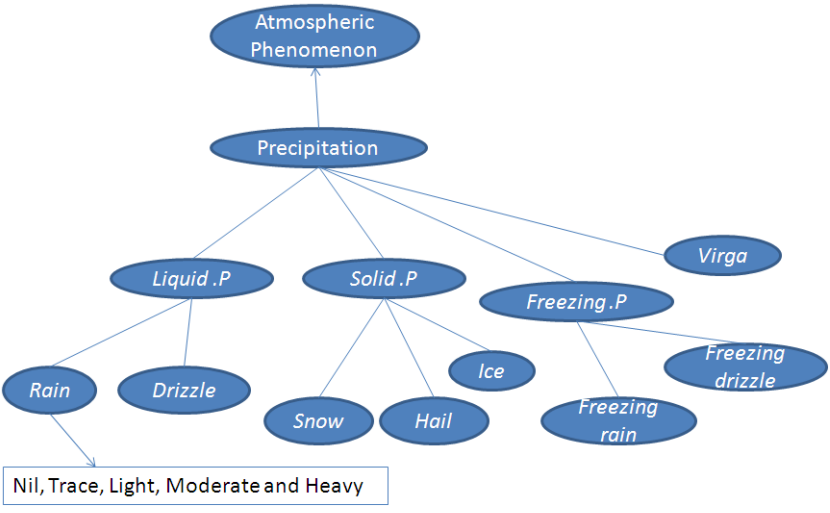


Fig. 7. Concepts of leaf level in Weather fuzzy Ontology are concrete fuzzy concepts

Conflicts between fuzzy ontologies may be considered on the following levels:

- Conflicts on concept level: The same concept has different fuzzy structures in different ontologies.

  The problem is addressed as follows: Let O_1 and O_2 be (A,V)-based ontologies. Assume that the concept c_1 belongs to O_1 and the concept c_2 belongs to O_2. We say that a conflict takes place in domain fuzzy concept level if $c_1 = c_2$ but $A^1 \neq A^2, V^1 \neq V^2$, or $f^1 \neq f^2$.
- Conflicts on relation level: The relations for the same concepts are different in different ontologies.

  The problem is addressed as follows: Let O_1 and O_2 be (A,V)-based ontologies. Let concepts c and c' belong to both ontologies. We say that an inconsistency takes place on relation level if $R_1^i(c, c') \neq R_2^i(c, c')$ for some $i \in 1, \ldots, m$.

These above mentioned problems are solved in our previous works by using consensus method.

7 Conclusions and Remarks

Weather forecast is a environment where there are inevitable uncertainties associated with weather phenomena. Even though, the analysis of language used in conventional forecasts are inherently and intentionally fuzzy such as rainfall is very heavy, it is mostly clear, and so on. Moreover, fuzzy logic is known to work in this domain. There have been many decision support systems for weather forecast using fuzzy logic to tackle difficult major in formulating forecast policy, and to cope up with vague and/or abnormal (chaotic) meteorological information. However, the current representation of meteorological information lacks of semantic or relationships among phenomena in weather

domain. The idea of integration fuzzy logic and ontology to represent meteorological knowledge in F-WIS are novel, necessary and effective for faciliating weather forecast and exchange knowledge with other agents. In this paper, we present the framework for building weather fuzzy ontology and demonstrate fuzzy values among in meteorological domain. The weather fuzzy inference system have been proposed, which takes fuzzy ontology and corresponding instances as its knowledge base.

Acknowledgement

This paper was partially supported by Polish Ministry of Science and Higher Education under grant no. N N519 407437.

References

1. Baader, F., Lutz, C., Sturm, H., Wolter, F.: Basic description logics. Journal of Logic and Computation (2003)
2. Zadeh, L.A.: Fuzzy sets. Inform. and Control 8, 338–353 (1965)
3. Calegari, S., Ciucci, D.: Integrating Fuzzy Logic in Ontologies. In: ICEIS (2), pp. 66–73 (2006)
4. Duong, T.H., Nguyen, N.T., Jo, G.S.: Fuzzy Ontology Integration Using Consensus Method. In: ICHIT 2010 Proceedings. ACM, New York (2010)
5. http://www.globalsecurity.org/wmd/library/policy/army/fm/3-6/3-6gl.htm
6. Hung, N.Q., Babel, M.S., Weesakul, S., Tripathi, N.K.: An artificial neural network model for rainfall forecasting in Bangkok, Thailand. Hydrol. Earth Syst. Sci. 13, 1413–1425
7. Lapedes, A., Farber, R.: Nonlinear signal processing using neural networks: prediction and system modeling. Technical Report LA-UR-87-2662, Los Alamos National Laboratory, Los Alamos, NM (1987)
8. Pedersen, T., Patwardhan, S., Michelizzi, J.: WordNet:Similarity-measuring the relatedness of concepts. In: Proceedings of NAACL (2004)
9. Mitra, A.K., Nath, S., Sharma, A.K.: Fog Forecasting using Rule-based Fuzzy Inference System. J. Indian Soc. Remote Sens. 36, 243–253 (2008)
10. Hansen, B., Riordan, D.: Fuzzy case-based prediction of cloud ceiling and visibility. In: 3rd Conference on Artificial Intelligence Applications to the Environmental Science. American Meteorological Society (2003)
11. Nguyen, N.T., Truong, H.B.: Consensus-based Method for Fuzzy Ontology Integration. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010. LNCS (LNAI), vol. 6422, pp. 480–489. Springer, Heidelberg (2010)
12. Taylor, J.W., Buizza, R.: Neural network load forecasting with weather ensemble predictions. IEEE Transaction of Power System 17(3), 626–632 (2002)
13. Lu, J., Li, Y., Zhou, B., Kang, D., Zhang, Y.: Distributed reasoning with fuzzy description logics. In: International Conference on Computational Science, vol. (1), pp. 196–203 (2007)

Cooperative Spectrum Sensing Using Individual Sensing Credibility and *Hybrid Quantization* for Cognitive Radio

Hiep Vu-Van and Insoo Koo*

School of Electrical Engineering, University of Ulsan
680-749 San 29, Muger 2-dong, Ulsan, Republic of Korea
iskoo@ulsan.ac.kr
http://mcsl.ulsan.ac.kr

Abstract. Cognitive radio (CR) has been considered as one of promising next-generation communication systems owning to its ability of sensing and making use of vacant channels that are currently unused by licensed user. Reliable spectrum sensing of licensed users is required for CR network to minimize interference to licensed users. However, due to the effects of the channel fading and shadowing, individual cognitive radio may not able to reliably detect the presence of a licensed user. Cooperative spectrum sensing has been proposed for improving sensing performance of CR network based on the cooperation of CR users in sensing signal from the license user. In this paper, we propose a cooperative spectrum sensing scheme using individual credibility and *hybrid quantization* for cognitive radios in order to improve sensing performance of global sensing performance and reduce communication traffic between CR users and the Fusion Center (FC). In the proposed scheme, few local sensing information that satisfies the credibility condition will be quantized using *hybrid quantization* scheme and further being reporting to the FC. The *hybrid quantization* scheme proposed in this paper can enhance sensing performance and reduce reporting bits of the sensing process based on the various numbers of quantized bits.

Keywords: Cognitive radio, cooperative spectrum sensing, valuation credibility, *hybrid quantization*.

1 Introduction

In recent years, due to the various applications basing on wireless communication technology, more bandwidth and higher bitrates are required in order to meet larger demand of the usage. According to the Federal Communications Commission's spectrum policy task force report [1], the utilization of licensed spectrum varies from 15% to 80%. Especially, in some cases, the utilization is only few percents of the full capacity. CR technology [2] has been proposed to solve the

* Corresponding author.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 528–537, 2011.

problem of ineffective utilization of spectrum bands. In cognitive radio network, both unlicensed and licensed users, termed the CR user and the Primary User (PU) respectively, operate. The scarcity of spectrum bands can be relieved by allowing some CR users to opportunistically access the spectrum assigned to the PU whenever the channel is free. Otherwise, CR users should vacate their frequency when the presence of a PU is detected. Therefore, reliable detection of the signal of the PU is a significant element of a CR network.

In order to ascertain the presence of a PU, CR users use one of several common detection methods such as the following: matched filter, feature, energy detections [2], [3]. If the CR user has limited information about the signals of the PU (e.g., only the local noise power is known), the optimal method would be energy detection [3]. In the energy detection, the frequency energy in the sensing channel is received in a fixed bandwidth W over an observation time window T to compare with the energy threshold and decide whether or not the channel is utilized. However, the received signal power may severely fluctuate due to multipath fading and the shadowing effect. Therefore, it is difficult to reach high reliable detection with only one CR user. Fortunately, we can obtain better usage detection by allowing some CR users to perform cooperative spectrum sensing [4].

In cooperative spectrum sensing, we rely on the variability of signal strength at various locations of CR users to improve sensing performance of the network with large number of CR users [4]. Cooperative spectrum sensing often takes the following three steps: sensing, reporting and decision making. The sensing step requires all CR users to perform spectrum sensing individually to make local decision. The reporting step transmits all of the local observations to the FC. The decision making step is accomplished by the FC using a data fusion rule to combine all of the local observations together into a global decision. There are some common data fusion rules such as AND, OR, Half Voting and Chair - Varshney rules [5] in which Chair - Varshney rule is the best combination rule with the lowest probability of error but needs knowledge about a priori information of the PU's activities, which is not available in practice. In order to solve the disadvantage of Chair - Varshney rule and further to implement Chair - Varshney rule in CR network without requirement information about PU's activities, Mansouri et al. have proposed an estimation algorithm based on counting rule [6].

Certainly, there are some disadvantages in the conventional cooperative spectrum sensing, that is, all CR users report local observations to the FC, which may result in collision in the control channel between CR users and the FC, and waste power consumption of CR nodes. Furthermore, if a low reliable local observation also contributes to make global decision at the FC, the reliability of the global decision will be decreased. On the other hand, all most of previous studies, the hard local decision (one bit decision) is applied because of the limited bandwidth of control channel while hard local decision always has lower sensing performance than that of soft decision (multi bits decision).

In this paper, we propose a cooperative spectrum sensing scheme using individual sensing credibility and *hybrid quantization* in order to improve global sensing performance of a CR network and maintain low traffic communication in the control channel between CR users and the FC. The *hybrid quantization* method offers various number of quantized bits, and further it can give more quantization level with less total bits of sending data through the control channel than those of the general quantization method. However, the *hybrid quantization* method requires more bits of sending data than that of hard decision. Therefore, individual sensing credibility was also proposed in the paper owing to its ability of decreasing the number of reports through the control. Basing on the individual credibility scheme, just few local sensing information that satisfies the credibility threshold will be report to the FC. Because of the disappearance of low reliable local decision at the FC, we can expect the improvement in the global sensing performance.

2 Cooperative Spectrum Sensing Using Individual Credibility with *Hybrid Quantization*

We consider a CR network including N CR users with various signal to noise ratios (SNR) between each CR user and the PU. We assume that all CR users individually perform spectrum sensing using the energy detection method.

In the proposed scheme, the local observation of each CR user will be checked for its credibility. The local observations which satisfy the credibility threshold will only be quantized and further being reported to the FC. The proposed scheme takes 3 steps as follows:

- *Step 1:* All CR users sense the signal from PU using energy detection method and determine the received signal power $E(i, j)$ where i is the index of the sensing time and j is the index of the CR users.
- *Step 2:* The credibility threshold will be exchanged for energy threshold. If the present local observation satisfies the energy threshold, it will be quantized using *hybrid quantization* and later being sent to the FC.
- *Step 3:* In the FC, all received quantization decisions will be combined to make a global decision using Chair - Varshney rule [5].

All above steps of the proposed scheme will be explained in more details by the following sub-sections:

2.1 Local Sensing with Energy Detector

All CR users receive signal energy in a fixed bandwidth W over an observation time window T. For the additive white Gaussian noise (AWGN) channel, the received signal power $E(i, j)$ on the j^{th} CR user at the i^{th} sensing time has the following distribution [6]:

$$E\left(i,j\right) \sim \begin{cases} H_0 : \chi^2_{2u} & \text{absence of PU's signal} \\ H_1 : \chi^2_{2u}\left(2\gamma_j\right) & \text{presence of PU's signal} \end{cases} \tag{1}$$

where χ^2_{2u} denotes a central chi-square distribution with $2u$ degrees of freedom and $\chi^2_{2u}(2\gamma_j)$ denotes a noncentral chi-square distribution with $2u$ degrees of freedom and a non-centrality parameter $2\gamma_j$. γ_j is the instantaneous SNR of the received signal at the j^{th} CR user and $u = TW$ is the time-bandwidth product.

For the hard-decision case, average probability of detection $P_d(j)$ and average probability of false alarm $P_f(j)$ of the j^{th} CR user can be calculated by following equations respectively [7]:

$$\begin{aligned} P_d(j) &= \text{Prob}\{E(i,j) \geq \lambda_j | H_1\} \\ &= Q_u\left(\sqrt{2\gamma_j}, \sqrt{\lambda_j}\right) \end{aligned} \tag{2}$$

$$P_f(j) = \text{Prob}\{E(i,j) > \lambda_j | H_0\} = \frac{\Gamma\left(u, \frac{\lambda_j}{2}\right)}{\Gamma(u)} \tag{3}$$

where λ_j denotes the energy threshold for local hard-decision of the j^{th} CR user, $\Gamma(a,x)$ is the incomplete gamma function which is given by $\Gamma(a,x) = \int_x^\infty t^{a-1}e^{-t}dt$. $\Gamma(a)$ is the gamma function, and $Q_u(a,b)$ is the generalized Marcum Q-function which is given by $Q_u(a,x) = \frac{1}{a^{u-1}}\int_x^\infty t^u e^{-\frac{t^2+a^2}{2}} I_{u-1}(at)dt$. $I_{u-1}(.)$ is the modified Bessel functions of the first kind and order $u-1$.

2.2 Evaluating the Credibility of the Local Sensing

In this step, the credibility of the local sensing information will be evaluated by comparing its probability of detection ($P_d(j)$) with the threshold β and by comparing its probability of false alarm ($P_f(j)$) with the threshold α where $P_d(j)$ and $P_f(j)$ are calculated by equations (2) and (3) respectively, and β and α are the double credibility thresholds for evaluating the credibility of CR user.

In order to make local decision according to credibility threshold, let's define double energy thresholds λ_{1j} and λ_{2j} corresponding to β and α respectively. Subsequently, we have

$$\begin{cases} \lambda_{1\text{j}} = \underset{\lambda_j}{\arg}\left(P_d(j) = \beta_j\right) \\ \lambda_{2\text{j}} = \underset{\lambda_j}{\arg}\left(P_f(j) = \alpha_j\right) \end{cases} \tag{4}$$

where the values of β and α are determined for satisfying the condition that λ_{2j} is not less than λ_{1j}.

To determine the range of double thresholds β and α for the credibility evaluating, let's define the optimal energy threshold $\lambda_{opt.j}$ for the case of one energy threshold which can minimize the error probability $P_e(j) = 1 - P_d(j) + P_f(j)$ of local sensing. Then, we have,

$$\lambda_{opt.j} = \underset{\lambda_j}{\arg}\left[\min\left\{(1 - P_d(j)) + P_f(j)\right\}\right] \tag{5}$$

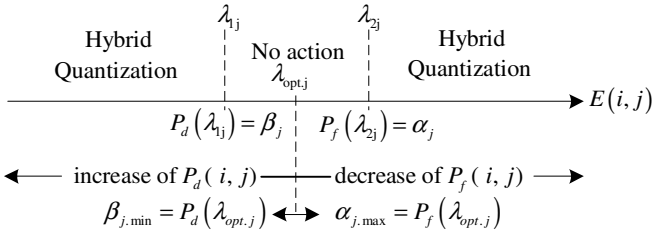


Fig. 1. Illustration of evaluating the credibility with *hybrid quantization*

Following that, we define the minimum value of β and the maximum value of α as β_{min} and α_{max} respectively where β_{min} and α_{max} are determined by the following equations:

$$\begin{cases} \beta_{\min} = \min\limits_{j=1,2,\ldots,N} (P_d(\lambda_{opt.j})) \\ \alpha_{\max} = \max\limits_{j=1,2,\ldots,N} (P_f(\lambda_{opt.j})) \end{cases} \tag{6}$$

Subsequently, the value of β and α will be selected in the range of $[\beta_{min}, 1)$ and $(0, \alpha_{max}]$ for the credibility checking respectively, and the corresponding λ_{1j} and λ_{2j} are determined by equation (4) in which λ_{2j} is alway larger than λ_{1j}

Fig.1 shows an illustration of evaluating the local credibility based on the double energy thresholds λ_{1j} and λ_{2j} where λ_{1j} and λ_{2j} can be changed according to the change of credibility thresholds β, α and γ_j.

As shown in the Fig.1, local sensing information of the j^{th} CR user at the i^{th} sensing time which satisfies the credibility threshold will be quantized using *hybrid quantization* scheme which will be described in the following subsection.

2.3 *Hybrid Quantization* Scheme for Local Sensing Information

In this step, *hybrid quantization* scheme will be used to quantize all local sensing information which satisfy the credibility threshold.

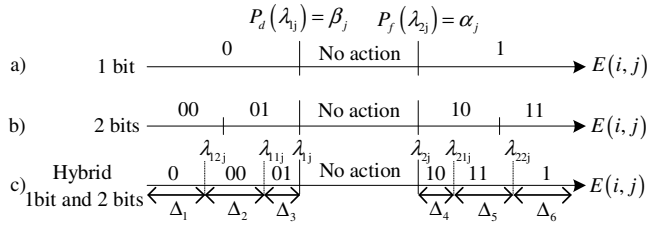


Fig. 2. Comparison of general quantization and *hybrid quantization*: a)Hard decision - 1 bit, b)General Quantization - 2 bits, c)*hybrid quantization* - 1 and 2 bits

It can be seen that the increasing number of quantized bits leads to improvement in sensing performance of CR network. However, more quantized bits spend more bandwidth of control channel, which may result in collision in communication between CR users and the FC. Therefore, in this paper we use *hybrid quantization* with two quantization bits, which is illustrated in Fig.2(c).

As shown in Fig.2(c), the proposed hybrid 2 bits-quantization scheme uses 1 bit or 2 bits to quantize sensing information of CR users. Consequently, the number of quantization levels can be enhanced to 6 levels compared with 4 levels of general 2 bits-quantization as well as the total number of bits sent through control channel will be decreased. As a result, the sensing performance at the FC can be improved while the collision in control channel is slightly reduced.

For the sake of comparing performance between *hybrid quantization* and general quantization, we assume that the occurring probability of each quantization level is the same and the number of sending bits in the case of hard decicion is k. Hence, we can show the table of performance comparison between general 2 bits-quantization and hybrid 2 bits-quantization as like the Table 1.

Table 1. Performance Comparison of Quantization Scheme

Method	No. of quantization level	Occurring probability of each quantization level	Number of sensing bits
Hard decision ($1\,bit$)	2	50%	k
General quantization ($2\,bits$)	4	25%	$2k$
Hybrid quantization ($1\,and\,2\,bits$)	6	17%	$1.7k$

In order to quantize the received energy $E(i,j)$, we define $u(i,j)$ as the quantization decision and $E(i,j)$ as the quantization input. The quantization process $Q(.)$ can be expressed as:

$$Q(E(i,j)) = u(i,j) \text{ if } E(i,j) \in \Delta_q, q = 1,2,...,6 \tag{7}$$

where $\Delta_q = [a_q, a_{q+1})$ is the quantization interval.

The quantization interval Δ_q can be illustrated as the Fig.2(c) where λ_{11j}, λ_{12j}, λ_{21j}, λ_{22j} can be defined as follows:

$$\begin{aligned}
\lambda_{11\mathrm{j}} &= \underset{\lambda_{\mathrm{j}}}{\arg}\left(P_d\left(j\right) = \left(\tfrac{2}{3}\beta_j + \tfrac{1}{3}\right)\right) \\
\lambda_{12\mathrm{j}} &= \underset{\lambda_{\mathrm{j}}}{\arg}\left(P_d\left(j\right) = \left(\tfrac{1}{3}\beta_j + \tfrac{2}{3}\right)\right) \\
\lambda_{21\mathrm{j}} &= \underset{\lambda_{\mathrm{j}}}{\arg}\left(P_f\left(j\right) = \tfrac{2}{3}\alpha_j\right) \\
\lambda_{22\mathrm{j}} &= \underset{\lambda_{\mathrm{j}}}{\arg}\left(P_f\left(j\right) = \tfrac{1}{3}\alpha_j\right)
\end{aligned} \tag{8}$$

Subsequently, the quantization decision will be made according to the value of received signal power $E(i,j)$ as following table:

Table 2. Quantization Decision

q	1	2	3	4	5	6
$[a_q, a_{q+1})$	$[0, \lambda_{12j})$	$[\lambda_{12j}, \lambda_{11j})$	$[\lambda_{11j}, \lambda_{1j})$	$[\lambda_{2j}, \lambda_{21j})$	$[\lambda_{21j}, \lambda_{22j})$	$[\lambda_{22j}, \infty)$
$u(i,j)$	0	00	01	10	11	1

2.4 Determining the Global Decision

All quantization decisions received in the FC will be combined to make a global decision by using the Chair-Varshney rule as follows:

$$\begin{cases} B(i) = 1 \text{ if } L = L_0 + \sum_{q=1}^{6}\sum_{j=1}^{n} L_{j,q} \geq 0 \\ B(i) = 0 \text{ otherwise} \end{cases} \tag{9}$$

where n is the number of quantization decision reported to the FC and

$$\begin{aligned}
L_0 &= \log\tfrac{P(H_1)}{P(H_0)} \\
L_{j,q} &= \log\tfrac{P(u(i,j)=q|H_1)}{P(u(i,j)=q|H_0)} \text{ if } u\left(i,j\right) = q
\end{aligned} \tag{10}$$

Due to the limited knowledge of PU's activities, we cannot calculate the exact values of L_0. Therefore, we have to use estimated one, we define a temporary global decision determined as function below:

$$\begin{cases} B_{tem}(i) = 1 \text{ if } 3n_1 + 2n_2 + n_3 \geq n_4 + 2n_5 + 3n_6 \\ B_{tem}(i) = 0 \text{ otherwise} \end{cases} \tag{11}$$

where n_q is the number of CR users who make the quantization decision $u(i,j) = q$ with $q = 1, 2, ...,6$.

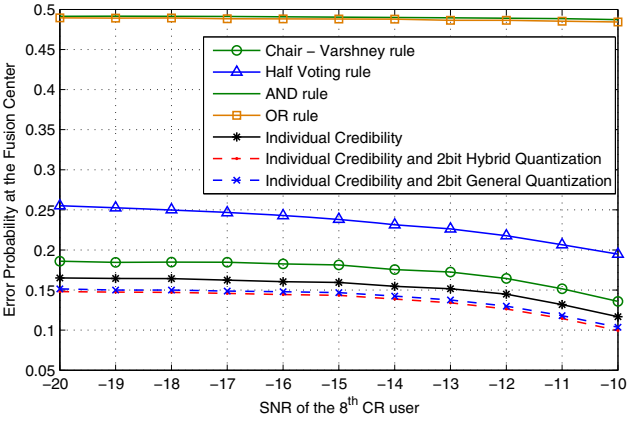


Fig. 3. Global Sensing Performance at the Fusion Center

We define c_{H_1} and c_{H_0} as the times that H_1 and H_0 occupy in history of temporary global decision B_{tem} respectively. As a result, the estimated value of L_0 is given as follows:

$$L_0 = \log \frac{P(H_1)}{P(H_0)} \approx \log \frac{c_{H_1}}{c_{H_0}} \tag{12}$$

Generally, we can determine value of $L_{j,q}$ by using following formula:

$$\begin{aligned} L_{j,q} &= \log \frac{P(\,u(i,j) = q|\,H_1)}{P(\,u(i,j) = q|\,H_0)} \\ &= \log \frac{P\left(a_q \le E\left(i,j\right) \le a_{q+1}\,|H_1\right)}{P\left(a_q \le E\left(i,j\right) \le a_{q+1}\,|H_0\right)} \\ &= \log \frac{P\left(E\left(i,j\right) \ge a_q\,|H_1\right) - P\left(E\left(i,j\right) \ge a_{q+1}\,|H_1\right)}{P\left(E\left(i,j\right) \ge a_q\,|H_0\right) - P\left(E\left(i,j\right) \ge a_{q+1}\,|H_0\right)} \\ &= \log \frac{Q\left(\sqrt{2\gamma_j}, \sqrt{a_q}\right) - Q\left(\sqrt{2\gamma_j}, \sqrt{a_{q+1}}\right)}{\Gamma\left(u, \frac{a_q}{2}\right)\frac{1}{\Gamma(u)} - \Gamma\left(u, \frac{a_{q+1}}{2}\right)\frac{1}{\Gamma(u)}} \end{aligned} \tag{13}$$

With the estimated values of L_0 and $L_{j,q}$, a global decision will be made according to the equation (9)

3 Simulation Results

In this section, the simulation results of the proposed scheme are analyzed in terms of the total number of bits sent through control channel, and sensing performance with comparison with other reference schemes such as AND rule, OR rule, Half Voting rule and Chair - Varshney rule [5].

For this simulation, we consider a CR network including *eight* CR users, each of which independently senses the presence of the PU by using energy

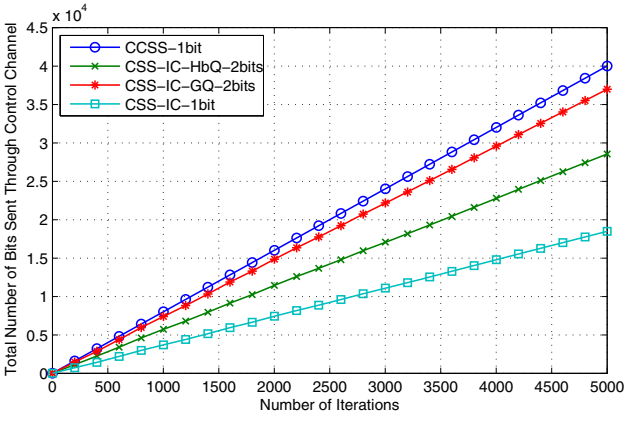


Fig. 4. Total Number of bits Sent through Control Channel

detection method. SNRs for the first seven users are given −20dB, −18.5dB, −16.0dB,−15.5 dB, −13.5dB, −11.5dB, −10dB respectively and the SNR of the 8^{th} CR user is changed from −20dB to −10dB. In order to evaluate sensing performance, we utilize the probability of error P_e which is defined as follows:

$$P_e = P_f P(H_0) + P_m P(H_1) \tag{14}$$

where P_f is probability of false alarm and P_m is probability of miss detection of the global decision.

For the sensing performance comparison at the FC, sensing performance of individual credibility method with hard decision at local CR users is also presented. Fig.3 shows the sensing performance at the FC and it is observed that according to SNR of the 8^{th} user, the individual credibility scheme with hard decision can achieve better performance comparison with Chair-Varshney rule. Undoubtedly, the combination of individual credibility and 2 bits-quantization scheme can reach better performance than both Chair-Varshney and individual credibility scheme. To recapitulate, the proposed scheme using hybrid 2 bits-quantization reaches better performance than the one of general 2 bits-quantization scheme and it is the best one at all.

Moreover, Fig.4 shows total number of bits sent through control channel according to four schemes: Conventional Cooperative Spectrum Sensing (CCSS-1bit), Cooperative Spectrum Sensing using Individual Credibility and Hybrid 2 bits-Quantization (CSS-IC-HbQ-2bits), Cooperative Spectrum Sensing using Individual Credibility and General 2 bits-Quantization (CSS-IC-GQ-2bits) and Cooperative Spectrum Sensing using Individual Credibility (CSS-IC-1bit). The figure shows about 50% reduction achieved in total number of bits sent through control channel by using CSS-IC-1bit comparison with CCSS-1bit. Likewise, the proposed scheme of CSS-IC-HbQ-2bits also represents about 25% reduction compared with CCSS-1bit and about 20% reduction compared with CSS-IC-GQ-2bits.

4 Conclusion

In this paper, we have proposed a cooperative sensing scheme using individual credibility and *hybrid quantization* in order to improve sensing performance of a CR network while reducing communication traffic through control channel between CR users and the FC. The simulation results prove that the proposed scheme can significantly reduce the error probability in global sensing decision at the FC and attain better performance than that of the optimal hard decision rule. Through the proposed scheme, the waste power consumption problem in the CR users and the collision in control channel can be reduced by the diminution of number of reporting bits of CR users.

Acknowledgement

"This work was supported by KRF funded by MEST (Mid-Carrier Researcher Program 2010-0009661)".

References

[1] Federal Communications Commission, "Spectrum Policy Task Force", Rep. ET Docket no. 02-135 (November 2002)
[2] Hur, Y., Park, J., Woo, W., Lim, K., Lee, C.-H., Kim, H.S., Laskar, J.: A wideband analog multi-resolution spectrum sensing (MRSS) technique for cognitive radio (CR) systems. In: Proc. IEEE Int. Symp. Circuit and System, Greece, May 21-24, pp. 4090–4093 (2006)
[3] Sahai, A., Hoven, N., Tandra, R.: Some fundamental limits on cognitive radio. In: Proc. Allerton Conf. on Communications, Control, and Computing, Monticello (October 2004)
[4] Ganesan, G., Li, Y.G.: Cooperative spectrum sensing in cognitive radio networks. In: Proc. IEEE Symp. New Frontiers in Dynamic Spectrum Access Networks (DySPAN 2005), Baltimore, USA, November 8-11, pp. 137–143 (2005)
[5] Chair, Z., Varshney, P.K.: Optimal data fusion in multiple sensor detection systems. IEEE Trans. Aerospace Electron. Syst., 98–101 (January 22, 1986)
[6] Mansouri, N., Fathi, M.: Simple counting rule for optimal data fusion. In: Proceedings of 2003 IEEE Conference on Control Applications, CCA 2003, June 23-25, vol. 2, pp. 1186–1191 (2003)
[7] Digham, F.F., Alouini, M.-S., Simon, M.K.: On the energy detection of unknown signals over fading channels. In: Proc. IEEE Int. Conf. Commun., Anchorage, AK, USA, May 11-15, pp. 3575–3579 (2003)

The Application of Fusion of Heterogeneous Meta Classifiers to Enhance Protein Fold Prediction Accuracy

Abdollah Dehzangi[1], Roozbeh Hojabri Foladizadeh[1], Mohammad Aflaki[2], and Sasan Karamizadeh[1]

[1] Faculty of Information Technology, Multi Media University, Cyberjaya, Selangor, Malaysia
[2] Faculty of Engineering, Multi Media University, Cyberjaya, Selangor, Malaysia
dehzangi@cse.shirazu.ac.ir, roozbeh.hojabry.fo08@mmu.edu.my, mohammad.aflaki06@mmu.edu.my, sasan.karamizadeh09@mmu.edu.my

Abstract. Protein fold prediction problem is considered as one of the most challenging tasks for molecular biology and one of the biggest unsolved problems for science. Recently, varieties of classification approaches have been proposed to solve this problem. In this study, a fusion of heterogeneous Meta classifiers namely: LogitBoost, Random Forest, and Rotation Forest is proposed to solve this problem. The proposed approach aims at enhancing the protein fold prediction accuracy by enforcing diversity among its individual members by employing divers and accurate base classifiers. Employed classifiers combined using five different algebraic combiners (combinational policies) namely: Majority voting, Maximum of Probability, Minimum of Probability, Product of Probability, and Average of probability. Our experimental results show that our proposed approach enhances the protein fold prediction accuracy using Ding and Dubchak's dataset and Dubchak et al.'s feature set better than the previous works found in the literature.

Keywords: Protein Fold Prediction Problem, Fusion of Heterogeneous Classifiers, Random Forest, Rotation Forest, LogitBoost, Feature Extraction, Majority Voting, Prediction Performance.

1 Introduction

Protein is considered as one of the most important macromolecules that plays a vital role in most of the biological reaction. Prediction of the tertiary structure of a protein from its primary structure is still remains as an unsolved issue for bioinformatics and molecular biology. Among the employed approaches, pattern recognition-based approaches attained promising results to address this problem [1]. In most of the previous studies, the protein fold prediction enhancement achieved by extracting new features [2-5]. However, most of the previously extracted features had high dimensionality which makes them inappropriate and impractical to be use for large protein databanks.

Despite critical role of feature extraction approaches on enhancing protein fold prediction accuracy, it is not the solitary approach to solve this problem. The employed classifier also plays a crucial role to solve the protein fold prediction problem. In this

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 538–547, 2011.

past two decades, varieties of classifiers such as *Bayesian- Based Classifiers*, *K-Nearest Neighbor* (*KNN)*, *Support Vector Machine (SVM)*, *Artificial Neural Network (ANN)*, Meta Classifiers, and *Hidden Markov Model (HMM)* [5-14] had been used to address this problem. Among all the employed classifiers, Fusion methods attained better results compared to use of an individual classifiers [3-5].

In 2002, Bologna and Appel proposed an ensemble of *Discretized Interpretable Multi Layer Perceptron (DIMLP)* to address this problem and achieved better results than its previous works [15]. DIMLP is an ANN-based classifier which use the staircase activation function (as its activation function) while its neurons are not fully connected in its first layer. DIMLP showed better performance than *Multi-Layer Perceptron (MLP)* (the most popular ANN-based classifiers) to address this problem. Despite enhancing the protein fold prediction performance, they could not achieve better than 61.1% prediction accuracy. In the following of Bologna and Appel, Chen et al. [10] proposed *Ensemble of Probabilistic Neural Network (EPNN)* another ANN-based classifier to tackle this problem. They achieved 63.1% prediction accuracy. But, the reported results achieved in a dramatically high time consumption.

In 2008, Krishnaraj and Reddy used AdaBoost.M1 and LogitBoost to tackle the protein fold prediction problem [13]. They achieved to their best results using LogitBoost (60.3%). Despite they could not enhance the prediction accuracy of this task; they achieved significant prediction accuracy in remarkably low time consumption. Recently, Random Forest and Rotation Forest, two recently proposed Meta classifiers were employed by Dehzangi et al., to tackle this problem [11, 12]. Random forest and Rotation Forest attained promising performances (62.7%, and 62.4% respectively) in a lower time consumption than AdaBoost.M1 and LogitBoost.

In this study, a fusion of three Meta classifiers namely: Random Forest, Rotation Forest, and LogitBoost proposed to challenge the protein fold prediction problem. The proposed approach aimed at enhancing the diversity among the classifiers used as its members by employing three diverse, consistent, and accurate classifiers which attained promising results to solve this problem. Our experimental results showed that the proposed method attained better results compared to previous works found in the literature using Ding and Dubchak benchmark [16] and Dubchak et al. feature set [17]. We achieved 65.3% protein fold prediction accuracy in lower time consumption than Chen et al [10].

2 Methods and Tools

Fusion method aims at enhancing the prediction performance of the classification task better than using a single classifier by constructing a collection of diverse and yet accurate classifiers. It undertakes statistical, computational, and representational issues that are more likely to occur for an individual classifier and affect its prediction performance [18].

The employed classifiers (Random Forest, Rotation Forest, and LogitBoost) used as a part of proposed fusion method were chosen based on their individual prediction accuracy reported by previous works [11-13], distribution of the classified proteins

for each fold individually, the number of proteins in each fold [11, 12], and the experimental studies conducted by Dehzangi et al. [6, 11, 12]. In the following sections, each method is introduced in brief.

2.1 Random Forest

Random Forest is a recently proposed classifier aims at building a diverse Meta classifier based on bagging and feature selection approaches [19]. Similar to bagging [20], Random Forest takes a bootstrap sample of training data multiple times and train a group of base learners (B base learners) called weak learner. In final, it combines all the weak learners using unweighted majority voting. But, instead of all feature set to train each base learners in bagging, in Random Forest each weak learner trains with selected features among a set of F randomly chosen features between T total features contained in feature vector (that's why it is called Random) to encourage diversity between base classifiers. The best features are selected based on the Gini Index between the F selected features [19]. This modification enforces the diversity in Random Forest better than in Bagging without really compromising the prediction accuracy of its base learner [21].

Random Forest is considered as an appropriate model to handle imbalance dataset. It is able to provide an empirical approach to trace variable interactions. In this study, Random Forest implemented in data mining toolkit WEKA (Waikato Environment for Knowledge Analysis) version 3.6.2 was used for classification [22]. For Random Forest, the number of base learners set to 80. It was shown by Dehzangi et al. that applying 80 base learners for Random Forest attains its highest performance comparing to other number of base learners in the range between 30 and 600 [11].

2.2 Rotation Forest

Another classifier used in this study was Rotation forest. Similar to Random Forest, Rotation Forest is a novel Meta classifier based on bagging approach aims at building a diverse Meta classifier. It designed to modify the Random Forest to enhance diversity within its base learners using feature extraction instead of feature selection [23].

In Rotation Forest, the feature vector (totally T features) is randomly split to F subsets which F is the parameter of the algorithm. Then it applied *Principle Component Analysis (PCA)* to each subset separately and transforms them to a new feature space. In the final, it combines all the transforms feature set and combines them to a new feature set (PS features which $PS \leq T$) to train the base learners. Rotation Forest consider as a diverse Meta classifier which is capable to be used with varieties of base learners [24]. In this study, Rotation Forest implemented in WEKA using J4.8 [22] (designed version of C4.5 [25] decision tree for WEKA) with 100 base learners was used. C4.5 is an algorithm used to generate a decision tree which is an extension of Quinlan's earlier ID3 algorithm [26]. It was shown by Kuncheva and Rodriguez that C4.5 is an appropriate base classifier to be used for Rotation Forest [24]. In addition, it was shown by Dehzangi et al. that using 100 base classifiers for Rotation Forest attained the best results among other number of base learners in the range between 10 and 200 [12].

2.3 LogitBoost

Boosting is a sequential algorithm in which each new base predictor (also called weak learner) is built based on the performance of the previously generated predictors [27]. In the Boosting-based methods, firstly, a base predictor is applied to the training dataset [27]. It updates the weights of the training data which is initially equal. In the following iterations, the training data with updated weights will be given as the input to the base learner. LogitBoost as a kind of boosting-based approach was originally proposed by Friedman et al. [27] to address the AdaBoost.M1 inefficiencies while dealing with noisy data. To be more robust than AdaBoost.M1 in noisy environments, LogitBoost tries to minimize the logistic lost function instead of minimizing the exponential lost function using in AdaBoost.M1 [28]:

LogitBoost aims at maximizing the likelihood by fitting an additive symmetric logistic model using Newton step. LogitBoost is relatively fast (depends on its base learner) and robust classifier in noisy environments [27]. Despite better performance of the AdaBoost.M1 than LogitBoost for varieties of studies, it was shown that LogitBoost attained better results than AdaBoost.M1 for the protein fold prediction task [13]. In this study, LogitBoost classifier implemented in WEKA using decision stump as its base learners was used. It was shown by Krishnaraj and Reddy that 100 base learner is an appropriate number of base learners to be use for LogitBoost to tackle the protein fold prediction problem (which is used in study as well)[13].

3 Dataset and Features

In this study, to compare our results with the previously achieved results, Ding and Dubchak's benchmark was used [16]. This benchmark is considered as one of the most popular benchmarks using pattern recognition-based approaches to address the protein fold prediction problem. This benchmark comprises of train and test datasets. The train dataset contains 311 proteins with less than 35% sequential similarities with each other extracted from *Protein Data Bank (PDB)*. These proteins belong to 27 most populated fold in PDB. The test dataset has 383 proteins with less than 40% sequential similarities extracted from *Structural Classifications of Proteins (SCOP)* [29].

We also used one of the most popular feature sets introduced by Dubchak et al [17]. This feature set contains six feature groups namely: *Amino Acid Composition (C), Predicted Secondary Structure Based on Normalized Frequency of α-Helix Residue (S), Hydrophobicity (H), Normalized Van Der Waals Volume (V), Polarity (P), and Polarizability (Z)*. The first feature group was extracted based on syntactical properties of the amino acids and present composition percentage of amino acids and contain 20 features. The rest of the feature groups were extracted based on physicochemical properties of the amino acids. In each feature groups, amino acids categorized into three groups based on their specific attributes and then 21 features based on: the composition of amino acids in each group along the amino acid sequence of protein (three features); transition frequencies between groups (group one to group two, two to three, and one to three); and the distribution amino acids in each group (where the first residue of a given constituent is located, and where 25%, 50%, 75%, and 100% of that constituent are contained) along the amino acid sequence (fifteen features) [17].

4 Results and Discussion

In this study, standard Q percentages accuracy measurement as popular accuracy measurement applied to make it easier to compare our results to the results achieved by previous works [16].

To combine employed classifiers and to avoid further complication, five linear *algebraic combiners (combinational policies)* namely: *Average of Probabilities, Product of Probabilities Maximum of Probabilities, Minimum of Probabilities, and Majority Voting (MV)* were used [26, 30]. Employed algebraic combiners were chosen based on three main factors namely: complexity, simplicity, and popularity. Employed algebraic combiners have been widely used for different tasks and attained better performances compared to the other simple, linear algebraic combiners [30].

The structure of our proposed fusion method is shown in Fig.1. To find the best combinational policy among the employed combinational policies, fusion of employed method using each combinational policy applied to different combinations of feature groups. In this study, to investigate the effectiveness and discriminatory information contain in each employed feature groups, as well as finding better and more effective combination of feature groups, eleven different combinations of feature groups instead of six combinations of feature groups used by Ding and Dubchak [16] were used. Employed combination of feature groups were chosen based on the discriminatory information contains in each feature groups and previous works found in the literature [6, 16]. The achieved results are shown in Table 1.

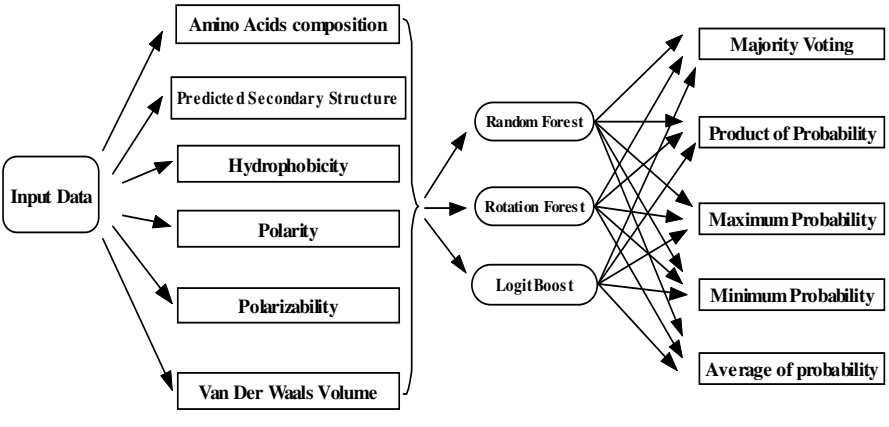


Fig. 1. The structure of our proposed fusion of heterogeneous Meta classifiers

As shown in Table 1, using Majority voting as the algebraic combiners obtained highest prediction performance compared to use of other employed algebraic combiners employed in this study. Despite merits of using fusion classifier other than an individual classifier, it also has two main insufficiencies. First, using fusion methods increase the computational complexity of classification task higher than using a single

classifier. Second, despite enhancing the classification performance by boosting diversity among the classifiers, it also boosts noise among its employed classifiers which is dramatically affect prediction performance.

Table 1. Results achieved (%) for combinations of the proposed fusion of Meta classifiers using five different combinational policies, and eleven different combinations of feature groups.

Voting policy	C	CX	CXV	CXZ	CXP	CXH	CXHV	CXHP	CXHPV	CXHPZ	CXHPZV
Maximum Probabilities	51.4	57.7	57.2	56.9	55.4	58.2	56.4	54.8	56.1	55.4	59.0
Minimum Probabilities	55.1	59.0	57.2	58.5	58.2	60.3	57.4	57.7	57.7	56.7	59.8
Majority Voting	56.9	61.6	60.6	62.9	61.4	61.4	60.1	62.1	61.1	61.6	***65.3***
Product of Probabilities	53.8	59.0	59.8	62.4	60.8	61.4	59.8	61.1	62.4	61.4	59.8
Average of Probabilities	53.0	58.5	59.3	59.3	58.2	60.1	57.4	58.0	58.0	57.2	60.6

Table 2. Comparison of the our results (%) with the previous reported results found in the literature using Ding and Dubchak [16] dataset, and Dubchak et al., feature set [17] (in each row, and from left to right are the reference number, the method, the combination of features they used, and in final the prediction accuracy they achieved)

[16]	AvA(SVM)	C+X+H+P	56.4
[13]	AdaBoost.M1	C+X+H	58.2
[13]	LogitBoost	C+X+H+P+V	60.3
[15]	DIMLP	C+X+H+P+Z+V	61.1
[9]	RS1_HKNN_K125	C+X+H+P+Z+V	60.0
[9]	RS1_KHNN_K25	C+X+H+P+Z+V	60.3
[1]	MOFASA	C+X+H+P+Z+V	60.0
[10]	EPNN	C+X+H+P+Z+V	63.1
[8]	ALH	C+X+H+P+Z+V	60.8
[7]	RBF Majority Voting Fuse	C+X+H+P+Z+V	49.7
[7]	RBF Bayesian Fuse	C+X+H+P+Z+V	59.0
[5]	OET-KNN	C+X+H+P+Z+V+ (a hundred features)	62.1
[4]	Data Fusion, NN+ HLA	C+X+H+P+Z+V	56.4
[11]	Random Forest (80 base learner)	C+X+H+P+Z+V	62.7
[12]	Rotation Forest (100 base learner)	C+X+H+P+Z+V	62.4
[31]	Kernel Combinational Methodology	C+X+H+P+Z+V	58.6
[32]	SHMM	Amino Acids Sequence Based Feature	51.6
[33]	OvO SVM	C+X+H+P+Z+V	58.3
This Paper	Rotation Forest + LogitBoost (Fusion)	C+X+H+P+Z+V	63.7
This paper	Our Proposed classifier	C+X+H+P+Z+V	**65.3**

To address the first issue, relatively fast with low computational complexity classifiers use to build a fusion method. To address the second issue, several approaches have been proposed in the literature. One of the most effective approaches is to use appropriate algebraic combiners which do not boost noise in the way that it boosts the prediction

performance of the fusion method. Among algebraic combiners, voting policies show better performance to undertake this issue. By combining final decision of each individual classifier used in fusion method, voting approaches technically isolate employed classifiers in fusion methods from each other. In this way, the effectiveness of noise on each employed classifier in fusion method will not affect the other classifiers performance. At the same time, by combining the final results of employed classifiers in fusion method, it boosts the diversity in fusion classifier [30].

As shown in Table 1, by using fusion of our employed classifiers combined with majority voting algebraic combiner, we achieved 65.3% accuracy using combination of all employed feature groups. Comparison of the achieved results with previously reported results (Table 2) showed that we achieved better results than Chen et al., [10] who used *Ensemble of Probabilistic Neural Network (EPNN)* using less classifier and dramatically lower time consumption (63.1%); and Bologna and Appel [15], who used *Descretized Interpretable Multi Layer Perceptron (DIMLP)* with better time and space consumption (61.1%). As it mentioned earlier, while several studies such as [5-9], achieved better performance than our achieved prediction accuracy by introducing new features (more than a thousand features); our method outperformed even their proposed classifiers using same set of features (Dubchak et al., feature set [17]) and same benchmark (Ding and Dubchak benchmark [16]) which have been widely used by previous works (Table 2).

To study the effectiveness of our employed classifiers to the achieved results, each classifier, individually excluded from the combination of the two remaining classifiers. Due to better prediction performance using majority voting policy, it used to combine the remaining classifier with together. The results achieved by this experiment are shown in Table.3. As shown in this table, the highest result achieved by combining LogitBoost, and Rotation Forest using combination of the all employed feature groups (63.7%). The achieved result showed the importance and effectiveness of each employed classifier in the proposed fusion method. Omitting each classifier reduced the prediction performance of the fusion method compared to the use of all employed classifiers simultaneously.

Table 3. Results (%) achieved using the proposed method with majority voting as combinational policy excluded each of the classifiers which mentioned in the first column

Methods	C	CX	CXV	CXZ	CXP	CXH	CXH V	CXH P	CXH PV	CXH PZ	CXH PZV
The proposed Method	56.9	61.6	60.6	62.9	61.4	61.4	60.1	62.1	61.1	61.6	***65.3***
Our Method except LogitBoost	57.7	60.1	60.6	60.8	60.3	62.4	59.5	61.4	61.1	63.2	63.5
Our Method except Random Forest	54.3	59.0	59.3	60.1	58.2	60.1	59.3	59.8	59.0	60.6	**63.7**
Our Method except Rotation Forest	55.1	60.1	60.1	61.9	58.5	60.8	58.5	58.2	59.5	60.8	61.4

5 Conclusion

In this study, fusion of heterogeneous classifiers namely: Random Forest (using 80 base learners), Rotation Forest (using 100 C4.5 decision trees as its base learner), and

LogitBoost (using 100 decision stump as its base learner) proposed to tackle the protein fold prediction problem. Employed classifiers were chosen based on their individual prediction accuracy reported by previous studies, distribution of the classified proteins depends on each fold individually, the number of proteins in each fold, and the experimental studies that were conducted by the Dehzangi et al [12]. To avoid further complication and maintain simplicity of the proposed method, five simple linear algebraic combiners namely: *Average of Probabilities, Product of Probabilities Maximum of Probabilities, Minimum of Probabilities, and Majority Voting* were used to combine employed classifiers.

Our experimental results showed that the majority voting is capable of achieving better prediction performance compared to the other employed algebraic combiners. Excluding each classifier individually from the fusion of remain classifiers which achieved lower prediction performance than using all employed classifiers simultaneously as the part of proposed fusion method showed the importance of all employed classifiers to enhance the protein fold prediction accuracy. Using proposed fusion of heterogeneous classifiers, we achieved 65.3% prediction accuracy better than previously reported results using same dataset (Ding and Dubchak's benchmark [16]), and same set of features (Dubchak et al., feature groups [17]) which shows the effectiveness of the proposed classifier.

References

1. Shi, S.Y.M., Suganthan, P.N., Deb, K.: Multi class protein fold recognition using multiobjective evolutionary algorithms. In: IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, pp. 61–66 (2004)
2. Dehzangi, A., Khosravi, B.G.: Introducing Novel Physicochemical Based Features to Enhance Protein Fold Prediction Accuracy. In: Proceeding in: IEEE International Conference on Computer Design and Applications, pp. 592–596 (2010)
3. Ghanty, P., Pal, N.R.: Prediction of Protein Folds: Extraction of New Features, Dimensionality Reduction, and Fusion of Heterogeneous Classifiers. IEEE Transaction on Nanobioscience 8(1), 100–110 (2009)
4. Lin, K.L., Li, C.Y., Huang, C.D., Chang, H.M., Yang, C.Y., Lin, C.T., Tang, C.Y., Hsu, D.F.: Feature Selection and Combination Criteria for Improving Accuracy in Protein Structure Prediction. IEEE Transactions on Nanobioscience 6(2), 186–196 (2008)
5. Shen, H.B., Chou, K.C.: Ensemble Classifier for Protein Fold Pattern Recognition. Bioinformatics 22(14), 1717–1722 (2006)
6. Dehzangi, A., Amnuaisuk, S.P., Ng, K.H.: Investigating the Influence of Combined Features to Classifiers' Performance: A Comparison Study on a Protein Fold Prediction Problem. In: 6th IEEE International Conference on Information Technology in Asia, pp. 213–217 (2009)
7. Hashemi, H.B., Shakery, A., Naeini, M.P.: Protein Fold Pattern Recognition Using Bayesian Ensemble of RBF Neural Networks. In: International Conference of Soft Computing and Pattern Recognition SOCPAR, pp. 436–441 (2009)
8. Kecman, V., Yang, T.: Protein Fold Recognition with Adaptive Local Hyper plane Algorithm. In: 6th Annual IEEE conference on Computational Intelligence in Bioinformatics and Computational Biology, Nashville, Tennessee, USA, pp. 75–78 (2009)

9. Nanni, L.: Ensemble of classifiers for protein fold recognition. In: New Issues in Neurocomputing: 13th European Symposium on Artificial Neural Networks, pp. 850–853 (2006)
10. Chen, Y., Zhang, X., Yang, M.Q., Yang, J.Y.: Ensemble of Probabilistic Neural Networks for Protein Fold Recognition. In: 7th IEEE International Conference on Bioinformatics and Bioengineering, pp. 66–70 (2007)
11. Dehzangi, A., Amnuaisuk, S.P., Dehzangi, O.: Using Random Forest for Protein Fold Prediction Problem: An Empirical Study. Journal of Information Science and Engineering 26(6) (2010)
12. Dehzangi, A., Amnuaisuk, S.P., Manafi, M., Safa, S.: Using rotation forest for protein fold prediction problem: An empirical study. In: Pizzuti, C., Ritchie, M.D., Giacobini, M. (eds.) EvoBIO 2010. LNCS, vol. 6023, pp. 217–227. Springer, Heidelberg (2010)
13. Krishnaraj, Y., Reddy, C.K.: Boosting methods for Protein Fold Recognition: An Empirical Comparison. In: IEEE International Conference on Bioinformatics and Biomedicine, pp. 393–396 (2008)
14. Lampros, C., Papaloukas, C., Exarchos, K., Fotiadis, D.I., Tsalikakis, D.: Improving the protein fold recognition accuracy of a reduced state-space hidden Markov model. Computers in Biology and Medicine 39(10), 907–914 (2009)
15. Bologna, G., Appel, R.D.: A comparison study on protein fold recognition. In: Proceedings of the Ninth International Conference on Neural Information Processing, pp. 2492–2496 (2002)
16. Ding, C., Dubchak, I.: Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17(4), 349–358 (2001)
17. Dubchak, I., Muchnik, I., Kim, S.K.: Protein folding class predictor for SCOP: approach based on global descriptors. In: Proceedings in the 5th International Conference on Intelligent Systems for Molecular Biology, pp. 104–107 (1997)
18. Dietterich, T.G.: Ensemble methods in machine learning. In: Kittler, J., Roli, F. (eds.) MCS 2000. LNCS, vol. 1857, pp. 1–15. Springer, Heidelberg (2000)
19. Breiman, L.: Random Forest. Machine learning 45(1), 5–32 (2001)
20. Breiman, L.: Bagging Predictors. Machine Learning 24, 123–140 (1996)
21. Livingston, F.: Implementation of Breiman's Random Forest Machine Learning Algorithm, Machine Learning. ECE591Q (2005)
22. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1) (2009)
23. Rodríguez, J.J., Kuncheva, L.I., Alonso, C.J.: Rotation forest: A new classifier ensemble method. IEEE Transactions 28(10), 1619–1630 (2006)
24. Kuncheva, L.I., Rodríguez, J.J.: An experimental study on rotation forest ensembles. In: Haindl, M., Kittler, J., Roli, F. (eds.) MCS 2007. LNCS, vol. 4472, pp. 459–468. Springer, Heidelberg (2007)
25. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco (1993)
26. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification. Wiley, New York (2001)
27. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting. Annals of Statistics 28(2), 337–407 (2000)
28. Freund, Y., Schapier, R.E.: A Short Introduction to Boosting. Journal of Japanese Society for Artificial Intelligence 14(5), 771–780 (1999)
29. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C.: SCOP: a structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology 247, 536–540 (1995)

30. Bauer, E., Kohavi, R.: An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Machine Learning 36, 105–139 (1999)
31. Damoulas, T., Girolami, M.A.: Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection. Bioinformatics 24, 1264–1270 (2008)
32. Bouchaffra, D., Tan, J.: Protein Fold Recognition using a Structural Hidden Markov Model. In: 18th International Conference on Pattern Recognition, pp. 186–189 (2006)
33. Shamim, M.T.A., Anwaruddin, M., Nagarajaram, H.A.: Support Vector Machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs. Bioinformatics 23(24), 3320–3327 (2007)

A Single Machine Scheduling Problem with Air Transportation Decision

P.S. You[1], Y.C. Lee[2], Y.C. Hsieh[3], and T.C. Chen[4]

[1] Graduate Institute of Marketing and Logistics/Logistics,
National ChiaYi University
580 Sin-Min Road, Chia-Yi 60054, Taiwan
psyuu@mail.ncyu.edu.tw
[2] Department of Security Technology and Management, WuFeng University, Taiwan
117, Sec. 2, Jianguo Rd., Minsyong Township, Chiayi 621, Taiwan
yclee@mail.efc.edu.tw
[3] Department of Industrial Management, National Formosa University
Huwei, Yunlin 632, Taiwan
yhsieh@nfu.edu.tw
[4] Department of Information Management, National Formosa University
Huwei, Yunlin 632, Taiwan
tchen@nfu.edu.tw

Abstract. Production scheduling and transportation strategies in manufacturing systems are two important topics in the fields of production or supply chain management. These two decisions are usually separately discussed in previous works. However, since a business may simultaneously face these two problems, it needs to develop a model to simultaneously deal with these two problems so as to minimize the total costs. This paper deals with a production scheduling and air-transportation problem with a single-machine and multi-delivery destinations. A heuristic approach is developed to deal with this problem. Computational results show that the heuristic approach can produce nearly optimal solutions for small-scale problem and can produce feasible solutions that commercial optimization software, Lingo, can not solve within a reasonable amount of time.

Keywords: assembly scheduling, transportation allocation, heuristic, Assembly delays.

1 Introduction

Upon receiving orders from the customer, many manufacturers have to deal with the assembly scheduling and the transportation allocation problems. As pointed by Li et al. [7-9], components in the PC industry are usually stored in warehouses. Upon receiving orders from customers, PC companies must arrange their orders' processing sequence on the assembly line to assemble the components for products. An order is then ready to be shipped to the customer after its assembly process is done. Usually, each order may come from a distinct customer and has a different due date. Thus, the orders' processing sequence and

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 548–556, 2011.

flight allocation decisions have significant impact on whether the orders can be shipped to their customers on time and if there will be products in their stores. In making these decisions, some business practices show that an order may be transported by commercial flights if the order is late from the assembly and has missed its schedule for the regular cargo flight's departure time. However, the commercial flights are usually much more expensive than regular cargo flights. From an economic point of view, a firm will try to hold as few inventories as possible to reduce inventory holding cost and delivery orders to customers on time to reduce earliness or tardiness penalties. Thus, the synchronization between assembly and air transportation plays an important role.

A number of works concerned with production or transportation scheduling have been widely studied [1-16]. Studies on these two problems usually focus on the processing sequence of orders and the vehicle routing schedules, respectively. Regarding the production scheduling problem, the goal of this research is similar to previous ones. On the other hand, regarding the transportation allocation problem, this paper focuses on the problem of how to allocate processed orders to various transport modes.

Zuo et al. [16] developed heuristics for production planning and distribution strategies for an agricultural production and distribution system. Ruiz-Torres and Tyworth [12] used the simulation approach to explore the combined problem of production scheduling and transportation in a logistic network. Chen [2] investigated the problem of integrating production and transportation scheduling in a make-to-order environment. This paper aims to minimize the total cost of transportation, tardiness penalties and overtime production. Garcia et al. [5] dealt with a production and delivery scheduling problem in a no-wait scenario with multiple production plants. The orders were required to be manufactured and delivered immediately via a homogenous vehicle fleet to customers. Tyan et al. [14] evaluated three freight consolidation strategies for a third party logistics provider that provided door-to-door distribution services for major notebook manufacturers in Taiwan.

The purpose of this paper is to simultaneously develop the decisions for production and transportation scheduling so as to minimize overall costs. This paper formulates a combined model for the problem, which is a mix integer programming problem. Since the assembly scheduling problem has been proven to be NP-hard, we developed a heuristic to solve this model.

The rest of this paper is organized as follows: Section 2 outlines all assumptions made and formulates the problem a constrained integer linear programming model. Section 3 then presents the solution methodology. Section 4 tests the performance of the proposed heuristic, using numerical examples to compare with the well-known commercial software, Lingo 10.0. Conclusions are finally drawn in Section 5.

2 Problem Formulation

Consider a manufacturer which sells a certain product in the make-to-order model. This manufacturer receives N orders from distinct customers. Order-i's

shipping destination, order due date and order size are G_i, d_i and Q_i, respectively. Each order needs to be processed by assembling components on a single machine. A processed order can not be split and allocated to more than carriages and delivered separately to its destination. The modes of the carriages include regular and commercial flights. The difference between these two modes is that the departure times of the regular cargo flights are predetermined while commercial flights are assumed to have more flexibility because orders can be transported immediately after assembly. The decision maker makes an initial schedule in which all orders are assigned to regular flights. However, an order is transported by non-regular flights once its assembly completion time is beyond the departure time of the scheduled regular flight. Inventory holding cost incurs if there is a waiting time before the completed order can be shipped by flights to its destination. We let denote the total number of regular cargo flights and D_f be the departure time of regular cargo flight f. The transportation cost per unit product, when allocated to regular cargo flight f, is dented by c_f^r. The transportation cost per unit product when allocated to a commercial flight is dented by c_f^s. In addition to the transportation cost, there exists a delivery earliness penalty cost if the arrival time of the order reaches its destination airport before its order due date, and a delivery tardiness penalty cost if the order reaches its destination airport beyond its order due date. We assume that the delivery earliness penalty cost and the delivery tardiness penalty cost per unit product and per unit time for order-i is α_i and β_i, respectively. Our objective is to minimize total cost, which consists of inventory holding cost, the transportation cost, the delivery earliness penalty costs, and the delivery tardiness penalty costs. Below, we will develop the mathematical model of the problem to find decisions in regard to assembly scheduling and transportation scheduling. Prior to formulating the model, we further make the following assumptions:

- The time and cost taken of transporting a completed order by local transportation from the manufacturer to an airport, together with the assembly setup time and cost of each order, are included in the assembly time and cost.
- An orders' assembly processing time is directly proportional to its order quantity.
- Business processing time and cost, as well as the load time and load cost of each carriage, are included in the transportation time and transportation cost.
- Order fulfillment is considered to be achieved when the order reaches its destination on time.
- Orders can not be split and allocated to more than one carriage and delivered separately.
- When an order cannot meet its scheduled regular carriage's departure time, it is immediately transported by an non-regular carriage at its completed time at a higher cost.

The combined model aims to minimize the overall cost by determining the orders' assembly sequence and allocating orders to existing carriages. The factors which are taken into account in the model include: (a) the number of available carriages for the distribution planning horizon, (b) the departure and arrival time of the carriages, (c) the designated capacity and its corresponding transportation cost, and (d) the possible capacity in each carriage with its corresponding carriage cost. The notation used to formulate the problem is summarized as follows:

i = the order index $i = 0, 1, \cdots, N+1$ where $i = 0$ and $i = N+1$ are aritificial orders,
G_f = the destination of regular flight f,
Q_i = the quantity of order i,
P_i = the processing time to assemble order i,
h_i = the per hour earliness penalty of order i,
D_f = the departure time of regular flight f,
A_f = the arrival time of regular flight f,
Cap_f = the capacity of carriage f,
Z_{if} = 1 if order i is planned to be allocated to regular flight f,
Z^r_{if} = 1 if order i is actually allocated to regular flight f,
Z^s_{if} = 1 if order i is transported by non-regular carriage immediately after assembly since it has missed its initial schedule for the regular carriage's departure time.
y^r_{if} = the quantity of the portion of order i transported by carriage f,
y^s_{if} = the quantity of the portion of order i transported by non-regular carriage immediately after assembly since it has missed its initial schedule for the regular carriage's departure time.
R_i = the release time to assemble order i,
C_i = the assembly completion time of order i,
PI_{ij} = 1 if the processing sequence of order i precedes processing sequence of order j on machine m, immediately, 0 otherwise.

$$\begin{aligned}\min_{\mathbf{R},\mathbf{Z},\mathbf{PI}} \quad & \sum_{i=1}^{N}\sum_{f=1}^{F_r} y^r_{if}(c^r_f + \alpha_i \max\{0, d_i - A^r_f\} \\ & +\beta_i \max\{0, A^r_f - d_i\} + h_i(D^r_f - C_i)) \\ & + \sum_{i=1}^{N}\sum_{f=1}^{F_r} y^s_{if}(c^s_f + \alpha_i \max\{0, d_i - A^s_{if}\} + \beta_i \max\{0, A^s_{if} - d_i\}). \quad (1)\end{aligned}$$

where

$$A^s_{if} = A^r_f + C_i - D^r_f. \quad (2)$$

According to the modeling sumption, we have the following constraints:

$$z_{if}(G_i - G^r_f)^2 \le 0, \quad \forall\ i, f, \quad (3)$$

$$\sum_{f=1}^{F_r} z_{0,f} = 0, \tag{4}$$

$$\sum_{f=1}^{F_r} z_{N+1,f} = 0, \tag{5}$$

$$\sum_{f=1}^{F_r} z_{if} = 1, \ \forall\, i, \tag{6}$$

$$z_{if}^r + z_{if}^s = z_{if}, \quad \forall\, i, f, \tag{7}$$

$$z_{if}^s M \geq z_{if}(C_i - D_f^r), \quad \forall\, i, f, \tag{8}$$

$$z_{if}^r M \geq z_{if}(D_f^r - C_i), \quad \forall\, i, f, \tag{9}$$

$$y_{if}^r = z_{if}^r Q_i, \quad \forall\, i, f, \tag{10}$$

$$y_{if}^s = z_{if}^s Q_i, \quad \forall\, i, f, \tag{11}$$

$$\sum_{i=1}^{N} y_{if}^r \leq Cap_f^r, \quad \forall\, f, \tag{12}$$

$$\sum_{i=1}^{N} y_{if}^s \leq Cap_f^s, \quad \forall\, f, \tag{13}$$

$$C_i = R_i + P_i, i = 0, 1, \cdots, N, N+1, \tag{14}$$

$$C_0 = 0, \tag{15}$$

$$R_0 = 0, \tag{16}$$

$$R_{N+1} \geq \sum_{i=1}^{N} P_i, \tag{17}$$

$$\sum_{j=1}^{N+1} PI_{ij} = 1, i \neq j, \ i = 0, 1, ..., N, \tag{18}$$

$$\sum_{j=0}^{N} PI_{ji} = 1, i \neq j, \ i = 1, ..., N+1, \tag{19}$$

$$PI_{i,0} = 0, \ i \leq N, \tag{20}$$

$$PI_{N+1,i} = 0, \ i \leq N, \tag{21}$$

$$C_i - C_j - MPI_{ji} \geq P_i - M, \ i, j = 0, 1, ..., N, N+1, \tag{22}$$

$$z_{if}, z_{if}^r, z_{if}^s, PI_{ij} \in \{0.1\}, \ i, j = 0, 1, .., N, N+1, \tag{23}$$

$$C_i, R_i, y_{if}^r, y_{if}^s \in \ integers. \tag{24}$$

The objective is to minimize the total cost, which consists of the total waiting cost between assembly and transportation, and the total transportation and delivery earliness/tardiness costs. Eq. (2) represents the non-regular flights' departure time of orders. Constraint (3) ensures that an order can be allocated a flight if the destinations of them are the same. Constraints (4) and (5) ensure

that artificial orders 0 and N+1 are not allocated to any flights. Constraint (6) ensures that an order only can be allocated to a flight. Constraint (7) ensures that an order is either transported by a regular or a commercial flight. Constraints (8) and (9) ensure that an order can be allocated into a flight if the flight's departure time is beyond the assembly completion time of that order. Constraint (10) determines the quantities of the portion of orders that are allocated into the regular cargo flight. Constraint (11) determines the quantity of the portion of orders that are allocated into commercial flight . Constraints (12) and (13) ensure that the capacities of the regular flight and the commercial flight are not exceeded. Constraint (14) represents the relationship between the release time, completion time and processing time of each order. Eq. (15) sets the completion time of artificial order to zero. Eq. (16) sets the release time of artificial order to zero. Constraint (17) sets the release time of the last job, RN'+1, to be larger or equal to the total processing time of all the jobs. Constraint (18) and (20) ensure that all the jobs should have a precedent job aside from the first job. Constraint (19) and (21) ensures that all the jobs should have a successive job aside from the last job. Constraint (22) represents the completion time relationship between any two jobs. Constraint (23) and (24) represents the binary and integer constraints on decision variables.

3 Solution Procedure

Due to the computational complexity of the model, no approach is guaranteed for solving the problem optimally within a polynomial time. To obtain a compromised solution within a reasonable CPU time, this paper presents a hybrid genetic-based heuristic to solve this problem. In this study, the GA consists of an initial randomly generated population that is evaluated for fitness using an objective function, and application of the GA operations of cloning, selection, crossover and mutation at each iteration. At each iteration, the values of the assembly sequence PI_{ij} and the idle times U_i are determined by GA. The solution procedure is shown as below.

1. Randomly generate initial population.
2. Determine decision variables C_i, R_i and Z_i based on assembly sequence PI_{ij} and the idle times U_i.
3. Solve partial problem to evaluate the remaining decision variables and evaluating fitness.
4. Genetic algorithm operations: cloning, selection, crossover and mutation.
5. Repeat steps 2-4 until terminal criterion is met.
6. Output results.

4 Computational Experience

The experiments were designed based on small and large scale problem categories in terms of the number of destinations L, number of regular carriages F,

number of orders N and time invertal T. The combinations of L, F, N and T for each problem size are fixed at $(3, 4, 8, 24)$ and $(3, 8, 30, 24)$ for the two problem categories. For each problem category, ten instances were generated. The parameters of all the instances in all problem categories were randomly generated. We use symbol GH to stand for our heuristic and LH to stand for Lingo approach. All instances were performed on a Pentium 3.2 GHz processor with 2 GB RAM. The computational time limit was set to 5 hours.

The criteria of performance considered was the quality of the cost used. The gap, defined as 100 (feasible solution obtained by GH - best feasible solution obtained by LH) / feasible solution obtained by LH, is used to evaluate the quality of solution for all instances in all problem categories. In addition, we use symbol N/A to present the situation in which no feasible solution found when solver time limit (5h) was reached. Numerical results are reported in Tables 1-2 and 3 for problem categories 1 and 2, respectively. We observe that LH can produce optimal solution for small scale problem. However, from Table 3, we can observe that Lingo approach can not produce feasible solution for large scale problem within 5 CPU hours. From the last column of Table 2, we can find that the heuristic solutions are close to the optimal solutions founded by Lingo solver. The gaps between the heuristic solutions and optimal solutions are no more than 2.0 percent. We also found that the deviations of the heuristic solutions of the ten instances are very small. This combines with the solution gaps with optimal solutions reflect that the quality and stability of the proposed heuristic are robust for small-scale problem. For large scale problem, we can observe that the deviations of the heuristic solutions are still small. Therefore, we can predict that our heuristic is a stable approach. Since the proposed heuristic can produce nearly optimal solution for small scale problem and can solve larger scale problems for which the Lingo solver can not solve, the proposed approach can be considered as a suitable approach to deal with this problem.

5 Conclusion

This study investigated a problem in which assembly scheduling and transportation scheduling are simultaneously considered. In this problem,A processed order can not be split and allocated to more than carriages and delivered separately to its destination. The problem was formulated as a costrained integer programming problem. A GA based heuristic approach was developed to deal with this problem in reasonable computational time. The chromosome in GA was expressed in terms of idle time variables in stead of the decision variables. The heuristic was evaluated by comparing its performance with well-known commercial software Lingo solvers on 20 randomly generated problems. The presented heuristic is shown to perform well compared with well-known commercial software. Form the numerical experiences, we found that the presented method can produce better solutions than the Lingo solvers.

Acknowledgment

The authors would like to thank anonymous referees for their helpful comments and suggestions that greatly improved the presentation of this paper. This research is partially supported by National Science Council, Taiwan, under grant NSC 97-2221-E-415-007-MY3.

References

1. D.E. Blumenfeld, D.E. Burns, L.D., Daganzo, C.F.: Synchronizing production and transportation schedules. Transportation Research B. 25, 23-27 (1991)
2. Chen, P.: Integrating production and transportation scheduling in a make-to-order environment. Ph.D. thesis, New York, Cornell University (2000)
3. Chen, Z.L., Vairaktarakis, G.L.: Integrated scheduling of production and distribution operations. Management Science 51(4), 614–628 (2005)
4. Fumero, F., Vercellis, C.: Synchronized development of production, inventory and distribution schedules. Transportation Science 33, 330–340 (1999)
5. Garcia, J.M., Lozano, S., Canca, D.: Coordinated scheduling of production and delivery from multiple plants. Robotics and Computer-Integrated Manufacturing 20(3), 91–198 (2004)
6. Lee, C.Y., Chen, Z.L.: Machine scheduling with transportation considerations. Journal of Scheduling 4, 3–24 (2001)
7. Li, K.P., Sivakumar, A.I., Mathirajan, M., Ganesan, V.K.: Solution methodology for synchronizing assembly manufacturing and air transportation of consumer electronics supply chain. International Journal of Business 9(4), 361–380 (2004)
8. Li, K.P., Sivakumar, A.I., Ganesan, V.K.: Synchronized scheduling of assembly and multi-destination air transportation in a consumer electronics supply chain. International Journal of Production Research 43(13), 2671–2685 (2005)
9. Li, K.P., Ganesan, V.K., Sivakumar, A.I.: Scheduling of single stage assembly with air transportation in a consumer electronic supply chain. Computers & Industrial Engineering 51(13), 264–278 (2006)
10. Li, K.P., Sivakumar, A.I., Ganesan, V.K.: Complexities and algorithms for synchronized scheduling of parallel machine assembly and air transportation in consumer electronics supply chain. Europear Journal of Operational Research 187, 442–455 (2008)
11. Roslof, J., Harjunkoski, I., Bjorkqvist, J., Karlsson, S., Westerlund, T.: An MILP-based reordering algorithm for complex industrial scheduling and rescheduling. Computers & Chemical Engineering 25(4-6), 821–828 (2001)
12. Ruiz-Torres, A.J., Tyworth, J.E.: Simulation based approach to study the interaction of scheduling and routing on a logistic network. In: Proceedings of the 29th conference on Winter Simulation, Atlanta, Georgia, United States, December 7-10, pp. 1189–1194 (1997)
13. Sarmiento, A.M., Nagi, A.: A review of integrated analysis of production-distribution systems. IIE Transactions 31, 1061–1074 (1991)
14. Tyan, J.C., Wang, F.K., Du, T.C.: An evaluation of freight consolidation policies in global third party logistics. Omega 31, 55–62 (2003)
15. Wan, G.H., Yen, B.M.P.C.: Tabu search for single machine scheduling with distinct due windows and weighted earliness/tardiness penalties. European Journal of Operational Research 142(2), 271–281 (2002)

16. Zuo, M., Kuo, W., McRoberts, K.: Application of mathematical programming to a large-scale agricultural production and distribution system. Journal of the Operational Research Society 42(8), 639–648 (1991)

Appendix

Table 1. Computational results for problem category 1

	Heuristic solution					Lingo solution		Sol.
no	max	min	mean	sigma	time	Sol.	Time	Gap
1	2336	2325	2330	1.0	258	2298.5*	253	-1.62%
2	2027	2027	2037	2.3	250	2002.8*	890	-1.18%
3	491	491	493	0.8	258	484.2*	292	-1.28%
4	520	520	528	1.8	260	517.5*	278	-0.50%
5	1429	1429	1438	2.5	262	1408.6*	148	-1.44%
6	965	965	967	0.3	250	964.4*	967	-0.01%
7	1467	1467	1473	2.7	262	1450.6*	240	-1.12%
8	1310	1310	1327	2.2	273	1296.2*	160	-1.07%
9	1847	1847	1850	1.4	270	1828.6*	334	-0.98%
10	834	834	837	0.6	258	829.9*	129	-0.47%

Symbol * means that the solution found is optimal.

Table 2. Computational results for problem category 2

	Heuristic solution					Lingo solution	
no	max	min	mean	sigma	time	Sol.	Time
1	2618.2	2567.3	2590.2	6.9	160.5	N/A	
2	5880.7	5822.9	5862.3	5.2	161.0	N/A	
3	2244.8	2203.0	2210.0	3.8	160.6	N/A	
4	3294.2	3261.6	3277.2	3.7	160.6	N/A	
5	3563.6	3502.8	3528.9	4.7	160.1	N/A	
6	3693.9	3660.9	3678.3	2.8	160.4	N/A	
7	3879.5	3873.5	3876.9	0.6	161.5	N/A	
8	2921.8	2914.6	2918.9	0.9	161.6	N/A	
9	2424.5	2384.8	2395.7	3.3	161.3	N/A	
10	2995.0	2960.4	2970.0	2.9	161.3	N/A	

N/A means that no feasible solution found when solver time limit (5h) was reached.

An Integrated BPM-SOA Framework for Agile Enterprises

Nan Wang and Vincent Lee*

Clayton School of IT, Monash University
Wellington Road, Clayton Campus, Victoria 3800
{nan.wang,vincent.cs.lee}@monash.edu

Abstract. We propose an integrated framework which aims to maximize the increase in agility for a given business process in the "real time" enterprise. The proposed framework is derived from holistic view of incorporating four groups of business process dynamics (business content, business intelligence, internal and external environment risks) that a real time decision support system can be implemented in a dynamic and fluid business process. It enables the capturing of agility that adapts to quantitative, qualitative, structured and unstructured information. Through examining the theoretical and conceptual foundations of agility drivers, we reveal ample innovation opportunities to integrate business process and IT services design methodology for increasing the enterprise's business agility.

Keywords: holistic view, process dynamics, agile enterprise, business process management, service-oriented architecture.

1 Introduction

Enterprise agility is the ability of a firm to adapt rapidly and cost efficiently in response to changes in its operating environments. Agility is therefore a concept that incorporates the ideas of flexibility, balance, adaptability, and coordination within one integrated process. In today's "real time" enterprise, enhancing agility is not a bonus but a necessity for gaining sustainable growth. Traditionally IT has been used as an enabling tool to help increase of competitive advantages in an enterprise with independent implementation of business process management. Firms that compete in today's global knowledge economy using the traditional approach, besides not optimally aligned IT strategy with business strategy, will not operationally optimize their rate of change of the agility. Hence intuitively there are needs to seamlessly integrate business process dynamics with IT services in a dynamic transaction processing platform.

To understand how the integration should be implemented, insight into context specific business process management (BPM) model is inevitable. BPM is a set of management discipline that aims to achieve continuous optimization of the efficiency of the process tasks and activities through automation [1]. Essentially, BPM provides support for the governance of business environment toward the purpose of enhancing agility and operational performance by blending incremental and transformative

* Corresponding author.

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 557–566, 2011.

methods. The rapidly changing business environments require suitable optimizations and appropriate adaptation. BPM does not, however, support decentralization, acquisition and outsourcing activities which are widely adopted in e-commerce industry of today. In [2] and [3], which alluded that business process activities can be fulfilled through service-oriented enterprise IT architecture (SOA). SOA is an IT-driven, architectural approach to software development on loosely-coupled for distributed services. SOA offers potentials for services to be integrated in order to create flexible, extensible and responsive enterprise.

The approach to reduce the lack of effective control and management of business process in ERP system, particularly through utilizing the process visualization capabilities provided by BPM was examined in [4]. The literature raised an important view that Business Intelligence (BI) delivers information source when forge with BPM, not just provides improved process management, but also enhanced decision making, particularly helpful for decision-intensive processes [5]. The functionalities of a BI have provided real-time data for computation of credit rating score. Thus Business Rule and Business Rule Engine are critical drivers for enterprises to rapidly react to changes in external business environmental conditions [6]. An architecture which enables people-centric BPM (business users, process designers and developers) to bridge the business and IT objectives' gaps in enterprise processes was studied in [7]. In [8], how a SOA is vital platform to enabling the success of BPM and proposed a conceptually combined architecture of BPM and SOA without adaptation to maximizing an enterprise's agilities. In [9], an integrated BPM (I-BPM) framework for a given IT architecture was proposed. Such I-BPM framework does not lead to enterprise's effective control of its agility. Aligning business process and IT architecture is required [10] in order to improve an enterprise's agility. A framework for semantic business process integration based on ontology alignment for a given services oriented IT architecture was examined in [11]. Business alliance formation in SOA for dealing with scalability problem has been discussed in [12]. Decision makers in enterprises should consider the alignment of business processes with business and IT strategies before systematically integrate BPM into long-term business objectives. Successful BPM requires the correct choice of management methods (e.g. Balanced Score Card, benchmarking, TQM, Six Sigma, etc.) in order to support continuous process improvement. Hence strategic and operational decision makers in an enterprise should have clear understanding on process-based risks and the possibility of changes in social, economic and legal requirements.

This paper proposes an integrated framework which aims to maximize the increase in agility for a given business process embedded in a "real time" enterprise. The framework with six interactive layers is derived via holistic view of incorporating four groups of business process dynamics - business content, business intelligence, internal and external environment risks confronting either machine learner or human decision maker in a given business process. The proposed framework enables the capturing of agility that is needed to respond to quantitative, qualitative, structured and unstructured information.

The proposed integrated framework differs from published BPM-SOA frameworks in three aspects. Firstly, both top-down and bottom-up information flows are made possible simultaneously via the dual-path adaptation which is indispensible for all agile "real time" enterprises. Secondly, changes in both internal and external business

environments are incorporated into endogenous factors. Many frameworks have implicitly regarded external business environment as an exogenous factor in their BPM life cycle. Thirdly, the proposed framework is non-process specific and hence it is scalable for multi-divisional and multi-product firms across the enterprise.

The paper is organized as follows. Section 2 discusses the semantic and ontology considerations to formulate the proposed framework. Section 3 examines the detailed steps for implementing the proposed framework. Discussions also devote to heighten the ample innovation opportunities in the integration of business process and IT services design methodology for increasing the enterprise's business agility. An example of implementation is given in Section 4. Our concluding remarks and future research directions are given in Section 5.

2 Semantic and Ontology Consideration

Under the bounded rationality, in the quest of improving an enterprise's agility need to adequately link its business processes with the computing capability via a service-oriented IT architecture (SOA). The basis of SOA concept can be represented as the Triangular SOA operational Model [13], which contains three important roles: service provider, service registry and service requestor. In this model, a service provider offers services; a service requestor invokes services which can be discovered from a service registry. A service registry aims to help service providers to publish services. SOA is a conceptual architecture that needs to be implemented by IT technologies, such as Web Services. Web services have created significant impact on SOA by providing a best practice standard-based approach to inter-operability between services. It establishes a mindset towards to reusable Service-Oriented design and addressed a bit of flexibility. However, the nature of Web Services is still a point-to-point architecture, without any intermediate intelligence. Thus the interconnection between service requesters and providers remains a complex unsolved puzzle.

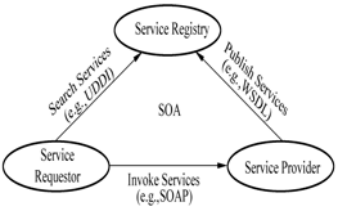


Fig. 1. Triangular SOA Operational Model

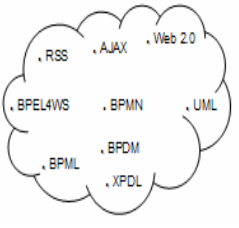


Fig. 2. BPM core components standards

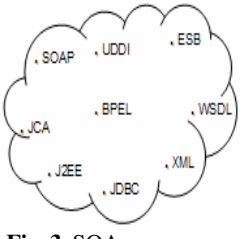


Fig. 3. SOA core components

As can be seen from Figures 2 and 3 above, the foundation standards of SOA and BPM are entirely different. Their own standard protocols and tools are not all compatible among each others. Because these standards are developed, promoted and supported by different institutions and vendors, some of these standards are competing against each other in each or across the field. In the middle of the two clouds are

BPEL and BPMN, respectively. Business Process Execution Language (BPEL) and Business process Modeling Notations (BPMN) are widely recognized as the orchestration language for service composition and chorography process modeling. BPMN and BPEL were not originally designed to work together [3], these two technologies can be used as the linkage between SOA and BPM.

In order to construct a scalable dynamic business process management in agile enterprise, new innovation must be incorporated when situation arises. Innovation can be incorporated via five fundamental steps (design, model, execute, monitor and optimize) which can be implemented through Business Process Management Suits (BPMS) [1]. This is a new type of management system that allows enterprises to model, execute and monitor its business processes that span numerous enterprise departments, systems and business partners. BPMS facilitate modeling the business processes by utilizing Business Process Modeling Notation (BPMN), and then these graphic images as blueprints can be reused for executable Business Process Execution Language (BPEL) processes. The difference between BPMN and BPEL is that BPMN is used when designing and improving the business process, whereas BPEL is used when implementing it.

Several significant benefits, offered by BPM through its lifecycle, can be concluded into three stages [14]. In the first stage of BPM initiative, it provides productivity benefit, which means reduce costs, quick improvement in processes, such as eliminating redundant activities and low-value tasks. In this stage, an enterprise focuses narrowly on specific and/or local processes. Stage two, organization gains process visibility benefit by creating transparency of entire value chain, increasing agility, coordination of resources and systems. The benefit of fifth stage is innovation power derived from integrating BPM and SOA which focuses on customers' satisfaction, develops corporate strategies, and optimizes core processes.

SOA as a technical concept shares similarity with BPM: i) SOA creates business agility which means loosely coupled services can be reused by different service requestor, whereas BPM supports reuse of processes and rules; ii) BPM and SOA overlaps as they both engage process management or orchestration; iii) They also organize services or processes into repositories; iv) They both accommodate monitoring and controlling functionality for processes or services [8]; and v) the important similarity is that both BPM and SOA meet the criterion of Real-time Enterprise by nature Event-driven. Beside the similarities, the different nature of BPM and SOA is also revealed in many other aspects: i) BPM is a business-driven initiative whereas SOA is an IT-driven initiative [3]; ii) BPMS is composed of tools and applications for automating business processes, whereas SOA is about IT architecture; iii) BPM represented a top-down approach refers to the procedure of recursively decomposing a business process into sub-processes or tasks, while SOA uses a bottom-up approach; and iv) SOA addresses the technology need for agility and flexibility, while BPM addresses both the business and the technology needs. However, each of BPM and SOA standalone has their limitations which justify the need for integration of BPM-SOA.

Traditional BPM approach alone exhibits its deficiencies when the firm grows either through acquisition of new business or entry into new markets.

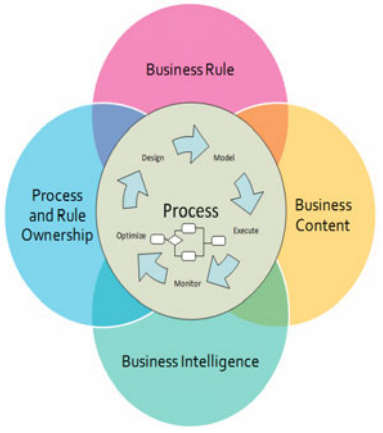


Fig. 4. Core elements in I-BPM- SOA

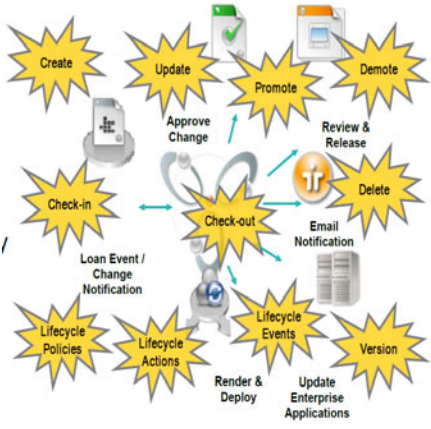


Fig. 5. Content + Process = "Active Content"

The complexities arose from expansion of new hard- and soft-resource supports which are unavailable without the reinforcement from SOA. The creation of BPM-SOA solution is therefore a cross-disciplinary process of concretizing abstractions and both must be intelligently integrated. Before identifying the gaps and obstacles that impeding fusion of BPM-SOA, the initiative is to focus on their foundation standards and technologies. Hence effective integration of BPM and SOA provides a truly agility to business, which create agile (reusable, configurable, extendable) processes, business rules and services to respond efficiently in this dynamic world. Integrating BPM and SOA requires innovative ideas to be put in practice by refining the scope and scale of the functionalities of each component to reduce overlapping. In a nutshell, Figure 4 depicts the core elements in the new BPM domain.

3 The Proposed Integrated BPM-SOA Framework

Six layers are needed in order to integrate BPM-SOA for timely extraction of data/information flow, business content, business rule and process. The functionalities and purposes of each layer, using bottom-up arrangement, are described as below.

3.1 Enterprise Resource Layer

The first premier layer from the bottom of this framework is enterprise operational resources layer. This layer consists of the underlying databases, data warehouses, information server, legacy mainframe, 3rd party applications and services. In general, all operational and legacy resources are implemented at this bottom layer.

3.2 SOA Adaptor Layer

This second SOA layer aims to managing various services, applications, and IT assets. Disparate technologies (J2EE, Web2.0, .NET, etc) and dissimilar services (IBM, SAP,

Oracle, and so forth) exist in an enterprise. In this layer, all services are classified under three categories: data service, adoption, and web service. This design will enhance the performance. The data service wrapper specifically designed for direct data access and manipulation in the database and data warehousing. The focus on this data service wrapper is speed. The business service wrapper is mainly used for integrating information and logic between legacy systems inside enterprise Intranet. Customization and reliability are the key goals of this business service wrapper.

Finally the web service wrapper is to enable application and process integration over the Internet. The biggest feature of this wrapper is the changeable service providers. Many services supported from this wrapper maybe only use once in one year. Generally, this layer provided a platform to enables the migration of inflexible IT services into merged, loosely coupled, reusable and on-demand services.

3.3 Enterprise Service Bus (ESB) Layer

The ESB layer deals with Invocation, integration, message routing, security, monitoring, mediation, process choreography, service orchestration, complex event processing and quality of service between data, business and web services. The concept of ESB [15] paves a new approach to service integration that can provide the implementation of backbone for a loosely coupled, event–driven SOA with a highly distributed network. It has been expressed as a key component of the SOA infrastructure. Mediation is the vital missing piece of the SOA layer. This intermediary layer unifies message oriented, event driven and service oriented approaches for integrating applications and service. It enables communication between different applications, processes and simplifies combination of business units, links heterogeneous software platforms and IT environments. The biggest advantage of implementing ESB arises from its integration logic is on using configuration against coding, which means business change prompting the modification in integration logic can be rapidly realized by reconfiguration rather than altering hard codes inside software. This heavily enhances the business agility ability [16].

3.4 Business Engine Layer

The business engine layer consists of 5 engines: they are Business Rule Engine, Business Intelligence Engine, BPEL Engine, Business Risk Management Engine and Business Content Engine. This complex Engine layer enables enterprise to become more dynamic. First, BPEL Engine takes outputs from upper layer as input. It then selects the process to be used for reprocessing procedures. All of the business processes are categorized such that each process as if is a decision related process the gears on the left hand begin to turn. A suitable rule is selected from Business Rule Engine. A Business Rule Engine can be simply thought as a sophisticated interpreter of “IF-THEN” statements. If the conditions of this rule are variables, then the data flow goes into Business Intelligence Engine. The source data extracting from Business Intelligence Engine aggregated as a real-time decision variable for the rule equation gained from Business Rule Engine. After the decision variables inserted into the Business Rule, a completed rule is then exported to Business Rule Management which is a BRMS that can be integrated into the Business Process Platform.

On the other hand, if each process is a content related process, the gears will turn the other way around via Business Content Engine and Risk Management Engine then come into Business Content Management, and finally goes back to the Business Process Platform. Risk management Engine consists of Analytic Hierarchy Process (AHP) [17], fuzzy mathematics, and possibly other statistical components that integrated to produce a coherent aggregated risk calculation. Criteria appropriate for the combination of weighted risk index and their priorities can vary widely, depending on the type of business content on hand and the internal and external environment at the time.

The quantitative and qualitative variables in the environment are also sometimes not mutually independent. Hence feedback loops are incorporated to modify the original AHP to yield the Analytic Network Process (ANP) [18] for analyzing unstructured decision and interdependent criteria based problems involving quantitative and qualitative measures. It not only permits multiple criteria, but also easily accommodates conflicting, multidimensional, incommensurable and incomparable objectives and judgments associated with complex decisions. Input to the Risk Management Engine includes both static and dynamic data from internal databases as well as external sources. Static data refers to systemic data representing relative contents, such as regulation and policy. Dynamic data refers to real-time or near-real-time information such as current economic conditions and current company conditions.

3.5 Process Platform Layer

The Process Platform layer enables an end-to-end automation and optimization of the business process. By separating the process, content and rule logic from the underlying enterprise resource layer, the transactions happen in the core systems while the processing such as rejects, approvals, exception handling, decision making is made in the Process Platform layer. Therefore, this process-aware architecture can seamlessly process choreography between disparate business applications across the enterprise. Design, modelling and simulation of business process and rules are done in this layer.

Business Process Platform (BPP) offers a shared platform of collection, management, communication, and action on business processes and events that flow from the application layer. The platform features a process modeler for constructing and adapting business processes utilizing an intuitive GUI, and a process simulator and monitor for dynamically simulating and tracking business processes. It offers configurable process to enable easily adaption on changing business requirements whilst considerably reducing time-consuming. A business process can have rules, contents and services or without them, or any combination of them. These all can be done in this layer which gives the enterprise the most flexibility of their assets – process.

3.6 Application Layer

The Application layer is the top layer of the process-aware enterprise framework. This layer provides the process access entrance. It gives different user interfaces to different roles in the enterprise. Each user interface may have different authorizations of service control. Some services could be Dashboard, Event box, and Report.

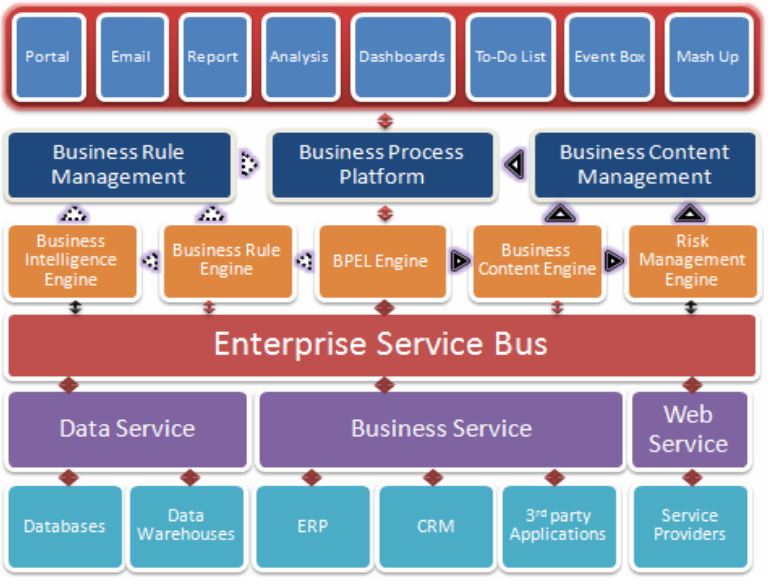


Fig. 6. The Proposed Integrated BPM-SOA Operational framework

Table 1. Business Process categories

Sign	Process Category	Process subcategory	Example
	(1)decision related process	(1.1)simple decision-making process (1.2) intelligent decision-making process	Pass to higher level manager (e.g. A to B).If within variance then approve.
	(2)content related process	(2.1)simple content related process (2.2) risky content related process	Upload financial report. Review financial report
	(3) decision content related process	(1.1)&(2.1), (1.1)&(2.2), (1.2)&(2.1), (1.2)&(2.2)	…
	(4) analysis related process	(4.1)enterprise performance analysis (4.2)external risk analysis	View dashboard View industrial risk index
	(5) service related process	(5.1) data service related process (5.2) business service-related process (5.3) web service related process	Access database, ERP Transaction, and Mash up

Table 2. Example of risk index

Risk Index	Example
External Environment	Financial situation at time A
Industries	Australia Mining industry (lowest rank risk at time A
Company rank	Rio Tinto has the 0.1 risk factor at time A
Company health	Company performance in current week compared with the same week 3 years ago
Company repayment	Rio Tinto's last 3 years financial situation
Etc	…

4 An Example of Implementation

We consider Business Loans to SMEs, which is an important source of revenue for commercial banks to demonstrate the scalablity of the proposed I-BPM-SOA framework. Loan approval process, one of the functionalities of the risk management engine is to coherently compute various risk indexes imbedded inside of the business contents.

The final risk index calculated from Risk Management Engine will then be transmitted to the strategic management level or decision makers. At the same time, the content and final risk index will be exported to Business Content Management which is an ECMS that can be incorporated into Business Process Platform. If the business process subcategory defined as (1.2) & (2.2) shown in Table 1, then the most knowledgeable process is created. Besides that, the business process also can be as process category (4) or (5). In that case, after process outputs from BPEL Engine it will skip through Business Rule Engine and Business Content Engine and directly goes into Business Intelligence Engine and Risk Management Engine respectively, the output will straight interact with ESB.

BPEL Engine also enables enterprises to orchestrate disparate applications and Web services into business processes. The Business Rule Engine comprises fact types, rule sets, rules, decision tables, and decision services. A fact type is a model that can be used for modeling rules which the rule dictionary is populated with. A rule set can be described as a service that likes a rules container used as a grouping mechanism for rules. One rule set is produced within the rule dictionary. A decision table is basically a group of rules with the same fact type model elements that can be visualized in a tabular format. A decision service is created and then exposed as a service called by business processes.

Risk Management Engine is essential to corporate governance. In order for an enterprise to stay alive and preserve a competitive advantage, it must take risks factoring in with profit and growth. Potential risks index can be allocated to specific process, where they are in turn related to controls. Company can focus on proactive rather than reactive risk management. The Business Content Engine is an object-oriented metadata repository, an important component of business engine layer, manages a full range of structured and unstructured data, business related information. Business intelligence Engine can provide source data to business rule engine for predefined business processes as well as simple task based on BI only.

5 Conclusion

This paper has formulated an integrated BPM-SOA framework for use by agile enterprises facing global competition and hence technology must be deployed effectively in real time enterprise environments. Through examining the theoretical and conceptual foundations of agility drivers, we reveal ample opportunities for enterprise using the proposed integrated framework as a platform to continuously implement the process innovation with optimized agility for sustainable growth. The proposed framework is scalable and therefore can be adapted for use by global and collaborative network enterprises in highly competitive production economy. Further research is underway to devise various measures needed to coherently capture the quantitative, qualitative, structured and unstructured information and tacit knowledge to optimize the enterprise agility.

References

[1] Finger, H., Smith, P.: Business Process Management. In: The Third Wave, Tampa, Florida. Meghan-Kiffer Press
[2] Malinverno, P., Hill, J.B.: BPM and SOA are Better Together, Gartner Research, ID number: G00145586 (February 2007)
[3] Colleen, F., News, W.: Special Report: BPM inside the belly of the SOA. Whale press (2006)
[4] JinPing, L.: Applied analysis of BPM and ERP system. Market Modernization 454, 51–52 (2006)
[5] Research, V.: Business Intelligence Meets Business Process Management - Powerful technologies can work in tandem to drive successful operations (2006)
[6] Halie, B.V., Goldberg, L.: The Business Rule Revolution - Runing Business the Right Way Happy About (2006)
[7] Todor, S., Stefan, S.: An Architecture for End-User Driven Business Process Management. In: Proceedings of 12th IEEE International Conference Enterprise Distributed Object Computing Conference, SAP AG, SAP Research CEC Darmstadt, pp. 53–62 (2008)
[8] Bajwa, I.S., et al.: SOA and BPM Partnership: A paradigm for Dynamic and Flexible Process and I.T. Management. World Academy of Science, Engineering and Technology (2008)
[9] Romero, D., Molin, A.: VO breeding environment & virtual organizations integral business process management framework. In: Information System Frontier, pp. 569–597 (2009)
[10] Neubauer, T.: An empirical study about the status of business process management. Business Process Management Journal 15(2), 166–183 (2009)
[11] Jung, J.J.: Service Chain-based Business Alliance Formation in Service-oriented rchitecture. Expert Systems with Applications 38(3), 2206–2211 (2011)
[12] Jung, J.J.: Semantic Business Process Integration Based on Ontology Alignment. Expert Systems with Applications 36(8), 11013–11020 (2009)
[13] Zhang, L.J., et al.: Services computing. Tsinghua University Press/Springer, Beijing/Berlin/New York (2007)
[14] Snabe, J.H., Zimniak, F.: Business process management the SAP roadmap, 1st edn., 411 p. Galileo Press, Boston (2009)
[15] Chappell, D.A., Hendrickson, M. (eds.): Enterprise Service Bus. O'Reilly Media, Inc., Sebastopol (2004)
[16] Mulik, S.: Using Enterprise Service Bus (ESB) for connecting corporate functions and shared services with business divisions in a large enterprise. In: Proceedings IEEE Asia-Pacific. APSCC (2009)
[17] Saaty, T.L.: The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation. McGraw-Hill, New York (1980)
[18] Saaty, T.L.: The Analytic Network Process: Decision Making with Dependence and Feedback. RWS Publications (1996)

Author Index