Lecture Notes in Artificial Intelligence 7803

Subseries of Lecture Notes in Computer Science

Ali Selamat Ngoc Thanh Nguyen
Habibollah Haron (Eds.)

Intelligent Information and Database Systems

5th Asian Conference, ACIIDS 2013
Kuala Lumpur, Malaysia, March 18-20, 2013
Proceedings, Part II

Springer

Volume Editors

Ali Selamat
Universiti Teknologi Malaysia
Faculty of Computing, Department of Software Engineering
81310 UTM Skudai, Johor, Malaysia
E-mail: aselamat@utm.my

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

Habibollah Haron
Universiti Teknologi Malaysia
Faculty of Computing, Department of Computer Science
81310 UTM Skudai, Johor, Malaysia
E-mail: habib@utm.my

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-36542-3 e-ISBN 978-3-642-36543-0
DOI 10.1007/978-3-642-36543-0
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2013930842

CR Subject Classification (1998): I.2.1, I.2.3-4, I.2.6-8, I.2.11, H.3.3-5, H.2.8, H.4.1-3, H.5.3, I.4.9, I.5.1-4, K.4.4

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 2013 was the fifth 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 2013 was to provide an internationally respected forum for scientific research in the technologies and applications of intelligent information and database systems. ACIIDS 2013 was co-organized by Universiti Teknologi Malaysia (Malaysia) and Wroclaw University of Technology Poland (Poland) in co-operation with Nguyen Tat Thanh University (Vietnam) and took place in Kuala Lumpur (Malaysia) during March 18–20, 2013. The first two events, ACIIDS 2009 and ACIIDS 2010, took place in Dong Hoi City and Hue City in Vietnam, respectively. The third event, ACIIDS 2011, took place in Daegu (Korea), while the fourth event, ACIIDS 2012, took place in Kaohsiung (Taiwan).

Submissions came from 20 countries from all over the world. Each paper was peer reviewed by at least two members of the International Program Committee and International Reviewer Board. Only 108 papers with the highest quality were selected for oral presentation and publication in the two-volumes proceedings of ACIIDS 2013.

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-business 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 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, Zaini Ujang (Universiti Teknologi Malaysia, Malaysia) and Tadeusz Więckowski (Rector of Wroclaw University of Technology, Poland) for their support.

Our special thanks go to the General Co-chair, Program Co-chairs, all Program and Reviewer Committee members, and all the additional reviewers for their valuable efforts in the review process that 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 contributed to the success of the conference.

We also would like to express our thanks to the Keynote Speakers (Hoang Pham, Naomie Salim, Mong-Fong Horng, Sigeru Omatu) for their interesting and informative talks of world-class standard.

We cordially thank our main sponsors, Universiti Teknologi Malaysia (Malaysia), Wroclaw University of Technology (Poland), and Nguyen Tat Thanh University (Vietnam). Our special thanks are due also to Springer for publishing the proceedings, and to 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: Habibollah Haron (Organizing Chair) 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.

March 2013

Ngoc Thanh Nguyen
Ali Selamat

Conference Organization

Honorary Chairs

Zaini Ujang	Vice Chancellor of Universiti Teknologi Malaysia, Malaysia
Tadeusz Wieckowski	President of Wroclaw University of Technology, Poland

General Co-chairs

Ngoc Thanh Nguyen	Wroclaw University of Technology, Poland
Mohd Aizaini Maarof	Universiti Teknologi Malaysia, Malaysia

Program Chairs

Ali Selamat	Universiti Teknologi Malaysia, Malaysia
Shyi-Ming Chen	National Taiwan University of Science and Technology, Taiwan

Organizing Chair

Habibollah Haron	Universiti Teknologi Malaysia, Malaysia

Session Chairs

Andri Mirzal	Universiti Teknologi Malaysia, Malaysia
Bogdan Trawinski	Wroclaw University of Technology, Poland
Jason J. Jung	Yeungnam University, Republic of Korea

Organizing Committee

Roliana Ibrahim	Universiti Teknologi Malaysia, Malaysia
Nor Erne Nazira Bazin	Universiti Teknologi Malaysia, Malaysia
Dewi Nasien	Universiti Teknologi Malaysia, Malaysia
Mohamad Shukor Talib	Universiti Teknologi Malaysia, Malaysia

Steering Committee

Ngoc Thanh Nguyen	Chair, Wroclaw University of Technology, Poland
Longbing Cao	University of Technology Sydney, Australia
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	University Paul Verlaine – Metz, France
Toyoaki Nishida	Kyoto University, Japan
Leszek Rutkowski	Technical University of Czestochowa, Poland
Dickson Lukose	Knowledge Technology Cluster at MIMOS BHD, Malaysia

Keynote Speakers

Hoang Pham	Rutgers, The State University of New Jersey, USA
Naomie Salim	Universiti Teknologi Malaysia, Malaysia
Mong-Fong Horng	National Kaohsiung University of Applied Sciences, Taiwan
Sigeru Omatu	Osaka Institute of Technology, Japan

Special Sessions Organizers

1. *International Workshop on Engineering Knowledge and Semantic Systems (IWEKSS 2013)*

Jason J. Jung	Yeungnam University, Korea
Dariusz Krol	Wroclaw University of Technology, Poland

2. *Modelling and Optimization Techniques in Information Systems, Database Systems and Industrial Systems (MOT-ACIIDS 2013)*

Le Thi Hoai An	University of Lorraine, France
Pham Dinh Tao	INSA-Rouen, France

3. *Intelligent Supply Chains (ISC 2013)*

Arkadiusz Kawa	Poznan University of Economics, Poland
Paulina Golińska	Poznan University of Technology, Poland
Milena Ratajczak-Mrozek	Poznan University of Economics, Poland

4. *Intelligent Systems for Medical Applications (ISMA 2013) Information Systems and Industrial Engineering (MOT-ISIE)*

Uvais Qidwai	Qatar University, Qatar

5. *Innovations in Intelligent Computation and Applications (ICA 2013)*

Shyi-Ming Chen	National Taiwan University of Science and Technology, Taiwan

6. *Computational Biology and Bioinformatics (CBB 2013)*

Mohd Saberi Mohamad	Universiti Teknologi Malaysia, Malaysia

7. *Multiple Model Approach to Machine Learning (MMAML 2013)*

Tomasz Kajdanowicz	Wroclaw University of Technology, Poland
Tomasz Łuczak	Wroclaw University of Technology, Poland
Grzegorz Matoga	Wroclaw University of Technology, Poland

8. *Intelligent Recommender Systems (IRS 2013)*

Adrianna Kozierkiewicz-Hetmańska	Wroclaw University of Technology
Ngoc Thanh Nguyen	Wroclaw University of Technology

9. *Applied Data Mining for Semantic Web (ADMSW 2013)*

Trong Hai Duong	Quang Binh University, Vietnam
Bay Vo	Information Technology College, Vietnam

International Program Committee

Abdul Rahim Ahmad	Universiti Tenaga Nasional, Malaysia
Abdul Samad Ismail	Universiti Teknologi Malaysia, Malaysia
Abdul Samad Shibghatullah	Universiti Teknikal Malaysia, Malaysia
Adrianna Kozierkiewicz-Hetmańska	Wrocław University of Technology, Poland
Alex Sim	Universiti Teknologi Malaysia, Malaysia
Alvin Yeo	Universiti Malaysia Sarawak, Malaysia
Amir Shafie	International Islamic University Malaysia, Malaysia
Annabel Latham	The Manchester Metropolitan University, UK

Antoni Wibowo	Universiti Teknologi Malaysia, Malaysia
Arkadiusz Kawa	Poznan University of Economics, Poland
Aryati Bakri	Universiti Teknologi Malaysia, Malaysia
Azlan Mohd Zain	Universiti Teknologi Malaysia, Malaysia
Azurah Abu Samah	Universiti Teknologi Malaysia, Malaysia
Bay Vo	Information Technology College, Vietnam
Behnam Rouzbehani	Islamic Azad University, Central Tehran Branch, Iran
Bing-Han Tsai	National Taiwan University of Science and Technology, Malaysia
Bjoern Schuller	Technische Universität München, Germany
Bogdan Trawinski	Wroclaw University of Technology, Poland
Boguslaw Cyganek	AGH University of Science and Technology, Poland
Cheng-Yi Wang	National Taiwan University of Science and Technology, Taiwan
Dariusz Barbucha	Gdynia Maritime University, Poland
Dariusz Frejlichowski	West Pomeranian University of Technology, Poland
Dariusz Krol	Bournemouth University, UK
Dongjin Choi	Chosun University, Korea
Dragan Simic	University of Novi Sad, Serbia
El-Houssaine Aghezzaf	Ghent University, Belgium
Elżbieta Kukla	Wrocław University of Technology, Poland
Eric Pardede	La Trobe University, Australia
Faisal Zaman	Kyushu Institute of Technology, Poland
Fan Wang	Microsoft, USA
Geetam Tomar	Malwa Institute of Technology and Management, Gwalior, India
Gia-An Hong	National Taiwan University of Science and Technology, Taiwan
Gordan Jezic	University of Zagreb, Coatia
Hae Young Lee	ETRI, Korea
Halina Kwasnicka	Wroclaw University of Technology, Poland
Hoai An Le Thi	University of Lorraine, France
Huey-Ming Lee	Chinese Culture University, Taiwan
Huynh Binh	University of Science and Technology, Vietnam
Hyon Hee Kim	Dongduk Women's University, Korea
Imran Ghani	Universiti Teknologi Malaysia, Malaysia
Ireneusz Czarnowski	Gdynia Maritime University, Poland
Iskandar Ishak	Universiti Putra Malaysia, Malaysia
Jafar Razmara	Universiti Teknologi Malaysia, Malaysia

Jaroslaw Jankowski	West Pomeranian University of Technology in Szczecin, Poland
Jason Jung	Yeungnam University, Korea
Jerome Euzenat	INRIA, France
Jesús Alcalá-Fdez	University of Granada, Spain
Jose Norbeto Mazon	University of Alicante, Spain
Kamal Zamli	Universiti Malaysia Pahang, Malaysia
Kang-Hyun Jo	University of Ulsan, Korea
Katarzyna Grzybowska	Poznan University of Technology, Poland
Kazuhiro Kuwabara	Ritsumeikan University, Japan
Khairuddin Omar	Universiti Kebangsaaan Malaysia, Malaysia
Lian En Chai	Universiti Teknologi Malaysia, Malaysia
Manh Nguyen Duc	ENSTA Bretagne, France
Marcin Hajdul	Institute of Logistics and Warehousing, Poland
Maria Bielikova	University of Technology in Bratislava, Slovakia
Md. Nazrul Islam	Universiti Teknologi Malaysia, Malaysia
Michał Woźniak	Wroclaw University of Technology, Poland
Milena Ratajczak-Mrozek	Poznan University of Economics, Poland
Minh Le Hoai	University of Lorraine, France
Mohd Helmy Abd Wahab	Universiti Tun Hussein Onn Malaysia, Malaysia
Mohd Murtadha Mohamad	Universiti Teknologi Malaysia, Malaysia
Mohd Ramzi Mohd Hussain	International Islamic University Malaysia, Malaysia
Mohd Saberi Mohamad	Universiti Teknologi Malaysia, Malaysia
Moon II Chul	Korean Institute Science and Technology, Korea
Muhammad Khan	King Saud University, Saudi Arabia
Muhamad Razib Othman	Universiti Teknologi Malaysia, Malaysia
Mohd Salihin Ngadiman	Universiti Teknologi Malaysia, Malaysia
Muhammad Shuaib Karim	Quaid-i-Azam University, Malaysia
Muhammad Suzuri Hitam	Universiti Malaysia Terengganu, Malaysia
Nghi Do Thanh	Telecom-Bretagne, France
Ngoc Thanh Nguyen	Wroclaw University of Technology, Poland
Niels Pinkwart	Clausthal University of Technology, Germany
Nojeong Heo	Dongyang University, Korea
Noorfa Haszlinna Mustaffa	Universiti Teknologi Malaysia, Malaysia
Nor Azizah Ali	Universiti Teknologi Malaysia, Malaysia
Nor Haizan Mohamed Radzi	Universiti Teknologi Malaysia, Malaysia
Nor Hawaniah Zakaria	Universiti Teknologi Malaysia, Malaysia

Norazah Yusof	Universiti Teknologi Malaysia, Malaysia
Norazman Ismail	Universiti Teknologi Malaysia, Malaysia
Olgierd Unold	Wroclaw University of Technology, Poland
Ondrej Krejcar	University of Hradec Kralove, Czech Republic
Trong Hai Duong	Inha University, Korea
Bernadetta Mianowska	Wroclaw University of Technology, Poland
Michal Sajkowski	Poznan University of Technology, Poland
Robert Susmaga	Poznan University of Technology, Poland

Table of Contents – Part II

Tools and Applications

Intelligent Recommender Systems

Multiple Model Approach to Machine Learning

Engineering Knowledge and Semantic Systems

Computational Biology and Bioinformatics

Computational Intelligence

Modelling and Optimization Techniques in Information Systems, Database Systems and Industrial Systems

Intelligent Supply Chains

Applied Data Mining for Semantic Web

Semantic Web and Ontology

Integration of Information systems

Conceptual Modeling in Advanced Database Systems

Table of Contents – Part I

Innovations in Intelligent Computation and Applications -1

Innovations in Intelligent Computation and Applications -2

Intelligent Database Systems -1

Intelligent Database Systems -2

Intelligent Information Systems -1

Intelligent Information Systems -2

Intelligent Information Systems -3

Detection of Noise in Digital Images by Using the Averaging Filter Name COV

Janusz Pawel Kowalski[1], Jakub Peksinski[2], and Grzegorz Mikolajczak[2]

[1] Pomeranian Medical University, Basic Computer Science, Rybacka 1, 70-204 Szczecin, Poland
janus@pum.edu.pl
[2] West Pomeranian University of Technology, Faculty of Electrical Engineering, 26 Kwietnia 10, 71-126 Szczecin, Poland
{jpeksinski,grzegorz.mikolajczak}@zut.edu.pl

Abstract. One of the significant problems in digital signal processing is the filtering and reduction of undesired interference. Due to the abundance of methods and algorithms for processing signals characterized by complexity and effectiveness of removing noise from a signal, depending on the character and level of noise, it is difficult to choose the most effective method. So long as there is specific knowledge or grounds for certain assumptions as to the nature and form of the noise, it is possible to select the appropriate filtering method so as to ensure optimum quality. This chapter describes several methods for estimating the level of noise and presents a new method based on the properties of the smoothing filter.

Keywords: noise estimation, smoothing filters.

1 Introduction

The dynamic development of computer techniques that has been observed over the past twenty years and the development of digital algorithms for signal processing accompanying it, allows for significant improvement of the quality of obtained images and purposeful interference in the image structure for bringing out certain qualities. Improvement of image quality makes it possible to obtain a significantly greater amount of useful information and also to create a better aesthetic impression. A significant practical matter is the search for methods of improvement of image quality and removal of distortions being the effect of noise.

Effectiveness of filtering, expressed e.g. by the noise reduction coefficient, is a function of many factors, including: the selected filtering algorithm, certain information with noise qualities, and also certain information about the model image [12]. Of special significance is information on the qualities of noise – random or determined, the distribution of power spectral density, variance, etc. In most cases, it is not possible to obtain full data on the noise and attempts at estimation are undertaken – assessment of the level of noise expressed by variance through analysis of image data. Using the information on the level of noise in the image allows for obtainment of an

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 1–8, 2013.

optimal filtering quality, especially for realization of problems of image reconstruction, edge detection, and others. It is also among the information necessary for the creation and operation of adaptive filtering algorithms.

The applications of noise level estimation in images are very wide and include, among others: removal of noise from astronomical photographs [8]; image reconstruction [2]; edge detection [3],[10, 13]; image segmenting [11, 13]; image smoothing [1]; reduction of noise in photographs made using magnetic resonance technology (MRI) [4].

One of the significant problems in digital signal processing is the filtering and reduction of undesired interference. Due to the abundance of methods and algorithms for processing signals characterized by complexity and effectiveness of removing noise from a signal, depending on the character and level of noise, it is difficult to choose the most effective method. So long as there is specific knowledge or grounds for certain assumptions as to the nature and form of the noise, it is possible to select the appropriate filtering method so as to ensure optimum quality. E.g. the moving average filter has greater noise reduction coefficient than the medium filter with the same mask size, so it will be more suited to removing “large” noise. However, the median is better for maintaining edges and interferes in the signal structure in a lesser degree, which, with a smaller noise reduction coefficient, is more suitable for filtering signals with “low” noise.

The problem is significantly more complex when the character of a given signal cannot be determined. Without additional information, it is often difficult to assess the level and "type of noise”, and sometimes, it is not possible to state whether a random or deterministic course is being dealt with. Due to this, at this point, a test analysis of the properties of smoothing filters and their influence on noise level in the output signal will be conducted.

2 Noise Level Estimation Using Exponential Smoothing Filter

The values of variances of noise signals for individual smoothing methods are derived below. Exponential smoothing of the signal is given by the formula:

$$y_m = (1-a)\cdot y_{m-1} + a\cdot x_m \qquad where \quad y_0 = (1-a)\cdot x_0 \tag{1}$$

With the acceptance of the following assumptions:

$$x = s + n \qquad V(N) = \sigma_n^2 \tag{2}$$

where: x – disrupted signal; s – useful signal; n – noise. Taking into account that the noise and useful signal are not correlated, the variance of the input signal is equal to:

$$V(X) = \sigma_s^2 + \sigma_n^2 \qquad where: \quad V(S) = V_s = \sigma_s^2 \tag{3}$$

Assuming that smoothing reduces only the noise variance, a dependency of the output signal variance can be written as:

$$V(Y) = \sigma_s^2 + \frac{a}{2-a}\sigma_n^2 \tag{4}$$

By designating: $V(Y)=V_a$ – variance after exponential smoothing, V_0 – signal variance without smoothing, the following is obtained:

$$V_0 = \sigma_s^2 + \sigma_n^2 \Rightarrow \sigma_s^2 = V_0 - \sigma_n^2 \tag{5}$$

By substituting (5) to (4), after transformations, the dependency for noise variance (6) is obtained, determined on the basis of knowledge of the variance of the disrupted signal and the variance of the signal after exponential smoothing.

$$\sigma_n^2 = \frac{2-a}{2-2a}(V_0 - V_a) \tag{6}$$

The method of determining noise variance presented above (6) is characterized by a need for knowledge of only the variance of the input and output signal, the value of which is known. This makes it possible to easily determine the noise level in the case of smoothing method. In general, the idea of the method is based on the knowledge of the noise variance reduction coefficient, the value of input and output signal variance, and on the assumption that, during filtering, only the noise in the output signal is attenuated [10,11,12]. The above establishments find their reflection in dependency (7), which is the basis for estimation of the noise level in the analyzed signal.

$$\sigma_n^2 = \frac{V(Y) - V_q(Y)}{1-q} \tag{7}$$

where: $V(Y)$ – signal variance without filtering; $V_q(Y)$ – signal variance after filtering; q - noise reduction coefficient.

3 Estimation Noise Variance Using Averaging Filter

For a two-dimensional weighted average filter with a 3x3 mask:

$$\begin{bmatrix} a & a & a \\ a & 1 & a \\ a & a & a \end{bmatrix} \quad where \quad 0 < a \leq 1 \tag{8}$$

the noise reduction coefficient is expressed by formula:

$$q = \frac{1+8a^2}{(1+8a)^2} \tag{9}$$

Due to the fact that the assumption pertaining to attenuation of only noise is not completely fulfilled, that is, the filter also interferes in the signal structure, small weight values are to be selected (*a~0.01*), and the area for estimation should be characterized by low variability of the useful signal. Table 1 presents the results of noise level estimation for areas marked on figure. 1, presenting image with 256 shades of gray and a size of 256x256 pixels, – Bridge_SF.

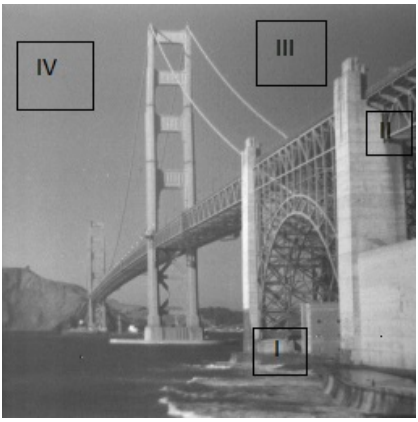


Fig. 1. Test image Bridge_SF I-IV – areas from which noise variance was estimated

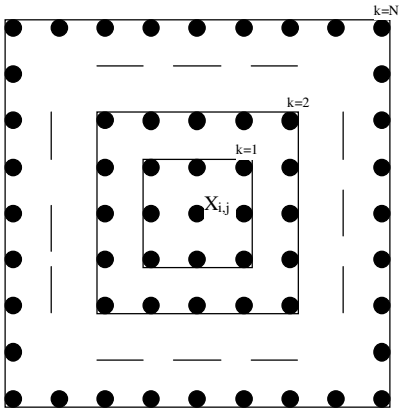


Fig. 2. Layout of points, with reference to which we calculate auto-covariance

The entire image was subjected to additive noise with normal distribution *N(0,σ)*. During the selection of areas, the leading consideration was that the assumption pertaining to the removal of only noise from the filtered signal was inaccurate. This is shown by the results of noise estimation in areas I and II. These areas contain a large amount of small details and a large range of gray levels, which makes estimation in these areas ineffective. However, in areas III and IV, the estimated variance value shows good conformance with the variance of the noise to which the test image was subjected. This is an effect of the fact that the selected areas exhibit high uniformity of gray levels and a lack of small details.

Table 1. Results of noise variance estimation in the Bridge_SF test image for individual areas

Area	I		II		III		IV.	
σ^2	a=0.01	a=0.02	a=0.01	a=0.02	a=0.01	a=0.02	a=0.01	a=0.02
25	110.1	114.0	300.1	310.3	30.4	30.6	29.4	28.3
100	172.2	169.5	280.5	354.9	99.7	97.3	95.4	96.8
225	279.3	288.4	456.6	472.4	215.1	220.1	216.7	219.1

The results shown confirm the correctness of the accepted assumption, on the basis of which the formula for noise level estimation was derived (7). The only source of doubts is the fact that selection of the area is done subjectively, that is, the area that seems the most appropriate for evaluating noise level is selected. A conclusion can be made, that a method for finding an area from which noise level estimation would give satisfactory results should be elaborated.

4 Finding the Area for Noise Level Estimation

It is proposed for selection of the area to be made based on image analysis due to the correlation coefficient determined based on auto-covariance, because the correlation between individual pixels of distinguished "good" areas is small. In the case of searching for an area with the lowest values of correlation coefficients, the following procedure was used:

- The average value – m of disrupted pixels in a given area is determined by moving an *NxN* window over the analyzed image.
- Next, for each pixel from the given area, the auto-covariance coefficient is calculated relative to pixels found in the *k-th* layer (10).

$$(C_{i,j})_k = \frac{1}{8k}(x_{i,j} - m)\left[\sum_{r=-k}^{k}(x_{i-k,j+r} - m) + \sum_{r=-k}^{k}(x_{i+k,j+r} - m) + \ldots \right.$$
$$\left. \ldots + \sum_{r=-k+1}^{k-1}(x_{i+r,j-k} - m) + \sum_{r=-k+1}^{k-1}(x_{i+r,j+k} - m)\right] \tag{10}$$

The placement of individual layers is shown on figure 2. The variance of pixels of the k-th layer is defined as:

$$(V_{i,j})_k = \frac{1}{(2k+1)^2}\sum_{r=-k}^{k}\sum_{s=-k}^{k}(x_{i+r,j+s} - m)^2 \tag{11}$$

- On the basis of the obtained auto-covariance (10) and variance (11) values in the analyzed area, the correlation coefficients ρ can be determined for each layer:

$$(\rho_{i,j})_k = \frac{(C_{i,j})_k}{(V_{i,j})_k} \tag{12}$$

- Next, the obtained values of correlation coefficients are averaged for individual points relative to successive layers. In this way, values of correlation coefficients R_k are obtained in the given area for a given layer:

$$R_k = \frac{1}{N^2} \sum_{r=1}^{N} \sum_{s=1}^{N} (\rho_{r,s})_k \tag{13}$$

- The selection of the area for estimation of the noise level is done based on the lowest value of the slope of the straight line determined from linear approximation (linear regression) carried out for correlation coefficients R_k.

5 Experiment and Discussion of Results

The proposed method for estimating the level of noise (7) has been subjected to verification during the experiment. A level of noise with a normal distribution (Gaussian) was estimated, in the range of change σ_n=1÷15, on test images Bridge_SF. An area for estimation of various sizes was searched for (10)-(13). The results of estimation, in the form of error calculated from average values of standard deviations obtained during individual tests were presented in table 2 and on figure 3.

Table 2. Results of estimation, in the form of error calculated from average values of standard deviations obtained of noise variance estimation in the Bridge_SF test image

	Av1	**Med.**	**Av2**	**Blok**	**Grad**	**Pyr**	**Cov**		
σ_n							9x9	19x19	29x29
1	6,130	4,759	0,406	1,521	2,504	1,128	0,891	1,160	1,108
2	2,665	1,989	0,134	0,501	1,110	0,350	0,215	0,362	0,391
3	1,541	1,118	0,060	0,234	0,579	0,217	0,026	0,108	0,173
4	1,007	0,714	0,033	0,107	0,430	0,238	0,021	0,059	0,102
5	0,702	0,490	0,019	0,029	0,363	0,163	0,030	0,035	0,076
6	0,512	0,357	0,012	0,008	0,298	0,093	0,182	0,003	0,059
7	0,383	0,265	0,007	0,032	0,261	0,065	0,080	0,003	0,069
8	0,295	0,201	0,039	0,051	0,269	0,053	0,110	0,040	0,009
9	0,231	0,160	0,003	0,054	0,205	0,035	0,110	0,021	0,008
10	0,183	0,124	0,001	0,077	0,194	0,032	0,072	0,036	0,030
11	0,144	0,102	0,001	0,090	0,132	0,015	0,139	0,054	0,014
12	0,115	0,080	0,001	0,089	0,208	0,009	0,152	0,066	0,004
13	0,087	0,069	0,002	0,097	0,187	0,000	0,098	0,079	0,022
14	0,070	0,059	0,003	0,103	0,096	0,005	0,081	0,036	0,007
15	0,057	0,046	0,003	0,110	0,109	0,021	0,136	0,0145	0,004

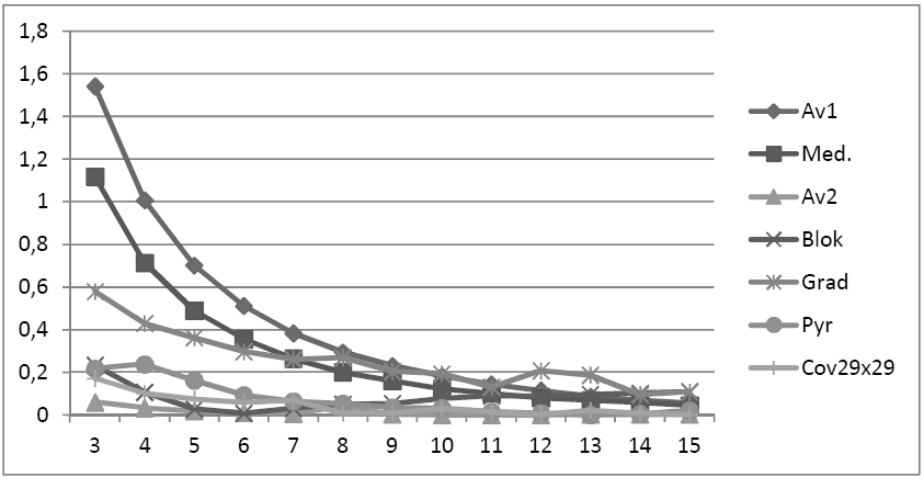


Fig. 3. Diagrams showing relative error of estimation σ_n in function of interference, methods: Cov (area 29x29), Av1, Av2, Med, Block,Grad, Pyr

The described method was compared to the methods shown in work [9]. Individual methods have been marked as:

- Av1 – estimation using an averaging filter, where σ_n was calculated from all pixels [9],
- Med. – as above, only using a median filter [9],
- Av2 – similar to Av1 only 10% of pixels with the lowest values were used for estimation [9],
- Block – σ_n calculated from averaging 10% of the variances calculated in areas of 7x7 pixels [5], [6],
- Gradient – on the basis of the gradient histogram [4] ,[13],
- Pyramid – through calculation of variances in individual blocks, with sizes $2^l x 2^l$, for l=1,2,…,n, where $2^n x 2^n$ – image size [7],
- Cov – on the basis of the method described by formula (7) and (10)-(13).

The obtained test results shown in table 2, show, that the most accurate results - 16 (the lowest errors) were obtained using the method designated as Av2, especially for σ_n<8. However the method shown by (10-13) placed second, mainly for a 29x29 area. For a 9x9 area, no advantage of this method over the others was visible.

The conclusion that comes to mind is such, that an even bigger area should be selected for estimation. However there exists a risk that in the case of a lack of a flat area of such a large size, there will be many small details in it, which may worsen the results.

References

1. Amer, A., Schroder, H.: A New Video Noise Reduction Algorithm using Spatial Sub-Bands. In: IEEE Proceesings of International Conference on Electronics, Circuits and Systems, Rodos, Greece, vol. 1, pp. 45–48 (1996)
2. Ranham, M.R., Katsaggelos, A.K.: Digital Image Restoration. IEEE Signal Processing Magazine 3, 24–41 (1997)
3. Canny, J.: A Computational Approach to Edge Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 9, 679–698 (1986)
4. Dixon, R.L.: MRI: Acceptance Testing and Quality Control - The Role of the Clinical Medical Physicist. Medical Physics Publishing Corporation, Madison (1988)
5. Lee, J.S.: Refined Filtering of Image Noise using Local Statistics. Computer Vision, Graphics and Image Processing 15, 380–389 (1981)
6. Mastin, G.A.: Adaptive Filters for Digital Noise Smoothing: An Evaluation. Computer Vision, Graphics and Image Processing 31, 103–121 (1985)
7. Meer, P., Jolion, J., Rosenfeld, A.: A fast Parallel Algorithm for Blind Estimation of Noise Variance. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(2), 216–223 (1990)
8. Murtagh, F., Starck, J.L.: Image Processing Through Multiscale Analysis and Measurement Noise Modeling. Statistics and Computing Journal 10, 95–103 (2000)
9. Olsen, S.I.: Estimation of noise in images: An evaluation. Graphical Models and Image Processing 55(4), 319–323 (1993)
10. Pęksiński, J., Mikołajczak, G.: Differential Approximation of the 2-D Laplace Operator for Edge Detection in Digital Images. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part III. LNCS, vol. 6423, pp. 194–199. Springer, Heidelberg (2010)
11. Pęksiński, J., Mikołajczak, G.: Generation of FIR Filters by Using Neural Networks to Improve Digital Images. In: 34th International Conference on Telecommunications and Signal Processing (TSP), Budapest, Hungary, pp. 527–529 (2011)
12. Pęksiński, J., Mikołajczak, G.: The Synchronization of the Images Based on Normalized Mean Square Error Algorithm. Advances in Intelligent and Soft Computing 80, 15–25 (2010)
13. Rosin, P.: Thresholding for Change Detection. In: IEEE Proceedings of International Conference on Computer Vision, Bombay, India, pp. 274–279 (1998)
14. Vorhees, H., Poggio, T.: Detecting Textons and Texture Boundaries in Natural Images. In: Proceedings of the 1. International Conference on Computer Vision, London, England, pp. 250–258 (1987)

k-Means Clustering on Pre-calculated Distance-Based Nearest Neighbor Search for Image Search

Jing Yi Tou and Chun Yee Yong

Faculty of Information and Communication Technology (FICT)
Universiti Tunku Abdul Rahman (UTAR),
Jalan Universiti, Bandar Barat, 31900 Kampar, Perak, Malaysia
toujy@utar.edu.my, loveshe6@hotmail.com

Abstract. Content-based image retrieval (CBIR) would be an important future trend in search engines. This paper proposed a nearest neighbor search (NNS) method that uses *k*-means clustering and pre-calculated distances on a known set of image samples to be used for performing image queries within the set. The proposed algorithm adds a clustering step prior to the rest on an existing algorithm and uses the nearest clusters only for the NNS. The distance between the query images to the cluster is determined by using twice the standard deviation for the clusters to estimate the boundary of each cluster. The feature used is grey-level co-occurrence matrices (GLCM). This reduces both the samples explored by 25.21% and execution time by 26.62% for 16 chosen clusters within 23 clusters and a search radius of 0.2. The experimental results had shown an improvement in time complexity but on the same time sacrifices the hit rate that had dropped from 100% in the previous method that explores all potential samples but the proposed method only manage to achieve 70.77%.

Keywords: nearest neighbor search, content-based image retrieval, *k*-means clustering, grey level co-occurrence matrices.

1 Introduction

Content-based Image Retrieval (CBIR) is the process to retrieve images by comparing the similarity of the image contents rather than the current dominant text-based searches. CBIR involves two major steps, first a feature extraction step where a set of features that describes the images are retrieved and extracted for the next step where the features are used to match up with other images to find the best similarity match [1]. Multimedia technology is growing rapidly around the world today resulting in the existence of countless images over the World Wide Web. CBIR would be an efficient and effective tool for the users to find the desired images from the large pool of sources. Introduced by Kato in the early 1990's, CBIR mainly retrieves the image based on its visual feature such as color, texture and shape [2]. This is more efficient than the search based on keywords that are either found in the file name, metadata, image caption or main text, but are often leading to inaccurate results that are poorly labeled [3]. Google Goggle is a recent example of CBIR being applied in mobile

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 9–18, 2013.

phones as an application, it offers search capability based on captured images from the camera but is currently only restricted to the recognition of certain objects, such as logos, characters and etc [5].

In the work of Tou *et al.*, a nearest neighbor search (NNS) algorithm is proposed to use pre-calculated distances between a set of images to reduce the need of thoroughly exploring all training samples available during an image query. The work shown that only an average of 27.13% from the 1024 training samples for the Brodatz texture dataset are explored during the query process as compared to the *k*-nearest neighbor (*k*-NN) that requires all training samples to be compared against [5]. However, the exploration of training samples could have been further reduced.

The main objectives of this paper are:

- To propose the implementation to cluster the training samples prior to the NNS.
- To reduce the time duration for each image query.
- To reduce the number of images explored to find the target images of a query.

The following sections of this paper include: Section 2 to describe the proposed algorithm of the NNS with *k*-means clustering; Section 3 to describe the experimental materials and settings used in the experiments; Section 4 to discuss the experiments conducted, its results and analysis; Section 5 to draw a conclusion for the paper.

2 Proposed Methodology

In this paper, the proposed algorithm is a NNS algorithm based on the work of Tou *et al.* [5] but added with a clustering process before it in order to further reduce the exploration of samples during the image query. The algorithm is divided into two steps; 1) preparation stage; and 2) query stage. The first step segments the samples into clusters, calculates the features and distances of the training samples. The second step is to obtain the nearest neighbors for the query image based on a determined search radius and the information calculated in the first step.

2.1 Step 1: Preparation Stage

For the preparation stage of the training samples, the clustering will be first conducted to place samples into clusters. The textural features will then be calculated. After that, the distances between all the training samples will be calculated based on the features. These calculated distances are required to be used in the query stage.

***k*-Means Clustering.** The k-means clustering method is a popular clustering method that automatically groups a collection of samples into natural clusters where the number of clusters is determined by a positive integer value *k* [6]. After deciding a *k*, the centroids of the *k* clusters will be randomly initialized, the centroids will be recalculated and shifted accordingly until it converges [7]. The standard deviation σ of each cluster will be calculated for the estimation of the cluster boundaries *b* by doubling the standard deviation such as:

$$b = 2 \times \sigma \tag{1}$$

Textural Feature. Different features could be used for CBIR, e.g. textures [2]. Textural feature is used for the experimentation of the NNS algorithm proposed in this paper. A simple textural feature, the grey level co-occurrence matrices (GLCM) is selected for its simplicity of computation. The GLCM is introduced by Haralick *et al.* in 1973 where it remained popular when it comes to texture classification [8]. In this paper, the raw GLCM is used instead of generating the second order textural features [9]. In this paper, four GLCMs are generated for spatial distance of one pixel, eight grey levels and four directions, i.e. 0°, 45°, 90° and 135°. This produces 256 features for each sample [10].

Distances Calculation. Euclidean distance is selected as the distance metric in this paper, because it is easy and fast to compute for it [5]. The distances are calculated for each training samples against every other training samples based on the calculated GLCM features. For each training samples, the calculated distances are sorted in ascending order. The distance between the training sample and itself will always be zero and are not stored. The distance between one training sample A and another training sample B would also be identical for the reversed situation between B and A. Therefore, with the number of training samples n_T, the number of distances n_D that are required to be computed is [5]:

$$n_D = \left(n_T - 1\right)/2 \tag{2}$$

2.2 Query Stage

During the query of an image, the GLCM features will be calculated from the query image as described in Section 2.1. The calculated features will be used to compare against the training samples that falls within the search range in order to search for the final set of training samples that are nearest to the query image. The training samples are regarded as candidates during the search.

Radius Criterion. A search radius r defines the maximum distance allowed between the query image and the training sample retrieved. The r can also be defined as a constant value [5].

Selecting Nearest Clusters. Two comparison criterions are used in this paper for the selection of the nearest clusters. The first criterion is to compare the query image against the cluster mean which is the centroid of the clusters. The n clusters with the smallest distance will be selected as the nearest cluster. The second criterion is to compare the query image against the estimated cluster boundary that is achieved using Eq. (1). The distance d_B is determined by:

$$d_B = c - b \tag{3}$$

After that, clusters with the smallest d_B will be selected as the nearest clusters.

Selecting Nearest Neighbors. The query image will be compared against one of the training samples to obtain a distance *d*. A set of potential candidates will be selected from the selected nearest clusters if their pre-calculated distances to that particular training sample are fallen within the range of [*d* - *r*, *d* + *r*]. The process will be repeated within each selected candidate until every candidate in the potential list is fully examined using the same criteria. During this process, the candidate list generated for that particular candidate will be compared against the current candidate list and only those existing on both lists will remain in the current candidate list for the next processes until all candidates in the list are tested to fall within the range of [*d* - *r*, *d* + *r*] from the query image [5].

3 Experimental Settings and Materials

This section described the development tools and dataset used for the experiments that are described in Section 4.

3.1 Tools

The development tool used in this paper is MATLAB because it provides a number of useful toolboxes such as Image Processing Toolbox for GLCM and Bioinformatics Toolbox for *k*-means clustering and etc. MATLAB is also a helpful tool in fast development and testing of prototype for the evaluation of the proposed algorithm. The computer used for the experiment is a PC with an AMD Athlon 64 × 2 Dual Core Processor 6000+ with 3.01 GHz, 2 GB of RAM which is running on Windows XP Professional Service Pack 3.

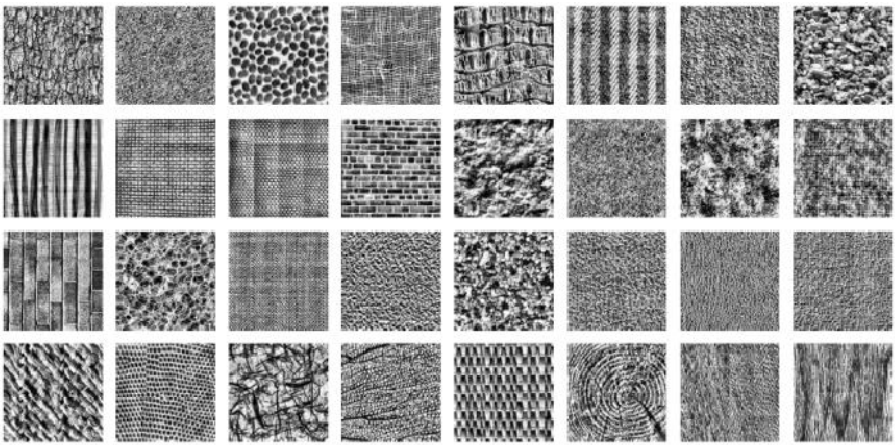


Fig. 1. The thirty-two textures selected from the Brodatz texture dataset [5]

3.2 Dataset

The Brodatz texture dataset is used for experiments in this paper. Out of the 112 textures, thirty-two are selected to be used [11]. Each sample of the textures are segmented into sixteen segments, each segments are rotated, scaled, as well as both rotated and scaled. Each sample has 64 × 64 pixels. Eight of the sixteen segments and their respective variations are randomly selected as the training samples and the remaining for testing samples. In this paper, ten sets of training and testing samples are randomly selected and the average results are shown in the experimental results [5]. The thirty-two textures are shown in Figure 1.

4 Experiments and Results

Three experiments are conducted to test and compare the proposed method in this paper against the previous method by Tou *et al.* [5]. The first experiment is conducted to compare the performance difference between the use of cluster mean and estimated cluster boundary for the distance calculation. The second experiment is conducted to identify the suitable parameter for the proposed method and the final experiment to compare the time performance of the proposed method against the previous method. The k for the k-means clustering is selected to be 23 in this paper.

4.1 Experiment 1: Cluster Mean and Estimated Cluster Boundary

This experiment is conducted to test the performance difference between the use of cluster mean and estimated cluster boundary that is obtained using the cluster standard deviation for distance measurements. The hit rate will be the performance measurement used for the evaluation. The number of selected clusters is represented by n for value of 1 to 12 and the radius is represented by r with the value of 0.2. The experimental results are as shown in Table 1.

Table 1. Comparison of hit rate (%) between calculations of distance against cluster mean and estimated cluster boundary for n for 1 to 12

n	Cluster Mean (%)	Estimated Cluster Boundary (%)
1	5.94	5.84
2	10.32	10.86
3	14.73	15.40
4	18.76	19.00
5	22.49	22.88
6	26.77	26.62
7	31.34	30.58
8	35.82	35.55
9	39.62	39.78
10	43.03	43.46
11	46.65	47.21
12	50.63	51.28

From the experimental results, the proposed method to compare based on the estimated cluster boundary using the standard deviation rather than the mean of the clusters had shown a slight improvement in the hit rate, with an overall improvement of 0.88%. This shows that by using the estimated cluster boundary, the experimental results may be better and could be improved, but even when more than half of the 23 clusters (12 clusters) are used for the experiment, the hit rate is still at an unsatisfied 51.28% when n is 12, where only half of the nearest samples would be extracted from the query. Therefore, the value of n needs to be further increased but shall not be too close to 23 or else it would defeat the original purpose of implementing the clustering.

The use of the cluster mean to find the nearest clusters would draw an issue when the cluster means are located closer to the query sample than other clusters that are actually having samples closest to the query image but were not selected because the clusters are naturally bigger and will have a further cluster mean to the query image as shown in Figure 2 where the star represents a query sample and the two clusters with cluster means closest to the query image is selected but the third cluster has samples closest to the query sample but is not selected because the cluster is bigger and naturally creates a larger distance from the query sample to its cluster mean.

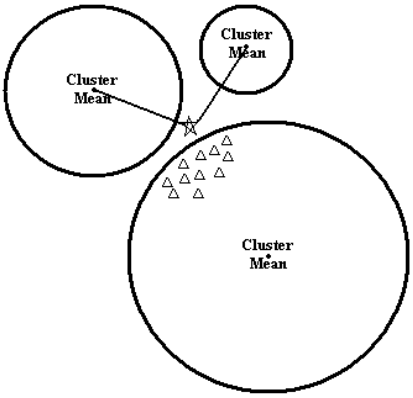


Fig. 2. Selecting nearest clusters based on comparison against the cluster means

To overcome the potential down fall of this issue, the cluster boundary is a better target for comparison than the cluster mean, but the cluster boundary is hard to be computed, therefore a simple estimation based on the standard deviation of the clusters is used. This would ensure the clusters with the closest boundaries would be selected as shown in Figure 3.

4.2 Experiment 2: Proposed Method Using Estimated Cluster Boundary

In this experiment, the n is selected to be from 13 to 20 to examine the performance of higher n and further tested for a range of r from 0.1 to 0.5. The comparison of the hit rate against the n and r are shown in Table 2 where the horizontal header represents the r.

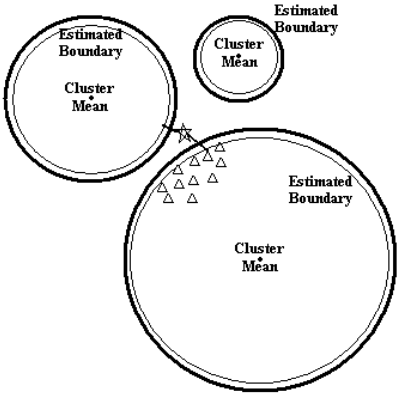


Fig. 3. Selecting nearest clusters based on comparison against the estimated cluster boundaries based on the cluster standard deviation

Table 2. Comparison of hit rate (%) for the proposed method using estimated cluster boundary for *n* of 13 to 20 and *r* of 0.1 to 0.5

n	0.1	0.2	0.3	0.4	0.5
13	57.69	55.63	55.71	55.67	55.66
14	62.65	60.79	60.98	60.9	60.90
15	65.84	64.47	64.49	64.43	64.44
16	71.3	70.77	70.97	70.91	70.92
17	75.24	74.42	74.51	74.55	74.59
18	78.41	77.57	77.66	77.77	77.83
19	83.04	81.97	82.03	82.1	82.15
20	84.72	85.48	85.66	85.62	85.65

The experimental results show that the differences between the uses of different *r* cast only little effect to the hit rate but the *n* plays a larger role. However, the larger the *n*, it means that more cluster needs to be explored and could potentially bring down the computational time and therefore, the selection of *n* should not be too close to *k*. The rates for the percentage of nearest neighbor (NN) selected against the percentage of sample explored are shown from Table 3 to Table 7 for *r* of 0.1 to 0.5 respectively.

From the experimental results, it is shown that the situation is similar to the findings of the work of Tou *et al.* [5] where the results with *r* is 0.1 show a very low selected neighbor due to the small distance and the larger the *r*, the larger proportion of the training samples would be explored and selected, where 74.72% of the samples are selected when the *r* is 0.5 and *n* is 20. In this paper, *r* is 0.2 is selected as a better choice as it only explored up to 31.22% of samples and select 26.79% as nearest neighbors within these explored samples, when *n* is 20. A target hit rate of at least 70% is aimed and therefore the *n* is selected to be 16 since it produces hit rate of 70.77% during this *n* value.

Table 3. The NN selected (%), sample explored (%) and the percentage of NN selected against the sample explored (%) for *n* of 13 to 20 and *r* of 0.1

n	NN Selected	Sample Explored	(NN Selected) / (Sample Explored)
13	0.21	5.12	4.04
14	0.22	5.50	4.08
15	0.24	5.76	4.10
16	0.26	6.20	4.12
17	0.27	6.45	4.18
18	0.28	6.66	4.22
19	0.30	6.94	4.29
20	0.30	7.18	4.23

Table 4. The NN selected (%), sample explored (%) and the percentage of NN selected against the sample explored (%) for *n* of 13 to 20 and *r* of 0.2

N	NN Selected	Sample Explored	(NN Selected) / (Sample Explored)
13	5.44	21.13	25.76
14	5.95	22.94	25.92
15	6.31	24.13	26.14
16	6.92	26.32	26.31
17	7.28	27.50	26.48
18	7.59	28.54	26.59
19	8.02	29.99	26.75
20	8.36	31.22	26.79

Table 5. The NN selected (%), sample explored (%) and the percentage of NN selected against the sample explored (%) for *n* of 13 to 20 and *r* of 0.3

n	NN Selected	Sample Explored	(NN Selected) / (Sample Explored)
13	22.31	38.78	57.52
14	24.42	43.33	57.68
15	25.82	44.72	57.74
16	28.42	49.09	57.88
17	29.83	51.51	57.92
18	31.10	53.66	57.95
19	32.85	56.54	58.10
20	34.30	58.92	58.21

Table 6. The NN selected (%), sample explored (%) and the percentage of NN selected against the sample explored (%) for *n* of 13 to 20 and *r* of 0.4

n	NN Selected	Sample Explored	(NN Selected) / (Sample Explored)
13	39.47	48.04	82.20
14	43.20	52.50	82.29
15	45.70	55.52	82.32
16	50.30	61.03	82.42
17	52.88	64.10	82.50
18	55.17	66.83	82.55
19	58.24	70.50	82.61
20	60.73	73.50	82.63

Table 7. The NN selected (%), sample explored (%) and the percentage of NN selected against the sample explored (%) for *n* of 13 to 20 and *r* of 0.5

n	NN Selected	Sample Explored	(NN Selected) / (Sample Explored)
13	48.56	52.86	91.87
14	53.13	57.81	91.91
15	56.22	61.15	91.94
16	61.88	67.23	92.03
17	65.07	70.67	92.08
18	67.90	73.72	92.11
19	71.67	77.77	92.15
20	74.72	81.10	92.14

4.3 Experiment 3: Execution Time

For this experiment, the selected parameter of *n* is 16 and *r* is 0.2 is used to compare the previous method in Tou *et al.* [5] and the proposed method with comparison against the estimated cluster boundaries. The comparison of the execution time for a single query, the percentage of sample explored and the hit rate is shown in Table 8.

From the experimental results, it is shown that the proposed method had improved the search speed by nearly 25.21% with 26.62% less samples required to be explored. However the hit rate was lowered to 70.77% because the method focuses on exploring fewer samples without the need to return all samples as the previous method. The proposed method therefore had shown a decrease of computational time by about 25% but at the same time lost nearly 30% of hit rate for the results falling in the nearest distance under the selected radius.

Table 8. Comparison of execution time (ms), sample explored (%) and hit rate (%) for the previous method [6] and the proposed method with k-means clustering and the use of estimated cluster boundaries

	Execution Time (ms)	Sample Explored (%)	Hit rate (%)
Previous Method	2118	35.87	100.00
Proposed Method	1584	26.32	70.77

5 Conclusion

The experimental results showed that the proposed method would be able to reduce both the samples explored and the execution time for each query by 25.21% and 26.62% respectively but in the mean time, lost 29.23% of hit rate for *r* of 0.2 and *n* of 16. This is because the proposed method is not focused on searching for all nearest neighbors within the defined search radius but only selects 16 out of the 23 clusters for comparison and selection of best neighbors and therefore reduces the hit rate.

For the proposed method, this paper selects to use the estimated cluster boundary instead of the cluster mean due to potential downfall that could happen due to the different sizes of clusters.

In the future, work shall be carried out to ensure that the ratio of NN selected against the samples explored shall be further reduce, the current ratio is as low as 26.31% indicating that most examined samples are not selected. To solve this problem, a better criterion should be determined for the selection of the samples to be explored based on higher probability to be a nearest neighbor than other potential candidates but such work had not been carried out.

References

1. Niblack, W., Barber, R., Equitz, W., Flickner, M., Glasman, E., Pektovic, D., Yanker, P., Faloutsos, C., Taubin, G.: The QBIC project: querying images by content using color, texture, and shape. In: Storage and Retrieval for Image and Video Databases, San Jose, CA (1993)
2. Yoshitaka, A., Ichikawa, T.: A Survey on Content-Based Retrieval for Multimedia Databases. IEEE Transactions on Knowledge and Data Engineering 11(1), 81–93 (1999)
3. Veltkamp, R.C., Tanase, M.: Content-Based Image Retrieval Systems: A Survey. Technical Report UU-CS-2000-34 (2000)
4. Google Goggles, `http://www.google.com/mobile/goggles/`
5. Tou, J.Y., Tay, Y.H., Lau, P.Y.: Exploiting Pre-Calculated Distance In Nearest Neighbour Search on Query Images For CBIR. In: Proceedings International Workshop on Advanced Image Technology 2010 (2010)
6. Frahling, G., Sohler, C.: A Fast k-Means Implementation using Coresets. In: Proceedings of the Twenty-Second Annual Symposium on Computational Geometry, SoCG (2006)
7. Mount, M.D., Kanungo, T., Nethanyahu, N.S.: An Efficient k-Means Clustering Algorithm Analysis and Implementation. IEEE Transaction of Pattern Analysis and Machine Intelligence 24, 881–891 (2002)
8. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural Features for Image Classification. IEEE Transactions on Systems, Man and Cybernatics, 610–621 (1973)
9. Tou, J.Y., Khoo, K.K.Y., Tay, Y.H., Lau, P.Y.: Evaluation of Speed and Accuracy for Comparison of Texture Classification Implementation. In: Proceedings International Workshop on Advanced Image Processing 2009 (2009)
10. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image Retrieval: Ideas, Influences, and Trends of the New Age. ACM Computing Surveys 40(2) (2008)
11. Ojala, T., Pietikainen, M., Kyllonen, J.: Gray level Cooccurrence Histograms via Learning Vector Quantization. In: Proc. 11th Scandinavian Conference on Image Analysis, pp. 103–108 (1999)

A New Approach for Collaborative Filtering Based on Mining Frequent Itemsets

Phung Do[1], Vu Thanh Nguyen[1], and Tran Nam Dung[2]

[1] University of Information Technology, Ho Chi Minh City, Vietnam
{phungdtm,nguyenvt}@uit.edu.vn
[2] University of Natural Science, Ho Chi Minh City, Vietnam
trannamdung@yahoo.com

Abstract. As one of the most successful approaches to building recommender systems, collaborative filtering (*CF*) uses the known preferences of a group of users to make recommendations or predictions of the unknown preferences for other users. In this paper, we first propose a new CF model-based approach which has been implemented by basing on mining frequent itemsets technique with the assumption that "*The larger the support of an item is, the higher it's likely that this item will occur in some frequent itemset, is*". We then present the enhanced techniques such as the followings: bits representations, bits matching as well bits mining in order to speeding-up the algorithm processing with CF method.

Keywords: Collaborative Filtering, mining frequent itemsets, bit matching, bit mining.

1 Introduction

As one of the most successful approaches to building recommender systems, collaborative filtering (CF) uses the known preferences of a group of users to make recommendations or predictions of the unknown preferences for other users. There are three main categories of CF techniques: memory-based, model-based, and hybrid CF algorithms (that combine CF with other recommendation techniques). The memory-based CF methods are deployed into commercial systems such as http://www.amazon.com/ and Barnes and Noble, because they are easy-to-implement and highly effective [4, 5]. The main drawback of memory-based methods are the requirement of loading a large amount of in-line memory. Well-known memory-based CF techniques include neighbor-based CF (item-based/user-based CF algorithms with Pearson/vector cosine correlation) [6, 13] and item-based/user-based top-N recommendations [14]. The problem is serious when rating matrix becomes so huge in situation that there are extremely many persons using system. Computational resource is consumed much and system performance goes down; so system can't respond user require immediately. Model-based CF techniques use the pure rating data to estimate or learn a model to make predictions [6]. The model can be a data mining or machine learning algorithm. Well-known model-based CF techniques include Bayesian belief nets (BNs) CF models [6–8], clustering CF models [9, 10], latent semantic CF models [5], MDP (Markov decision process) - based

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 19–29, 2013.

CF [11] and CF using dimensionality reduction techniques, for example, SVD, PCA [12]. Model-based CF achieved real-time response when inference speed much faster than calculated on the entire data in memory, but the time for building the model is slowly. In this paper, we proposed a model-based CF approach based on mining frequent itemsets and bit matching technique in order to speeding-up the algorithm processing with CF methods. In addition, the proposed approach is to increase the usefulness of recommendations to present those items that user interest by discovering the user's purchasing patterns. In section *2* we propose an idea for the model-based CF algorithm based on mining frequent itemsets. The heuristic algorithm is discussed carefully in the section *3*. We propose an abstract architecture along with a framework which assists them in implementing and evaluating CF algorithm in the section *4*. Section *5* is the evaluation. Section *6* is the conclusion. Note that terms such as "rating matrix", "dataset" and "database" have the same meaning in this paper.

2 A New CF Algorithm Based on Mining Frequent Itemsets

With the following given rating vectors, for instance, *u* = (*item 1 = 3, item 2 = 5, item 3 = 2*). It means that user *u* rated on *item 1, item 2, item 3*. Where, their values are *3*, *5* and *2*, respectively. Then the concept of creating this new CF algorithm, based on mining frequent itemsets is to consisting of two following processing steps:

- Modeling process: A set of frequent itemsets *S* is mined and it is performed in offline process mode.
- Recommendation process: whenever user *u* requires to get recommended items, a frequent itemset *s* is chosen from *S* so that *s* contains items *1*, *2* and *3*, for instance, *s* = (*item 1, item 2, item 3, item 5, item 7*). The additional items *5* and *7* are then recommended to user. This meant recommendation isn't like the modeling process; instead it's an on-line process.

Although the modeling process does consume much more time than that of the recommendation one. But, it is executed in offline mode. Therefore it won't be causing any negative-time consuming impact on recommendation process. However, there would be a serious problem that could be raised, when the frequent itemset *s* = (*item 1, item 2, item 3, item 5, item 7*) didn't give any indication that which rating values items *1, 2 ,* and *3* have been assigned. It is obvious to know that items *1, 2* and *3* are rated by the values of *3, 5* and *2*, respectively in rating vector *u*. This means the rating vector *u* and the frequent itemset *s* don't match exactly. This eventually causes another hazard which is impossible when launching an attempt to compute predictive values, with missing ratings for rating vector. Now please pay attention, this problem will be eliminated or solved by using the technique so-called *bit transformation*. Note that the terms "*bit*" and "*binary*" have the same meaning.

For instance, a rating matrix, where its rows indicate users, its columns indicate items and each cell is the rating which user has given to item. With the ratings are in a range from *1* to *5* {*1…5*}, then, the sample rating matrix is shown as what will be seen in *Table 1*. The value 5 indicates the most preference.

Each item is "stretched" into *5* sub-items which are respective to *5* possible rating values {*1*…*5*}. Each sub-item is symbolized as *item_j_k* carrying two binary-states *1* and *0*, which indicates whether user rates on item *j* with concrete value *k*. For example, the bit *item_2_5* getting state *1* shows that user gave rating value *5* on item *2*. Now the rating matrix is transformed into bit rating matrix in which each cell is the rating of bit sub-item. Supposing that, empty cell has shown such cell get valued by *0*; it means that there is no one to give a rating on the cell yet. The bit rating matrix is shown in *Table 2*.

Table 1. Rating matrix table

	Item 1	Item 2	Item 3	Item 4
User 1	*3*	*5*	*2*	*1*
User 2	*3*	*5*	*2*	*1*
User 3	*1*	*5*	*4*	

Table 2. Bit rating matrix table

	User 1	User 2	User 3
Item_1_1	*0*	*0*	*1*
Item_1_3	*1*	*1*	*0*
Item_2_5	*1*	*1*	*1*
Item_3_2	*1*	*1*	*0*
Item_3_4	*0*	*0*	*1*
Item_4_1	*1*	*1*	*0*

Each frequent itemset, that has been extracted from bit rating matrix and, it will carry a so-called bit form, s = ($item\_j_1\_k_1$, $item\_j_2\_k_2$,…). Where, each component *item_j_k*, has been defined as bit sub-item. After that, rating vector u is also to be transformed into bit rating vector u = ($item\_j_1\_k_1$, $item\_j_2\_k_2$,…). It's so easy to find that matching the twos, bit frequent itemset, to bit rating vector is completely simple. For instance, if using the shown previous examples; where, the rating vector u = (*item1 = 3, item 2 = 5, item 3 = 2*) is to be transformed into u = (*item_1_3, item_2_5, item_3_2*). While the frequent itemsets are s_1 = (*item_1_3, item_2_5, item_3_2, item_4_1*) and s_2 = (*item_1_1, item_2_5, item_3_4*). We also find that itemset s_1, has been matched at the most, to u and so, the item *4* is recommended to user with predictive value *1*.

Now the previous mentioned problem is solved but our algorithm should be enhanced for a little more better. Suppose that the number of frequent itemsets is huge and even each itemset has also a lot of items. When we are to match the rating vector and frequent itemset, there will be a boom of combinations that may cause computer system collapsed or consumed an even a whole lot more of processing time. Therefore, we propose an enhancement method for matching purpose, based on the technique, called *bit matching*.

2.1 Bit Representation and Bit Matching

Suppose there are *4* items and each item has *5* possible rating values, we use the bit set whose length is *4 * 5 = 20* bits (so-called *20*-length bit set) to represent rating vectors and frequent itemsets. The bit set is divided into many clusters or groups, for example, if each item has *5* possible rating values then each cluster has *5* bits. So each cluster represents a sub-item and the position of a bit in its cluster indicates the rating

value of corresponding sub-item. If a cluster contains a bit which is set, its corresponding sub-item is rated with the value which is the position of such set bit. Following is an example of bit set:

Table 3. Bit representation

0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
Cluster 1 (item 1 = 3)					Cluster 2 (item 2 = 5)					Cluster 3 (item 3 = 2)					Cluster 4				

For example, rating vector u = (*item1 = 3, item 2 = 5, item 3 = 2*) is transformed into u = (*item_1_3, item_2_5, item_3_2*) which is represented as u = (*00100 00001 01000 00000*) having four clusters. The frequent itemset s_1 = (*item_1_3, item_2_5, item_3_2, item_4_1*) which is represented as s_1 = (*00100 00001 01000 10000*). The frequent itemset s_2 = (*item_1_1, item_2_5, item_3_4*) which is represented as s_2 = (*10000 00001 00010 00000*). In order to match s_1 (or s_2) with u, all we need is to do AND bit-operation between s_1 (or s_2) and u.

- If s_1 *AND* $u = u$ then s_1 matches with u
- If s_1 *AND* $u \neq u$ then s_1 doesn't match with u

When s_1 get matched with u, we do *AND – NOT* operation, as to extract items which are recommended to users. Suppose, the recommended item is denoted *r_item*:

$$r\_item = s_1 \; AND \; (NOT \; u) = (00000\ 00000\ 00000\ 10000)$$

From this bit set, it is easy to recognize that item *4* is recommended with predict value is *1* because the first bit of 4^{th} cluster is set.

As a result, our algorithm will consist of 3 steps:

- Step 1: Rating matrix is transformed into bit rating matrix.
- Step 2: Bit rating matrix is mined, as well as to extract frequent itemsets.
- Step 3: Rating vector and frequent itemsets are represented as bit sets. Bit matching operations are performed in order to find out the appropriate frequent itemset which is matched with rating vector. Basing on such frequent itemset, it is possible to determine which items are recommended. Moreover missing values of recommended items can be also predicted.

2.2 Pseudo-code Like C for New CF Algorithm

Let *D, B, S* be rating matrix, bit rating matrix and the set of frequent itemsets, respectively. Let *matched_itemset* and *r_item* be matched itemset and recommended item, respectively. Let *bitset*(…), *count*(…) be functions that transforms item into bit set and counts the number of bit *1* (s) in bit set. Let *bit_transform* be the function which transforms rating matrix into bit rating matrix. Let *mining_frequent_itemset* be the mining function which extracts frequent itemsets from bit rating matrix (see *sections 3.1, 3.2*). Following is the pseudo-code like C for our CF algorithm:

```
B = bit_transform(D)
S = mining_frequent_itemset(B)
matched_itemset = null
max_count = -1
For each s ∈ S
   bs = bitset(u) AND bitset(s)
   If bs = bitset(u) && count(bs) > max_count then
      matched_itemset = s
      max_count = count(bs)
   End If
End For
r_item = bitset(matched_itemset) AND (NOT bitset(u))
```

The second step is the most important, since it's very often, ones are to asking that. Such as, there is a question: "How frequent itemsets are extracted from rating matrix". This question is, then answered in the next section about mining frequent itemsets.

3 Mining Frequent Itemsets

Our mining frequent itemsets method is based on the assumption: "*The larger the support of an item is, the higher it's likely that, this item occurs in some itemset*". In other words, items with the high support tend to combine together so as to form a frequent itemset. So our method is the heuristic algorithm so-called *Roller* algorithm. The basic idea is similar to that of a white-wash task. Suppose you imagine that there is a wall and there is the dataset (namely, rating matrix) containing all items. Such dataset is modeled as this wall. On the wall, all items are shown in a descending ordering of their supports; it means that the higher frequent item is followed by the lower frequent item. Moreover, we have a roller and we roll it on the wall, from item to item, with respect to the descending ordering. If an item is found, satisfied at a minimum support (*min_sup*), it is, then added to the frequent itemset and the rolling task is continued to keep moving on, until there is no item that meets minimum support. The next time, all items in this frequent itemset are removed from the meant wall and the next rolling task will be performed to find out new frequent itemset.

Our algorithm includes four following steps:

- Step 1: Computing the supports of all items and arranging these items on the wall, according to the descending ordering of their supports. Note that all items whose supports don't meet minimum support are removed from this descending ordering. The kept items are called the frequent items.
- Step 2: The i^{th} itemset is initialized by the first item in this descending ordering. The support of i^{th} itemset is initialized as the support of this first item. The current item now is the first item and it is removed from descending ordering.
- Step 3: If there is no item in descending ordering, the algorithm will be terminated. Otherwise:
  - 3.1. If the current item is the last one, in descending ordering, then all items in the i^{th} itemset are removed from the descending ordering and the number i is increased by *1* ($i = i + 1$). Go to step 2.
  - 3.2. If the current item is NOT the last in descending ordering, then, the next item is picked and so the current item now is the next item.

- Step 4: Checking the support of current item:
  - The support of current item satisfies *min_sup*: the support of the i^{th} itemset *support*(i^{th} *itemset*) is accumulated by current item; it is the count of total transactions that contains all items in both the i^{th} itemset and current item. If *support*(i^{th} *itemset*) is equal to or larger than *min_sup*, this item is added to the i^{th} itemset.
  - Go back step *3*.

It is easy to recognize that step *3* and *4* are similar to that of a white-wash task which "rolls" the i^{th} itemset modeled as the roller. After each rolling (each iteration), such itemset get thicker with more items.

Let $I = (i_1, i_2,..., i_m)$ and S be a set of item and a set of frequent itemset, respectively. Let $O = (o_1, o_2,..., o_n)$ be the list of items whose supports are sorted according to descending ordering, $O \subseteq I$. Let s_i be the i^{th} itemset. Let c be the current item. Let *support*(…) be the function calculating the support of item or itemset. Let *sort*(…), *first*(…), *next*(…), *last*(…) be sorting, getting first item, getting next item, getting last item functions, respectively. Following is the pseudo-code like C for Roller algorithm (function *mining_frequent_itemset*):

```
O = sort(I)
i = 1
While (true)
   c = first(O)
   s_i = s_i ∪ {c}
   O = O / {c}
   If O = ∅ then return S
   While (true)
      If c = last(O) then
         S = S ∪ s_i
         O = O / S
         i = i +1
         break
      Else
         c = next(O, c)
         If support(c) < min_sup continue
         b = bitset(S) AND bitset(c)
         If count(b) ≥ min_sup then
            s_i = s_i ∪ {c}
         End If
      End If
   End While
End While
```

Although the Roller algorithm may ignore some frequent itemsets but it runs much faster than traditional mining frequent itemsets methods. Especially our algorithm can be enhanced by using a so-called technique of bit mining.

3.1 Bit Mining

When rating matrix (dataset) is transformed into bit rating matrix, item and itemset become cluster (sub-item) and bit set (see *section 2*). The support of item or itemset are the number of bits whose values are 1 (*s*) in bit set. Given bit rating matrix as

above table (*see table 2*). In step *1*, sub-items are sorted according to descending ordering and some sub-items not satisfying *min_sup* are removed given the *min_sup* is *2*. Now sub-items are represented as bit cluster: *Item_2_5* = (*111*), *Item_1_3* = (*110*), *Item_3_2* = (*110*), *Item_4_1* = (*110*).

In step *2*, the first itemset s_1 is initialized as *Item_2_5*

s_1 = (*111*) and *support*(s_1) = *count* (111) = *3*

Where *count* (…) indicates the number of bits whose values are *1* (s) in bit set (…).

In step *3* and *4*, sub-items (clusters) such as *Item_1_3, Item_3_2, Item_4_1* are picked in turn and all of them satisfy *min_sup*.

- Picking *Item_1_3*: $s_1 = s_1$ *AND Item_1_3*=(*111*) *AND* (*110*) = (*110*) → *support*(s_1) = *2*.
- Picking *Item_3_2*: $s_1 = s_1$ *AND Item_3_2* = (*110*) *AND* (*110*) = (*110*) → *support*(s_1) = *2*.
- Picking *Item_4_1*: $s_1 = s_1$ *AND Item_4_1* = (*110*) *AND* (*110*) = (*110*) → *support*(s_1) = *2*.

Finally, the frequent itemset is s_1 = (*110*) which include *Item_2_5, Item_1_3, Item_3_2, Item_4_1*. We recognize that the bit set of frequent itemset, named s_1 is accumulated by frequent item after each iteration. This make algorithm runs faster. The cost of counting bit set and performing bit operations isn't significant.

3.2 The Improvement of Roller Algorithm

Roller algorithm may lose some frequent itemsets because there is a case in that some frequent items don't have so high a support (they are not excellent items) and they are in the last of descending ordering. So they don't have many chances to join to frequent itemsets. However they really contribute themselves into some frequent itemset because they can combine together to build up frequent itemset, but they don't make the support of such itemset decreased much. It is difficult to discover their usefulness. In order to overcome this drawback, the Roller algorithm is modified so that such useful items are not ignored.

So in step *3*, instead of choosing the next item as the current item, we can look up an item whose support is *pseudo-maximum* and choose such item as the current item. In step *3.2*, if the current item is NOT the last in descending ordering, we look up the item which is combined (AND operation) with i^{th} itemset so as to form the new itemset whose support is maximum. Such item as being the so-called *pseudo-maximum support item* is chosen as the current item.

The improved Roller algorithm take slightly more time than normal Roller algorithm for looking up *pseudo-maximum support item* in step *3* but it can discover more frequent itemsets. So its accuracy is higher than normal Roller algorithm.

4 Algorithm Implementation in General Architecture

The new CF is implemented and evaluated according to the general architecture interpreted by UML language has 4 basic interfaces and classes: *Algorithm*, *KBase*, *Dataset* and *Evaluator*. Such interfaces and classes are considered as the software-engineering standard for CF algorithm. Researcher will conform to such standard when they apply this framework into writing a new algorithm.

- Interface *Algorithm* represents abstract algorithm. The main task that researchers does is to realize this interface according to their goals when they invent a new algorithm. In most cases, they implement directly two classes *MemoryBasedCF* and *ModelBasedCF* which are derived from *Algorithm*. *MemoryBasedCF* and *ModelBasedCF* represent memory-based *CF* algorithm and model-based *CF* algorithm, respectively.
- Interface *KBase* represents knowledge base which associates with a model-based algorithm *ModelBasedCF*. The structure of *KBase* is very flexible and it depends on ideas and purposes of algorithm.
- Class *Dataset* is composed of a rating matrix and personal profiles. Each row of rating matrix is represented by class *RatingVector*. Personal profile is represented by class *Profile*. Framework is responsible for manipulating *Dataset*.
- Class *Evaluator* is used by framework in order to evaluate algorithm according to criterions so-called *measures* such as time, precision and recall. Such measures are defined inside *Evaluator*. *Evaluator* reads and feeds dataset on algorithm *Algorithm*. Finally, it evaluates such algorithm by calculating *Measures* based on result of executing algorithm.

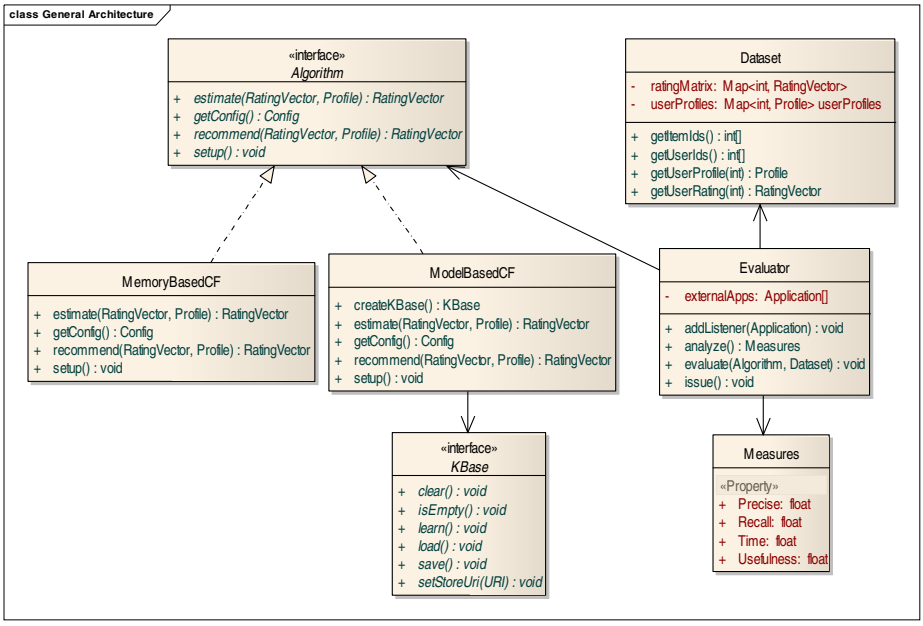


Fig. 1. General architecture for CF algorithms

The CF algorithm in this paper is realized as following steps:

- Interface *Algorithm* is impemented according to the goal and schema of this CF
- Maximum frequent itemsets are represented as *KBase*. So Roller algorithm which finds frequent itemsets is implemented as method *KBase::learn*().
- This CF is executed by the *Evaluator* in such framework
- The evaluation measures on this CF are computed by such framework after this execution.

5 Evaluation

Database *Movielens* [1] including *100,000* ratings of *943* users on *1682* movies is used for evaluation. Database is divided into 5 folders, each folder includes one training set over 80% whole database and one testing set over 20% whole database. Training set and testing set in the same folder are disjoint sets.

Table 4. Evaluation result

	Our method	**Cosine**	**Pearson**	**Simple**
MAE	0.21459	0.37278	0.45747	0.33391
MSE	0.10285	0.23887	0.31175	0.24616
RMSE	0.32041	0.48485	0.51458	0.47705
Precision	0.10554	0.00064	0.00046	0.00077
Recall	0.04042	0.00024	0.00018	0.00028
F1	0.05758	0.00035	0.00026	0.0004
Spearman	0.12540	-1	0	0
Time	0.00491	1.2117	2.10073	0.11006

The system setting includes: Processor Pentium(R) Dual-Core CPU E5700 @ 3.00GHz, RAM 2GB, Available RAM 1GB, Microsoft Windows 7 Ultimate 2009 32-bit, Java 7 HotSpot (TM) Client VM. Our CF method is compared to three other methods: *simple method* – simplest memory-based CF algorithm, c*osine method* – memory-based CF algorithm in which the cosine measure is used, *Pearson method* – memory-based CF algorithm in which the Pearson measure is used and Green Fall method – model-based CF using mining frequent itemsets technique.

There are 8 metrics used in this evaluation: *MAE, MSE, RMSE, recall, precision, F1, Spearman correlation* [4] and *time*. Note that all metrics except time metric are normalized in range [0, 1] and time metric is calculated in seconds. Table 4 is the evaluation result.

Our method is much better than cosine, Pearson, simple methods in all aspects: less error ratio via *MAE*, *MSE* and *RMSE* metrics; more accuracy via recall, precision and

F1 metrics; higher correlation via Spearman metric. The best thing is that it runs much faster than other methods.

6 Conclusion

Our CF approach is different from other model-based CF methods when trying to discover user interests. The mining technique is important for extracting frequent itemsets considered as patterns of user interests. However traditional mining algorithms consume much more time and resources. So we proposed a new mining method, a so-called Roller algorithm. Based on evaluation measures, Roller is proved as reliable algorithm with high performance, fast speed, high usefulness and consuming less time and resources. Its sole drawback is that it may ignore some user patterns because of heuristic assumption. However this drawback is alleviated by taking advantage of enhancement techniques such as bit mining, the concept of *pseudo-maximum support*.

In the future, we will propose another new model-based CF method which uses Bayesian network in inferring user interests. Such method based on statistical mechanism will be compared to the method in this paper so that we have an open and objective viewpoint about mining technique and statistical technique.

References

1. Movielens dataset 2011. Home page is `http://www.movielens.org`, Download dataset from `http://www.grouplens.org/node/12`
2. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Elsevier Inc. (2006)
3. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating Collaborative Filtering Recommender Systems. ACM Transactions on Information Systems 22(1), 5–53 (2004)
4. Linden, G., Smith, B., York, J.: Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Computing 7(1), 76–80 (2003)
5. Hofmann, T.: Latent semantic models for collaborative filtering. ACM Transactions on Information Systems 22(1), 89–115 (2004)
6. Breese, J., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, UAI 1998 (1998)
7. Miyahara, K., Pazzani, M.J.: Collaborative filtering withthe simple Bayesian classifier. In: Proceedings of the 6th Pacific Rim International Conference on Artificial Intelligence, pp. 679–689 (2000)
8. Su, X., Khoshgoftaar, T.M.: Collaborative filtering for multi-class data using belief nets algorithms. In: Proceedings of the International Conference on Tools with Artificial Intelligence, ICTAI 2006, pp. 497–504 (2006)
9. Ungar, L.H., Foster, D.P.: Clustering methods for collaborative filtering. In: Proceedings of the Workshop on Recommendation Systems. AAAI Press (1998)

10. Chee, S.H.S., Han, J., Wang, K.: RecTree: an efficient collaborative filtering method. In: Proceedings of the 3rd International Conference on Data Warehousingand Knowledge Discovery, pp. 141–151 (2001)
11. Shani, G., Heckerman, D., Brafman, R.I.: An MDP-based recommender system. Journal of Machine Learning Research 6, 1265–1295 (2005)
12. Billsus, D., Pazzani, M.: Learning collaborative information filters. In: Proceedings of the 15th International Conference on Machine Learning, ICML 1998 (1998)
13. Sarwar, B.M., Karypis, G., Konstan, J.A., Riedl, J.: Analysis of recommendation algorithms for E-commerce. In: Proceedings of the ACM E-Commerce, Minneapolis, Minn, USA, pp. 158–167 (2000)
14. Sarwar, B.M., Karypis, G., Konstan, J.A., Riedl, J.: Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International Conference on World Wide Web, WWW 2001, pp. 285–295 (May 2001)

Reduction of Training Noises for Text Classifiers

Rey-Long Liu

Department of Medical Informatics, Tzu Chi University, Hualien, Taiwan
rlliutcu@mail.tcu.edu.tw

Abstract. Automatic text classification (TC) is essential for the archiving and retrieval of texts, which are main ways of recording information and expertise. Previous studies thus have developed many text classifiers. They often employed training texts to build the classifiers, and showed that the classifiers had good performance in various application domains. However, as the training texts are often inevitably unsound or incomplete in practice, they often contain many terms not related to the categories of interest. Such terms are actually *training noises* in classifier training, and hence can deteriorate the performance of the classifiers. Reduction of the training noises is thus essential. It is also quite challenging as training texts are unsound or incomplete. In this paper, we develop a technique TNR (Training Noise Reduction) to remove the possible training noises so that the performance of the classifiers can be further improved. Given a training text d of a category c, TNR identifies a sequence of consecutive terms (in d) as the noises if the terms are not strongly related to c. A case study on the classification of Chinese texts of disease information shows that TNR can improve a Support Vector Machine (SVM) classifier, which is a state-of-the-art classifier in TC. The contribution is of significance to the further enhancement of existing text classifiers.

1 Introduction

Information is often expressed in textual form and classified into categories to facilitate archiving and retrieval. Automatic text classification (TC) is thus essential. It aims at two goals: (1) *classifying* relevant documents into suitable categories, and (2) *filtering out* those documents that are not relevant to any of the categories of interest. The former goal is to determine the proper categories of *in-space* documents, which are those documents that belong to the category space of interest. The latter goal is to filter out all *out-space* documents, which are those documents that fall out of the category space of interest. Filtering of out-space documents is particularly essential as text classifiers are often built for a limited application domain and hence most real-world documents are actually out-space documents for the domain.

Previous studies often integrated the two goals in a seamless way. Each category is associated with a classifier that can autonomously make a decision of accepting or rejecting a document. Therefore, a document can be classified into zero, one, or several categories, and when the document is not classified into any category, it is actually filtered out by the classifiers. One popular way to build the text classifiers is

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 30–39, 2013.

to train the classifiers using a set of training texts that have been tagged with proper category labels. However, as the training texts are often inevitably unsound or incomplete in practice, they often contain many terms that are not really related to the categories of the texts. Such terms are actually *training noises* in classifier training, and hence can deteriorate the performance of the classifiers.

1.1 Problem Definition and Motivation

In this paper, we develop a technique TNR (Training Noise Reduction) to remove the possible training noises for the text classifiers. Reduction of the training noises is essential. By proper reduction of the noises, the classifiers can be trained more properly and hence the classifiers can have better performance in classifying in-space documents and filtering out out-space documents.

Technically, reduction of training noises is challenging since training texts are inevitably unsound or incomplete. In response to the challenge, TNR employs *term proximity* information as the key evidence to identify the noises in the training texts. Given a training text d of a category c, TNR identifies a sequence of consecutive terms (in d) as the noises if the terms are not strongly related to c. The idea is motivated by the observation: those terms (in d) that have many neighboring terms not related to c may simply happen to appear in d and hence are likely to be irrelevant to c (and hence are likely to be the training noises in d). By removing the possible training noises, the classifiers can be trained more properly.

1.2 Major Challenge and Related Work

Major challenge of developing TNR is the fusion of relatedness scores of consecutive terms to determine whether the terms are training noises. To our knowledge, no previous studies tackled the challenge.

Given training texts that are tagged with category labels, previous TC studies have developed and tested many feature (term) selection techniques that estimated the relatedness of a term with respect to each category, and accordingly selected those features that were capable of discriminating the categories (e.g., [8][11]). The selected features served as the basis on which text classifiers were built. Many excellent text classifiers have been developed, including the Support Vector Machine (SVM) classifiers that have been routinely employed and shown to be one of the best classifiers. TNR aims at serving as a front-end processor of the feature selection and classification techniques. It reduces the noises in the training texts for the techniques.

TNR employs term proximity information as the evidence to identify training noises. Previous studies have noted term proximity as important information as well. However the term proximity information was employed to improve the ranking of text retrieval [4][10][12] or feature selection for text classification (e.g., considering multiple consecutive terms [9], nearby terms in a fixed order, and co-occurring terms in whatever order and location [3]). TNR employs term proximity to reduce training noises. The training texts with noises removed by TNR can be used as the training data for the previous proximity-based feature selection and classification techniques.

TNR identifies a sequence of consecutive irrelevant terms as the training noises. The sequence of terms can be viewed as a passage, and hence TNR actually removes certain passages from training texts. Previous passage identification techniques often aimed at detecting the existence or transition of topics in texts (e.g., [2]), retrieving textual parts for a given query (e.g., [1]), and extracting certain textual parts to be classified by text classifiers [6][7]. The previous techniques extracted textual parts from testing documents, while TNR extracts and removes textual parts from the *training* documents so that noises in the training documents can be reduced to improve the training process of text classifiers.

Section 2 presents TNR. To empirically evaluate TNR, section 3 reports a case study on the classification of Chinese texts of disease information. The result shows that TNR can improve SVM, which is a state-of-the-art technique for TC. The contribution is of significance to the enhancement of existing text classifiers for classifying in-space documents and filtering out out-space documents.

Table 1. Analysis of possible kinds of terms: For a category *c*, false positive (FP) and true negative (TN) terms in a training document *d* are actually *noises* that should be removed (from *d*) so that the text classifier can be improved

Case	Characteristics of the terms		Cause of the case	Proper actions for the case
	Actual **Correlation type to *c***	***Estimated*** **correlation type to *c***		
TP: *True Positive*	Positive	Positive	Training data is proper	Keeping the terms in *d*
FN: *False Negative*	Positive	Negative	(1) Terms have too high frequency in categories other than *c*, (2) Terms have too low frequency in *c*	Keeping the terms in *d*
FP: *False Positive*	Negative	Positive	(1) Terms have too low frequency in categories other than *c*, (2) Terms have too high frequency in *c*	Removing the terms from *d*
TN: *True Negative*	Negative	Negative	Training data is proper	Removing the terms from *d*

2 Removal of Noises from Training Texts

Table 1 summarizes an analysis that serves as the basis for the development of TNR. Given a term *t* in a training documnt *d* of category *c*, we are concerned with two types of correlation between *t* and *c*: *positive* correlation and *negative* correlation. We say that *t* is positively correlated to *c* if occurrence of *t* increases the possibility of classifying *d* into *c*; otherwise *t* is negatively correlated to *c*. Therefore, depending on the *actual* and the *estimated* correlation types of *t* to *c*, we have four possible kinds of terms: TP (True Positive) terms, FN (False Negative) terms, FP (False Positive) terms, and TN (True Negative) terms. Obviously TNR should treat FP and TN terms

as the training nosies as their occurrences in d may mislead the training process of the txt classifier for c. It is also interesting to note that, as the correlation types of a term is estimated across all categories, the removal of FP and TN terms for a category may reduce the number of FN terms for another category.

Table 2. Algorithm of TNR

Procedure: *TNR*(d,c), where d is a training document of category c
Output: d': The content of d with training noises for c removed
Method:

// Initialization of variables
(1) $d' \leftarrow d$;
(2) $NetS \leftarrow 0$;
(3) $NoiseStart \leftarrow$ NULL;
(4) $CurrentPos \leftarrow 1$;
// Sequentially scan all terms in d to remove all training noises in d
(5) While $CurrentPos \leq$ Length of d, do
  (5.1) $t \leftarrow$ The term at position $CurrentPos$ of d;
  (5.1) $\chi^2(t,c) \leftarrow$ Chi-square correlation strength of t with respect to c;
  // Revise the correlation strength of t to c
  (5.2) If t is positively correlated to c and $\chi^2(t,c) \geq$ CorThreshold,
    $RevisedS \leftarrow \text{Log}_2(1+\chi^2(t,c)-\text{CorThreshold})$;
  (5.3) Else if t is positively correlated to c, // t is weakly positively correlated to c
    $RevisedS \leftarrow -1\times\text{Log}_2(1+\text{CorThreshold}-\chi^2(t,c))$;
  (5.4) Else $RevisedS \leftarrow -1\times\text{Log}_2(1+\text{CorThreshold}+\chi^2(t,c))$. // t is negatively correlated to c
  // Identify the start and end of the possible training noise
  (5.5) If $NetS = 0$ and $RevisedS < 0$ // The start of a possible training noise is found
    $NoiseStart \leftarrow$ Position of t in d;
  (5.6) $NetS \leftarrow NetS + RevisedS$;
  (5.7) If $NetS \geq 0$, // Identify the end of the possible training noise
    (5.7.1) $NetS \leftarrow 0$;
    (5.7.2) If $NoiseStart <>$ NULL and more than three terms after $NoiseStart$ have $RevisedS < 0$, // A training noise is found and hence should be removed
      (5.7.2.1) $NoiseEnd \leftarrow$ Position of the term at which $NetS$ is minimized;
      (5.7.2.2) $d' \leftarrow d'$ with the textual part from $NoiseStart$ to $NoiseEnd$ removed;
      (5.7.2.3) $CurrentPos \leftarrow NoiseEnd$;
    (5.7.3) Else $NoiseStart \leftarrow$ NULL; // Not a training noise, and thus no removal is done
  (5.8) $CurrentPos \leftarrow CurrentPos + 1$;
// Remove the final noise, if there is one before the end of d
(6) If $NoiseStart <>$ NULL and more than three terms after $NoiseStart$ have $RevisedS < 0$,
  (6.1) $NoiseEnd \leftarrow$ Position of the term at which $NetS$ is minimized;
  (6.2) $d' \leftarrow d'$ with the textual part from $NoiseStart$ to $NoiseEnd$ removed;
(7) Return d';

However, as training texts are inevitably unsound or incomplete, precise identification of FP and TN terms is challenging. TNR thus employs term proximity

information as the evidence to tackle the challenge. Those terms (in *d*) that have many neighboring terms with negative or low correlation strengths to *c* may simply happen to appear in *d* and hence are likely to be the training noises in *d*.

More specifically, Table 2 defines the algorithm of TNR. Given a training document *d* of a category *c*, TNR identifies certain sequences of terms as the training noises for *c* and accordingly returns *d'* whose content is copied from *d* with the noises removed (ref. Output in Table 2). TNR sequentially scans each term *t* in *d* to detect the noises (ref. Step 5). To estimate the types and strengths of term-category correlations, TNR employs the χ^2 (chi-square) technique (ref. Step 5.1), which is a state-of-the-art technique routinely employed and shown to be one of the best feature scoring techniques [11]. For a term *t* and a category *c*, $\chi^2(t,c)$ is estimated by equation 1, where *N* is the total number of training texts, *A* is the number of training texts that are in *c* and contain *t*, *B* is the number of training texts that are not in *c* but contain *t*, *C* is the number of training texts that are in *c* but do not contain *t*, and *D* is the number of training texts that are not in *c* and do not contain *t*.

$$\chi^2(t,c) = \frac{N \times (A \times D - B \times C)^2}{(A+B) \times (A+C) \times (B+D) \times (C+D)} \quad (1)$$

If $A \times D > B \times C$, *t* is more likely to appear in documents of *c* (than in documents not of *c*) and hence is said to be *positively correlated* to *c*; otherwise it is *negatively correlated* to *c*. To tackle the problem of false positive terms (FP, recall Table 1), TNR employs a chi-square strength threshold (CorThreshold in Step 5.2 in Table 2) to determine whether *t* is *strongly* positively correlated to *c* (likely to be really positively correlated to *c*). The threshold is set to the top 70% chi-square strength of the terms that are estimated to be positively correlated to *c*, and hence 70% of the positively correlated terms are said to be strongly positively correlated to *c*. TNR then revises the correlation strength of *t* to *c* by considering the threshold and employing logarithm to reduce the effect of dramatic change of the $\chi^2(t,c)$ scores (ref. Steps 5.2 ~ 5.4). Only those terms that are strongly positively correlated to *c* have positive revised strengths.

To identify the start and the end of the textual part that is believed to be a training noise, TNR maintains *NetS*, which is the net sum of the revised strengths (ref. Step 5.6). The start of the training noise *NoiseStart* should be at the position at which *NetS*=0 and the revised strength of the current term *t* is negative (since in this case *t* is the first possible FP term, ref. Step 5.5). After scanning *t*, NetS becomes negative (as the revised strength of *t* is negative, ref. Step 5.6). Once *NetS*≥0 again at some later position (ref. Step 5.7), TNR triggers the removal of the training noise whose end is the position at which *NetS* is minimized (i.e., *NoiseEnd*, ref. Steps 5.7.2.1 and 5.7.2.2). This is because the textual part spanning from *NoiseStart* to *NoiseEnd* actually contains more negatively correlated terms and hence is more likely to contain FP and TN terms that should be removed. To properly employ term proximity to identify training noises, a training noise should have at least three terms whose revised strengths are negative (ref. Step 5.7.2). Finally, if the end of *d* is encountered but it is still possible to identify a training noise, TNR employs the same way to identify and remove the noise (ref. Step 6).

3 A Case Study on Classification of Disease Information

A case study on thousands of Chinese texts of disease information was conducted to empirically evaluate the contribution of TNR. Automatic classification of disease information is a fundamental task for the dissemination of medical information for medical decision support and disease management.

3.1 Experimental Data

In the case study, we tested medical texts about top-10 fatal diseases and top-20 cancers in Taiwan, resulting in 28 diseases of interest. Chinese texts about the 28 diseases were collected from several reliable sources, including National Taiwan University Hospital[1], Department of Health in Taiwan[2], and many medical associations and hospitals. We were also interested in 5 aspects of the diseases: *etiology*, *diagnosis*, *treatment*, *prevention*, and *symptom*, which are critical aspects in clinical practice and disease management. There were 2 diseases for which we could not find medical texts about the aspect of prevention, and hence we totally had 138 categories (=28×5-2) in which there were 4669 medical texts. We randomly selected 2300 of the documents as the training data. The remaining 2369 documents were used as the in-space testing data (i.e., the testing data belonging to the space of the 138 categories). The data was used to measure how text classifiers performed in classifying in-space documents.

On the other hand, to measure how text classifiers performed in filtering out out-space documents (those that belonged to none of the 138 categories), we collected 446 medical texts about other 15 diseases. As noted in Section 1, filtering of out-space documents is particularly essential as most real-world documents are actually out-space documents for the domain in which text classifiers are trained.

3.2 Underlying Text Classifier

We employed SVM as the underlying classification technique. SVM is a popular technique in TC, and previous studies have shown that SVM is one of the best classification techniques. To implement the SVM classifier, we employed SVM^{Light} that is publicly available[3] [5] and has been tested in many previous studies. SVM required a feature set, which was built using the training data. The features were selected based on their maximum χ^2 scores with respect to all categories [11]. We reported the contributions of TNR to SVM under different feature set sizes.

3.3 Evaluation Criteria

We employed different criteria to evaluate the two goals of TC: classification of in-space test documents and the filtering of out-space test documents. For the former, we

[1] Available at `http://www.ntuh.gov.tw/en/default.aspx`.
[2] Available at `http://www.doh.gov.tw/EN2006/index_EN.aspx`.
[3] Available at `http://www.cs.cornell.edu/People/tj/svm%5Flight/old/svm_light_v5.00.html`.

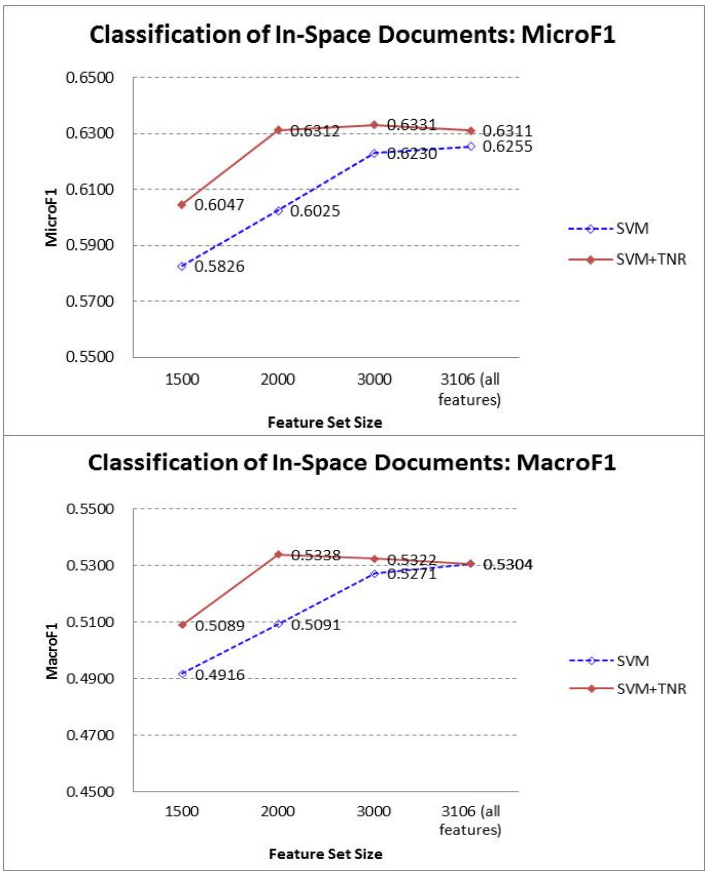


Fig. 1. Classification of in-space medical documents: TNR helps SVM to achieve better performance, especially when fewer features are available for SVM

employed F_1, which is a popular criterion measured by 2×P×R/(P+R), where P (precision) is [total number of correct classifications / total number of classifications made] and R (recall) is [total number of correct classifications / total number of correct classifications that should be made]. There are two ways to compute average performance in F_1: *micro-averaged* F_1 ($MicroF_1$) and *macro-averaged* F_1 ($MacroF_1$). In measuring $MicroF_1$, P and R are computed by viewing all categories as a system, and the resulting P and R are used to compute $MicroF_1$. $MacroF_1$ is simply measured by averaging F_1 values on *individual* categories.

For the filtering of out-space test documents, we employed two evaluation criteria: *filtering ratio* (FR) and *average number of misclassifications* (AM). FR is [number of the out-space documents that are successfully rejected by all categories / number of the out-space documents] and AM is [number of misclassifications for the out-space documents / number of the out-space documents]. A system should filter out as many out-space documents as possible (i.e., higher FR) and avoid misclassifying the out-space documents into many categories (i.e. lower AM).

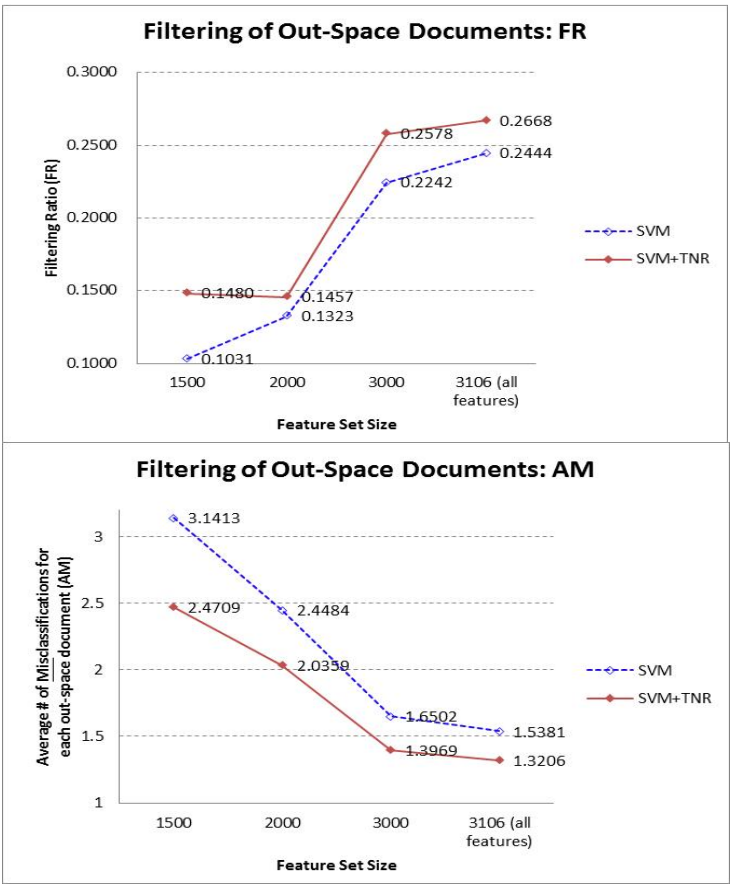


Fig. 2. Filtering of out-space medical documents: TNR successfully helps SVM to filter out a higher percentage of out-space documents (i.e., achieving higher FR) and reduce the average number of misclassifications for the out-space documents (i.e., achieving a lower AM)

3.4 Result and Discussion

Figure 1 shows the performance of SVM and SVM+TNR (SVM with TNR as the front-end processor to remove training noises) in classifying in-space medical documents. The results show that TNR can help SVM to achieve better performance, especially when fewer features are available for SVM, indicating that for those application domains in which it's difficult to get complete training data, noise reduction provided by TNR is particularly helpful for SVM to achieve better performance. In the case study, TNR helps SVM to achieve better and stable performance when only 2/3 of the features (i.e., 2000 features) are employed. The result confirms the contribution of training noise reduction to the classification of in-space documents.

Figure 2 shows the performance of SVM and SVM+TNR in filtering out out-space medical documents. The results show that TNR can help SVM to filter out a higher

percentage of out-space documents (i.e., higher FR achieved by SVM+TNR), and when an out-space document is misclassified into some categories TNR helps SVM to reduce the number of the misclassifications (i.e. lower AM achieved by SVM+TNR). As most real-world testing documents are actually out-space documents for the classifiers, the contribution of TNR is of practical significance. Based on the results in Figure 1 and Figure 2, contributions of TNR to both the classification of in-space documents and the filtering of out-space documents are confirmed.

4 Conclusion and Future Work

Previous studies have developed many techniques to build excellent text classifiers using training texts, which are tagged with category labels. However, as the training texts are often inevitably unsound or incomplete in practice, they often contain many training noises, which are those terms not related to the categories of interest. Reduction of the training noises is helpful to promote the performance of text classifiers. We thus present a technique TNR to remove the possible training noises. Given a training text *d* of a category *c*, TNR identifies and removes the terms (in *d*) that are likely to be negatively correlated to *c*. As the identification of such terms is challenging due to the unsoundness and incompleteness of training data, TNR employs term proximity as the evidence to identify such terms. It identifies a sequence of consecutive terms (in *d*) as the noises if the terms are not strongly related to *c*. A case study on thousands of Chinese texts of disease information shows that TNR can improve SVM in both the classification of in-space documents (those that belong to some categories of interest) and the filtering of out-space documents (those that belong to none of the categories). The contribution is of significance to text classification studies.

Although SVM has been a state-of-the-art classification technique, we are applying TNR to other kinds of text classification techniques so that we may further measure the contributions of TNR to various kinds of text classifiers. It is also interesting to extend the medical case study reported in the paper. An interesting extension is the classification of texts without disease names, as healthcare professionals and consumers often like to enter a textual description without disease names (e.g., descriptions of symptoms and risk factors) and ask for relevant diseases. The text classifiers enhanced with TNR can serve as the disease classifier so that more information of the relevant diseases can be disseminated and shared among the healthcare professionals and consumers.

Acknowledgment. This research was supported by the National Science Council of the Republic of China under the grant NSC 100-2221-E-320-004-MY2.

References

1. Abdul-Jaleel, N., Allan, J., Croft, W.B., Diaz, F., Larkey, L., Li, X., Metzler, D., Smucker, M.D., Strohman, T., Turtle, H., Wade, C.: UMass at TREC 2004: Notebook. In: Proceedings of the 13th Text Retrieval Conference. National Institute of Standards and Technology, Gaithersburg (2004)

2. Chen, C.C., Chen, M.C.: TSCAN: A Novel Method for Topic Summarization and Content Anatomy. In: Proceedings of SIGIR 2008, Singapore, pp. 579–586 (2008)
3. Cohen, W.W., Singer, Y.: Context-Sensitive Mining Methods for Text Categorization. In: 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1996), Zurich, Switzerland, pp. 307–315 (1996)
4. Gerani, S., Carman, M.J., Crestani, F.: Proximity-Based Opinion Retrieval. In: Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2010), Geneva, Switzerland, pp. 403–410 (2010)
5. Joachims, T.: Making Large-Scale SVM Learning Practical. In: Schölkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning. MIT Press (1999)
6. Kim, J., Kim, M.H.: An Evaluation of Passage-Based Text Categorization. Journal of Intelligent Information Systems 23(1), 47–65 (2004)
7. Mengle, S., Goharian, N.: Passage Detection Using Text Classification. Journal of the American Society for Information Science and Technology 60(4), 814–825 (2009)
8. Mladeniá, D., Brank, J., Grobelnik, M., Milic-Frayling, N.: Feature Selection Using Linear Classifier Weights: Interaction with Classification Models. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 234–241 (2004)
9. Peng, F., Schuurmans, D.: Combining Naive Bayes and n-Gram Language Models for Text Classification. In: Sebastiani, F. (ed.) ECIR 2003. LNCS, vol. 2633, pp. 335–350. Springer, Heidelberg (2003)
10. Svore, K.M., Kanani, P.H., Khan, N.: How Good is a Span of Terms? Exploiting Proximity to Improve Web Retrieval. In: Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Geneva, Switzerland, pp. 154–161 (2010)
11. Yang, Y., Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization. In: Proceedings of the 14th International Conference on Machine Learning (1997), Tennessee, Nashville, pp. 412–420 (1997)
12. Zhao, J., Yun, Y.: A Proximity Language Model for Information Retrieval. In: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2009), Boston, USA, pp. 291–298 (2009)

Prediction of Relevance between Requests and Web Services Using ANN and LR Models

Keyvan Mohebbi[1], Suhaimi Ibrahim[2], and Norbik Bashah Idris[2]

[1] Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia (UTM), Malaysia
[2] Advanced Informatics School (AIS), Universiti Teknologi Malaysia (UTM), Malaysia
mkeyvan2@live.utm.my, suhaimiibrahim@utm.my, norbik@scan-associates.net

Abstract. An approach of Web service matching is proposed in this paper. It adopts semantic similarity measuring techniques to calculate the matching level between a pair of service descriptions. Their similarity is then specified by a numeric value. Determining a threshold for this value is a challenge in all similar matching approaches. To address this challenge, we propose the use of classification methods to predict the relevance of requests and Web services. In recent years, outcome prediction models using Logistic Regression and Artificial Neural Network have been developed in many research areas. We compare the performance of these methods on the OWLS-TC v3 service library. The classification accuracy is used to measure the performance of the methods. The experimental results show the efficiency of both methods in predicting the new cases. However, Artificial Neural Network with sensitivity analysis model outperforms Logistic Regression method.

Keywords: Web service discovery, Semantic Web service matchmaker, Relevance prediction, Logistic Regression, Artificial Neural Network.

1 Introduction

With the invention of Service Oriented Architecture (SOA), Web services have gained more popularity. A Web service is "a software system identified by a URI, whose public interfaces and bindings are defined and described using XML. Its definition can be discovered by other software systems. These systems may then interact with the Web service in a manner prescribed by its definition, using XML based messages conveyed by Internet protocols" [1]. Nowadays, the number of Web services provided by the various organizations is increasing dramatically. Thus, how to effectively and efficiently find the Web services which meet the user's requirements is a challenge. Web service discovery is defined as "the act of locating a machine-processable description of a Web service that may have been previously unknown and that meets certain functional criteria. It involves matching a set of criteria with a set of Web service description. The goal is to find an appropriate Web service" [2]. Semantic approaches describe the properties and capabilities of Web

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 40–49, 2013.

services in an unambiguous and machine-understandable form. This will ease the way to automate usage tasks such as discovery, selection, composition, mediation, and invocation of Web services. Specifically, Semantic Web service discovery attempts to make the process of finding Web services run automatically. During this process which is often called matchmaking, the formalized description of a service request and that of a service advertisement need to be compared with each other in order to recognize common elements in these descriptions. Semantic Web service discovery approaches are mainly classified into three categories: *Logic-based*, *Non-logic-based*, and *Hybrid* [3]. While *Logic-based* approaches rely on logic inferences for the matchmaking, *Non-logic-based* matchmakers exploit semantics that are implicit in patterns or relative frequencies of terms in service descriptions. They rely on techniques such as graph matching, linguistics, data mining, or information retrieval. Finally, *Hybrid* approaches combine techniques from both of the previous matchmakers.

In this paper, we propose a *Non-logic-based* semantic matchmaker. Several *Non-logic-based* approaches have been proposed to solve the problem of Web service discovery. They receive a request as input and return as output a list of Web services ordered by their similarity to the request. Usually, the similarity is a value between 0 and 1. Determining a threshold for the similarity value is a challenge. However, the current approaches share a common weakness, as they disregard such challenge. To address this problem, we propose the use of classification methods to predict the relevance of requests and Web services. In this approach, the classification method rely either on the past user's experiences or on a repository of known pairs of request and Web services to predict the relevance of a new pair of service descriptions.

The classification methods used in this study are Logistic Regression and Artificial Neural Network. These methods provide the same functionality, but follow different approaches. This raises the question if one method has better performance than the other. To investigate this question, we compared the performance of these methods using the same set of data.

The remainder of this paper is structured as follows. In Section 2 we explain our proposed *Non-logic-based* semantic matchmaker. Section 3 describes the used classification methods. In Section 4 the application of the classification methods for predicting the relevance of requests and Web services is explained. This is followed by the evaluation process and the obtained results. We conclude in Section 5.

2 Non-logic-based Semantic Matchmaker

To determine the similarity of services, *Non-logic-based* approaches generally exploit techniques other than reasoning on logical expressions. Our proposed *Non-logic-based* matchmaker relies on semantic similarity measuring techniques to calculate the matching level between the signatures of a request and a Web service.

The signature of a request or a Web service can be defined in terms of the inputs and the outputs included in its offered description. The inputs of a service as well as its outputs are described as set of concepts defined in some domain-specific ontology.

The *Non-logic-based* matchmaker determines the similarity level of a request and a Web service by comparing their respective signatures.

Several approaches have been proposed to measure the semantic similarity and relatedness among a pair of given concepts with respect to their relationship in the reference ontology [4], [5]. Based on the evaluation of similarity measures performed in the literature [6], intrinsic Information Content-based approaches provide higher accuracy in comparison with other measures. They also keep the complexity level of computations low. These features are useful when implementing a generic approach for Semantic Web service matchmaking. In our approach, we adopt the intrinsic IC-based measure proposed by Lin [7].

The selected measure is based on the notion of Information Content (IC) in the information theory. The IC of the concept c is computed by the negative logarithm of its appearance probability, $p(c)$ in a taxonomy.

$$IC(c) = -\log p(c) \tag{1}$$

According to Lin, the similarity between a pair of terms A and B in a taxonomy can be measured as the ratio between the amount of information needed to state their commonality and the information needed to fully describe what they are . This measure is computed as follows:

$$sim_{Lin}(A,B) = \frac{2 \times IC\big(LCS(A,B)\big)}{IC(A) + IC(B)}, \tag{2}$$

where LCS denotes the Least Common Subsumer for a pair of terms in an ontology.

To obtain the maximum matching value based on the selected similarity measure, we apply the same principle to the set of advertised components A and the set of requested components Q. Suppose that p_{qa} is a binary variable indicating whether concepts $q \in Q$ and $a \in A$ have been paired and their similarity is determined by a function $sim_{Lin}: (Q, A) \rightarrow [0..1]$. Then, function $sim_{max}: (Q, A) \rightarrow [0..1]$ is defined as follows:

$$sim_{max}(Q,A) = \frac{1}{|Q|}.\sum_{q \in Q} p_{qa}.sim_{Lin}(q,a) | a \in A\,. \tag{3}$$

That means the maximum similarity of two component sets is the average of the maximum similarities found for components in A for each component in Q.

To compute the signature-level similarity value of a request $\mathcal{R}$ and a Web service $\mathcal{W}$, the maximization function in Equation 3 is applied to the sets of input (or output) concepts of $\mathcal{R}$, respectively, $\mathcal{W}$. Thus, two distinct matching values are obtained for input and output parts of their signatures. Finally, the overall similarity will be defined as the average of the above values:

$$sim(\mathcal{R},\mathcal{W}) = \frac{sim_{max,in}(\mathcal{R},\mathcal{W}) + sim_{max,out}(\mathcal{R},\mathcal{W})}{2}. \tag{4}$$

3 Classification Methods

3.1 Logistic Regression

Logistic Regression (LR) [8] is a well known classification method in the field of statistical learning. LR is a type of regression analysis used to find the best model to describe the relationship between an outcome (dependent) variable and a set of predictor (covariate) variables. Binary LR refers to the case where the observed outcome has two possible types (e.g. "yes" vs. "no", "male" vs. "female", or "pass" vs. "fail").

Unlike ordinary regression, what we want to predict from the knowledge of covariates and coefficients is not a numerical value of a dependent variable, but rather the probability that it belongs to a particular group [9]. To accomplish this goal, let the conditional probability that the outcome is present be denoted by $P(Y = 1|x) = \pi(x)$. The logit function (i.e. the logistic transformation) of the LR model is given by the equation:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \,, \tag{5}$$

where β_0 is the constant of the equation, $x_1 \ldots x_p$ are predictor variables and $\beta_1 \ldots \beta_p$ are their respective coefficients. Predictor variables may be either numerical or categorical. The general method of estimation of logit models is maximum likelihood, in which we want to maximize the probability of getting the observed results given the fitted regression coefficients. The LR model is:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}. \tag{6}$$

Here, e is the base of natural logarithm [8].

Ease of use and the ability to interpret the model parameters are the most important advantages of LR. The drawback of LR is that it is not suitable for modeling complex interactions since it uses linear combination of variables. LR is used extensively in numerous disciplines such as business management [10], civil and environment [11], medical and health [12], social sciences [13], and image processing applications [14].

3.2 Artificial Neural Network

Artificial Neural Network (ANN) is a computer modeling technique based on the observed behaviors of biological neurons [15]. A neural network is similar to humans in that it takes previously solved problems to gain knowledge for constructing a new model. This model is able to makes decisions, classifications, and predictions [16]. ANN is a complex and flexible nonlinear system with properties not found in other modeling systems. These properties include robust performance in dealing with noisy or incomplete input patterns, high fault tolerance, and the ability to generalize from the input data [17].

ANN has proven its predictive power through comparison with other statistical techniques using real data sets. The data sets of past experiences include corresponding input and output variables. Supervised neural networks such as multi layer perceptron (MLP) and radial basis function (RBF) use training and testing data to build a model. The training data is a portion of data set that is used to learn the ANN model in predicting the known output. The validation of the model is conducted by testing data. ANN is aimed to predict the output for any given input [18]. Fig. 1 shows an example of a typical MLP model [19].

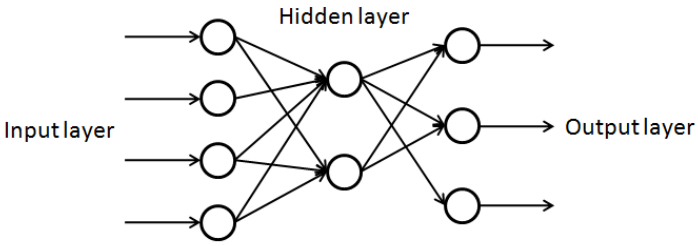


Fig. 1. Typical Multi Layer Perceptron Model

As can be seen, this simple neural network consists of input, hidden layer, and output layers**.** In each layer there are one or more nodes as processing elements (PEs). PEs simulate the neurons in the brain. They have connections to each other in various layers. The connections have weights determined during the training phase. In this feed forward neural network, information only passes in the forward direction with no feedback loops [19].

There is not any standard method to construct the architecture of ANN models, therefore their developing is difficult. Besides, in contrast to the regression models, the standardized coefficients and odd ratios associated with each variable in ANN cannot be easily computed. One of the major drawbacks of ANN models is the difficulty of interpreting the weights of the predictors generated by the neural network analysis. Another disadvantage of ANN is its complexity [20, 21].

ANN has been applied in many disciplines, including biology, psychology, statistics, mathematics, medical science, and computer science [22]. It has also been applied to a variety of business areas such as accounting and auditing, finance, management and decision making, marketing and production. Recently, ANN models become very popular and have been applied to diagnose diseases and to predict the survival ratio of the patients [15], [21], [23], and [19].

4 Predicting the Relevance of Requests and Web Services

Our proposed Non-logic-based matchmaker calculates the similarity between a request and a Web service by comparing their descriptions. At the core of this matchmaker is a function that returns a similarity value in the range of [0,1] for a pair of service signatures. The higher the result of this function, the higher the similarity

between two services is. This similarity value should be used to predict whether the request and the Web service are relevant or not. In our approach, we aim to automate this prediction by means of the classification methods such as those discussed in Section 3.

First, the selected classification method should be trained using a number of requests and Web services for which the relevancy is known. After training to predict the relevance of a Web service to a given request the classification method receives their similarity value and classifies the Web service in one of two possible classes, namely: relevant or not relevant. This process must be repeated to compare all Web services available in the repository with the given request. Due to the large amount of the data, the training might takes a high computation time. So this phase should be performed off-line.

4.1 Evaluation

The validation of our approach was performed with OWLS-TC. This is the most prominent service retrieval test collection that has been applied since 2006 in the international contest series on Semantic Service Selection (S3) [24]. It should be noted that we do not utilize an artificially created test data set and this collection is completely independent from the work at hand which makes the results of the evaluations more reliable.

We have selected version 3.0 revision 1 of OWLS-TC. This version consists of 1007 Semantic Web services written in OWL-S 1.1 from seven application domains that are education, medical care, food, travel, communication, economy, and weapon. It also provides a set of 29 test queries. The measurement of the performance of semantic service retrieval requires any test collection to include a set of relevant services for each query in the collection [24]. These sets of relevant services were determined for OWLS-TC by more than a dozen users according to the standard TREC working definition of *binary relevance* [25], i.e. each Web service is judged either *relevant* or *not relevant* to a query. The semantic annotations of all services base on references to 34 OWL ontologies in total.

In order to use the classification methods LR and ANN for analyzing the data, the first step is to define a model. For this we assume that the similarity of a request and a Web service on the signature level (the input and the output features) predicts the membership of that Web service to the request's relevance set. Therefore, our model consists of a dependent variable $rel^{r,\mathcal{W}}$ that denotes the relevancy of a request r and a Web service $\mathcal{W}$, and an independent variable $sim^{r,\mathcal{W}}$ that denotes the similarity between r and $\mathcal{W}$. As mentioned earlier, in the test collection OWLS-TC each Web service is defined as either relevant or not relevant to a request. To consider these discrete relevance grades as the dependent variable, they should be mapped to numerical values such as the following:

$$rel^{r,\mathcal{W}} = \begin{cases} 1, & r \text{ and } \mathcal{W} \text{ are relevant} \\ 0, & \text{else} \end{cases} \tag{7}$$

For the training phase, we have selected all test queries from the OWLS-TC repository. Each pair of a request r_i and a Web service $\mathcal{W}_j$ is represented by a vector as follows:

$$Pair(r_i, \mathcal{W}_j) = \langle sim^{r_i,\mathcal{W}_j}, rel^{r_i,\mathcal{W}_j} \rangle, \tag{8}$$

in which the similarity of r_i and $\mathcal{W}_j$ and their relevancy are the dimensions of the vector. For the pair of r_i and $\mathcal{W}_j$, the Non-logic-based matchmaker matches them and also retrieves their predefined relevancy. The computed similarities as well as the relevancy are stored in a vector. All the vectors $\{Pair(r_i, \mathcal{W}_j)\}$ are considered as the training set. These vectors constitute a matrix where each pair of request and Web service yields one row with two columns. An exemplary matrix is depicted in the following:

$$M = \begin{bmatrix} sim^{r_1,\mathcal{W}_1} & rel^{r_1,\mathcal{W}_1} \\ sim^{r_1,\mathcal{W}_2} & rel^{r_1,\mathcal{W}_2} \\ sim^{r_1,\mathcal{W}_3} & rel^{r_1,\mathcal{W}_3} \\ sim^{r_2,\mathcal{W}_1} & rel^{r_2,\mathcal{W}_1} \\ \vdots & \vdots \end{bmatrix} = \begin{bmatrix} 0.19 & 0 \\ 0.80 & 1 \\ 0.30 & 0 \\ 0.26 & 0 \\ \vdots & \vdots \end{bmatrix}, \tag{9}$$

where the first four rows contain the data for the pair of requests and Web services $(r_1, \mathcal{W}_1)$, $(r_1, \mathcal{W}_2)$, $(r_1, \mathcal{W}_3)$, and $(r_2, \mathcal{W}_1)$ respectively. It can also be seen that r_1 and $\mathcal{W}_2$ are relevant while other three pairs are not relevant according to the mapping shown in Equation 7.

Variables $rel^{r,\mathcal{W}}$ and $sim^{r,\mathcal{W}}$ from the aforementioned model enter in both LR and ANN methods as respective categorical output and continuous input, using the data of the above matrix. The ANN used in this study is a multi layer perceptron (MLP) model with a three-layer topology (input, hidden, and output layers). The input layer consists of one neuron corresponding to $sim^{r,\mathcal{W}}$, the hidden layer consists of eight neurons transforming the input to a 8-dimensional space. Finally, the output layer has only one neuron, representing two possible states: relevant or not relevant.

After training the model, the next step is evaluating the classification methods by making predictions against the test set. For evaluation, we use the same set of training data. Since this data set already contains relevance set of each request, it is easy to determine whether the method's guesses are correct.

Applying LR results are depicted in the Table 1.

Table 1. Classification Results of Logistic Regression

Observed		Predicted		
		$rel^{\mathcal{G},\mathcal{W}}$		Percentage Correct
		0	1	
$rel^{\mathcal{G},\mathcal{W}}$	0	21372	0	100.0
	1	2571	5260	67.2

For each case, the predicted response is 1 if that case's model-predicted probability is greater than the cutoff value. In this test, the default cutoff of 0.5 is used. Of the cases, all of the not relevant, and 5260 of 7831 relevant Web services are classified correctly. Overall, 91.2% of the cases are classified correctly. For ANN, the training performance of the MLP model is 92.5%. Apart from the above results, in the following we adopt another approach to evaluate independently the performance of the aforementioned methods.

The Receiver Operating Characteristics (ROC) graph is a useful way to evaluate the performance of the classifiers that categorize cases into one of two groups [26]. A ROC graph demonstrates the accuracy of a classifier in terms of two measures, namely: *Specificity* and *Sensitivity*. *Specificity* is the probability that a negative case (i.e. a not relevant Web service) is correctly classified. *Sensitivity* is the probability that a positive case (i.e. a relevant Web service) is correctly classified. The most commonly used measures of discriminatory power of the classifiers are the area under the ROC curve (AUC), *Sensitivity*, and *Specificity*. While *Sensitivity* and *Specificity* of a classifier are reported for a single threshold, the AUC represents a common measure of *Sensitivity* and *Specificity* over all possible thresholds.

For every pair of request and Web service, their similarity $sim^{r,w}$ is calculated. Then LR and ANN methods are applied separately to predict the relevance of that request and Web service using $sim^{r,w}$. After processing all pairs of requests and Web services, the ROC graph is plotted for each method using the predicted and the actual retrieved relevancies. A single point is produced for each method corresponding to its overall false positive and true positive rates. The result is illustrated in Fig. 2.

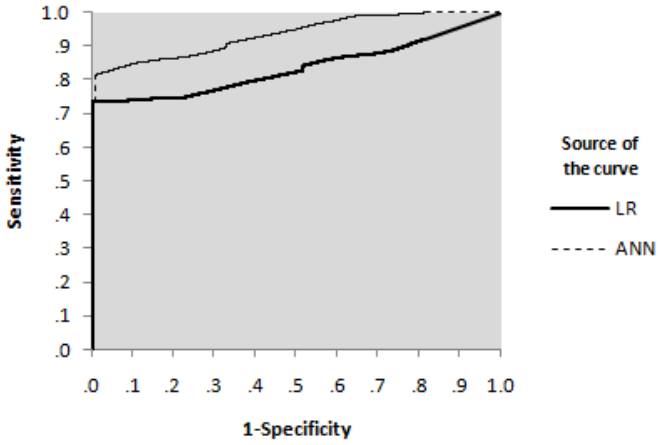


Fig. 2. ROC Graph for LR and ANN Methods

As can be seen in Fig. 2, the ROC curves of the aforementioned methods indicate that LR and ANN methods have similar accuracies, however, ANN outperforms LR. The respective AUCs are 89.8% versus 83.6%.

The neural network development software used in this study was *STATISTICA*, version 7 (StatSoft, Inc., Tulsa, OK, USA). The statistics analysis was performed using *PASW Statistics* version 17 (SPSS, Inc., Chicago, IL, USA).

5 Conclusion

In this paper we have presented an approach to solve the discovery of Semantic Web services. The designed *Non-logic-based* matchmaker relies on semantic similarity measuring techniques to calculate the matching level between both inputs and outputs elements included in the OWL-S service description. It receives a request and a collection of Web services as input and produces as output a list of Web services ordered by their level of similarity to the request. Furthermore, we proposed a mechanism to predict the relevance of requests and Web services by means of a classification method. Applying this mechanism on Web services of the ordered list determine each one either as relevant or not relevant to the given request. This will assist the requester in selection among the discovered Web services.

Two classification methods, namely: Logistic Regression and Artificial Neural Network have been applied and compared. For the comparison, these methods have been evaluated against the OWLS-TC version 3.0 revision 1 repository. Both methods (LR and ANN) presented results of high precision in predicting the relevance of the cases, however ANN outperformed LR.

Comparing two predictive methods in general, ANN provide more accurate results in prediction whereas LR could also identify the effect of factors on the classification [15]. The complexity of ANN makes it difficult to relate their output to input. ANN model allows the inclusion of a large number of variables [27]. Another advantage of ANN is that there are not many assumptions (such as normality) that need to be verified before the models can be constructed.

Acknowledgments. This research is supported by Universiti Teknologi Malaysia (UTM) under the Vot. 00H74. The authors would like to thank UTM and Ministry of Higher Education (MOHE) Malaysia.

References

1. Web Services Architecture Requirements, http://www.w3.org/TR/wsa-reqs/
2. Web Services Architecture, http://www.w3.org/TR/ws-arch/
3. Mohebbi, K., Ibrahim, S., Idris, N.B., Tabatabaei, S.G.H.: A Guideline for Evaluating Semantic Web Service Discovery Approaches. Advances in Information Sciences and Service Sciences (AISS) 4, 330–346 (2012)
4. Budanitsky, A., Hirst, G.: Evaluating WordNet-based Measures of Lexical Semantic Relatedness. Computational Linguistics 32, 13–47 (2006)
5. Tsatsaronis, G., Varlamis, I., Vazirgiannis, M.: Text Relatedness Based on a Word Thesaurus. Journal of Artificial Intelligence Research 37, 1–39 (2010)
6. Batet, M., Sánchez, D., Valls, A.: Deliverable D3: State of the art of clustering algorithms and semantic similarity measures (2010)

7. Lin, D.: An Information-Theoretic Definition of Similarity. In: Fifteenth International Conference on Machine Learning (ICML 1998), pp. 296–304 (1998)
8. Hosmer, D.W., Lemeshow, S.: Applied Logistic Regression. John Wiley & Sons (2000)
9. Burns, R., Burns, R.: Logistic Regression. Business Research Methods and Statistics Using SPSS, pp. 568-588. SAGE Publications (2008)
10. Ramayah, T., Ahmad, N.H., Halim, H.A., Zainal, S.R.M., Lo, M.C.: Discriminant analysis: An illustrated example. African Journal of Business Management 4, 1654–1667 (2010)
11. Alkarkhi, A.F.M., Easa, A.M.: Comparing Discriminant Analysis and Logistic Regression Model as a Statistical Assessment Tools of Arsenic and Heavy Metal Contents in Cockles. Journal of Sustainable Development 1 (2008)
12. Antonogeorgos, G., Panagiotakos, D.B., Priftis, K.N., Tzonou, A.: Logistic Regression and Linear Discriminant Analyses in Evaluating Factors Associated with Asthma Prevalence among 10- to 12-Years-Old Children: Divergence and Similarity of the Two Statistical Methods. International Journal of Pediatrics 2009 (2009)
13. Dattalo, P.: A Comparison of Discriminant Analysis and Logistic Regression. Journal of Social Service Research 19, 121–144 (1995)
14. Liu, C., Wechsler, H.: Gabor Feature Based Classification Using the Enhanced Fisher Linear Discriminant Model for Face Recognition. IEEE Transactions on Image Processing 11, 467–476 (2002)
15. Kazemnejad, A., Batvandi, Z., Faradmal, J.: Comparison of artificial neural network and binary logistic regression for determination of impaired glucose tolerance/diabetes. Eastern Mediterranean Health Journal 16, 615–620 (2010)
16. Terrin, N., Schmid, C., Griffith, J., D'Agostino, R., Selker, H.: External validity of predictive models: A comparison of logistic regression, classification trees, and neural networks. Journal of Clinical Epidemiology, 721–729 (2003)
17. Patterson, D.: Artificial Neural Networks: Theory and Applications. Prentice Hall, Englewood Cliffs (1996)
18. Cerny, P.A.: Data mining and Neural Networks from a Commercial Perspective. In: 36th Annual ORSNZ Conference (2001)
19. Raghavendra, B.K., Srivatsa, S.K.: Evaluation of Logistic Regression and Neural Network Model With Sensitivity Analysis on Medical Datasets. International Journal of Computer Science and Security (IJCSS) 5, 503–511 (2011)
20. Baxt, W.: Application of artificial neural networks to clinical medicine. Lancet, 1135–1138 (1995)
21. Eftekhar, B., Mohammad, K., Ardebili, H.E., Ghodsi, M., Ketabchi, E.: Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. BMC Medical Informatics and Decision Making 5 (2005)
22. Leondes, C.: Neural network systems, techniques, and applications. Academic Press, San Diego (1998)
23. Song, J.H., Venkatesh, S.S., Conant, E.A., Arger, P.H., Sehgal, C.M.: Comparative Analysis of Logistic Regression and Artificial Neural Network for Computer-Aided Diagnosis of Breast Masses. Academic Radiology 12, 487–495 (2005)
24. Klusch, M.: The S3 Contest: Performance Evaluation of semantic web services. In: Blake, M.B., Cabral, L., König-Ries, B., Küster, U., Martin, D. (eds.) Semantic Web Services: Advancement through Evaluation. Springer (2012)
25. English Relevance Judgements, `http://trec.nist.gov/data/reljudge_eng.html`
26. Swets, J.: Measuring the accuracy of diagnostic systems. Science 240, 1285–1293 (1988)
27. Bishop, C.: Neural networks for pattern recognition. Oxford University Press (1995)

A Ruled-Based Part of Speech (RPOS) Tagger for Malay Text Articles

Rayner Alfred[1], Adam Mujat[1], and Joe Henry Obit[2]

[1] School of Engineering and Information Technology, Universiti Malaysia Sabah, Jalan UMS, 88400, Kota Kinabalu, Sabah, Malaysia
[2] Labuan School of Informatics Science, Universiti Malaysia Sabah, Labuan, Malaysia
ralfred@ums.edu.my, adammujat@gmail.com, joehenryobit@yahoo.com

Abstract. The Malay language is an Austronesian language spoken in most countries in the South East Asia region that includes Malaysia, Indonesia, Singapore, Brunei and Thailand. Traditional linguistics is well developed for Malay but there are very limited resources and tools that are available or made accessible for computer linguistic analysis of Malay language. Assigning part of speech (POS) to running words in a sentence for Malay language is one of the pipeline processes in Natural Language Processing (NLP) tasks and it is not well investigated. This paper outlines an approach to perform the Part of Speech (POS) tagging for Malay text articles. We apply a simple Rule-based Part of Speech (RPOS) tagger to perform the tagging operation on Malay text articles. POS tagging can be described as a task of performing automatic annotation of syntactic categories for each word in a text document. A rule-based POS tagger generally involves a POS tag dictionary and a set of rules in order to identify the words that are considered parts of speech. In this paper, we propose a framework that applies Malay affixing rules to identify the Malay POS tag and the relation between words in order to select the best POS tag for words that have two or more valid POS tags. The results show that the performance accuracy of the ruled-based POS tagger is higher compared to a statistical POS tagger. This indicates that the proposed RPOS tagger is able to predict any unknown word's POS at some promising accuracy.

Keywords: Rule-Based POS Tagger, Computational Linguistic, Malay Affixing Rules, Malay Word Relation.

1 Introduction

In Malaysia, the Malay language is officially known as Bahasa Malaysia, which translates as the "Malaysian language". The total number of speakers of Standard Malay is about 18 million. There are also about 170 million people who speak Indonesian, which is a form of Malay. Malay language is used as a national language for Malaysia and Indonesia and ranked fourth after Spanish for the most widely spoken languages on earth. Nevertheless, it is one of the least studied and known about, to the extent that it is even left out of rank orders of the world's major languages. Traditional linguistics is well developed for Malay but there are very limited resources and tools that

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 50–59, 2013.

are available or made accessible for computer linguistic analysis of Malay language. For example, the part of speech (POS) tagging for Malay text articles is one of the limited tools for computer linguistic analysis. POS is a process of tagging a text into corresponding part of speech tag based on the word definition and relation. A post of speech (POS) tagger for Malay language has some end product applications. Firstly, POS tagger for Malay language can be used as a grammar checker that identifies word relation based on word class, by checking the word class before and after the word. Next, a POS tagger for Malay language can also be used to classify question by identifying question focus [6] (e.g., a noun and verb after the interrogative word and keyword can be used to identify the question focus).

For English language, a simple rule-based POS tagger was first introduced by Eric Brill [1]. In his work, he has illustrated that a rule-based tagger for English language can perform as good as taggers based upon probabilistic or statistical models. Statistical tagging for English text articles has been widely applied into tagged corpora using various approaches. Among the early technique was Hidden Markov Model (HMM) algorithm [12] which achieved the accuracy of more than 96% for English text articles. For Malay language, a statistical POS tagger using trigram Hidden Markov Model for tagging Malay text articles has been designed but only achieved the accuracy of 67.9%. The efforts in statistical POS tagger initiatives are mainly focused on European languages like English, German, Spanish etc [7,8,9,10,11]. The development of this research is mainly contributed by the availability of their language resources such as dictionaries and annotated corpus. Minority languages such as Malay language still need more supports in term of researches conducted in order to assist the development of tools for computer linguistic analysis of Malay language. In this work, a framework of a rule-based POS tagging for Malay language will be outlined, since Malay language has a very limited POS tagged corpus accessible for Malay language researchers.

This is paper is organized as followed. Section 2 explains the background of the POS tagger for Malay language. Section 3 outlines the ruled-based POS tagger framework for Malay language articles. Section 4 describes the experimental design setup and discusses the experimental results. Section 5 discusses the results obtained from the experiments and finally, this paper is concluded with future works in Section 6.

2 Part of Speech Tagging for Malay Language

Part of Speech (POS) tagging is a process of tagging a text into corresponding word class or part of speech, based on word definition and word relation. A simple rule-based POS tagger for Malay language applies a POS tag dictionary and affixing rules in order to identify the Malay word definition. The POS tag dictionary is manually extracted from the Malay thesaurus and stored in a text format [2]. Fig. 1. illustrates a snapshot of the Malay POS tag dictionary. Table 1 shows the list of POS tags for Malay language words.

adunan	NN
agak	GUT
agaknya	PEN
mengagak-agak	VB
teragak	VB
teragak-agak	VB

Fig. 1. Malay POS tag dictionary snapshot

All the affixing rules that are applied in the proposed approach are studied and manually extracted from the *Tatabahasa dewan edisi ketiga* [3]. The derived word relations are based on the word types where some word types co-occur with words other word types (see Table 2). For instance, given a phrase in Malay language as follows,

saya suka makan → saya (NN) suka (JJ/VB/RB) makan (VB)

where *saya* is a noun that co-occurs with the word *suka* which is classified as an adjective, a verb and an adverb. However, *makan* is a verb and only an adverb that is allowed to co-exist with the word *makan*. Thus, we will have the following word relations

saya suka makan → saya (NN) suka (RB) makan (VB)

Table 1. POS tag list for Malay

Word Type (English language)	**Subtype (English language)**	**Subtype (Malay language)**	**Tag**
Noun			NN
	Proper noun		NNP
Verb			VB
Adjective			JJ
Function	Conjunction	Kata hubung	CC
	Interjection	Kata seru	UH
	Interrogative	Kata tanya	WP
	Command	Kata perintah	CO
		Kata pangkal ayat	PNG
	Auxiliary (Amplifier)	Kata bantu	AUX
		Kata penguat	GUT
	Particles	Kata penegas	RP
	Negation	Kata naïf	NEG
		Kata pemeri	MER
	Preposition	Kata sendi nama	IN
		Kata pembenar	BNR
	Direction	Kata arah	DR
	Cardinal number	Kata bilangan	CD
		Kata penekan	PEN
		Kata pembenda	BND
	Adverb	Adverb	RB

Table 2. Word Type Relation

Word Type	Valid Sequences of Word Types
Noun (NN)	adjective (JJ), adverb (RB),verb (VB),noun (NN),preposition (IN)
Verb (VB)	auxiliary (AUX), adverb (RB), noun (NN), penekan (PEN), pembenda (BND)
Adjective (JJ)	penguat (GUT), preposition (IN)
Adverb (RB)	verb (VB), preposition (IN), adjective (JJ), noun (AUX)
Direction (DR)	noun (NN), preposition (IN)
Preposition (IN)	noun (NN), verb (VB), adjective (JJ)
Auxiliary (AUX)	adjective (JJ), verb (VB), preposition (IN)
Cardinal number	noun (NN)
Penekan (PEN)	adverb (RB), noun (NN), conjunction (CC)
Pembenda (BND)	conjunction (CC), noun (NN)
Conjunction (CC)	noun (NN), verb (VB), preposition (IN), adjective (JJ)
Penguat (GUT)	adjective (JJ)
Interrogative (WP)	noun (NN), verb (VB)
Pangkal ayat (PNG)	noun (NN)

Most Malay POS tagging systems apply a POS tag dictionary and affixing rules acquisition for POS (see section 3), because of the unavailability of resources such as tagged POS tag corpus.

2.1 Analysis of Affixed Word

Bali analyzes Malay affixed words by identifying affixed words, segmenting them and finally interpreting the affixed words in Malay language [4]. In Malay, the form of words can be simple or complex. Affixed words are complex words generated by a morphological process called affixation that includes *prefixation*, *suffixation*, *circumfixation*, and *infixation*. *Prefixation* is the process of adding a prefix at the left side of the base and *suffixation* is the adding of a suffix to the right side of the base (See Fig. 2.). *Circumfixation* is the simultaneous adding of a discontinuous morphological unit called circumfix at the left and right sides of the base [4]. A circumfix is a combination of a prefix and a suffix treated as a single morphological unit. In Malay language, *infixation* is the insertion of an infix just after the first consonant of the base.

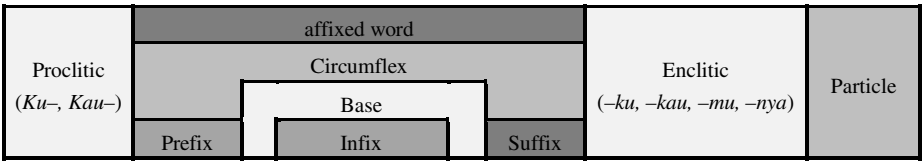


Fig. 2. Clitics, Affixing and Particle in Malay word

In Bali's work, she has identified the affixing words, the clitics and particles and their relations. A word containing clitics and particle cannot be affixed but affixed word may have clitic and particle. In Fig. 2, it is shown that an affixed word can be

the host of one and only one clitic and/or one and only one particle. A clitic attached before the base is called proclitic and a clitic attached after the base is called enclitic.

Fig. 2 shows the structure of an affixed word in Malay with the addition of a clitic (proclitic or enclitic) and particle [4]. In Malay language, there are two proclitics, four enclitics and three particles. *Ku–* and *Kau–* are two proclitics that generate passive word. On the other hand, *–ku*, *–kau*, *–mu* and *–nya* are four enclitics that are functioning as an object pronoun of active verb and a possessive adjective. In addition to that, the enclitic *–nya* is also functioning as a subjective pronoun of passive verb and a definite article. Finally, the Malay particles include *–kah* and *–tah* that generate question marker, and *–lah* that generates imperative and predicative marker.

3 A Rule-Based Part of Speech (RPOS) Tagger for Malay Texts

In this paper, the proposed rule-based POS tagger for Malay language applies three general tagging convention of the Penn Treebank [6] that includes

a) the part of speech tags are defined based on their syntactic distribution rather than their semantic function and
b) the tagger capitalizes words tagged as proper noun and
c) the tagger tags the abbreviations and initials.

In addition to that, the proposed rule-based POS tagger for Malay language has additional POS tags which are not included in the Penn Treebank tags [3]. These tags include

a) *kata perintah* – command (CO)
b) *kata pangkal ayat* (PNG)
c) *kata bantu* – auxiliary (AUX)
d) *kata penguat* (GUT)
e) *kata naïf* – negation (NEG)
f) *kata pemeri* (MER)
g) *kata pembenar* (BNR)
h) *kata arah* – direction (DR)
i) *kata penekan* (PEN)
j) *kata pembenda* (BND)

In this paper, we outline a simple rule-based POS tagger for Malay language. The rules involve the affixing and word relation rules [3]. Malay language affixing has a prefix, infix, suffix and combination, in this paper only the prefix, suffix and combination are considered. This is because infix is not a productive affixing and it can cause ambiguity in the POS tagging as a similar infix may exist in the noun, verb and adjective. The affixing rules consist of a noun (as shown in Table 3), a proper noun, an adjective (as shown in Table 4), a verb (as shown in Table 5), *pembenda*, *penegas* and *penekan*.

The *penegas* rule includes a sequence of characters ending in *–kah*, *–lah* and *–tah*. The *pembenda* rule includes a non noun root word and ending with *–nya*. Finally, the *penekan* rule includes a noun root word and ending with *–nya*. In addition to the affixing rules, we also include the word type relation rules. The word type relation rule is a rule used for selecting the base POS tag to represent the word if the word has more than one POS tags. This is done by checking the validity of the word type relation before and after the word as explained in the Section 2. The word type relation list, shown in Table 2, is not an exhaustive list which is extracted from *Tatabahasa Dewan Edisi Ketiga* [3].

Table 3. Noun Affixing Identification Rules

Rules	Prefix	Next character	Sequences of character	Suffix	
				May end with	
1a	pe	ny, ng, r, l and w	a-z	an	-
1b	pem	b and p	a-z	an	-
1c	pen	d, c, j, sy and z	a-z	an	-
1d	peng	g, kh, h,k and vowel	a-z	an	-
1e	penge	-	a-z (3 to 4 character)	an	-
1f	pel or ke	-	a-z	an	-
1g	juru, maha, tata, pra, swa, tuna, eka, dwi, tri, panca, pasca, pro, anti, poli, auto, sub, supra	-	a-z	-	-
1h	not started with me, meng, mem, menge, ber, be, di, diper	-	a-z	-	an, at, in, wan, wati, isme, isasi, logi, tas, man, nita, isme, ik, is, al

Table 4. Adjective Affixing Identification Rules

Rules	Prefix	Next character	Sequences of character	Suffix	
				May end with	
2a	ter, se, bi	-	a-z	-	-
2b	ke	-	a-z	an	-
2c	not starting with di and men	-	a-z	-	in, at, ah, iah, sequences of vowels then wi and sequences of consonants end ending with i

Table 5. Verb Affixing Identification Rules

Rules	Prefix	Next Character	Sequences of character	Suffix	
				May end with	
3a	me	ny, ng, r, l, w, y, p, t , k, s	a-z	-	-
3b	mem	b, f, p and v	a-z	kan or i	-
3c	men	d, c, j, sy, z, t and s	a-z	kan or i	-
3d	meng	g, gh, kh, h, k and vowel	a-z	-	-
3e	menge	-	a-z (3 to 4 character)	-	-
3f)	memper or diper	-	a-z	kan or i	-
3g)	ber	not r	a-z	kan or an	-
3h)	bel	-	a-z	-	-
3i)	Ter	not r	a-z	-	-
3j)	Ke	-	a-z	-	An
3k)	-	-	a-z	-	i or kan
3l)	di or diper	-	a-z	kan or i	-

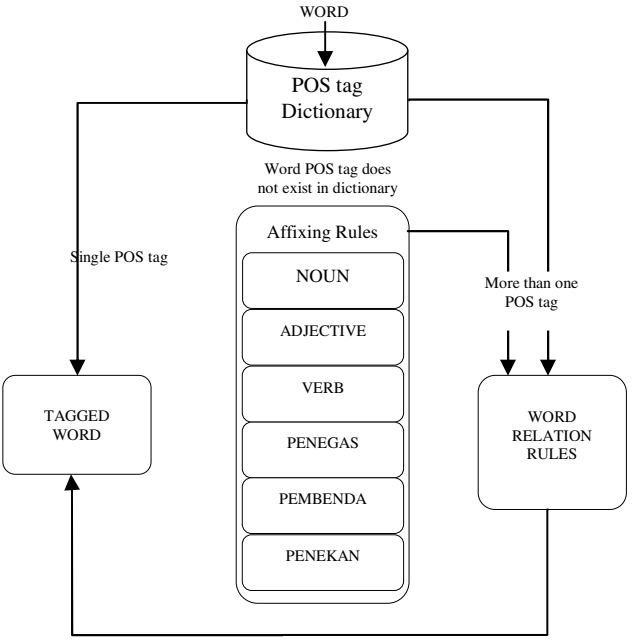


Fig. 3. The Rule-Based POS Tagger Framework for Malay Text Articles

Fig. 3 illustrates the framework of the proposed rule-based POS (RPOS) tagger for Malay text articles which consists of a POS tag dictionary, a set of affixing rules and word relation rules. The POS tag dictionary consists of Malay words with their POS tags and these Malay words are extracted manually from *Thesaurus Bahasa Melayu*

that has more than 8,700 tagged words [2]. The Malay language affixing generates a new word and meaning and in this paper we apply affixing characteristics in order to identify POS tags only for the noun, adjective, verb, *penegas*, *pembenda* and *penekan*.

First, the rule-based POS tagger starts by checking the existence of the word POS tag in the POS tag dictionary. If the word exists in the POS tag dictionary and has only one tag then the word tagging is completed. If the word exists in the dictionary and has more the one possible tagging name, identifying valid word type relation will be performed to select the proper POS tag name. Otherwise, if the word does not exist in the POS tag dictionary, the word will be processed in line with the affixing rules before it is processed in the tagging process again.

4 Experimental Setup and Evaluations

In this experiment, we have extracted ten sets of news article from the Malay online news and ten sets of biomedical articles from the Malaysian Journal of community health (http://161.142.92.97/). We performed the rule-based POS tagging process on these sets of news and biomedical articles based on the affixing and word type relation rules. We then compared the results with the actual tags. We have performed the process of tagging the words manually in order to evaluate the accuracy of our proposed algorithm. Table 6 shows the percentage accuracies of the rule-based POS tagger performance against the manually tagging process for both the news and biomedical articles. In Table 6, the total token represents the actual number of word found in the test sets. The counted token represents the number of words actually used for POS tagging.

Table 6. Experiment Results for Rules based POS tagging for Malay language

Test Set	News Articles		Biomedical Articles		Accuracy (%) News Articles	Accuracy (%) Biomedical Articles
	Total token	Counted token	Total token	Counted token		
1	386	296	456	399	94	86
2	302	229	436	400	91	89
3	199	167	501	440	85	89
4	185	132	490	444	88	82
5	177	141	434	390	86	88
6	189	136	453	379	90	79
7	149	107	477	420	92	83
8	175	136	411	380	92	93
9	304	225	300	250	88	87
10	434	340	231	187	86	85
				Average	**89**	**86**

5 Discussions

The results show that the proposed rule-based Malay POS tagger achieves 89 percent accuracy for the Malay news articles and 86 percent accuracy for the Malay biomedical articles. The result of the rule-based Malay POS tagger for Malay biomedical

articles is lower due to the existing of some borrowed words in Malay from the English Language.

Based on our experiment results, for the news articles, we also have identified some of the words POS tags that the rule-based POS tagger for Malay language has failed to identify. These words POS tags include the words *kopersai* (NN), *berniaga* (VB), *selepas* (RB), *waktu* (NN/AUX), *bertugas* (VB), *selepas*(RB) and *waktu* (AUX).

On the other hand, for the biomedical articles, it shows that the rule-based POS tagger for Malay language have failed to identify some words POS tags that include words which are borrowed from the English language such as *antropometri* (anthropometry – a noun), *dialysis* (dialysis – a noun), *inflamasi* (inflammation – a noun), *komplikasi* (complication – a noun), *vascular* (vascular – a noun or adjective), *nefropati* (nephropathy – a noun), *neuropati* (neuropathic – a noun), *retinopati* (retinopathy – a noun), *infarksi* (infarction – a noun), *myocardium* (myocardium – a noun), *amputasi* (amputation – a noun) and *superfisial* (superficial – a adjective).

6 Conclusion

In this paper, we have outlined the framework for a simple Rule-based Part of Speech (RPOS) tagger for Malay text articles. Based on our experiment results, the performance of the proposed rule-based POS tagger is acceptable compared to performance of a statistical POS tagger reported earlier. This indicates that a ruled-based POS tagger for Malay language is able to predict any unknown word's POS at some promising accuracy. The performance of the proposed rule-based POS tagger for Malay language can be improved by adding more word type relations and more POS tags into the POS tag dictionary. By improving the word type relations, more sentence formats can be handled.

References

1. Brill, E.: A simple rule-based part of speech tagger. In: HLT 1991: Proceedings of the Workshop on Speech and Natural Language, pp. 112–116. Association for Computational Linguistics, Morristown (1992)
2. Thesaurus Bahasa Melayu, New Edition Kuala Lumpur, Dewan Bahasa dan Pustaka (2008) ISBN 983628558X
3. Karim, N.S., Onn, F.M., Musa, H.H., Mahmood, A.H.: Tatabahasa Dewan Edisi Ketiga. Dewan Bahasa dan Pustaka, Kuala Lumpur (2008)
4. Ranaivo-Malancon, B.: Computational Analysis of Affixed Words in Malay Language. In: The 8th International Symposium on Malay/Indonesian Linguistics (ISMIL8), Penang, Malaysia (2004)
5. Purwarianti, A.: Developing Cross Language Systems for Language Pair with Limited Resource-Indonesian-Japanese CLIR and CLQA, Phd. thesis, Toyohashi University of Technology (2007)
6. Santorini, B.: Part-of-Speech tagging guideline for the Penn Treebank Project, 3rd Revision, 2nd Printing (1990)

7. Merialdo, B.: Tanging English Text with a Probabilistic Model. Computational Linguistics 20(2), 155–171 (1994)
8. Elworthy, D.: Does Baum-Welch Re-estimation Help Taggers? In: Proceedings of the 4th ACL Conference on Applied Natural Language Processing, ANLP (1994)
9. Banko, M., Moore, R.C.: Part of Speech Tagging in Context. In: Proceedings of the 8th International Conference on Computational Linguistics, COLING (2004)
10. Wang, Q.I., Schuurmans, D.: Improved Estimation for Unsupervised Part-of-Speech Tagging. In: Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering, IEEE NLP-KE (2005)
11. Biemann, C., Giuliano, C., Gliozzo, A.: Unsupervised Part-Of-Speech Tagging Supporting Supervised Methods. In: Proceedings of RANLP 2007, Borovets, Bulgaria (2007)
12. Jurafsky, D., Martin, J.H.: Speech and language processing. Prentice Hall, New Jersey (2000)

Viable System Model in Capturing Iterative Features within Architectural Design Processes

Roliana Ibrahim[1], Khairul Anwar Mohamed Khaidzir[2], and Fahimeh Zaeri[2]

[1] Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia, UTM, Johor Bahru, Malaysia
[2] Faculty of Built Environment, Universiti Teknologi Malaysia
UTM, Johor Bahru, Malaysia
roliana@utm.my

Abstract. In project management, iteration can be seen as an undesirable characteristic that increases risk and lengthen the cycle time. However, in design management, iteration is the key feature in designing. Iteration can also be manifest as different types which give particular characteristics to different stages of the design process. However, there are no existing methods that could capture and model the iterative activities of designers and support the analysis and design of designers' process management systems. The design structure matrix is one method that has the capability to capture iterative activities. However, this method does not seem suitable to support the development of a design process management system as it does not highlight the functional features within iterative activities. Therefore, the aim of this paper is to discuss the potential use of viable system model in capturing functional features and requirements within iterative activities of the architectural design process. This paper also highlights an example of a previous study which adapted viable system model in the diagnosis of complex processes.

Keywords: design process, design thinking, Viable System Model, complex processes, systems engineering.

1 Introduction

Creative architectural design processes can be considered as complex processes. One of the basic features in a design process that makes design a complex process is known as iteration [1], which commonly refers to a cyclical process. It is believed that the design concepts emerge and become complete through iterations of analysis, synthesis and evaluation. During the design process, designers iteratively explore problems, generate ideas and operationalize concepts for better solutions. In architectural design, design activity is a formal search through a problem space for objects that satisfy multiple constrains as well as uncertainty surrounding the design process [2]. Design processes are modeled as mechanistic interactions between activities.

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 60–69, 2013.

[3] describe six perspectives of iteration in engineering design processes. These perspectives are exploration, convergence, refinement, rework, negotiation and repetition. Repetition is considered to differ from exploration, convergence, negotiation, rework and refinement because it involves re-visiting similar design activities to achieve a different goal, rather that re-visiting a goal using potentially different methods. In computer science, iteration can be defined as the repetition of a process. As an activity, it can be a goal directed and non-linear process that utilizes heuristic designing process and strategies [2]. For project managers, iterations can be seen as an undesirable characteristic which increases risk and lengthens cycle time. However, from the designers' perspectives, iteration is a key feature in designing. Iterations can also be manifest in different types which give particular characteristics to the different stages of the design process. Since, there are the two different perspectives of project managers and designers in design processes, this study aims to explore the iterative activities within the architectural design processes, particularly focusing on capturing the functional and requirements of each activity. Architects are designers rather than project managers and the viability of iterative activities in architectural design processes could determine the success of design for architectural designers.

The models in existing research involves that address complex iterations in design processes are task-based, actor-based and information based [3]. However, in [3], the modeling of iterative activities in their study is for new product development processes, and the nature and priorities of those processes are different from architectural design processes. An example of a previous study in modeling the complex iterations of architectural processes is the design structured matrix [4]. Unfortunately, it is unclear how they extracted the parameters in terms of functionalities to support the systems design of architectural process management system. The method they used is not a functional based method appropriate for systems design. Therefore, this paper is discusses the potential use of an alternative model for capturing the functional features and requirements within iterative activities of the architectural design process. The paper is organized as follows : Section 2 and 3 briefly explain iterations in the context of design processes and the methods to capture them. Section 4 and 5 explains the viable system model (VSM) and provide a case study example.

2 Capturing Iteration Activities

In order to capture iterative activities, we need to diagnose the architectural design processes. For this task we use a viable system model as our modeling approach. A VSM is a systems thinking approach, introduced by Stafford Beer [24], which regards the system in focus as an organization. It is also regarded as a functional model, which can be used to support systems design. Many researchers have used VSM as a template and framework for diagnosis [6]. The VSM is built upon three principles: cybernetics, the law of requisite variety and recursion [5]. The cybernetics principle explains how systems are connected as one component sharing resources. The law of

requisite variety enables systems to be self-sustaining by accommodating changes from the environment. Recursion shows how a viable system is contained; within a recursion, another viable system is contained and there are some components to monitor the feedback loop in VSM model.

Some related works on the use of the VSM are [5] and [6]. These works focuses on diagnosing the processes within an organization, but did not clearly show how they established the processes or reveal the information flow within the organization and use this information in the system design phase for systems development. Acknowledging that iteration is fundamental and necessary for designers and that it is vital for design processes to proceed consistently in a stable state, this study aims to explore the viability of the iteration activities in architectural design processes using the viable system model approach.

3 Overview of Iteration

Iteration is considered to be an integral part of the design process. It is a natural feature of a designer's competency. As an activity, iteration can be a goal-directed and non-linear process that utilizes heuristic designing process and strategies [7].

The term "iteration" is used in a variety of disciplines and in different contexts. Iteration commonly refers to a cyclical process. The iterative method describes a problem-solving methodology in many fields. These iterative methods share the description of techniques that use successive approximations to obtain more accurate solutions at each step [8]. Two fundamental forms of iteration in the field of system analysis and design process are as follows [9]:

a. Iterating cognitive processes which take place in the minds of the developers, often through interactions with representations. Since this kind of iteration occurs in the mind of the designer, it is less explicit and is described as mental iteration.
b. Iterations of representational artifacts that are used by designers and others during the design.

A number of studies advocate that better problem scoping leads to better performance [10] or that the use of opportunistic or flexible methodical approaches to design problem-solving are a function of expertise [11],[12]. Moreover they have found that instructional inventions have a positive effect on the number of transitions between steps in the design process, the number of criteria considered and the number of alternatives generated. In a large scale study of design problem-solving behavior among freshman and senior students, it was found that seniors performed more transitions in between design steps. It shows a higher number of transitions per minute, and that the transition behavior for both groups of students related positively to the quality of the final design[13]. The seniors students were also found to exhibit more problem scoping behaviors, and the number of constraints was considered to be correlated with the quality and number of alternative solutions generated [20]. According to Radcliffe and Lee, "the adoption of a systematic, iterative and logical design sequence correlated with the effectiveness of the design and the efficiency of

the designer's process"[14], [20]. Overall, we can propose that all systems analysis and design depends on in the thoughts of designers. Therefore this cognitive activity occurs in an iterative fashion, where some form of mental looping operation takes place to guide the design [20]. In other word, "mental iteration in engineering design is a repetition of cognitive activities that occur in designer's thinking process" [15].

4 Iteration in the Design Problem-Solving Process

In design-problem solving the problem formulation phase is not static. The process starts with searching for information to understand and solve the problem. This cyclical process continues until the designer can describe a final solution. In fact, there isn't any stopping point in the design problem-solving process unlike science problem- solving. A design activity is "a formal search through a problem space for objects that satisfy multiple constraints" [16]; and the boundaries of this space "are mutable and open to interpretation, and evolve through a process of iteration" [17], [18].

According to [17], for engineering design problems, "there are often multiple solutions that can satisfy design requirements. The presence of multiple solutions means that design solutions are a result of design decisions that emerge unpredictably as iterations". During design, designers iteratively explore problems for better understanding, generate ideas and operationalize concepts for better solutions. Therefore iteration as a design decision model has an effect on understanding the problems, and the modification or generation of solutions. We conclude this open-ended process through following steps:

*Step 1 : Selecting a starting point :*In this step, the designer should identify dominant need and problem and then s/he selects a start point. There are many problems in selecting the exact starting point. Therefore it is better for the designer to start with iteration instead of spending lots of time to find the accurate starting point.
Step 2 : *Identifying and understanding problems* :In fact, the designer starts to gain insight into the problems. S/he tries to explore the problems in this step.
Step 3 : *Deepening and broadening the insight* :Making decisions, modifying, operationalizing and generating the solutions occur in this step.
Step 4: *Refining or selecting the next need or problem.*

Step 1 is to select a starting point. To select a starting point, it's more important that we start with the iteration than to spend a lot of time trying to find the most ideal starting point. After step 1, the iteration starts with step 2 to gain insight. The next iteration is prepared in step 4 refining or selecting the next need or problem. As [20] explained, during this iteration, "continuous effort is required to communicate with the stakeholders to keep them up to date, to consolidate in simple models that are used during analysis and discussions and to refactor the documentation to keep it up to date with the insight obtained".

Transition is described as "a goal-directed progression between steps of the design process that may contribute to a change in the design state" [6]. In this context we can describe iteration as a goal-directed problem-solving process that is associated with the sequence of transition behaviors between information processing and decision making. These behaviors may be reasoning processes that utilize information processing activities to guide decision activities. These behaviors are similar to transforming classes of inputs to classes of outputs in information processing theories. Inputs are considered as information processing activities which are used to gather and filter information about the problem and solutions, and outputs are considered as decisions to change or elaborate the problem representation or possible solutions. During the process of design problem- solving, developing and refining the formulation of a problem and ideas about a solution are caused with constant iteration of analysis and evaluation processes between the problem space and solution space [19].

5 Viable System Model (VSM)

The VSM was introduced by Beer [24,25], with the focus on understanding organizations, redesigning them and supporting the management of change [28]. The main VSM concept is recursion. The organizational structure of the VSM is not the traditional hierarchical structure but a recursive structure. It is believed that organizations with recursive structures operate differently from those having hierarchical ones. Organizations with a hierarchical structure operate according to a top-down command structure. In a recursive structure, the organization is perceived as containing living systems, with each one comprising defined sub-systems. Each of these individual sub-systems contains self-organized and self-regulated characteristics, which make each of them a viable system in itself. A system is considered to be viable if it is able to maintain its separate existence [25].

This means that the sub-systems identified in the organization with a recursive structure are considered as autonomous units and are able to function independently and adapt to changes in the environment. [26] describes autonomous units as the procedural activities of an organization. These procedural activities are also the main primary activities operating within the organization. There are two mechanisms that support viability. These mechanisms are cohesion and adaptation [27]. Both mechanisms are equally important for maintaining the organization as a whole. It is important to take into account the cohesiveness between functionally decentralized, autonomous units. To achieve cohesion between each autonomous unit, each one should be able to communicate and regulate effectively so that, whatever the primary activities undertaken by those units, they are able to provide services to the organization. Achieving cohesion also means providing the autonomous units with true or local autonomy within an integrated framework. Cohesion in an organization is achieved by showing the necessary supporting links between individual or autonomous units. To achieve adaptation, the autonomous units as a whole should not only be cohesive but, as a whole, they must also be adaptive to changes within their environment. This is the criterion which transforms the collective into an organization.

In order to investigate whether the autonomous units become a cohesive whole and are adaptive to any changes within their environment, the VSM concept incorporates cybernetics concepts. Cybernetics is considered suitable and useful in dealing with the complexities of any kind of system. A system is anything that contains many parts which are connected together [28]. A system is considered complex when interconnectivity and interrelationships between parts or elements within the system are difficult to unravel. Cybernetics offers a way of investigating a system where complexity is outstanding in a scientific manner. The VSM concept separates the organization into two parts. These parts are the meta-system and the operational system. The VSM prescribes five functions for diagnosing whether those autonomous units exist within the organization as a whole, and at the same time have the capacity to maintain a separate existence. These functions are known as systems one to five. Figure 1 illustrates the schematic diagram of the VSM.

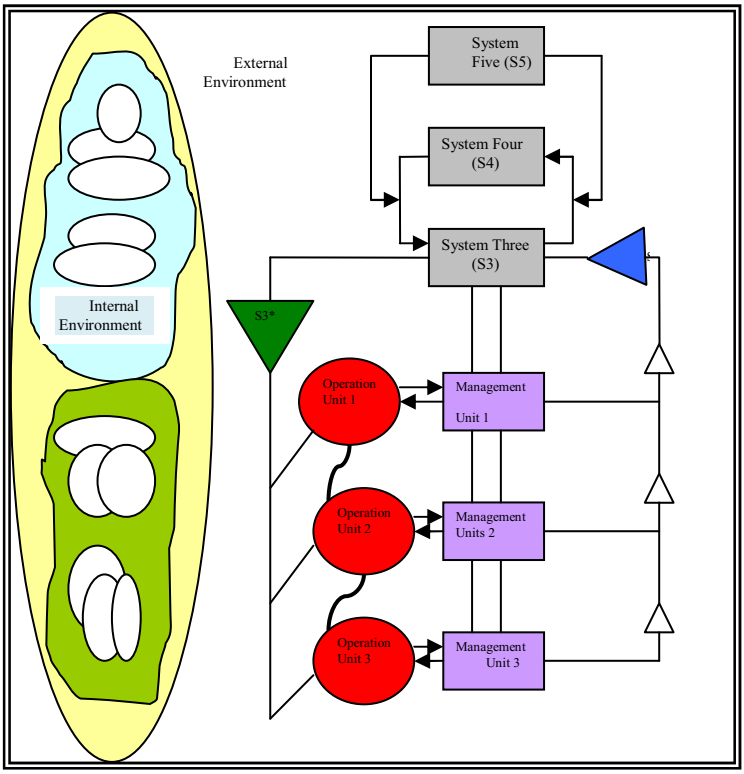


Fig. 1. Schematic Diagram of the viable system model

6 Case Study Example

The reasons for choosing VSM in this study are mainly due to its recursive and cybernetics concepts. We believe that, having these concepts, we would be able to

capture the iterative activities of designers in architectural design processes. In these iterative activities we would focus on the functional requirements that would achieve cohesion and adaptive characteristics within the design process. In our proposed iteration capturing framework, we would determine the constraint and criteria parameters within the iteration activities as collective elements for us to design a cohesive and adaptive information model which could support designers decision making within the architectural design process. We would also investigate the transition behavior between one iterative activity to another state of activity.

Below we give an example of our existing work which adapted VSM in diagnosing complex processes. The example shows the use of VSM to diagnose the process of key performance index (KPI) delivery focusing on the KPI of scholarly publications and citations. The recursive nature of VSM will enable the researcher to identify the unsustainable parameters that affect the KPI delivery and also suggest a sustainable tool that will help in achieving the targets based on monitoring of the KPI delivery. The concept of using VSM as diagnosis tool follows the work of [22] and [23].

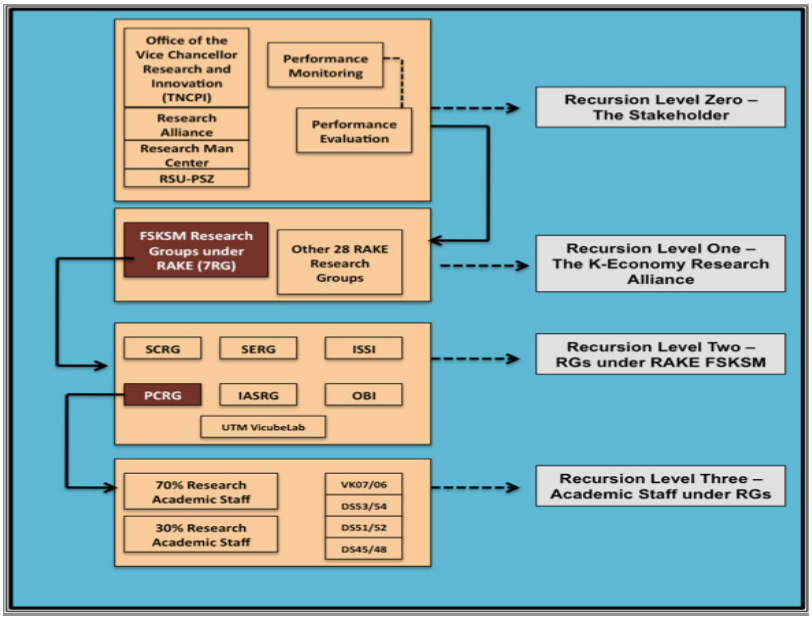


Fig. 2. Adaptation of VSM for diagnosing complex processes

Recursion level 0 is the stakeholder group, which constitutes the major players in the KPI delivery process on scholarly publications and citations. This recursion level forms the "whole" which have other levels of recursions: for example, all the research alliances form another recursion level that also has its own VSM sub-systems. This goes further to another recursion levels such as the research groups, and these are derived from the upper level of recursion (research alliance). At the last recursion level (recursion level three), there are the academic staff group, which handles all the

information provision or data generation on scholarly publications and citations. These groups are categorized into (A): 70% Research and 30% Teaching; and (B) 30% Research and 70% Teaching. Each category also belongs to a research group (RG), and there are also different cadres of appointments from "VK07 Professor" down to "DS54 academic staff with PhD". As mentioned earlier (Section 3), the VSM is comprised of five systems whereby system 1, system 2 and system 3 are regarded as sub-system and system 4 and system 5 are the meta-system.

Due to limited space, explanation of the VSM in this case study is in the context of iterative behavior, and cybernetics is explained at Recursion Level 0. At recursion level 0, there are two system 1 divisions (Research Alliances and Research Support Unit of Library) as the sub-system to this recursion level. The function of system 1 is to implement the KPI delivery process and to ensure that all sources of publication data from other sources such as the Web of Science and Scopus and others are collected, confirmed and continually reported. This task involves the role of system 3 which monitors the feedback process. This is the step whereby the VSM of the KPI delivery process would be monitored and the constraint and criteria parameters within this process were identified. The feedback process in this VSM can be considered as the iterative behavior of the KPI delivery process.

7 Conclusion

This paper highlighted the importance of iteration in architectural design processes, the nature of iterations in design processes and the existing method for capturing iteration activities. One objective of this study is to develop an iteration capture framework for the architectural design process. The purpose of this framework is to enable researchers to analyse the iterative activities within architectural design process. In addition, the purpose is to determine the functionalities within these activities for the systems engineering of a designers' process management system. Thus, this paper suggested the use of the viable system model which consists of five systems. The justification for adopting the viable system model is, we believe that the VSM is able to capture the iterative features in a complex process due to its cybernetics elements which incorporate feedbacks and iterations.

Acknowledgment. This work is supported by the Ministry of Higher Education (MOHE) and Reseach Management Centre (RMC) at the Universiti Teknologi Malaysia (UTM) under Exploratory Research Grant Scheme (Vote No. R.J130000.7828.4L051).

References

1. Chusilp, P., Jin, Y.: Impact of Mental Iteration on Concept Generation. Transactions of the ASME 128, 14–25 (2006)
2. Idi, D.B., Zaeri, F., Khaidzir, K.A.M.: The function of creativity and Innovation in Architectural Design Management. In: International Conference on Construction and Project Management (ICCPM), Singapore, September 16-18 (2011)

3. Wynn, D.C., Eckert, C.M., Clarkson, P.J.: Modelling Iteration in Engineering Design. In: Internation Conference on Engineering Design Problem Solving (ICED 2007), Cite Des Sciences Et De L'industrie, Paris (2007)
4. Zaeri, F., Khaidzir, K.A.M., Idi, D.B.: The Applicability of the Design Structure Matrix (DSM) Method in Representing and Prescribing the Architectural Design Processes. In: Six International Conferences on Construction in the 21st century (CITC- VI), Kuala Lumpur, July 5-7 (2011)
5. Ibrahim, R., Sani, S.I., Selamat, A., Bakri, A.: Development of Sustainable Key Performance Indicator (KPI) Monitoring and Control System Using Viable System Model. In: Brunei International Conference on Engineering and Technology 2012 (BICET 2012), Bandar Seri Begawan, January 25-26 (2012)
6. Leonard, A.: Integrating Sustainability Practices Using Viable System Model. Journal of ISSS 25, 9 (2008); Jacobs, I.S., Bean, C.P.: Fine particles, Thin films and exchange anisotropy. In: Rado, G.T., Suhl, H. (eds.) Magnetism, vol. III, pp. 271–350. Academic, New York (1963)
7. Adam, R.S., Atman, C.J.: Cognitive Processes in Iterative Design Behaviour, San Juan, Puerto Rico (1999)
8. Barrett, R., Barry, M., Chan, T.F., Demmel, J., Donato, et al.: Templates for the Solution of Linear Systems - Building Blocks for Iterative Methods SIAM (1994)., http://www.netlib.org/templates/Templates.html
9. Berente, N., Lyytinen, K.: Iteration in Systems Analysis and Design: Cognitive Processes and Representational Artifacts Case Western Reserve University, USA. Working Papers on Information Systems 5(23) (2005), http://sprouts.aisnet.org/5-23
10. Sutcliffe, A.G., Maiden, N.A.M.: Analyzing the novice analyst: Cognitive models in software engineering International. Journal of Man-Machine Studies 36, 719–740 (1992)
11. Guindon, R.: Designing the design process: Exploiting opportunistic thoughts. Human-Computer Interaction 5, 305–344 (1990)
12. Ennis Jr, C.W., Gyeszly, S.W.: Protocol analysis of the engineering systems design process. Research in Engineering Design 3(1), 15–22 (1991)
13. Atman, C.J., Chimka, J.R., Bursic, K.M., Nachtman, H.L.: Comparison of Freshman and Senior Engineering Design Processes. Design Studies 20(2), 131–152 (1999)
14. Radcliffe, D.F., Lee, T.Y.: Design methods used by undergraduate engineering students. Design Studies 10(4), 199–207 (1989)
15. Yan, J., Chusilp, P.: Study of mental iteration in different design situations. Design Studies 27, 25–55 (2006)
16. Goel, V., Pirolli, P.: The structure of design spaces. Cognitive Science 16, 395–429 (1992)
17. Gero, J.S.: Design prototypes:A knowledge representation schema for design. AI Magazine 11(4), 26–36 (1990)
18. Hybs, I., Gero, J.S.: An evolutionary process model of design. Design Studies 13(3), 273–290 (1992)
19. Maher, M.L., Poon, J., Boulanger, S.: Formalising design exploration as co- evolution: a combined gene approach. In: Gero, J.S., Sudweeks, F. (eds.) Advances in formal design methods for CAD. Chapman and Hall, London (1996)
20. Grit, M.: Threads of Reasoning (2010), http://www.gaudisite.nl/
21. Flood, R.L., Zambuni, S.A.: Viable systems diagnosis. 1. Application with a major tourism services group. Systemic Practice and Action Research 3(3), 225–248 (1990)
22. Espejo, R.: The Viable System Model: A Briefing About Organizational\Structure, Chichester (2003)
23. Stafford, B.: The Heart Of The Firm. John Wiley & Sons, Chichester (1979)

24. Stafford, B.: Diagnosing The System For Organizations. John Wiley & Sons, West Susses (1985)
25. Ahmad, R., Yusoff, M.B.: A Viable System Approach To Tackle Complex Enterprise Situation For SISP. Malaysian Journal of Science 19(1), 87–94 (2006)
26. Espejo, R.: The Viable System Model: A Briefing About Organisational Structure (2003), http://www.syncho.com/pages/pdf/INTRODUCTION%20TO%20THE%20%SYSTEM%20MODELS3.pdf
27. Espejo, R., Gill, A.: The viable system model as a framework for understanding organisations (1997), http://www.phrontis.com
28. von Bertalanffy, L.: General System Theory: Foundations, Development, Applications. George Braziller, Inc., New York (1969)

Identifying Same Wavelength Groups from Twitter: A Sentiment Based Approach

Rafeeque Pandarachalil and Selvaraju Sendhilkumar

Dept. of Information Science and Technology
Anna University, Chennai-25, India
rafeeqpc@auist.net, thamaraikumar@cs.annauniv.edu

Abstract. Social scientists have identified several network relationships and dimensions that induce homophily. Sentiments or opinions towards different issues have been observed as a key dimension which characterizes human behavior. Twitter is an online social medium where rapid communication takes place publicly. People usually express their sentiments towards various issues. Different persons from different walks of social life may share same opinion towards various issues. When these persons constitute a group, such groups can be conveniently termed same wavelength groups. We propose a novel framework based on sentiments to identify such same wavelength groups from twitter domain. The analysis of such groups would be of help in unraveling their response patterns and behavioral features.

Keywords: Same wavelength group, Sentiment analysis, Behavioral analysis, Overlapping community, Homophily.

1 Introduction

Reacting to social issues or events through Online Social Networks (OSNs) has become a social habit. The rapid growth of communication technologies dramatically changed the way of expressing emotions, attitudes etc. A recent statistics[1] show that 76% of twitter users are active tweeters and 23% of facebook users check their account five or more times daily. This rich source of user generated content as attitude, opinions, comments etc. in the social media are of immense significance for the analysis of human behavior.

All OSNs follow the fundamental principle of homophily: similarity breeds connection [12]. People in the OSN may be connected to one another with regard to many sociodemographic, behavioral and interpersonal characteristics. Recent studies [2, 24] shows users in the same social circle are more likely to share same opinion. A person's sentiment towards a given issue is determined to a great extent by those of his or her neighbors. For instance, a person's propensity to purchase a commodity is heavily dependent on the kind of opinions likely to emanate from his friends. With this key observation, it is reasonable to state that

[1] http://www.socialnomics.net/2012/06/06/

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 70–77, 2013.

those who share same sentiments have a strong likelihood of falling into group of similar nature. Such groups would embody persons sharing same opinion on different issues. These persons can be grouped together to form subgroups which can be conveniently termed same wavelength groups. In other words they are the proverbial same feather birds.

Identifying such same wavelength communities online has multifaceted benefits. First, social scientists are enabled to analyse the responses of the group to a socio-political incident or an ethical issue. Second, online recommendation and targeted advertising system can be improved by deep assessment of the groups. Third, responses of the groups can be predicted when a new issue comes up.

Twitter is a micro blogging service in which people share their political, religious, business or personal views in 140 characters without constrained by space and time. Some of the recent works [5, 6, 14] observed that tweet sentiments are strong indicators to predict socio-economic fluctuations. But most of the recent works on twitter sentiments focused either on tweets or the user sentiments on existing groups. We propose a framework to identify same wavelength groups from the public based on the sentiments towards the trending issues or events. The analysis of such groups can unravel the behavioral features and response patterns in a more subtle and effective manner.

2 Related Works

Social scientists have studied extensively the sociodemographic, behavioral and interpersonal characteristics. They used the traditional mode of collecting the data through online, offline and mixed-mode surveys. But recently, the rich data from various OSNs has attracted significant attention from the research community.

Some of the previous work primarily focused on usage statistics and sequences of user activities in OSNs to analyse user behavior. Benevenuto et al. [3] used cickstream data to capture behavior of OSN users. They provided a click stream model and observed that silent interactions like profile browsing dominates other visible activities. Guo, Lei et al. [8] analysed users posting behavior of original content and observed that 20% users contribute 80% total content in the network. Jiang et al. [9] also analysed the latent and visible interactions in OSN. They conducted a study on Renren, a largest OSN available in China. They constructed a latent interaction graph to capture browsing activity among OSN users. They also observed that latent interactions dominate visible interactions. Lewis et al. [11] created a facebook dataset and they analysed how sociodemographic dimensions like gender, race and ethnicity are correlated with certain network activities. A recent work [13] examined the role of five dimensions (Openness, Conscientiousness, Extroversion, Agreeableness, Neuroticism) of personality on facebook usage and features. They observed that certain personality traits are correlated with facebook usage.

A recent work [16] examined how position in the network, activities and user preferences are correlated. They provided a new affinity measure based on distance and conducted studies on email graph and twitter mention graph. They identified the homophily in terms of demography, queries and tweets among the closely connected users.

Users in the OSNs can have multiple affiliations or dimensions. Analysing multiple social dimensions of users exposed to social network environment is known as collective behavioral analysis [20]. Tang et al. [19, 21] provided a novel framework to extract latent social dimensions based on network connectivity and constructed a discriminative classifier to determine relevant social dimensions. Behavioral prediction can be made from the learned data model. A recent work [17] used topics, social graph topology and nature of user interactions to discover latent communities in social graphs.

Twitter is micro blogging service to share interesting thought at each moment. Most of the recent work on sentiment analysis in twitter [4–7, 10, 14] has been done at the tweet level. Bollen et al. [5] analysed tweet sentiments and observed that tweets mood is a strong indicator for stock market prediction. O'Connor [14] et al. also mapped public opinion poll to tweet sentiments and found high correlation.

Some of the recent works [1, 18] also considered connected users in the twitter domain to study the behavioral correlation. Tan et al. [18] observed that the probability of sharing the same opinion is high if they are connected. Abbasi et al. [1] have selected an online community which resembles a real world community in terms of race, language, religion etc. They extracted tweets related with Arab Spring to analyse the mood before and after the event. They observed that Yemenis were more concerned about security whereas Egyptians were more concerned about revolution and freedom.

3 General Framework

Most of the previous works primarily focused either to analyse sentiments at the tweet level or to study the characteristics of tweeters in a connected environment. But people from different walks of social life may have same opinion on different issues and they need not be connected. We propose a framework to mine such groups (same wavelength groups) from twitter. Fig. 1 shows the general frame work for identifying and analysing same wavelength groups.

The tweet extraction phase extract relevant tweets with respect to the trending issues or events. Normalization is fundamental to all text mining task. Each extracted tweet may be cryptic and irregular in nature. Moreover tweet may be encoded with lot of sentiment informations like punctuation, emoticons, acronyms etc. So normalization phase is important for sentiment analysis.

Sentiment analyser will find sentiments of users towards each issue or event. Let $U = (u_1, u_2, \ldots, u_m)$ represent the set of users and $E = (e_1, e_2, \ldots, e_n)$ represent the set of events that users respond in a particular time period. Each user may express positive or negative sentiment towards each event. Users sharing

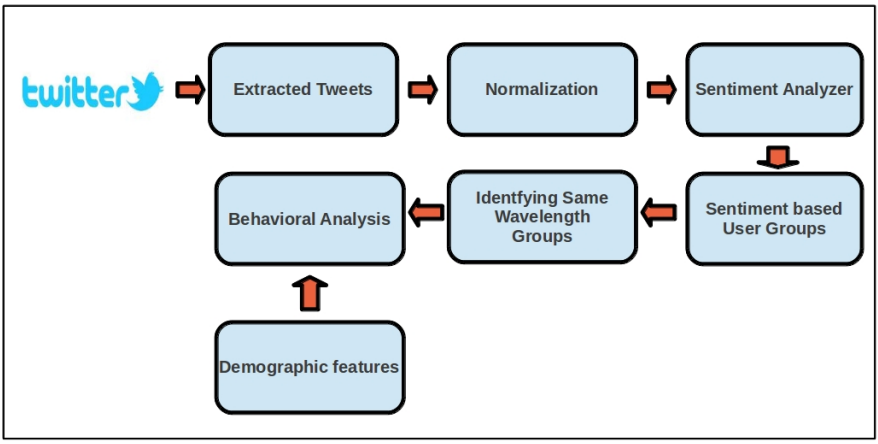


Fig. 1. General Framework

same opinion towards an event form a $k-clique$ (complete subgraph of size k) where k is the number of users shared the same opinion. For n such events $2n$ such $k-cliques$ will be formed (one for positive and other for negative). Table 1. shows the user-sentiment matrix, say $S_{m\times n}$, where each entry $S(i,j)$ represent the sentiment(positive(P) or negative(N)) towards each event.

Table 1. User-Sentiment matrix $S_{m\times n}$

Users	Event#1	Event#2	...	Event#n
$user_1$	P	P	...	P
$user_2$	P	N	...	N
$user_3$	P	P	...	N
.	...	...	...	...
$user_m$	...	...	...	...

Consider the toy example as shown in Fig. 2. Suppose there are three events (e_1, e_2, e_3) in which nine users $(u_1, u_2, \ldots, u_9)$ express their opinion. The nodes denote users and the edges denote the affiliation with respect to sentiments towards the event. The sets $(u_1, u_2, u_3, u_4, u_5), (u_2, u_3, u_5, u_6, u_7), (u_2, u_3, u_5, u_9)$ are positive sentiment groups and $(u_6, u_7, u_8, u_9), (u_1, u_4, u_5, u_9), (u_1, u_4, u_6, u_7, u_8)$ are negative sentiment groups. Each such group form a clique with various sizes. That is an edge $(u_i, u_j) \in clique$ if (u_i, u_j) share the same sentiment towards an event.

Different persons from different walks of social life may share same opinion towards various issues or events. The dotted line in the toy example shows the common users shared the positive opinion towards three events. if (c_1, c_2, c_3) are three cliques formed from the positive responses towards events (e_1, e_2, e_3) then

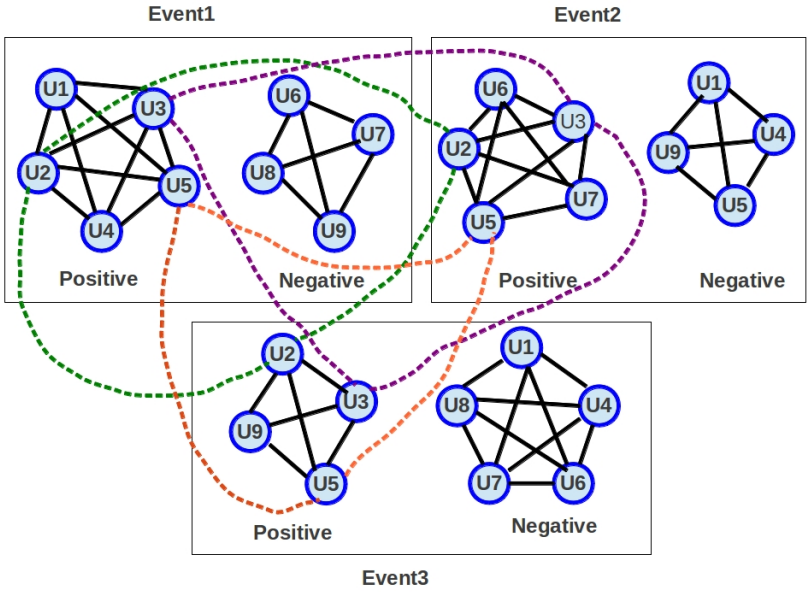


Fig. 2. A toy example. *Each clique represents the sentiments of users towards each event. Dotted line represents the set of users shared the positive opinion towards various events.*

$(u_2, u_3, u_5, c_1, c_2, c_3)$ form a group which constitutes users and subgroups share the same opinion. This group can be termed same wavelength group.

Formally, let $U = (u_1, u_2, \ldots, u_m)$ denote the set of m distinct users included in the cliques $c_i (1 \leq i \leq k)$ and $C = (c_1, c_2, \ldots, c_k)$ denote cliques generated based on the opinion towards n events. Now identifying same wavelength groups will reduce to an overlapping community identification problem [15, 23] from a bipartite graph $G(U, C, E)$, where U denote set of users and C denote set of groups(cliques) identified by the sentiment analyser phase. For instance, consider a bipartite graph with four users (u_1, u_2, u_3, u_4) and three groups (c_1, c_2, c_3) as shown in Fig. 3. The equivalent user-clique matrix is shown in Table 2 where each entry $C(i, j)$ represents the presence or absence of an user in a particular clique(group). Fig. 3 depicts how the same wavelength groups can be extracted from a bipartite graph. If the value of k (number of events) is two then two bi-cliques, $(u_1, u_2, u_3, c_1, c_2)$ and $(u_2, u_3, u_4, c_2, c_3)$ can be identified with maximum number of users. In a general bipartite graph, for a particular value of k, several such same wavelength groups may exist.

Human behavior is closely related with sociodemographic variables like age, sex, education, status etc. Analysing same wavelength group with sociodemographic features will give more insights about the evolution of such groups and hence will help to predict future activities. Moreover online social interactions are random and sometimes subtle in nature. Recently Wang, Chunyan and

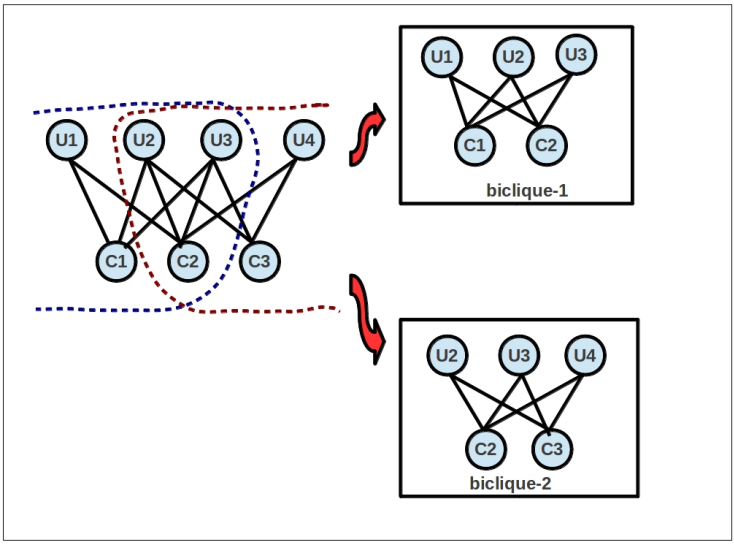


Fig. 3. Identification of same wavelength groups from bipartite graph

Huberman, Bernardo [22] observed that individual behavior is less predictable when becoming members of an explicit group. So identifying same wavelength group from the public is the more subtle way to analyse the behavioral features. Besides that recommendation systems and targeted advertising can be improved.

Table 2. User-Clique matrix $C_{4\times3}$

Users	c_1	c_2	c_3
u_1	1	1	0
u_2	1	1	1
u_3	1	1	1
u_4	0	1	1

4 Conclusions and Future Work

Opinions in OSNs have been identified as a strong dimension which induces homophily. In this paper we presented a novel framework for identifying same wavelength group from twitter domain. The idea is to determine groups of people from the public who share same opinion on various issues or events. This is one of the subtle ways to study the group responses and behavioral pattern. By using the other demographic features behavioral analysis within the groups and the shared groups can be done effectively. We have mapped the proposed framework to a graph theoretical model which will identify the cliques formed based on the

sentiments towards each issue and later determine the overlapping bicliques that share the same sentiments towards a set of issues. This work needs to be explored more using real time twitter data to evaluate the results and computational cost.

References

1. Abbasi, M.A., Chai, S.K., Liu, H., Sagoo, K.: Real-World Behavior Analysis through a Social Media Lens. In: Yang, S.J., Greenberg, A.M., Endsley, M. (eds.) SBP 2012. LNCS, vol. 7227, pp. 18–26. Springer, Heidelberg (2012)
2. Adams, J., Faust, K., Lovasi, G.S.: Capturing context: Integrating spatial and social network analyses. Social Networks 34(1), 1–5 (2012)
3. Benevenuto, F., Rodrigues, T.: Characterizing user behavior in online social networks. In: Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference, NY, USA, pp. 49–62 (2009)
4. Bifet, A., Frank, E.: Sentiment Knowledge Discovery in Twitter Streaming Data. In: Pfahringer, B., Holmes, G., Hoffmann, A. (eds.) DS 2010. LNCS (LNAI), vol. 6332, pp. 1–15. Springer, Heidelberg (2010)
5. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science 2(1), 1–8 (2011)
6. Bollen, J., Pepe, A.: Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. In: Fifth International AAAI Conference on Weblogs and Social Media, pp. 450–453 (2011)
7. Davidov, D., Tsur, O., Rappoport, A.: Enhanced sentiment learning using twitter hashtags and smileys. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, PA, USA, pp. 241–249 (2010)
8. Guo, L., Tan, E., Chen, S., Zhang, X., Zhao, Y.E.: Analyzing patterns of user content generation in online social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2009, pp. 369–378 (2009)
9. Jiang, J., Wilson, C., Wang, X., Huang, P., Sha, W., Dai, Y., Zhao, B.Y.: Understanding latent interactions in online social networks. In: Proceedings of the 10th Annual Conference on Internet Measurement, IMC 2010, pp. 369–382 (2010)
10. Jiang, L., Yu, M., Zhou, M.: Target-dependent twitter sentiment classification. In: 49th Annual Meeting of the Association for Computational Linguistics, Oregon, pp. 151–160 (June 2011)
11. Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., Christakis, N.: Tastes, ties, and time: A new social network dataset using Facebook.com. Social Networks 30(4), 330–342 (2008)
12. Mcpherson, M., Smith-lovin, L., Cook, J.M.: Birds of a Feather: Homophily in Social Networks. Annual Review of Sociology 27(1), 415–444 (2001)
13. Moore, K., McElroy, J.C.: The influence of personality on Facebook usage, wall postings, and regret. Computers in Human Behavior 28(1), 267–274 (2012)
14. O'Connor, B.: From tweets to polls: Linking text sentiment to public opinion time series. In: Proceedings of the International AAAI Conference on Weblogs and Social Media, Washington, DC, pp. 122–129 (2010)
15. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043), 814–818 (2005)

16. Panigrahy, R., Najork, M., Xie, Y.: How user behavior is related to social affinity. In: Proceedings of the Fifth ACM International Conference on WSDM 2012, Washington, pp. 713–722 (2012)
17. Sachan, M., Contractor, D., Faruquie, T.A., Subramaniam, L.V.: Using content and interactions for discovering communities in social networks. In: Proceedings of the 21st International Conference on World Wide Web, WWW 2012, pp. 331–340 (2012)
18. Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., Li, P.: User-level sentiment analysis incorporating social networks. In: ACM International Conference on Knowledge and Data Engineering (KDD 2011), California, USA, pp. 1397–1405 (2011)
19. Tang, L., Liu, H.: Relational learning via latent social dimensions. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2009, pp. 817–825 (2009)
20. Tang, L., Liu, H.: Toward Collective Behavior Prediction via Social Dimension Extraction. IEEE Intelligent Systems 25(6), 19–25 (2010)
21. Tang, L., Wang, X., Liu, H.: Scalable learning of collective behavior. Knowledge and Data Engineering 24(6), 1080–1091 (2012)
22. Wang, C., Huberman, B.A.: How Random are Online Social Interactions? Scientific Reports 2, 633–638 (2012)
23. Wang, X., Tang, L., Gao, H., Liu, H.: Discovering Overlapping Groups in Social Media. In: 2010 IEEE International Conference on Data Mining, pp. 569–578 (December 2010)
24. Yang, X., Steck, H., Liu, Y.: Circle-based recommendation in online social networks. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1267–1275 (2012)

An Improved Evolutionary Algorithm for Extractive Text Summarization

Albaraa Abuobieda[1,2], Naomie Salim[1], Yogan Jaya Kumar[1], and Ahmed Hamza Osman[1,2]

[1] Faculty of Computer Science and Information System, Universiti Teknologi Malaysia, 81310, Skudai, Johor, Malaysia
[2] Faculty of Computer Studies, International University of Africa, 2469, Khartoum, Sudan
{albarraa,ahmedagraa}@hotmail.com, naomie@utm.my, yogan@utem.edu.my

Abstract. The main challenge of extractive-base text summarization is in selecting the top representative sentences from the input document. Several techniques were proposed to enhance the process of selection such as feature-base, cluster-base, and graph-base methods. Basically, this paper proposed to enhance a previous work, and provides some limitations in the similarity calculation of that previous work. This paper proposes an enhanced mixed feature-base and cluster-base approaches to produce a high qualified single-document summary. We used the Jaccard similarity measure to adjust the sentence clustering process instead of using the Normalized Google Distance (NGD) similarity measure. In addition, this paper proposes a new real-to-integer values modulator instead of using the genetic mutation operator which was adopted in the previous work. The Differential Evolution (DE) algorithm is used for train and test the proposed methods. The DUC2002 dataset was preprocessed and used as a test bed. The results show that our proposed differential mutant presented a satisfied performance while the Genetic mutant proved to be the better. In addition, our analysis of NGD similarity scores showed that NGD was an inappropriate selection in the previous study as it performs successfully in a very big database such as Google. Our selection of Jaccard measure was fortunate and obtained superior results surpassed the NGD using the new proposed modulator and the genetic operator. In addition, both algorithms outperformed the standard baseline Microsoft Word Summarizer and Copernic methods.

Keywords: Differential Evolution, Evolutionary Algorithms, Text Summarization, NGD, Jaccard, Similarity measure, Genetic Mutant, DUC.

1 Introduction

The history of automatic text summarization (ATS) started in the late 1950s and 1960s[1,2,3].These works are concerned with designing features for sentence selection process. There are different approaches proposed for extractive-based summarization such as the graph-based, fuzzy approach and others.The following

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 78–89, 2013.

recent surveys cover most of the past, present, and future works of summarization [4,5,6]. In particular there are a number of methods which were built using evolutionary techniques such as Genetic Algorithm [7], Particle SwarmOptimization [8] and Differential Evolution [9].This paper proposed an enhanced method for generating summaries with high diverse sentence topics. The diversity in the summarization concerns with selecting sentences which cover most of document themes and skip falling into the redundancy problem.

To optimize the sentence clustering process, the Differential Evolution (DE) [10] algorithm had been used. This paper proposes enhancement factors to a previous work [9]. There are three main differences between our work and the previous one. First, in this paper we employed the Jaccard-Coefficient similarity measure [11] instead of using the Normalized Google Distance (NGD) [12]. The reason of employing the Jaccard-Coefficient will be explained later. Second, the feature-based approach was used to select a representative sentence from each cluster as opposed to the sentence centroid-based approach which was used in the previous work. The limitation of sentence centroid-based approach is in its disability of capturing the full relationship between a sentence and other sentences in a cluster or a document. Mainly, this problem occurred when the sentence scoring process is performed in an isolated manner from other sentences of the document [13]. Third, in this paper we designed a new real-to-integer modulator to adjust the sentence clustering process. In our proposed method we firstly compute the feature-score vector space, and then the DE is triggered to cluster all sentences using both the Jaccard similarity measure and our new real-to-integer modulator. Then from the optimized cluster, the top feature-scored sentences will be selected to be included in the summary. While in the previous work, the sentences will be clustered using the NGD and genetic mutant operator, and then the centroid sentence (i.e. A sentence that has high intersection with the centroid words) will be selected from each cluster to be represented in the summary. This paper is organized as follows. Section 2 presents the whole proposed methods. The experimental results and discussion are presented in Section 3. Last, section4 concludes the paper.

2 The Methodology

Our methodology part consists of seven Subsections (2.1 till 2.7). The first four subsections concern with configuring the DE algorithm. To achieve the sentence clustering process, a similarity measure is needed; thus, Subsection 2.5 illustrates the two similarity measures used. Subsection 2.6 exhibits the way the selected features are computed. Subsection 2.7 describes the data set and the evaluation measure used.

2.1 The Chromosome Representation

A chromosome representation concerns on how to represent and formulate a specific problem. In this study, the chromosome represents a full document and the

S1	S2	S3	S4	S5	S6	S7	S8	...	S25
C3	C1	C1	C1	C5	C2	C2	C2	...	C5

Fig. 1. A sample of a chromosome representation

number of genes in the chromosome represents the number of sentences of that document in the same order of the genes positions. A gene takes a value between [1, k] where k is the maximum number of required clusters. Figure 1 visualizes the chromosome representation and encoding. Where S# refers to the sentences number and C# is a cluster number. This chromosome is a representation of a document consists of 25 sentences. The figure tells us to place sentence number 4 together with sentence number 2 and 3 into a cluster number 1.

2.2 The Mutation Operation

This subsection discusses both the genetic mutation operator [9] and our new designed real-to-integer modulator.

The Genetic Mutant Operator

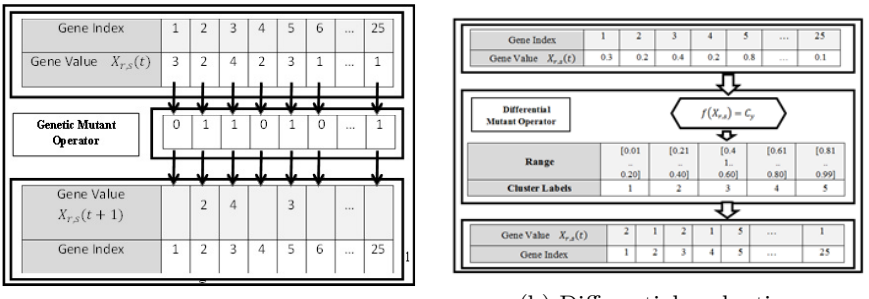
In this subsection, we discuss the mutation operation which was proposed by [9]. Unlike the traditional DE, the previous work [9] initializes the DE population using random integer-based values. The interval of these integer values (genes) ranges between[1,k]. An example of an integer chain of the population is $X_r(t) = [X_{r,1}(t), X_{r,2}(t), \ldots, X_{r,n}(t)]$ where the chromosome $X_{r,s}(t) = (1, 2, \ldots, k, r = 1, 2, \ldots, N)$ and $s = (1, 2, \ldots, n)$; N is the population size, and n is the number of sentences in the document. Consider the following example. A document $D = [S_1, S_2, \ldots, S_{20}]$ consists of 20 sentences of length n = 20, a number of required clusters k = 4, and the population size N = 4. Then a population can be represented as follows: $X_1(t) = [3, 4, \ldots, 1]$, $X_2(t) = [3, 1, \ldots, 2]$, $X_3(t) = [1, 4, \ldots, 2]$, and $X_4(t) = [2, 2, \ldots, 4]$. From the previous representation and the given inputs, chromosome $X_1(t)$ commanded the following: allocate sentence S_1 into cluster C_3, S_2 into C_4, ..., and S_{20} into C_1. The same goes with other chromosomes $X_2(t)$,$X_3(t)$,and $X_4(t)$. To generate the mutant vectors $Y_r(t+1)$, the DE needs to randomly select three individuals $X_{r1}(t)$,$X_{r2}(t)$ and $X_{r3}(t)$ where $r1 \neq r2 \neq r3$. Then computes the difference between $X_{r2}(t)$,and $X_{r3}(t)$, scales it using the control parameter F, sums the scaled result to $X_{r1}(t)$. Finally, composes the trail offspring $Y_r(t+1) = [Y_{r,1}(t+1), Y_{r,2}(t+1), .., Y_{r,n}(t+1)]$. In the previous work[9], the DE control-parameters were not explained; section 2.4 shows how did we assign them. To enable the DE dealing with this explained integer format as in the example above, a genetic mutation operator was adopted according to formula 1:

$$Z_{(r,s)}(t+1) = \begin{cases} 1 & \text{if } rand() < sigm(Y_{(r,s)}(t+1)) \\ 0 & otherwise \end{cases} \quad (1)$$

The adopted vector $Z_s(t+1)$ holds the required changes to move the particle $X_r(t)$ to the new position $X_r(t+1)$. Now there are two probabilities: either a gene on vector $Z_{r,s}(t+1)$ holds 0 or 1. If it includes 1, then copy and move gene $X_r(t)$ to the new position $X_r(t+1)$. Otherwise, the gene will be mutated. The next example shows in steps how a mutation can be occurred. Figure 2.a shows the genetic mutation operation. Consider the following vector $X_r(t)$ needed to move to the new position $X_r(t+1)$. Where:

1. $X_r(t) = [3, 2, 4, 23, 1, 4, 1]$
2. $Y_r(t+1)$:
   (a) $Y_{(r,s)(t+1)} = \begin{cases} X_{(r1,s)}(t) + F\left(X_{(r2,s)}(t) - X_{(r3,s)}(t)\right), & \text{if } rand() < CR \\ X_{(r,s)}(t), & otherwise \end{cases}$
   (b) $Z_{(r,s)}(t+1) = \begin{cases} 1, & \text{if } rand(s) < sigm\left(Y_{(r,s)}(t+1)\right) \\ 0, & otherwise \end{cases}$
3. $X_r(t+1)$:
   (a) If $Z_{(r,s)}(t+1) = 1$, then $X_{(r,s)}(t+1) = X_{(r,s)}(t)$
   (b) If $Z_{(r,s)}(t+1) = 0$, then $Z_{(r,s)}(t+1) = 0 = \{1, 4, 6, 7\}$ // genes addresses not genes values
   (c) $S^+ = \max S = \max\{1, 4, 6, 7\} = 7$
   (d) $S^- = \min S = \min\{1, 4, 6, 7\} = 1$
4. $X_{(r,s^-)}(t+1) = X_{(r,s^+)}(t) : X_{(r,1)}(t+1) = X_{(r,7)}(t) = 4$
5. $X_(r, s^+)(t+1) = X_(r, s^-)(t) : X_(r, 7)(t+1) = X_(r, 1)(t) = 3$
6. $S = \frac{S}{\{S^+, S^-\}} = \frac{\{1,4,6,7\}}{\{1,7\}} = 4, 6$
7. Then, go to step (2), until $S = \frac{S}{\{S^+, S^-\}} = \{\varnothing\}$

Note that $rand()$ is a uniform function responsible of generating random numbers between [0,1]. In addition, the CR parameter value was not declare in [9], thus we assigned as described in Section 2.4.



(a) Genetic Mutation (b) Differential evaluation

Fig. 2. Mutation Operation

The Proposed Real-to-Integer Modulator

The DE is a real population based algorithm. Staying dealing with real values may not consider a suitable solution for handling specific problem (e.g., clustering). In this paper, the need for modulating the real-coded values into discrete values (integer-coded values) is required. Equation 2 is proposed to modulate the real numbers into discrete (integer) numbers (cluster labels). The goal of the proposed modulator is to enable the DE algorithm working in it's principle real values environment, then those values are modulated into discrete integer values to handle the cluster problem assignment. Equation 2 is proposed to define the range interval R_{ch} of the dimension.

$$R_{ch} = \frac{m}{k} \tag{2}$$

Such that R_{ch} is a new proposed interval unit for DEs chromosome, m is the upper value boundary of DE generated values, "k" refers to the required number of clusters. The new interval format (Equation 2) gives the ability of partitioning the chromosome into equivalent intervals and grants for fair random sentences distributions. Suppose the number of required clusters $k = 5$ and the interval range is between $[0, 1]$. Then, the interval range of the five clusters will equal to 0.2. Figure 2b shows the general idea of the proposed real-to-integer modulator.

If, a gene holds a real value ranges in one of following interval: [0.01,0.20] then a cluster label 1 is assigned, or [0.21,0.40] then a cluster label 2 is assigned, or [0.41,0.60] then a cluster label 3 is assigned, or [0.61,0.80] then a cluster label 4 is assigned, or [0.81,1.00] then a cluster label 5 is assigned. Figure 2.b shows a gene (sentence) #1 in the first population $x_{1,1}$ holds the real value 0.3, then our proposed real-to-integer modulator receives this value in a function $f(x_{1,1})$ to assign a suitable cluster (C_2).

2.3 Objective Function

In order to adjust the quality of the partitional clustering, the same objective function that was proposed in[9] is used in this paper. The adapted objective function, which is a combination of two criterion functions, aims to balance both intra-cluster similarity and inter-cluster dissimilarity.

Intra-cluster Similarity: This criterion function is used to adjust the similarity degree between the grouped sentences in the given cluster. The maximum similarity score obtained, the one is much higher required. Equation (3) shows how to compute the intra-cluster similarity.

$$F1 = \sum_{l=1}^{k} |C_l| \sum_{S_i, S_j \epsilon C_l} sim_{NGD}(S_i, S_j) \rightarrow \max \tag{3}$$

where C is a cluster, l is the cluster number, k is the total number of clusters, sim is a similarity, NGD is the current similarity measure selected. S_i and S_j are two sentences currently selected to measure the similarity degree between them and to report the consistency of the cluster itself.

Inter-cluster Dissimilarity: This criterion function is used to adjust the dissimilarity degree between the cluster sentences in the given cluster. The minimum similarity score obtained, the one is much higher required. Equation(4) shows how to compute the inter-cluster dissimilarity.

$$F2 = \sum_{l=1}^{k-1} \frac{1}{|C_l|} \sum_{m=l+1}^{k} \frac{1}{|C_m|} \sum_{S_i \epsilon C_l} \sum_{S_j \epsilon C_m} sim_{NGD}(S_i, S_j) \rightarrow \min \tag{4}$$

where C is a cluster, l and m are the cluster, k is the total number of clusters, sim is similarity, NGD is the current selected similarity measure. "S_i " and "S_j" are two sentences currently selected to measure the dissimilarity degree between the two clusters.

Note that we used the compression ratio as the number of required clusters k. For instance, suppose a document contains 25 sentences and the required compression ratio is 20%, then $k = \frac{25 \times 20}{100} = 5$.

The following objective functions, as shown in Eq.(5), is designed to balance both inter-cluster dissimilarity and intra-cluster similarity.

$$F = (1 + sigm(F_1))^{F_2} \rightarrow max \tag{5}$$

where $sigm(n)$ is a sigmoid function used to bound and issuing differentiable real values within the range $[0, 1]$. The $sigm$ function can be computed as in Eq. (6).

$$sigm(n) = \frac{1}{1 + exp(-n)} \tag{6}$$

Accordingly, the DE algorithm assigns the computed objective function (Eq. 5) for each generated solution (chromosome) as a fitness value.

2.4 Parameters Setup

This section discusses how to assign the DE run-time parameters which are: F, CR, and NP. The F parameter is a scale factor used to adjust the mutation process, and is set to 0.9. The CR is a user input parameter used to increase the diversity of vectors' parameters, and is set to 0.5. The NP is the population size set to 100 chromosomes per generation. All values of these parameters are set according to a previous study concern on adapting DE control parameters [15].

2.5 The Selected Similarity Measures

The following subsections demonstrate the two similarity measures that were used in this study. The first similarity measure is the NGD [12], and the work found in [9] was set as a benchmark for the comparison study. The second similarity measure is the Jaccard coefficient [11]; we proposed to use Jaccard in order improve the process of diversity and to overcome the limitation of the NGD implementation.

The Normalized Google Distance (NGD)

The NGD was used to score similarity between a pair of $(document, query)$ in large database such as the Google database. The similarity between each two concepts is measured by counting the retrieved Google pages when querying them. The NGD similarity is calculated between each pair of terms as in Equation (7):

$$sim_{NGD}(t_i, t_j) = exp(-NGD(t_i, t_j)) \tag{7}$$

Where

$$NGD(t_i, t_j) = \frac{\max\{\log(m_i), \log(m_j)\} - \log(m_{ij})}{\log(n) - \min\{\log(m_i), \log(m_j)\}} \tag{8}$$

m_i is the number of sentences including term t_i, m_{ij} refers to the number of sentences including terms t_i and t_j, and n is the total number of sentences in the document. Finally, Eq.(9) is used to score the similarity between sentence S_k and S_l.

$$sim_{NGD}(S_k, S_l) = \frac{\sum_{t_i \in S_k} \sum_{t_j \in S_l} NGD(t_i, t_j)}{m_i m_j} \tag{9}$$

Jaccard Coefficient Similarity Measure

In this paper, the Jaccard similarity measure is used to improve the cluster performance and compare the results with ones obtained by the NGD. Between each two sample sets, the Jaccard measures the size of the intersection divided by the union of the sample sets. Equation (10) shows the Jaccard coefficient similarity measure, where $sim\_jaccard(S_k, S_l)$ is a similarity measure computed between S_k and S_l.

$$sim\_jaccard(S_k, S_l) = \frac{S_k \cap S_l}{S_k \cup S_l} \tag{10}$$

Table 1. Selected features and their calculations [13]. Where S_i indexes the i^{th} sentence, and t is the total number of sentences on a given document.

Feature	Calculation
Title	$\frac{\text{No. of}(S_i)\text{words matched title words}}{\text{No. of Title's words}}$
Sentence Length	$\frac{\text{No. of words in } S_i}{\text{No. of words in longest sentence}}$
Sentence Position	$\frac{(t-i)}{t}$
Numerical Data	$\frac{\text{No. of numerical data in } S_i}{\text{Sentence Length}}$
Thematic Words(TW)	$\frac{\text{No. of thematic words in } S_i}{\text{Max number of TW found in a sentence}}$

2.6 The Selected Features

One of the main differences of this work is using the feature-score concept to select the top and high relevance sentences to the document topic; unlike[9], the sentence centroid was used to select the relevance sentences from each cluster to represent summary sentences. The problem with this method is scoring the sentence relevance in an isolated way from other sentences. In this way, the method may not able to capture the full advantage of the relationship between a sentence and other sentences in the document or a cluster [13]. Table 1 shows the names and the computational equations of how the selected features were calculated.

2.7 Dataset and Evaluation Measure

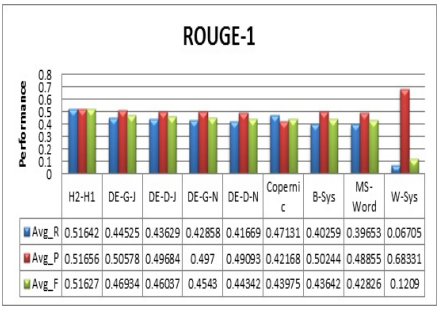
A set of 100 documents was collected from the standard summarization competition DUC 2002 [16]. The DUC submitted each news article document to two human experts to extract model summaries used for automatic system comparison and evaluation. In this experiment we assigned the first human summaries as reference summaries, while we compared the second human summaries against the first one (H2-H1) in order to evaluate the automatic systems performance with the human performance. The documents were preprocessed using the sentence segmentation, tokenization, stop-words removal and words stemming. The ROUGE [14] is a standard automatic evaluation system in summarization. It includes a set of measures used to evaluate both single and multi-document summarization. ROUGE (1, 2, and L) are being used here as they are the suitable measure when evaluate single document summarization [14]. The ROUGE toolkit produces "recall", "precision" and "F" measures for each generated summary; the harmonic mean average is computed for each one of these measures. It's worth mentioning that all evaluation scores extracted using ROUGE are statistically significant at the 95% confidence interval. For result analysis, we used the F-measure to evaluate the methods' performance as it balances both "recall" and "precision" scores.

3 Experiments Description, Results and Discussion

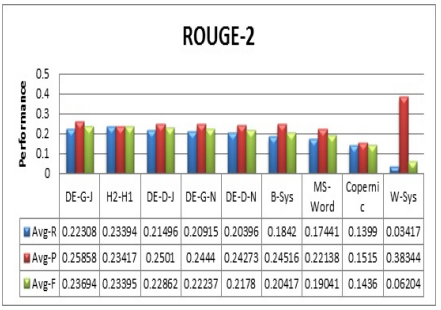
The main goal of this paper is to propose an enhanced cluster-base method to one proposed by [9]. Several issues have been proposed such as the use of feature-based approach consolidated with both a real-to-integer modulator and the use of Jaccard similarity measure. Moreover, the previous genetic mutation operator and the NGD similarity measure were also employed interchangeably with our proposed components in order to bring more competitive results. It is worth stating that, in all experiments the feature-based approach will be used for sentence selection process for each cluster instead of using the sentence centroid-based approach. Our proposed method, is first, computes the feature-scores vector for each sentence in the input document. Then, it groups the sentences together,

separately from feature-score approach, using the Jaccard similarity measure. Afterward, to test the validity of the clustering process, a summary is generated and evaluated using the ROUGE toolkit. Finally, the top scored sentence is selected from each cluster to represent the topic of the cluster.

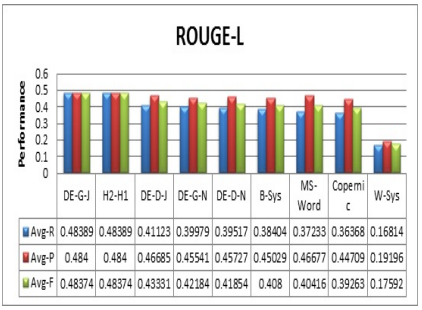
The following name-code (DE-X-Y) is used to mark out the four implemented methods, where the "DE" refers to the backbone algorithm (DE) used to solve the optimization problem, "X" states the selected mutation operator which is either "G" (genetic) or "D" (our proposed real-to-integer modulator), and last "Y" refers to the similarity measure used which is also either "N" (NGD) or "J" (Jaccard). The ROUGE-(1, 2, and L) measures are used to evaluate the system summaries against the human summaries as they proved the suitable measures for single document summarization [14]. All ROUGE evaluation scores are statistically significant at the 95%-Confidence.



(a) ROUGE-1



(b) ROUGE-2



(c) ROUGE-L

Fig. 3. Average Recall, Precision, and F-measure of all methods

Figures 3.a, 3.b, and 3.b show that our proposed implementation of Jaccard similarity measure gives better results than the results obtained using the NGD in all terms of ROUGE-(1, 2, and L) measures. In addition, the genetic mutation operator (DE-G-J) outperformed our proposed operator (DE-D-J) only when using the Jaccard; while ours outperformed the Genetic operator when

Table 2. Similarity scores extracted using NGD and Jaccard

	NGD				Jaccard			
S	1	2	3	4	1	2	3	4
1	1.000	0.645	0.682	0.720	1.000	0.076	0.000	0.047
2	0.6455	1.000	0.689	0.712	0.076	1.000	0.041	0.047
3	0.6827	0.689	1.000	0.782	0.000	0.041	1.000	0.000
4	0.7201	0.712	0.782	1.000	0.047	0.047	0.000	1.000

use the NGD (DE-G-N) is took place. In addition, our proposed modulator with the NGD (DE-D-N) outperformed all other rest of the methods which are the Microsoft-Word Summarizer (MS-Word), Copernic, Best DUC system (B_{Sys}) and Worst DUC System (W_{Sys}). Moreover, our proposed Jaccard similarity based method (DE-G-J) obtained superior result compared to human performance in terms of ROUGE-2 and ROUGE-L.

Carefully consider all ROUGE-1 methods scores; the human summaries are 51% similar to reference summaries. Based on the generalization of the results of ROUGE-1, the (DE-G-J) method generated summaries that are 47% similar to human performance; while the (DE-D-J) generated summaries that are 47% similar to human generated summaries. Also, MS-Word and Copernic generated summaries that are 44% and 43% similar to human generated summaries, respectively.

Our system was designed to display the similarity calculation of the current selected measure. Next we analyse the similarity scores of both measures to investigate the methods performance. Table 2.a and 2.b show the similarity score matrix between four sentences from a selected document using NGD and Jaccard measures. In this table, we only displayed the similarity scores for the first four sentences of the document (DUC-2002, Doc-Number: AP880428-0041) which has a total of 25 sentences. The left part of Table 2 shows that each sentence has high relationships with all the other sentences. Sentence-to-Sentence similarity score ranges between good to excellent mark [0.6, 0.9]. For example, study sentence number 3 and sentence number 4 from the same document:

S_3:"However,doctors learned that long inactivity did more harm than good."
S_4:"Patients got out of shape,developed blood clots and became demoralized."

Its easy for the reader to note that the semantic and character matching similarity scores of both sentences are "nil". After performing the pre-processing steps, the similarity score between those two sentences using NGD (Sim_{NGD} (S_4, S_3)) is 0.7829, which is not correct. This confirms what had been reported by [12]: "employing the NGD in small text corpora may probably lead to imprecise performance". Such scores will affect the sentence clustering performance, which could further influence the generation of a highly diverse summary. The NGD is a similarity measure most appropriate to score documents similarity in large scale database such as Google [12,17]. For example, at a time of writing this paper we

used the Google search engine to retrieve the number of pages contain the terms "Text", "Summarization" and "Text Summarization" which were 4,260,000,000, 13,100,000, and 221,000 pages, respectively. The same couldnt be imagined when dealing with single document summarization in which the document may contain an average of tenth to thirtieth of sentences. The effect is becoming clear in the logarithm normalization computation of the NGD Equation (8). On the other hand, the right part of Table 2 shows the similarity scores between each two sentences of the same document using Jaccard measure. The Jaccard measure presented precise similarity scores between the sentences. It could be observed that the obtained similarity scores are very different than those obtained by NGD. The following similarity score is for the same two previous sentences using Jaccard co-efficient similarity measure —$Sim_{Jaccard}(S_4, S_3)$ = 0.00. From this example we can easily induce how Jaccard could outperformed the NGD. Thus, experimental results proved that the use of the Jaccard similarity measure lead to generate summaries better than the ones generated by the NGD.

4 Summary

In this paper the evolutionary algorithm is introduced to a text summarization problem. The Differential Evolution algorithm is used as text clustering method to optimize the process of sentence clustering. There are two main contributions proposed in this study; the first is the use of Jaccard as an alternative similarity measure to NGD while the second is the new real-to-integer modulator. Since the study focuses on clustering problem, two similarity measures were used and investigated NGD and Jaccard Coefficient. The Jaccard measure outperformed the NGD measure. This study confirmed that, the NGD is unuseful when used for measuring similarity in small text corpora; NGD was successfully implemented in a large data set with high number of web pages (thousands, million, or billion) instead of a small number of sentences (for example a document with 9 sentences). The DUC 2002 data set was used as a test bed. The improper NGD sentence-to-sentence similarity scores led to weak topic diversity coverage compared to Jaccard measure. For comparison purpose, we also evaluated the performance of different summarization methods. The second issue, a real-to-integer modulator was proposed and compared against the genetic operator. Our proposed modulator is computationally simpler than the genetic operator in which we only divide the upper-value range interval by the required number of clusters. Our proposed real-to-integer modulator showed a powerful performance compared to other baseline methods. It can be concluded that, the proper selection of similarity measure plays an important role in determining the quality of the summary and our real-to-integer modulator can be successfully adopted for solving discrete problem optimization.

Acknowledgements. This work is supported by Ministry of Higher Education(MOHE) and Research Management Centre (RMC) at the UniversitiTeknologiMalaysia (UTM) under Research University Grant Category (VOT Q.

J130000.7826 .4F011). Deep thanks to Ameer Tawfik for his help in preparing the manuscript.

References

1. Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. Dev. 2, 159–165 (1958)
2. Baxendale, P.B.: Machine-made index for technical literature: an experiment. IBM J. Res. Dev. 2, 354–361 (1958)
3. Edmundson, H.P.: New Methods in Automatic Extracting. J. ACM 16, 264–285 (1969)
4. Lloret, E., Palomar, M.: Text summarisation in progress: a literature review. Artificial Intelligence Review 37, 1–41 (2012)
5. Nenkova, A., McKeown, K.: A Survey of Text Summarization Techniques Mining Text Data. In: Aggarwal, C.C., Zhai, C. (eds.), pp. 43–76. Springer, US (2012)
6. Saggion, H., Poibeau, T.: Automatic Text Summarization: Past, Present and Future. In: Poibeau, T., et al. (eds.) Multi-source, Multilingual Information Extraction and Summarization, pp. 3–21. Springer, Heidelberg (2013)
7. Kiani, A., Akbarzadeh, M.R.: Automatic Text Summarization Using Hybrid Fuzzy GA-GP. In: 2006 IEEE International Conference on Fuzzy Systems, pp. 977–983 (2006)
8. Binwahlan, M.S., Salim, N., Suanmali, L.: Swarm based features selection for text summarization. International Journal of Computer Science and Network Security IJCSNS 9, 175–179 (2009b)
9. Alguliev, R.M., Aliguliyev, R.M.: Evolutionary Algorithm for Extractive Text Summarization. Intelligent Information Management 1, 128–138 (2009)
10. Storn, R., Price, K.: Differential Evolution A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. J. of Global Optimization 11, 341–359 (1997)
11. Jaccard, P.: Ẽtude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin del la Sociẽtẽ Vaudoise des Sciences Naturelles 37, 547–579 (1901)
12. Cilibrasi, R.L., Vitanyi, P.M.B.: The Google Similarity Distance. IEEE Transactions on Knowledge and Data Engineering 19, 370–383 (2007)
13. Shen, D., et al.: Document summarization using conditional random fields. Presented at the Proceedings of the 20th International Joint Conference on Artifical Intelligence, Hyderabad, India (2007)
14. Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of summaries. Presented at the Proc. ACL Workshop on Text Summarization Branches Out (2004)
15. Rahnamayan, S., et al.: Opposition-Based Differential Evolution. IEEE Transactions on Evolutionary Computation 12, 64–79 (2008)
16. DUC, The Document Understanding Conference (DUC), `http://duc.nist.gov`
17. Wu, L., et al.: Flickr distance. Presented at the Proceedings of the 16th ACM International Conference on Multimedia, Vancouver, British Columbia, Canada (2008)

Hybrid-Learning Based Data Gathering in Wireless Sensor Networks

Mohammad Abdur Razzaque, Ismail Fauzi, and Akhtaruzzaman Adnan

FSKSM, Universiti Teknologi Malaysia, Skudai
JB, Malaysia
marazzaque@utm.my

Abstract. Prediction based data gathering or estimation is a very frequent phenomenon in wireless sensor networks (WSNs). Learning and model update is in the heart of prediction based data gathering. A majority of the existing prediction based data gathering approaches consider centralized and some others use localized and distributed learning and model updates. Our conjecture in this work is that no single learning approach may not be optimal for all the sensors within a WSN, especially in large scale WSNs. For, example for source nodes, which are very close to sink, centralized learning could be better compared to distributed one and vice versa for the further nodes. In this work, we explore the scope of possible hybrid (centralized and distributed) learning scheme for prediction based data gathering in WSNs. Numerical experimentations with two sensor datasets and their results of the proposed scheme, show the potential of hybrid approach.

Keywords: Wireless Sensor Networks, Data Compression, Learning, Collaborative Learning.

1 Introduction

Wireless Sensor Networks (WSNs) are critically resource constrained by limited power supply, memory, processing performance and communication bandwidth [1]. Due to their limited power supply, energy consumption is a key issue in the design of protocols and algorithms for WSNs. Typically, energy consumption is dominated by radio communication [2,3] and its energy consumption is directly proportional to the number of bits of data, i.e., data traffic, transmitted within the network [4]. Therefore, using compression to reduce the number of bits to be transmitted has the potential to drastically reduce communication energy costs and so increase network lifetime.

Statistical model based sensor data predictions (Predictive Coding) or estimations at the sink or base station are promising ways of compressing data and communications in WSNs [5]. In predictive coding (PC) the inherent temporal correlation between consecutive readings at an individual sensor is used to predict future observations at the sink based on the statistical model and recent measurements. For parametric statistical modeling (what most existing PC

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 90–99, 2013.

schemes [6,7,8,9,10] exploit) of signal or phenomenon of interest, it is necessary to learn (know) the statistical parameters related to it (e.g. mean, variance). Existing PC algorithms use either a centralized or distributed learning scheme. In a network, centralized learning is good for nodes closer to sink, while a distributed approach is better for more distant nodes. Hence, single scheme may not be optimal for all the nodes in a WSN. Hybrid of centralized and localized learning schemes can be a better choice. This is why the main objective of this work is to present a hybrid learning scheme for PC in WSNs. Finding the optimal scheme for part of a WSN is not always a trivial task, it may form an optimization problem. PSO (Particle Swarm Optimization) like optimizer can be used in finding the optimal learning scheme for PC. We will present numerical experimentations with two datasets and their results of the proposed algorithm in a WSN scenario.

Section 2 provides a brief overview of predictive coding along with their existing learning schemes. Section 3 presents the proposed hybrid learning scheme. Numerical experimentations of the proposed scheme and their results are presented in section 4. Finally section 6 concludes the work and points to areas of potential future work.

2 Predictive Coding

2.1 Overview of Predictive Coding

In predictive coding (PC) the inherent temporal correlation between consecutive readings at an individual sensor is used to predict future observations at the sink based on the statistical model and recent measurements. Depending on the nature of the sensor data, PC can use parametric modeling or non-parametric modeling. For parametric modeling it is necessary to know (or learn) the statistical parameters, such as mean and variance of the sensor data. On the other hand, non-parametric modeling utilizes regression to represent sensor data, requiring very little prior knowledge about the sensor data. The majority of existing PC schemes [6,7,8,9,10] are based on parametric modeling, where a predictive model is established for every sensor node during a training or learning phase, and the parameters of the model are passed to the sink. Thereafter, nodes only transmit updates to the sink whenever new data arrives or the difference between the model predicted value and the sensed value exceeds a threshold. Thus it reduces the number of communications between source nodes and the sink, providing communication level compression. A typical PC technique consists of the followings:

Statistical Model: The statistical model and its prediction accuracy are the heart of PC [10]. Key models are mainly autoregression based. Autoregressive (AR) models [9] are computationally simple and predict future observations as a weighted sums of previous measurements. Autoregressive Moving Average (ARMA) models [8] use a similar approach but the model is more complex, allowing higher accuracy in some situations, at the cost of greater computational

complexity. Autoregressive Integrated Moving Average models (ARIMA) [11] support modeling non-stationary data as well as stationary data but are even more computationally complex.

Learning Phase: During the learning phase the system determines the parameters of the statistical model, which can be centralized or distributed. In the centralized case [6], all sensor nodes send their readings to the sink, or central node, which determines the parameters of the prediction model and transmits them back to the nodes. In distributed case [8,9], each sensor node calculates their own model parameters and, if necessary, transmits them to the sink.

Model Update: This is done at the sink in one of two ways: (i) Pull: the sink requests updates as they are needed [6], and (ii) Push: the sensor sends an updates as they are needed or become available [7]. In lossless applications, sensors transmit all prediction errors, or residues. These prediction errors replace the raw observations and reduce the amount of transmitted data. In lossy applications, updates are only sent when the prediction error exceeds a pre-defined threshold. Clearly, the lossy approach allows for a greater reduction in the number of communications.

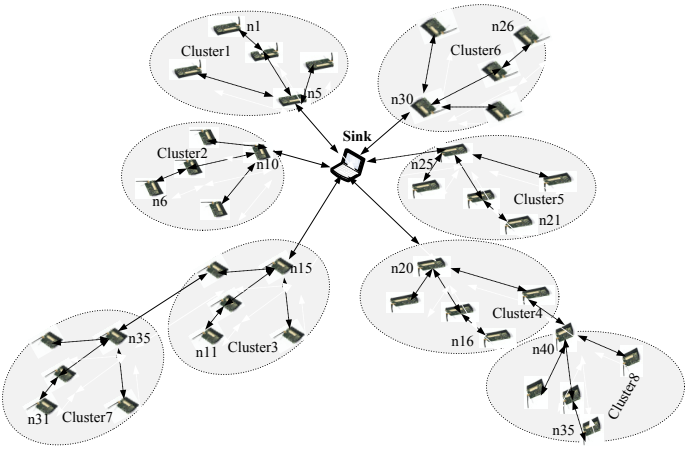


Fig. 1. Clustered Wireless Sensor Networks

2.2 Centralized vs. Localized Learning in PC

Early works [6] on PC exploited centralized learning scheme, where a centralized node a sink or base station collects all the sensors' training or learning dataset and determines the parameters of the prediction model and transmits them back to the nodes. In distributed and localized case [8,9,12], each sensor node or cluster-head locally and collaboratively collects corresponding training dataset and calculates their own model parameters and, if necessary, transmits them to the sink.

To briefly analyze both of the schemes, we exploit the clustered WSN showed in Figure 1. Network consists of $N = 40$ sensor nodes along with a sink and nodes are divided into 8 clusters of equal cluster size 5 (considered for simplicity), and cluster-head selection depends on the clustering scheme. In centralized scheme, all the sensor nodes $n1 - n40$ send training data to the sink and it calculates model parameters, and send them back to every node. On the other hand, in localized (clustered) scheme, every sensor node in each cluster send its training related data to the cluster-head and it calculates model parameters and send them back to each member node, and if needed, send them to the sink as well. For instance, in $Cluster1$ nodes $n1 - n4$ send their reading to node $n5$ and it calculates model parameters, and send them back to nodes $n1 - n4$ and sink. Unfortunately, none of these exploit the distance in their scheme, which can be exploited in selecting an efficient learning scheme for a network or part of a network. For instance, as shown in Figure 1 nodes $n5, n10, n25, n24, n30$ are very close to the sink and in case of localized learning, they can individually send their data to sink (rests follow localized scheme) and sink does the calculation for model parameters, and finally send them back to these node through a limited broadcast. Result of this analysis in normalized 1-hop communication costs is presented in Table 1[1].

Table 1. 1-Hop Communication Cost for Different Learning Schemes

Operations	Centralized	Localized	Hybrid
Training Dataset Gathering	100	40	38
Model Parameters Sending (Nodes)	7(3 Bc)	16 (8Bc)	12
Model Parameters Sending (Sink)	0	12	1 (Bc)
Overall	107	68	51

3 Proposed Hybrid Learning Scheme

Result presented in Table 1 is really inspiring for the hybrid learning approach. Hence we are proposing and formalizing it in this section.

3.1 Motivations

Motivations of the new scheme as follow:

– It is not hard to find real WSN applications, where distances between sink and its 1-hop neighbor sensor nodes are comparable to the distances between sink's 1-hop neighbor sensor nodes and their 1-hop neighbor nodes. If so, these nodes can pick sink as their model calculator. Moreover, in case of node-level temporal signal, each node can work independently and pick sink

[1] In the table Bc means broadcast, which includes Sink broadcast and localized cluster-level broadcast.

as its model calculator without the concern of others or cluster-head. Even, in spatial case, sink's 1-hop neighbor nodes can get model parameters, which is collected from other spatially related nodes and verified with their own data in the sink.
- Impact of the hybrid approach in communication overhead of learning is clear from Table 1. These savings can be significant in large-scale WSNs and environments, where frequent model updates may be needed.
- Typically, sink's 1-hop neighbor sensor nodes are the critical nodes in a WSN. If they work as model estimators, they may die earlier than other nodes as the complexity of executing a model parameter estimation process is $O(m^3 n_{ls})$, where m is the order of the model and n_{ls} the length of the data record [13] or learning samples, which is directly proportional to n. Moreover, these nodes are the bottleneck in WSNs and carry more loads than others. If they can get rid of the model estimation duty, they can help in load balancing in the network and improving network lifetime.

3.2 Overview of the Scheme

Main objective of the hybrid scheme is to take the advantages of centralized and localized scheme. As localized scheme works better for distant node and centralized for closer nodes in a WSN, hence we will consider closer nodes (at least sink's 1-hop neighbor sensor nodes) for the centralized approach and rests for the localized scheme. Selection of scheme will be primarily based on the distance between sink and the corresponding node, and their energy contribution. Objective is to find a combination of these two schemes, which minimizes or optimizes the overall energy cost of learning in PC. Sink will run the selection algorithm and send the selection results to the sensor nodes.

In finding the distance between sink and other sensor nodes, sink needs to know positions of the sensor nodes. We assume that position information of all the sensor nodes are available at sink, which can be done using GPS or other means and this not expensive in static WSN (what most existing applications exploit). We will also assume sink knows energy consumption profile of each sensor node's radio. Based on all these information, we can summarize the hybrid learning scheme as an algorithm shown in algorithm 1, where N is number of sensor nodes in the network, C_n is the communication range of sensor nodes (same for homogenous nodes), n_{s-1h} is the number of sink's 1-hop neighbor sensor nodes, d_{s-n} is the distance between sink and a sensor node, d_{nn-1h} is the distance between sink's 1-hop neighbor sensor node and its 1-hop neighbor nodes, n_{cl} list of sensor nodes, which exploit centralized learning, n_{ll} list of sensor nodes, which exploit localized learning, E_l is the cost of learning, and E_{nn-1h} is the learning cost contributed a node using d_{nn-1h}, and E_{s-1h} is the learning cost contributed a node using d_{s-1h}.

Algorithm 1. Hybrid Learning

```
Var: N, C_n,
for i = 1 to N do
    calculate d_{s-n}
    if d_{s-n} ≤ C_n then
        n_{s-1h} = n_{s-1h} + 1
        calculate d_{nn-1h}
        calculate E_{nn-1h} and E_{s-1h}
        if E_{s-1h} ≤ E_{nn-1h} then
            add to n_{cl}
        else
            add to n_{ll}
        end if
    end if
end for
Send n_{ll} and n_{cl} to sensor nodes.
```

4 Numerical Experiments

In this section we evaluate the effectiveness of hybrid learning scheme in reducing the learning cost, through numerical experiments. For the experimentations, we apply all the learning schemes to one synthetic dataset (one) and one real life sensor dataset (two) and do their numerical analysis. The WSN for dataset one consisted of 40 source nodes (TelosB [14]) and one sink (like Figure 1). For simplicity, a constant hop distance of 3 m was used. The environmental humidity is sampled by every node every 5 minutes. The deployment operated for a month. The total number of samples gathered was 8,640 per node and 345,600 for the whole network. Dataset two is from the Intel Lab Data [15] between February 28th and April 5th,2004 [15]. Mica2Dot [16] sensors with weather boards collected time stamped topology information, along with humidity, temperature, light and voltage values once every 31 seconds. In the span of 38 days, around 2.3 million readings were collected from these sensors. As shown in Figures 2 and 2, sensor readings in the datasets are highly temporally correlated. For dataset two network, there is no information available [15] on clustering, hence we clustered them into 10 clusters and cluster heads are circled (changeable in case of dynamic clustering). We also consider the sink or server is close to clusters 1, 2, 3, 4, 5, and 6.

For the learning phase of PC, we exploit 2 days/week (1 in weekdays and 1 in weekends) data, which means for the dataset one we need 8 days readings (80,640 samples) and for dataset two we need 12 days readings (726315 samples approximately).

The learning costs are calculated based on node characterization and sensor information contained in [14,17,16,18,16,5]. The results approximated for each learning algorithm are given in Table 2. Results presented in Table 2 show the potential of hybrid learning, especially in reducing communication overhead in each round of learning. We did not consider the model estimation (processing)

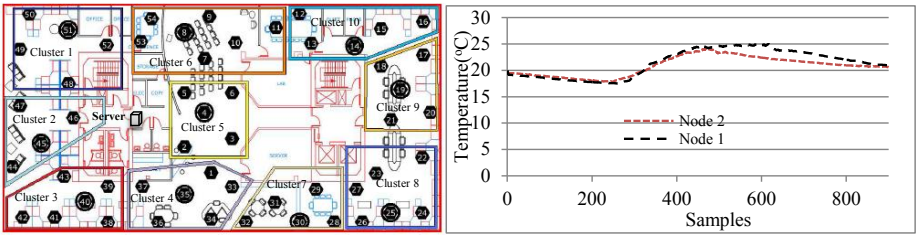


Fig. 2. Network used in dataset and Snapshot of temporal correlation in node 1 and 2

cost, which can be significant in large sample size and this will affect localized and hybrid learning schemes. However, hybrid learning still will be better than localized one. It is clear from Table 2, saving in hybrid scheme depends on the topology (as two datasets with two WSN topologies show different savings), especially on n_{s-1h}, the number of sink's 1-hop neighbor sensor nodes. It also depends on the type of sensor nodes used as they have different energy profile.

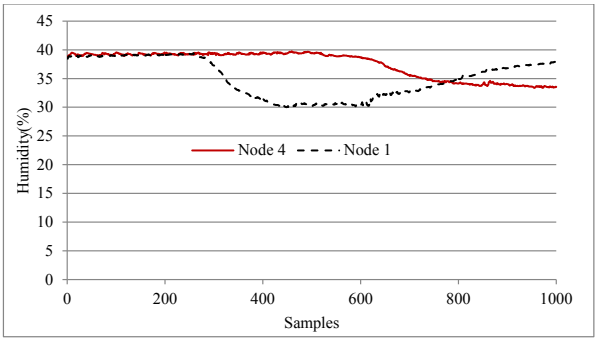


Fig. 3. Snapshot of temporal correlation in node 1 and 4 of the synthetic dataset

Figures 4, 5, 6, and 7 present the predicted outputs based on ARX (AR with external input) model along with their prediction errors. These prediction results are generated in MATLAB, which show 94% and 95.59% matching with the measured data, and errors are bounded within most applications requirement. As the impact of different learning schemes was not significant in predicted results, we have disregarded them here.

5 Related Works

So far we know, there is no work, which considers hybrid of centralized and distributed learning approaches (based on some network parameters) in estimating

Table 2. Numerical Experiments: Costs (communication) for Different Learning Schemes

Operations	Dataset	Centralized(mJ)	Localized(mJ)	Hybrid(mJ)
Training Dataset Gathering	one	2.37	.95	.902
Model Parameters Sending (Nodes)	one	.195	.247	.175
Model Parameters Sending (Sink)	one	0	.15	.02
Overall	one	2.57	1.35	1.07
Training Dataset Gathering	two	30.213	7.83	7.39
Model Parameters Sending (Nodes)	two	3.72	2.5	2.51
Model Parameters Sending (Sink)	two	0	5.4	5.4
Overall	two	33.92	15.74	15.31

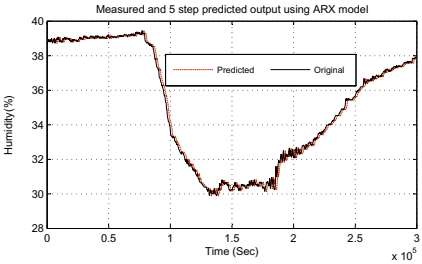


Fig. 4. Model-based data prediction for dataset one

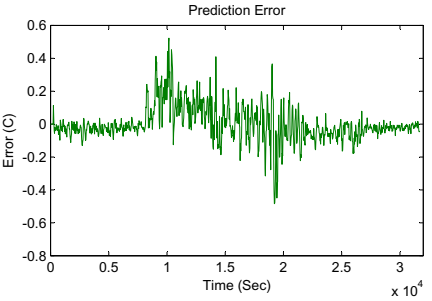


Fig. 5. Prediction error in humidity data

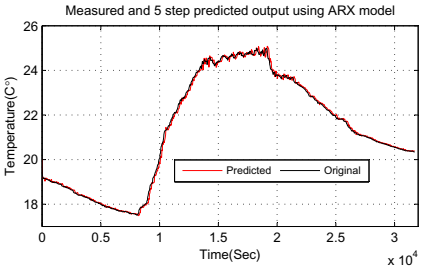


Fig. 6. Model-based data prediction for dataset two

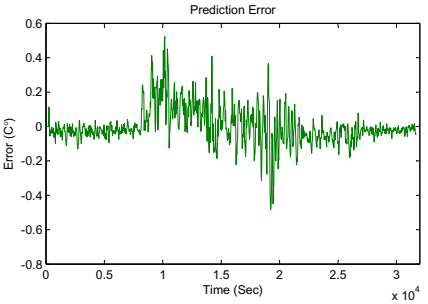


Fig. 7. Prediction error in temperature data

the model. Even works, which explicitly deal with the learning approaches of predictive coding in WSNs are limited [12] in number. Authors in [12] have presented a distributed learning method for nonparametric signal estimation or gathering in WSNs. This approach is suitable where data is sparse or prior knowledge is vague. On the other hand, there are some good works on predictive coding or statistical model based data gathering in WSNs, where authors considered centralized or distributed learning approach for their models. In [6], authors have presented a model based data acquisition in sensor network where

they exploited centralized learning approach. Sensor level autoregressive models building is the main concern in [9]. Based on local readings, nodes locally learn the model to be used in predicting data. Similarly, authors in [8] have exploited distributed learning approach for their model estimation.

6 Conclusion

Statistical model based sensor data prediction (Predictive Coding) or estimation is a key member of compression techniques in WSNs [5]. Parametric statistical modeling (what most existing PC schemes exploit) of signal or phenomenon of interest requires learning (know) the statistical parameters related to it (e.g. mean, variance). Existing parametric statistical modeling based PC algorithms use either a centralized or distributed learning scheme but single scheme may not be optimal for all the nodes in a WSN. Considering this, we have considered a hybrid of centralized and localized learning schemes. Results show the potential of saving communication energy in hybrid learning scheme. Along with the communication energy cost savings, it has the potential to minimize model processing costs compared to the localized learning, and load balancing in the network.

We have disregarded the model estimation (processing) cost in this work, which can be significant in large sample size and large WSNs. Investigation of the model estimation (processing) cost especially in localized and hybrid learning schemes will be one of our future works. This could form an optimization problem, which will need PSO like optimizer. So, integration of PSO or other suitable optimizing algorithm will be our future endeavor.

Acknowledgment. This work is fully supported by UTM, Malaysia under grant number PY/2012/00306.

References

1. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless sensor networks: a survey. Computer Networks 38(4), 393–422 (2002)
2. Pottie, G.J., Kaiser, W.J.: Wireless integrated network sensors. Commun. ACM 43, 51–58 (2000)
3. Barr, K.C., Asanović, K.: Energy-aware lossless data compression. ACM Trans. Comput. Syst. 24(3), 250–291 (2006)
4. Heinzelman, W.R., Ch, A., Balakrishnan, H.: Energy-efficient communication protocol for wireless microsensor networks. In: The Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, pp. 3005–3014 (2000)
5. Razzaque, M.A., Bleakley, C., Dobson, S.: Compression in wireless sensor networks: a survey and comparative evaluation. ACM Trans. on Sensor Networks (2012) (accepted)
6. Deshpande, A., Guestrin, C., Madden, S.R., Hellerstein, J.M., Hong, W.: Model-driven data acquisition in sensor networks. In: VLDB 2004: Proceedings of the Thirtieth International Conference on Very Large Data Bases, VLDB Endowment, pp. 588–599 (2004)

7. Chu, D., Deshpande, A., Hellerstein, J.M., Hong, W.: Approximate data collection in sensor networks using probabilistic models. In: ICDE 2006: Proceedings of the 22nd International Conference on Data Engineering, p. 48. IEEE Computer Society (2006)
8. Lu, J., Valois, F., Dohler, M., Wu, M.Y.: Optimized data aggregation in wsns using adaptive arma. In: International Conference on Sensor Technologies and Applications, pp. 115–120. IEEE Computer Society (2010)
9. Tulone, D., Madden, S.: PAQ: Time Series Forecasting for Approximate Query Answering in Sensor Networks. In: Römer, K., Karl, H., Mattern, F. (eds.) EWSN 2006. LNCS, vol. 3868, pp. 21–37. Springer, Heidelberg (2006)
10. Jun Xiao, J., Ribeiro, A., Quan Luo, Z., et al.: Distributed compression-estimation using wireless sensor networks. IEEE Sginal Processing Magazine 23, 27–41 (2006)
11. Liu, C., Wu., K., Tsao, M.: Energy efficient information collection with the arima model in wireless sensor networks. In: Proc. of IEEE Global Telecommunications Conference, pp. 5–10 (2005)
12. Predd, J., Kulkarni, S., Poor, H.: Distributed learning in wireless sensor networks. IEEE Signal Processing Magazine 23(4), 56–69 (2006)
13. Deng, K., Moore, A.W., Nechyba, M.C.: Learning to recognize time series: Combining arma models with memory-based learning. In: IEEE Int. Symp. on Computational Intelligence in Robotics and Automation, pp. 246–250. IEEE Press (1997)
14. Crossbow: Datasheet of TelosB (2011)
15. Berkeley, U.: Intel lab data (2004), `http://db.csail.mit.edu/labdata/labdata.html`
16. Mica2Dot: Mica2dot datasheet (2004), `http://www.cmt-gmbh.de/Mica2dot.pdf`
17. Morton, G., Venkat, K.: Msp430 competitive benchmarking. Technical report, TI (2006)
18. Sensirion: Datasheet sht1x. Technical report (2010), `http://datasheet.octopart.com/SHT11-Sensirion-datasheet-5323722.pdf`

Orienteering Problem Modeling for Electric Vehicle-Based Tour

Junghoon Lee and Gyung-Leen Park*

Dept. of Computer Science and Statistics
Jeju National University
Jeju-Do, Republic of Korea
{jhlee,glpark}@jejunu.ac.kr

Abstract. This paper presents the design and analyzes the performance of a tour planner for electric vehicles, aiming at overcoming their long charging time by computational intelligence. This service basically finds the maximal subset out of the whole user-selected tour spots and their visiting sequence not inducing waiting time for battery charging. For the schedule search belonging to the orienteering problem category, genetic algorithms are employed. It includes encoding a visiting sequence based on omission probability, defining a fitness function to count the number of visitable destinations, and tailoring genetic operators. For constraint processing, the waiting time estimator prohibits those schedules having non-permissible waiting time to be included in the population. The performance measurement result obtained from a prototype implementation discovers that the proposed service can include 95 % of selected spots in the final schedule on the typical tour scenario for the given inter-destination and stay time distribution.

Keywords: electric vehicle, tour planning, orienteering problem, genetic algorithm, visitable places.

1 Introduction

Along with the increasing attention on smart grid, the future transportation system pursues energy efficiency and greenhouse gas reduction by deploying electric vehicles, or EVs, from now on, to our daily lives [1]. This effort includes not just prompting personal ownership but also providing carsharing services on EVs. The rent-a-car service is a typical business model in carsharing, and EV rent-a-car companies also begin to appear nowadays. However, if renters drive long distance and visit many destinations, EV batteries must be charged en-route. Long charging time and short driving distance are definitely the most critical drawbacks of EVs, and those problems can be alleviated by computational

* Prof. Gyung-Leen Park is the Corresponding author.
This research was financially supported by the Ministry of Knowledge Economy (MKE), Korea Institute for Advancement of Technology (KIAT) through the Inter-ER Cooperation Projects.

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 100–108, 2013.

intelligence such as visit scheduling, until an innovative battery technology is commercialized and its product becomes commonly available [2].

Particularly, the driving distance is more likely to get longer for rent-a-cars in tour places, as tourists rent a car usually on daily basis and visits quite a lot of tour spots starting from their hotels [3]. Some destinations are equipped with EV charging facilities or devices, so the drivers can take their tours, while their EVs are being charged. In this case, EVs can earn driving credit in proportion to the stay time, avoiding the waiting time for EV charging. On the contrary, if the distance between two EV-chargeable destinations is too long to reach with current battery remaining, the renters may suspend their tours and wait until their EVs complete charging electricity enough to reach the next destination. Moreover, if no charging station is available between two destinations, the tourist must rebuild their tour schedule. Anyway, the renters cannot tolerate idle-waiting or inconvenience in tour planning. As a result, intelligent recommendation services for a visiting sequence are not just helpful but necessities in EV-based tours.

For an EV-based tour, the renters select the set of tourist attractions they want to visit and the tour planner or recommender decides the visiting order [4]. The tour planner can implement either optimal schedulers exploiting backtracking-based exhaustive searches or suboptimal schedulers exploiting genetic algorithms or simulated annealing techniques [5]. Basically, it belongs to the classic TSP (Traveling Salesman Problem), which mainly finds the visiting sequence minimizing the driving distance. However, in EV-based tours, it is essential to take into account the charging time in a visiting sequence. That is, a visiting sequence must be excluded in the search space, if it makes the tourists wait for charging. Here, for the set of tour spots the tourists selected, if some of them can be omitted not to violate the restriction posed on tour length and time, this scheduling is equivalent to an orienteering problem [6].

In this regard, this paper designs a tour planner for EVs to enrich the EV rent-a-car business, efficiently overcoming the inconvenience coming from long charging time and short driving distance. The corresponding service decides the visiting order, namely, tour schedule, considering stay time, and inter-destination distance, and tolerable waiting time. It tries to include tour spots as many as possible in the final visiting sequence. According to our simple estimation, the search space complexity is approximated to be $O(2^n \cdot n!)$, where n is the number of tour attractions the tourists submitted to the recommender. $O(2^n)$ accounts for the condition that not all of selected attractions are necessarily included in the final tour schedule, while $O(n!)$ the number of feasible visiting sequences. So, it is necessary to employ genetic algorithms to create a schedule within the reasonable time bound, beyond which the travelers refuses to use the planner.

This paper is organized as follows: After outlining the problem in Section 1, Section 2 introduces related work and background. Section 3 explains the proposed tour planner for EVs in rent-a-car business in detail. Section 4 discusses the performance measurement result, and Section 5 finally concludes this paper with a brief description of future work.

2 Related Work and Background

On the generalization of legacy TSP, each vertex, or destination, is associated with a profit and all vertices are not necessarily visited [7]. Its tour scheduler pursues maximizing the profit while reducing the travel cost. For a variety of such TSP applications, the classification depends on the way the two objectives are addressed. The first class defines an object function combining both of them. Second, the object function takes into account only profit maximization while the travel cost works as a constraint, that is, the solutions violating the cost constraint is excluded from the search space. The third class oppositely takes the profit as a constraint and its object function calculates the travel cost only. All of them are belonging to the NP problem, and greatly dependent on the number of destinations. The second class is called the orienteering problem or interchangeably, the selective TSP.

For the orienteering problem, [6] develops a genetic algorithm where a chromosome is a sequence of visited vertices. This approach allows non-feasible solutions, which violate the cost constraint, to be accepted in the population for better diversity, but with a penalty estimated by the distance from feasibility. In each chromosome, vertices are listed in the visiting order from the start to end points. For initial population, a list of n vertices is generated in random order. A vertex will be removed with the omission probability, making the corresponding entry zero in the list. With this initialization, regular genetic operators such as crossovers and mutations are applied. However, in each offspring, duplicated vertices will be replaced by disappearing ones. Its encoding scheme is comprehensive and robust, so our work will partly employ it, but with a different constraint on waiting time estimation.

As for the integration of charging and routing for EVs, [8] proposes a routing service for searching and reserving charging stations to reach the given destination. A query and routing protocol is defined between an EV and the routing service, through which EV drivers check the availability of charging stations and reserve a time slot. For the routing service, a broker selects the best charging station for each EV considering resource availability, location convenience, price signal change, and the like. In each charging station, controlled charging tries to reduce the peak load stemmed from multiple charging tasks concentrated in a small time period. Meanwhile, for scheduling of charging intervals, they build an integer linear formulation and subsequent local ratio heuristics. Even if this work addresses a charging-integrated routing mechanism to match energy supply and demand in EV penetration scenarios, it just focuses on the drive within urban area having sufficient charging facilities.

In the mean time, our research team has been developing an intelligent tour planning service for EV rent-a-car business, mainly targeting at Jeju City, which is an internationally famous tour place hosting one of world's largest smart grid testbeds. First, [9] develops a waiting time estimation model for a specific tour schedule according to the battery amount, driving distance, and stay time. Then, a backtracking-based exhaustive search finds the optimal schedule minimizing the waiting time, even with a driving distance loss. Its main motivation is to

make the overlap of *stay-and-charge* as much as possible. Next, some constraint processing is integrated to meet the user-issued requirements, for example, the tourists want to have supper at one of preferred restaurants around 6 PM [10]. Moreover, the search technique includes the simulated annealing scheme for a timely schedule generation even for the large number of destination.

3 Tour and Charging Scheduler

Among various suboptimization schemes, genetic algorithms are known to be an efficient search technique inspired by principles of natural selection and genetics [5]. They build an initial population and then iterate genetic operations either until an acceptable solution is found or by the given time bound. Genetic operators include selection, crossover, and mutation. For the given set of selected destinations, our scheduler finds the best schedule which has the maximum number of visitable tour spots and brings no charging delay during the tour. In applying genetic algorithms to our tour scheduler, it is necessary to encode a visiting sequence, define a fitness function, and tailor genetic operators. First, each schedule, or visiting sequence, is encoded into an integer-valued vector called a chromosome. Second, the fitness function counts the number of visitable tour spots in a schedule. Here, constraint processing is needed to eliminate a schedule which brings charging delay beyond the given bound. Third, crossover operations replace the duplicated tour points with disappearing ones.

Figure 1 illustrates the overall procedure of our recommendation service for EV-based tours. Given the set of selected tour points, the genetic scheduler begins with initial population having a predefined number of schedules. Large population can accommodate more diversity of chromosomes, but brings much computation overhead. Between two conflicting performance parameters, we experimentally set the number of schedules in population to 96. In orienteering problems, omission probability decides whether a selected tour point is included in the schedule or not. The omitted places are denoted by -1. Then, the actual visiting sequence can be obtained by removing -1 in the schedule. Hence, $(A, B, -1, C)$ and $(A, -1, B, C)$ are equivalent and considered to be duplicated. However, as they can mate and generate a completely different schedule, their co-existence in the population is allowed.

From the initial population, each evolutionary step creates the next-generation population consisting of better schedules, or visiting sequences. It mainly selects the best solution in a population by a fitness function and mates them to form the next generation. Out of existing and newly created schedules, only the fittest survive in the population, so the fitness value of the population gets better generation by generation. The genetic loop iterates selection by fitness, reproduction by crossover, and mutation within a chromosome. Selection is a method that picks parents by the fitness function. Following a kind of the tournament strategy, 5 random chromosomes are selected in the mating pool and the best one is selected. The reproduction procedure exchanges substrings of two mated chromosomes to generate offsprings. Some tour points appear more than once, while

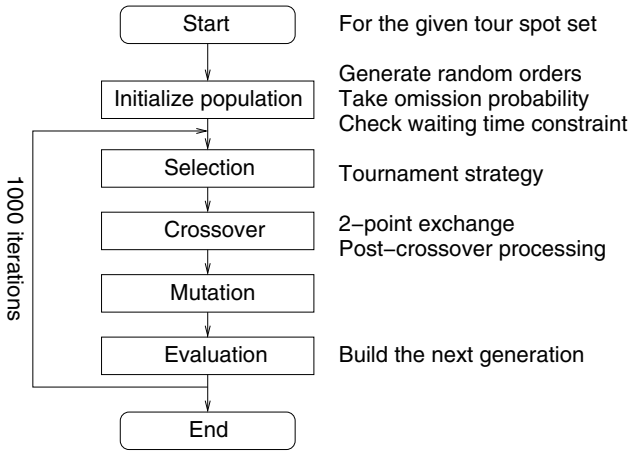


Fig. 1. Recommender service operation

others have disappeared in offsprings. So, for each crossover, the duplication is checked and replaced by disappearing one or -1.

To check the constraint on the permissible waiting time bound, each generation of a new chromosome evaluates the waiting time and eliminates if the schedule violates the constraint. Waiting time is the time interval the tourists just wait for their EVs to be charged without doing anything. The waiting time estimator follows the visiting sequence from the starting point, while the initial distance credit is set to the driving distance of a fully charged battery. Then, for each move, say, from A to B, if the distance credit is less than the distance between A and B, the waiting time is induced. Otherwise, the distance credit is reduced by the distance. In addition, it increases according to the stay time at B, but limited by the maximum battery capacity. In our design, a new schedule may make its waiting time beyond the given bound. It is experimentally found that such a schedule had better be eliminated from the population. Even though this restriction may narrow the search scope and stick to a local minimum, it helps the genetic loop to converge quickly. Otherwise, the search procedure may spend too much time in processing invalid chromosomes.

4 Performance Measurement

This section implements a prototype of the proposed recommendation service for EV-based tours using Visual C++ 6.0 to assess its performance. The experiment makes it run on the platform equipped with Intel Core2 Duo CPU and 3.0 GB memory as well as installing Windows Vista operating system. As for the genetic operation parameters of our implementation, the number of chromosomes in each population is set to 96, and the number of iterations is set to 1,000. Such a parameter selection is decided to generate a tour schedule within 1 second. The

performance parameters include the number of user-selected destinations, inter-destination distance, stay time, omission probability, and permissible waiting time, while the main performance metric is the number of visitiable destinations. Here, every tour spot is assumed to have charging facilities. Actually, a tour spot with no chargers has the effect of extending the distance between two chargeable spots. For each parameter setting, 30 destination sets are generated, and each waiting time is averaged.

The first experiment measures the number of visitable places according to the number of user-selected places ranging from 7 to 15. The inter-destination distance exponentially distributes with the given average. Figure 2 plots three curves for average inter-destination distances of 10, 15, and 20 km, respectively. Here, average stay time also distributes exponentially with the average of 30 min, omission probability is set to 0.6, and permissible waiting time to 0 min. If the total driving distance is less than the battery capacity, all of user-selected points can be included in the final schedule, namely, visitable. In this experiment, maximum battery capacity enables an EV to drive 90 km, and the EV is fully charged overnight. Moreover, the overlapped *charge-and-stay* at each destination can further increase the number of visitable destinations in tour recommendation. If the inter-destination distance is 10 and 15 km, almost every selected place can be visited. On the contrary, for the 20 km case, only 10.3 can be visited when the tourists select 15 destinations.

The next experiment measures the effect of average stay time for the number of selected destinations also ranging from 7 to 15. Average inter-destination distance is set to 15 km, while other parameters are set equally as in the previous experiment. The total driving distance linearly increases according to the increase in the number of selected destinations. For each destination, the distance credit, namely, the distance an EV can drive with current battery remaining, increases in proportion to the stay time. 1 hour charging earns the distance credit by 15 km. Hence, if the stay time is 30 or 45 min, almost every place can be included in the final schedule with an appropriate visiting schedule. As contrast, when stay time is 15 min, up to 30 % destinations need to be removed from the final schedule.

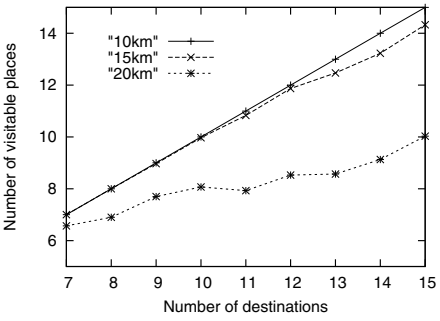


Fig. 2. Inter-destinations distance effect

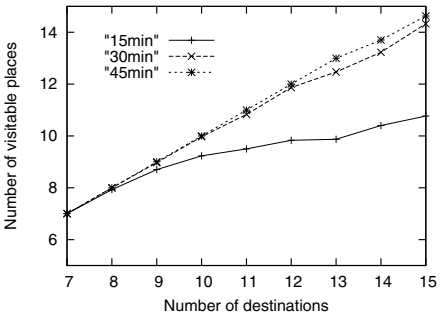


Fig. 3. Stay time effect

Figure 4 and Figure 5 measure the effect of permissible waiting time. First, Figure 4 plots the number of visitable places for the 3 cases when the tolerable bound is 0, 30, and 60 *min*. Here, average inter-destination distance and average stay time are set to 15 *km* and 20 *min*, respectively. Obviously, if tourists permit more waiting time, the number of visitable destinations increases. Until the number of selected spots becomes 10, they are rarely removed from the final schedule, as the total distance is not so long and can be covered by *stay-and-charge*. From this point, waiting time tolerance begins to invite more destinations to the schedule. As the impact of sequence rearrangement is not dependent on the number of destinations if it is sufficiently large, the performance gap between two different waiting time bounds is decided by average stay time.

Next, Figure 5 plots the actual waiting time for the 3 permissible tolerable bounds. For 15 destinations, 3 curves mark 0, 21, and 48 *min*, respectively, just like in Figure 4. If this tolerable bound is 0 *min*, actual waiting time is also 0 *min* for the whole range, as a schedule cannot be accepted if its waiting time is larger than 0 *min*. When the number of destinations is 13 and more, the actual waiting time gets stable just with an insignificant fluctuation. The scheduler is more efficient when this actual waiting time gets closer to the tolerable bound. According to Figure 5, the actual-to-permissible waiting time is 0.80 for the 60 *min* case, while that for the 30 *min* case is 0.71. Our scheme works more efficiently when the tolerable bound gets larger. Judging from the experiment result shown in Figure 4 and Figure 5, the proposed scheduler doesn't have anomalous behavior, for example, failing in finding an acceptable schedule on the overloaded condition.

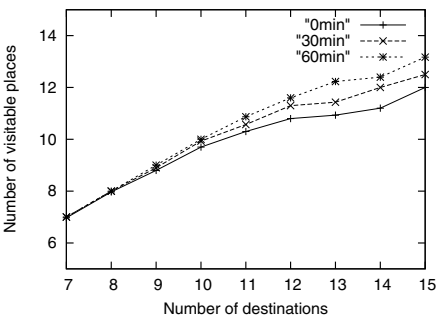


Fig. 4. Tolerable waiting time effect

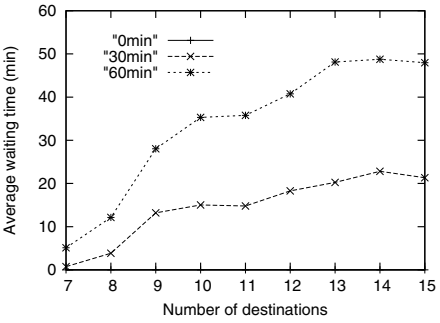


Fig. 5. Actual waiting time measurement

Figure 6 plots the effect of omission probability in generating a new schedule either in building initial population or replacing duplicated chromosomes. We plot 2 curves for 15 and 20 destination cases. Here, stay time is set to 30 *min*. For the 15 destination case, around 94.4 % of user-selected destinations remain in the final schedule, while 86.3 % $\sim$ 88.6 % remains in the 20 destination case. Actually, the omission probability closer to this ratio is more likely to find a

better schedule, as more schedules will be generated around the final answer. However, the experiment result shows that the number of visitable places is not so much affected by the omission probability. The misselection of this parameter just results in time extension in building initial population.

The last experiment measures the effect of the tolerable bound ranging from 0 to 60 *min* with the number of user-selected destinations fixed to 15 and 20, respectively. Stay time is set to 20 *min* for the consistency with Figure 4. As shown in Figure 7, for the case of 15 destinations, 12 spots can be visited when the tolerance waiting time bound is 0 *min*, and 13.17 spots when the bound is 60 *min*. The number of visitable places linearly increases according to the increase of the tolerable bound. This behavior is the same for the 20 destination case. 14.07 and 15.43 spots can be included in the final schedule when the tolerable bounds are 0 and 60 *min*, respectively. As two curves change with the same ratio, the gap in the number of visitable places keeps constant. This result confirms that the number of visitable places is mainly affected by stay time.

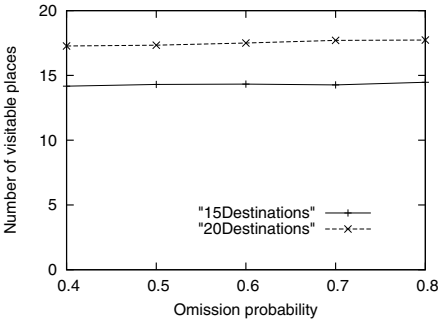


Fig. 6. Omission probability effect

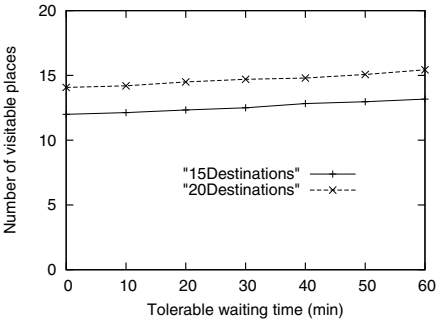


Fig. 7. Refined tolerance bound analysis

5 Conclusions

In this paper, we have designed a tour planner for electric vehicles by mapping the scheduling process to an orienteering problem model. The developed service can alleviate the long charging time and short driving distance of EVs, which are expected to replace gasoline-powered vehicles for smart and clean transportation. In the proposed service scenario, the tourists select destinations they want to visit and submit to the recommender, possibly through global connection such as the Internet, cellular networks, and vehicular communication mechanisms. This service basically finds the visitable destinations out of user-selected ones and their visiting sequence not inducing waiting time for EV charging. If users don't satisfy with the generated schedule, they will reselect the tour points and submit them to the recommender again. Another, possibly interactive, recommendation for the tour points will be also a very useful service for EV rent-a-car users.

For the sake of taking advantage of computational intelligence, genetic algorithms are employed by encoding a visiting sequence based on omission probability, defining a fitness function to count the number of visitable destinations, and tailoring genetic operators for duplicated entry replacement. During constraint processing, the waiting time estimator prohibits those schedules having non-permissible waiting time to be included in the population. The performance measurement result obtained from a prototype implementation discovers that 95 % of selected spots can be included in the final schedule with non-waiting visiting order for the typical tour scenario having less than 9 destinations for the given inter-destination and stay time distribution.

As future work, we are planning to develop a tour scheduler capable of combining a reservation mechanism for charging stations. According to our observation, different tours have many destinations in common. Hence, the recommender system is highly likely to assign similar visiting sequences to them. Then, in some tour spots, chargers cannot be available to some EVs. So, the distribution of charging requests over the shared route is needed.

References

1. Hermans, Y., Le Cun, B., Bui, A.: Individual Decisions and Schedule Planner in a Vehicle-to-Grid Context. In: International Electric Vehicle Conference (2012)
2. Kobayashi, Y., Kiyama, N., Aoshima, H., Kashiyama, M.: A Route Search Method for Electric Vehicles in Consideration of Range and Locations of Charging Stations. In: IEEE Intelligent Vehicles Symposium, pp. 920–925 (2011)
3. Garcia, A., Arbelaitz, O., Linaza, M., Vansteenwegen, P., Souffriau, W.: Personalized Tourist Route Generation. In: 10th International Conference on Web Engineering, pp. 486–497 (2010)
4. Ferreira, J., Pereira, P., Filipe, P., Afonso, J.: Recommender System for Drivers of Electric Vehicles. In: Proc. International Conference on Electronic Computer Technology, pp. 244–248 (2011)
5. Giardini, G., Kalmar-Nagy, T.: Genetic Algorithm for Combinational Path Planning: The Subtour Problem. Mathematical Problems in Engineering (2011)
6. Tasgetiren, M., Smith, A.: A Genetic Algorithm for the Orienteering Problem. In: Proc. Congress on Evolutionary Computing, pp. 1190–1195 (2000)
7. Feillet, D., Dejax, P., Gendreau, M.: Traveling Saleman Problems with Profits. Transportation Science 39, 188–205 (2005)
8. Bessler, S., Grønbæk, J.: Routing EV Users towards an Optimal Charging Plan. In: International Battery, Hybrid and Fuel Cell Electric Vehicle Symposium (2012)
9. Lee, J., Kim, H., Park, G.: Integration of Battery Charging to Tour Schedule Generation for an EV-Based Rent-a-Car Business. In: Tan, Y., Shi, Y., Ji, Z. (eds.) ICSI 2012, Part II. LNCS, vol. 7332, pp. 399–406. Springer, Heidelberg (2012)
10. Lee, J., Kim, H., Park, G., Lee, B., Lee, S., Im, D.: Tour Schedule Generation Integrating Restaurant Options for Electric Vehicles. To appear at International Conference on Ubiquitous Information Technologies and Applications (2012)

Integrating Social Information into Collaborative Filtering for Celebrities Recommendation

Qingwen Liu, Yan Xiong, and Wenchao Huang

School of Computer Science of University of Science and Technology of China
mrhead@mail.ustc.edu.cn, {yxiong,huangwc}@ustc.edu.cn

Abstract. With the exponential growth of users' population and volumes of content in micro-blog web sites, people suffer from information overload problem more and more seriously. Recommendation system is an effective way to address this issue. In this paper, we studied celebrities recommendation in micro-blog services to better guide users to follow celebrities according to their interests. First we improved the jaccard similarity measure by significant weighting to enhance neighbor selection in collaborative filtering. Second, we integrated users' social information into the similarity model to ease the cold start problem. Third we increased the density of the rating matrix by predicting the missing ratings to ease the data sparsity problem. Experiment results show that our algorithm improves the recommendation quality significantly.

Keywords: micro-blog, collaborative filtering, user similarity model, data sparsity.

1 Introduction

As the age of Web 2.0 comes, social media becomes more and more popular. Recently, the micro-blog web sites have shown a great charm, with millions of users joining in it. The micro-blog web sites fundamentally provide a public platform for their users to seek and share information, to communicate with others, and to build online friendships. It can be seen as a hybrid of email, instant messaging and news broadcasting systems. Unlike other social networks like Facebook or LinkedIn, the following relationship between users in micro-blog system is not necessarily reciprocal. For this reason, people can follow anyone they like without requiring acceptance. Building friendship in this way lowers down the cost of expanding one's network and allows some users to be followed by many users without following many themselves, effectively becoming celebrities or stars [1].

In the view of the exponential growth of micro-blog user population and volumes of content generated by them, it gets difficult for users to choose whom to follow and what to read. Users may easily be flooded with information streams. Personalized recommendation is an important way to address this issue and it has been well studied by both academia and industry recently. In this paper, we will study the problem of recommending celebrities or stars in micro-blog

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 109–118, 2013.

services. We are motivated by the rich social information to provide potential evidences for users' similarity computation and missing ratings prediction in collaborative filtering (CF). And we prosed a novel collaborative filtering framework which improve jaccard similarity measure by significant weighting and integrate social information to ease the data sparsity problem and enhance user similarity modeling.

The rest of the paper is organized as follows. Section 2 covers related works on collaborative filtering and social recommending systems. We describe a novel approach which integrates social information into collaborative filtering to address the cold-start problem and the data sparsity problem in Sec. 3. Evaluation metric and experiment results are demonstrated in Sec. 4. Finally, we conclude in Sec. 5.

2 Related Work

We will review related works from 2 different research areas: CF algorithms and the role of social features played in recommendation systems.

The fundamental assumption of CF is that if two users have rated some items similarly, or they have similar behaviors (e.g. watching, buying, listening), and hence they will rate or act on other items similarly [2]. One of the biggest challenges in CF is the data sparsity problem, which leads to the failure of finding similar users or items. The density of available ratings in commercial recommending systems is often less than 1% [3] and the density of our data set is 0.64%. Many algorithms have been proposed to overcome the data sparsity problem. In [4], a dimensionality reduction technique, Singular Value Decomposition (SVD), is employed to remove unrepresentative or insignificant users or items and map the rating space into a lower dimensional semantic space. However, some information about users or items may be discarded by SVD, thus resulting a decrease in the recommendation quality. P.Melville et al. proposed a hybrid model named content-boosted CF to address the data sparsity problem, in which external content information was used to produce predictions for new users and new items [5]. The result of this method is promising, in our paper rich social information is extracted from micro-blog web sites to enhance collaborative filtering. H.Ma et al. increased the density of the rating matrix by predicting the missing ratings using a user-based and item-based combined model [6]. No using of external information in this method will limit the recommendation quality, thus we propose to integrate the social information to address this issue.

To understand micro-blog usage, Akshay et al. showed how users with similar intentions connect with one another by analyzing the user intentions associated at a community level in [7]. The findings motivated us to use social information to discover similar users for CF based recommendation. In [8], different content-based recommending systems are built by using different types of social information to recommend URLs extracted from micro-blog content. Ido Guy et al. measured user similarity based on social features from two aspects: users' social network structure and users' content information [9]. The results of these two

papers both show that adding social features into traditional recommendation algorithms can significantly improve accuracy. Daly systematically studied how to measure the network effects of recommending social connections and how different social recommending algorithms differ [11]. His findings guide us to choose appropriate types of social information for celebrities recommendation. Chen et al. claimed in [10] that content information is more effective than other kinds of social information in people recommendation. Finally, Hannon evaluated a range of different user profiling and recommendation strategies in [12]. It found that a mixture profiling strategy which use both contents and social connections can produce better received recommendations.

3 Recommender System Description

3.1 Problem Definition

Formally, we will formulate our problems as follows. Let $\mathcal{U}$ be a user set and let $\mathcal{C}$ be a celebrity set in micro-blog web sites. The following relationships between users and celebrities are denoted by a $|\mathcal{U}| \times |\mathcal{C}|$ matrix, called user-item rating matrix. Every entry $r_{ui} \in \{-1, 0, 1\}$ represents the value that user $u \in \mathcal{U}$ rated item $i \in \mathcal{C}$ where 1 means user u followed item i, -1 means user u refused to follow item i and 0 means the user has not rate the item yet. Given a user u and an item i, let $\mathrm{P}(u, i)$ be a recommending function that measures the preference of user u on item i, i.e. $\mathrm{P} \in \{\mathcal{U} \times \mathcal{C} \to \mathbb{R}\}$. Then given a user u and an item list $\mathcal{L}$, we will rank the items in $\mathcal{L}$ according to $\mathrm{P}(u, i)$ and select top N items as the recommending items for u. More formally:

$$\forall u \in \mathcal{U} \quad \mathcal{S}_u = \underset{i \in \mathcal{L}}{\arg\mathrm{TopN}} \quad \mathrm{P}(u, i) \tag{1}$$

where $\mathcal{S}_u$ is the recommendation result.

3.2 Significant Weighting for Jaccard Similarity Measure

User similarity computation is a critical step for collaborative filtering algorithms. We claim that jaccard similarity is a more natural way to model similarity between two binary rating vectors than other similarity measures by its definition. In our problem, we define jaccard similarity of two rating vectors as (2).

$$sim(v_1, v_2) = \frac{\sum_{i=1}^{n} 1\{r_{1i} = r_{2i} \wedge r_{1i} \neq 0\}}{\sum_{i=1}^{n} 1\{r_{1i} \neq 0 \wedge r_{2i} \neq 0\}} \tag{2}$$

where $1\{*\}$ is an indicator function: $\{true, false\} \to \{1, 0\}$ and r_{1i} or r_{2i} is the ith rating of vector v_1 or v_2.

This definition has two disadvantages. First, positive ratings are much more informative than negative ratings in our problem, but (2) treats both identically.

Second, (2) will overestimate the similarity of users who happen to rate quite a few items identically but who may not have similar overall preference. The estimation is not reliable since too few co-ratings have no statistical significance. To address the first problem, we give different weights to positive ratings and negative ratings by using the following equation:

$$\mathrm{sim}'(v_1, v_2) = \frac{\sum_{i=1}^{n} 1\{r_{1i} = r_{2i} = 1\} + \lambda \sum_{i=1}^{n} 1\{r_{1i} = r_{2i} = -1\}}{\sum_{i=1}^{n} 1\{r_{1i} \neq 0 \wedge r_{2i} \neq 0\}} \tag{3}$$

where λ is a parameter between 0 and 1. To address the second problem, we follow the intuition that computing without enough supporting evidence (co-ratings) should be punished. Thus, a penalty function was introduced by (4).

$$\mathrm{pun}(v_1, v_2) = \frac{\min(\sum_{i=1}^{n} 1\{r_{1i} \neq 0 \wedge r_{2i} \neq 0\}, \phi)}{\phi} \tag{4}$$

where $\sum_{i=1}^{n} 1\{r_{1i} \neq 0 \wedge r_{2i} \neq 0\}$ means the number of items rated in common and ϕ is a threshold which is greater than 1. By applying this penalty function, we get the new similarity measure in (5).

$$\mathrm{jaccard\_sim}(v_1, v_2) = \mathrm{sim}'(v_1, v_2) \times \mathrm{pun}(v_1, v_2) \tag{5}$$

Equation (5) will devalue the similarity of v_1,v_2 if the number of co-rated items are smaller than ϕ and give different weights to positive and negative ratings.

3.3 Social Information Integrating for Neighbor Selection

Given a user, his/her neighbor set is composed of two parts. One is computed by jaccard similarity based on the rating matrix and the other is computed by social information. In micro-blog web sites, the social information can be classified into 3 types, which are content of posts, social connections and social activities. Accordingly, we will model users' social similarity from three aspects by (6).

$$\mathrm{social\_sim}(u, v) = \alpha * \mathrm{sim_c}(u, v) + \beta * \mathrm{sim_n}(u, v) + \gamma * \mathrm{sim_a}(u, v) \tag{6}$$

where α, β, γ are three parameters which determine the weights of different types of similarity. Following the approach in [8], we build a profile vector for each user from the words that were included in their posts. Each entry of the profile vector is weighted by the term-frequency inverse-user-frequency (TF-IUF) of the corresponding word. $\mathrm{sim_c}(u, v)$ is then computed as the cosine similarity between their profile vectors.

$$\mathrm{sim_c}(u, v) = \frac{\sum_{i \in W} w_{ui} w_{vi}}{\sqrt{\sum_{i \in W} {w_{ui}}^2} \sqrt{\sum_{i \in W} {w_{vi}}^2}} \tag{7}$$

where W is the set of words which are extracted from users' posts and w_{ui} is the weight of the ith word of user u.

In practice, if a user u follows another user v in micro-blog web sites, user u may be interested in user v as an information seeker or they might be friends in the real world. Motivated by this, we model the social connection similarity by a binary function as (8).

$$\mathrm{sim_n}(u,v) = \begin{cases} 1 & \text{u follows v} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Users have three types of social activities, mention, repost and comment. Given a user u, let $\mathcal{A}_u$ represent the set of users who has ever been interact with u in micro-blog web sites. Intuitively, more activities imply more intimate relationship and more intimate relationship implies more similar interests between users. According to this, we model the action similarity by (9).

$$\mathrm{sim_a}(u,v) = \frac{\#\mathrm{men}_{uv} + \#\mathrm{rep}_{uv} + \#\mathrm{com}_{uv}}{\max\limits_{v' \in \mathcal{A}_u} (\#\mathrm{men}_{uv'} + \#\mathrm{rep}_{uv'} + \#\mathrm{com}_{uv'})} \tag{9}$$

where $\#\mathrm{men}_{uv}$ is the number of times that user u mentioned user v, $\#\mathrm{rep}_{uv}$, $\#\mathrm{com}_{uv}$ is the number of times that user u reposted and commented on user v's posts.

Once the user similarity is modeled, we can select the neighbor set for users. H.Ma argued in [6] that the top-N neighbor selection method is misleading when a user actually has few neighbors and that selecting the ones whose similarity is greater than some threshold as neighbors results in more accurate recommendations. In our problem, for every user u, we generate two neighbor sets of u according to (10)(11).

$$\mathrm{T}_u = \{v | v \in \mathcal{U} \wedge \mathrm{jaccard\_sim}(u,v) > \varphi\} \tag{10}$$

$$\mathrm{S}_u = \{v | v \in \mathcal{U} \wedge \mathrm{social\_sim}(u,v) > 0\} \tag{11}$$

where S_u the neighbor set that is computed based on social information, T_u is the neighbor set that is computed based on the rating matrix and φ is a threshold between 0 and 1.

Now we have demonstrated the two methods for neighbor selection. To take advantage of both methods, we first make predictions using T_u and S_u respectively, and then combines the predictions linearly by (12).

$$\begin{aligned} \mathrm{P}(u,i) = \bar{r_u} + \theta \cdot \frac{\sum\limits_{v \in \mathrm{S}_u} (r_{vi} - \bar{r_v}) \cdot \mathrm{social\_sim}(u,v)}{\sum\limits_{v \in \mathrm{S}_u} \mathrm{social\_sim}(u,v)} + \\ (1-\theta) \cdot \frac{\sum\limits_{v \in \mathrm{T}_u} (r_{vi} - \bar{r_v}) \cdot \mathrm{jaccard\_sim}(u,v)}{\sum\limits_{v \in \mathrm{T}_u} \mathrm{jaccard\_sim}(u,v)} \end{aligned} \tag{12}$$

3.4 Effective Missing Ratings Prediction

Addressing the data sparsity problem is one of the most critical issues in collaborative filtering. A lot of methods have been proposed to deal with this problem as mentioned in Sec. 2. Missing ratings prediction is an intuitive, simple and effective way to increase the density of the rating matrix. A model which chooses to predict the missing ratings or not according to confidence is proposed by H.Ma in [6]. Significant improvement has been seen in this model. However, it only iterates the original rating matrix to produce a denser one without accessing external information, which will limit the predicting quality. To this end, we integrate social information to provide evidence for missing ratings prediction in our model. As illustrated in Sec. 3.3, we generate two neighbor sets from social information and rating information. Then, for a missing rating r_{ui}, we use the two neighbor sets to predict missing ratings by equation (12). In our problem the rating mode is binary, so the result produced by equation (12) can be viewed as the confidence for positive or negative ratings. And at last, we determine the prediction of r_{ui} by a parameter ζ as (13).

$$r_{ui} = \begin{cases} 1 & \mathrm{P}(u,i) > \zeta \\ -1 & \mathrm{P}(u,i) < -\zeta \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

where 1 represents a positive rating, -1 represents a negative rating and 0 represents a missing rating.

4 Evaluation Metric and Experimental Analysis

4.1 Evaluation Metric and Data Set

We are most interested in a system that can recommend items in a ranked list where the most user-interested items take top positions rather than a method that accurately predicts the numeric rating of every item. To analyze this, we use the predicted score to rank the recommended items, and apply the Mean Average Precision at N i.e. MAP@N to measure the recommendation quality. We evaluate our algorithm in the data set provided by Tencent Inc. for KDD CUP 2012, which represents a sampled snapshot of Tencent Weibo users' preferences for various items. The user-item rating matrix in this data set contains 42118498 distinct binary ratings rated by 1392873 users on 4710 items. The density of the matrix is 0.64%. In addition to the rating matrix, the data set contains rich social information about users and items. We divided the ratings into two parts: the ratings made before 22:16:00 5th November 2011 as training data, and the rest ratings as testing data. To set up the experiments effectively, we sampled 6000 users and their ratings randomly from the training data and built three training sets containing 1000, 2000, 3000 users respectively. Then we sampled 200 testing users and their ratings from the testing data accordingly.

4.2 Experiments and Analysis

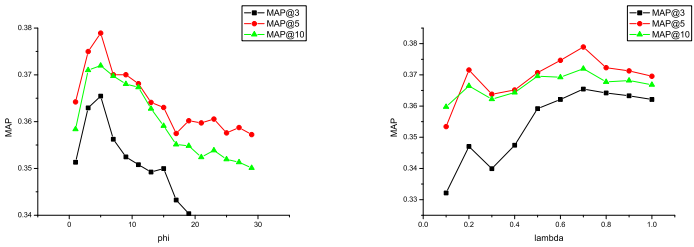
We have described how to improve jaccard similarity measure to fit our problem, how to integrate social information into neighbor selection process and how to predict the missing rating to make the rating matrix denser to improve recommendation quality. Accordingly, we will conduct several experiments to answer the following questions:

1. Does the improved jaccard similarity measure help to improve prediction accuracy? If it does, what is the effect of the parameter λ and ϕ?
2. Does the social similarity model help to improve prediction accuracy? If it does, how do the jaccard similarity model and social similarity model benefit each other?
3. We implemented $3 \times 2 = 6$ algorithms from the following two dimensions:
   (a) neighbor selection
      i. use social information only
      ii. use rating matrix only
      iii. combine both
   (b) missing ratings prediction
      i. predict missing ratings
      ii. not predict missing ratings

   Among all the 6 algorithms, which one performs the best?

Question 1. To answer Question 1, we build a model which does not incorporate the social information and the missing rating predicting process for clarity. First, we set ϕ to 5, and vary the range of λ from 0 to 1 with a step value of 0.1. Then we plot the MAP-λ curve to show the impact of λ. Fig. 1 shows how λ affects MAP@3, MAP@5, MAP@10 respectively. Setting λ to 1.0 means equally weighting positive and negative ratings and decreasing λ means reducing the influence of negative ratings. As we see in Fig. 1, MAP increases as we reduce λ from 1.0 to 0.7, which implies that reducing the influence of negative ratings does help to increase the recommendation accuracy. If we continue to reduce λ to 0, MAP decreases. So, we get the best performance when $\lambda = 0.7$ on our experiment data set.

To show the effects of ϕ, we set λ to 0.7, and vary the range of ϕ from 1 to 29 with a step value of 2. Then we plot the MAP-ϕ curve to show the impact of ϕ. Fig. 1 shows how ϕ affects MAP@3, MAP@5, MAP@10 respectively. The purpose of introducing ϕ is to devalue the similarity of users who have too few co-ratings and make the similarity computation more sensible. The larger the value of ϕ, the similarity of the users who have few co-ratings will be devalued more seriously. Setting the value of ϕ to 1 means computing user similarity normally. As we see in Fig. 1, MAP increases as we increase the value of ϕ from 1 to 5, which implies that introducing the penalty function to similarity computation does help to improve recommendation quality. If we continue to increase the value of ϕ to 29, we can see that MAP decreases on the overall trend. Thus, we get the best performance when $\phi = 5$ on our experiment data set.



(a) Impact of ϕ (b) Impact of λ

Fig. 1. Impact of significant weighting

Question 2. To answer Question 2, we combine the two neighbor selection methods to make predictions. For clarity, we remove the missing rating prediction step. Parameter θ balances the effect of social features and the effect of ratings. It takes advantages of these two neighbor selection methods. If $\theta = 0$, we only use the rating matrix to compute neighbor set for users, and if $\theta = 1$, we only use the social features to compute neighbor set for users. In other cases, we combine the predictions based on the two neighbor sets to get the final predictions. To show how the two neighbor selection methods benefit each other, we first set λ to 0.7 and set ϕ to 5 respectively, and then vary the range of θ from 0 to 1 with a step value of 0.1 and plot MAP-θ curve.

Observed from Fig. 2, we draw the conclusion that combination of the two neighbor selection methods does help to improve prediction accuracy significantly. Figure 2 shows that as the value of θ increases from 0 to 0.3, MAP increases. As the value of θ continues to increase, MAP decreases on overall trend. We get the best performance at $\theta = 0.3$, which may indicate that the rating information is more important than social information.

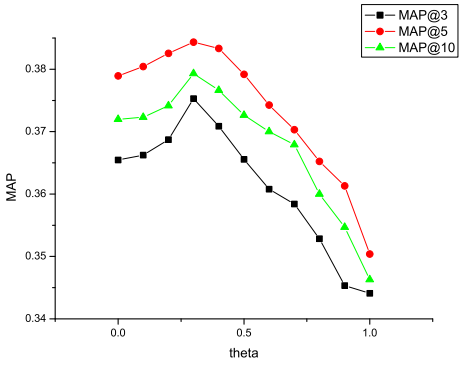


Fig. 2. Impact of θ

Question 3. To answer Question 3, we build a model that makes predictions using neighbor selection, missing rating prediction combined as a single factor. Thus, we can compare all the 6 algorithms individually side by side. In these algorithms, we set the parameters to the best values according to the previous experiment results, i.e. $\lambda = 0.7, \phi = 5, \alpha = 0.2, \beta = 0.2, \gamma = 0.6, \varphi = 0.1$, $\theta = 0.3$ and $\zeta = 0.5$ (the tuning process of $\alpha, \beta, \gamma, \varphi$ and ζ is not included in this paper due to space limitation). Table 1 illustrates the performance of the 6 algorithms. The result suggests that the algorithm integrating social information for neighbor selection and predicting missing ratings outperform other algorithms.

Table 1. Comparison of Algorithms

		Predict missing	Not predict missing
MAP@3	SNS	0.349	0.344
	REG	0.373	0.365
	COM	**0.387**	0.375
MAP@5	SNS	0.354	0.350
	REG	0.384	0.379
	COM	**0.395**	0.384
MAP@10	SNS	0.347	0.346
	REG	0.378	0.372
	COM	**0.392**	0.379

5 Conclusion and Future Work

In this paper, we studied the celebrities or stars recommendation problem on micro-blog web sites. First, we improved jaccard similarity by significant weighting to make the similarity measure more reasonable. Second, we integrated social information for neighbor selection. Third, we predicted missing ratings to enhance collaborative filtering. The experiment results showed that our approach improves the recommendation quality significantly. We claim that our recommendation framework is easy to be generalized to fit other collaborative filtering problems, which are provided with external information about users. However, domain-specific properties may have great impact on the effectiveness of the algorithms and more specific user similarity models need to be developed.

Further study may explore more social features to deepen our understanding on user similarity modeling. For example, we may use the sequential information such as time stamp of ratings to make session analysis to find similar patterns for users as the evidence for similarity computation. In addition to the users' social information, items' social information is valuable to leverage to enhance item based collaborative filtering.

Acknowledgement. We would like to acknowledge Tencent Inc. for providing the exciting and comprehensive data set. This work was supported in part by the National Natural Science Foundation of China (No.61170233, No.61232018,

No.61272472, No.61202404, No.61272317)and China Postdoctoral Science Foundation (No.2011M501060).

References

1. Brzozowski, M.J., Romero, D.M.: Who Should I Follow? Recommending People in Directed Social Networks. In: AAAI Conference on Weblogs and Social Media (2011)
2. Goldberg, K., Roeder, T., Gupta, D., Perkins, C.: Eigentaste: a constant time collaborative filtering algorithm. Information Retrieval 4(2), 133–151 (2001); Advances in Artificial Intelligence, 1–20 (2009)
3. Sarwar, B., Karypis, G., Konstan, J., Reidl, J.: Item based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International Conference on World Wide Web, pp. 285–295 (2001)
4. Billsus, D., Pazzani, M.: Learning collaborative information filters. In: Proceedings of the 15th International Conference on Machine Learning (1998)
5. Melville, P., Mooney, R.J., Nagarajan, R.: Contentboosted collaborative filtering for improved recommendations. In: Proceedings of the 18th National Conference on Artificial Intelligence, pp. 187–192 (2002)
6. Ma, H., King, I., Lyu, M.R.: Effective missing data prediction for collaborative filtering. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 39–46 (2007)
7. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis (2007)
8. Chen, J., Nairn, R., Nelson, L., Bernstein, M., Chi, E.H.: Short and Tweet: Experiments on Recommending Content from Information Streams. In: Proceedings of the 28th Conference on Human Factors in Computing Systems, pp. 1185–1194 (2010)
9. Guy, I., Zwerdling, N., Carmel, D., Ronen, I., Uziel, E., Yogev, S., Ofek-Koifman, S.: Personalized recommendation of social software items based on social relations. In: Proceedings of the 3rd Conference of Recommender Systems, pp. 53–60 (2009)
10. Chen, J., Geyer, W., Dugan, C., Muller, M., Guy, I.: Make new friends, but keep the old - Recommending people on social networking sites. In: Proceedings of the 27th Conference on Human Factors in Computing Systems, pp. 201–210 (2009)
11. Daly, E.M., Geyer, W., Millen, D.R.: The network effects of recommending social connections. In: Proceedings of the 4th ACM Conference on Recommender Systems (2010)
12. Hannon, J., Bennett, M., Smyth, B.: Recommending twitter users to follow using content and collaborative filtering approaches. In: Proceedings of the 4th ACM Conference on Recommender Systems (2010)

A Semantically Enhanced Tag-Based Music Recommendation Using Emotion Ontology

Hyon Hee Kim

Department of Statistics and Information Science, Dongduk Women's University
23-1 Wolgok-Dong, Sungbuk-Gu, Seoul, South Korea
heekim@dongduk.ac.kr

Abstract. In this paper, we propose a semantically enhanced tag-based approach to music recommendation. While most of approaches to tag-based recommendation are based on tag frequency, our approach is based on semantics of tags. In order to extract semantics of tags, we developed the emotion ontology for music called UniEmotion, which categorizes tags into positive emotional tags, negative emotional tags, and factual tags. According to the types of the tags, their weights are calculated and assigned to them. After then, user profiles using the weighted tags were generated and a user-based collaborative filtering algorithm was executed. To evaluate our approach, a data set of 1,100 users, tags which they added, and artists which they listened to was collected from last.fm. The conventional track-based recommendation, the unweighted tag-based recommendation, and the weighted tag-based recommendation are compared in terms of precision. Our experimental results show that the weighted tag-based recommendation outperforms other two approaches in terms of precision.

Keywords: music recommendation, collaborative tagging, emotion ontology.

1 Introduction

Collaborative tagging has attracted attention as a powerful tool for users to present their opinion about web resources in social web sites. It allows a user to add keywords called tags which are freely chosen by himself to not only his resource but also other users'. The whole set of tags, which is called folksonomy, has become an emerging classification scheme of web resources. Since the tags do not have any pre-defined terms or hierarchies of them, a user's set of tags represents his preference and interests explicitly. Therefore, research on applying folksonomy to generate user profiles in a recommender system has been extensively done [1, 2].

In the conventional recommender systems, rating systems which allow users to rate items with numbers, for example from 1 to 5, are used to generate user profiles. Rating systems suffer from the well-known problem of data sparsity [3]. Usually, the number of ratings users evaluated is much smaller than the number of ratings that need to be predicted. To overcome the problem, users' implicit information, such as clickstream, is used as user profile information. However, collecting implicit information is time-consuming and, sometimes, contains noisy data.

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 119–128, 2013.

Collaborative tagging systems offer users an alternative way to represent their opinion about items. Since tags are originally created to organize resources by users' own way, they contain meaningful concepts to users. Therefore, they serve as a bridge representing semantic relationships between users with items. In addition, while the number of clicks in web pages does not always mean interests or preferences about the items, frequently used tags by users certainly do. From this point of view, tag-based user profiles can improve performance of a recommender system.

Especially, tags in social music sites, such as last.fm, play an important role rather than other social sites [4]. In general, users in the music sites tend to listen to favorite music items repeatedly and continuously instead of rating them. Therefore, users' preferences can be better captured by implicit information, i.e., listening habits rather than explicit rating systems. However, collecting users' listening behavior takes time. Moreover users do not listen to all of their favorite music items through the music sites. Rather, tags are added to their favorite music items without listening to them, and contain users' opinion and sentiment about the music items. Therefore, using tags for user profiling could be more important for music recommendation.

The aim of this study is to provide a novel approach to music recommendation using semantically enhanced tag-based user profiles. In particular, we emphasize that different from the conventional tag-based music recommendation, which uses tag frequency for user profiles, semantics of the tags are considered. There are largely two types of tags in folksonomy. One is describing facts on the music item, for example, genre, region, year, nationality, etc. and the other is describing users' emotion on the music items, for example, excellent, cool, disgusting, etc. Here, we realized that the latter one represents more concrete and direct users' opinion. However, in recent tag-based recommendation research which uses tag frequency, semantics of tags are less considered.

To capture semantics of tags, we developed the emotion ontology called UniEmotion. Recently, to improve semantic value of folksonomy, ontologies have been developed, and called tag ontology [5]. In the tag ontology, the class of tags can have several properties representing semantic conceptualization of tags. For example, property "equivalent" is defined for equivalent tags, and property "related" is for relevant tags. In the same way, spelling variants or acronyms are also defined to resolve semantic ambiguity of them. However, those tag ontologies do not consider definition and practical use of tags representing human's emotion, such as opinion or sentiment about the items.

In this paper, we present a semantically enhanced tag-based music recommender system using semantics of tags. First, we propose the UniEmotion ontology which defines emotional tags, and categorizes them as positive, negative, and factual ones in the domain of a social music site. Second, we show an algorithm of recommending music items using the tag-based user profiles. Finally, to validate our system, we collected a data set of 1,100 users from last.fm and executed our recommendation algorithm. Our approach is compared with the conventional music recommendation approach which uses both track-based user profiles and tag-based user profiles by tag frequency.

The contribution of this paper is in the development and validation of the architecture of a tag-based music recommender system that considers semantics of folksonomy for user profiling. We have defined emotional tags and calculated their weights. By exploiting the weighted tags, user profiles are generated and music items are recommended. We have implemented a prototype music recommender system and conducted an empirical study to evaluate the effect of tag-based user profiling considering semantics of tags in music recommendation. Our experimental results show that our approach outperforms the conventional track-based recommendation and tag-based recommendation in terms of precision.

The remainder of this paper is organized as follows. In Section 2, we mention related work, and In Section 3, we describe the overview of the system. Section 4 explains the UniEmotion ontology and shows the proposed music recommendation method. Section 5 provides the experimental evaluation, and finally Section 6 gives concluding remarks.

2 Related Work

In this Section, we briefly present some of the research literature related to tag-based recommender systems and music recommendations. We also present research on ontology for emotion and emotion-based music recommendation.

Recent studies have focused on exploiting folksonomy as user information for user profiling. To improve tag-based recommendations, classifying tags into content-based, context-based, subjective, and organizational categories has been done [6]. To remove semantic heterogeneity, such as acronym, misspelling, or compound nouns, each tag is preprocessed, and then mapped to YAGO ontology, which is an ontological knowledge bases containing information from WordNet and Wikipedia [7]. Once a mapping is found, the YAGO ontology finds a subcategory of tags, and a recommendation algorithm considering tags is executed. Cantador et al. show that the tag-based recommendation improves recommendation performance.

In particular, folksonomy plays a central role in music information retrieval and music recommendation [8]. MusicBox is a personalized music recommender system based on social tags [9]. To capture the 3-way correlations between users, tags, and music items, 3-order tensors model is used. Nanopoulos et al. show that the proposed method improves the recommendation quality. Kim et al. proposed a method for tag-based user profiling in a social music site [10]. To evaluate the proposed approach, K-means clustering algorithm is executed on tag-based user profiles. Their experimental result shows that the proposed tag-based approach clusters similar users more efficiently than the conventional track-based approach.

Although folksonomy has attracted attention as emergent user information, there is no semantics and agreement of the tags added by users. Therefore, it is necessary to define semantic concepts of those tags and relationships among them. Ontology, which is an explicit specification of a conceptualization, can be enabling technology for conceptualization of folksonomy. Gruber formed the basis of a tag ontology, which is represented by an object, a tag, and a tagger [11]. Recently, tag ontologies

have been developed to define the concept of a "tagging" and to resolving semantic ambiguity of tags [5]. However, in the tag ontologies, definition of tags representing users' emotion, opinion, or sentiment has been seldom considered.

OntoEmotions [12] is an ontology of emotional categories covering the basic emotions: *Sadness, Happiness, Surprise, Fear* and *Anger*. Arsmeteo, the art portal, allows user to add tags and to retrieve artworks via the tags. A tag in Arsmeteo folksonomy is mapped to a concept of OntoEmotions, and new relations can be inferred by reasoning on the ontology of emotions. COMUS ontology [13] describes music related information about genre, mood such as angry and happy, location, time and events in daily life. Using the ontology, users' desired emotion state is evaluated and appropriate music is recommended. However, both ontologies do not consider the weight of emotional tags used in the UniEmotion ontology.

3 Overview of the System

In this section, we present the architecture of the recommender system. The system is largely composed of five components: folksonomy databases, UniTag ontology, UniEmotion ontology, user profile generator, and music recommender engine.

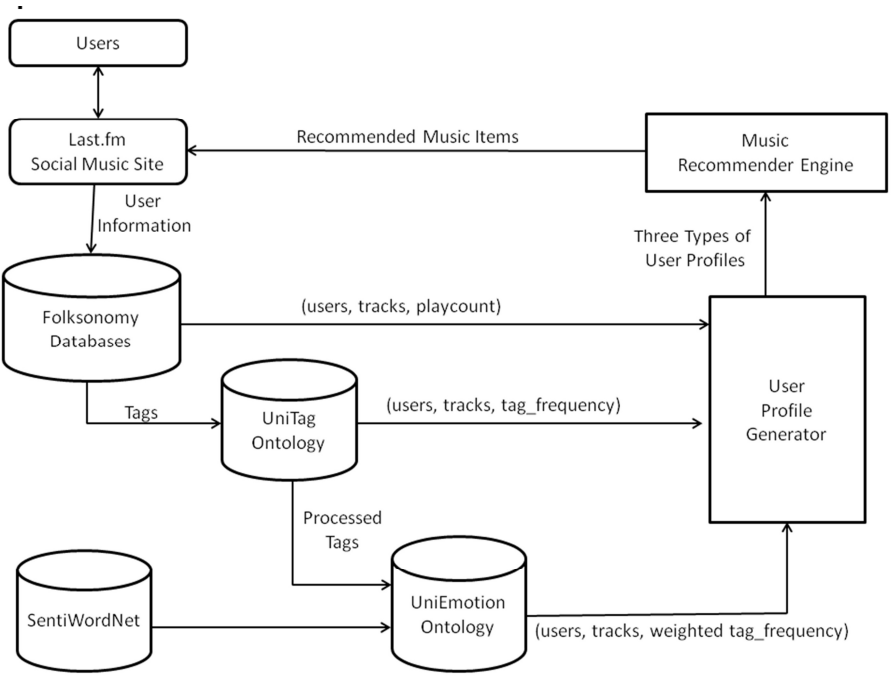


Fig. 1. System Architecture

In the last.fm, users listen to music and add tags to the music items. Using open API, users, track information, and tag information are extracted and stored into the folksonomy databases shown in the left side of Figure 1. First, UniTag ontology preprocesses the set of tags. It resolves semantic ambiguity of tags such as synonym, homonym, and acronym. The details of the UniTag ontology are described in [10], and will not be mentioned. After preprocessing the tags, the weight of the tags are assigned by UniEmotion ontology shown at the bottom of the Figure. UniEmotion ontology categorizes the tags into factual tags and emotional tags. The factual tags represent facts about the music items, such as musical genre, artists, age, nationality, region, etc., while the emotional tags represent users' opinion and sentiment such as cool, excellent, my favorites, etc. The emotional tags are classified into positive tags and negative tags. Based on the SentiWordNet ontology, the intensity of the emotional tags is also defined. UniEmotion ontology assigns the weighted values to the set of tags.

The user profile generator shown in the right side of Figure 1 creates three types of user profiles: track-based profiles, tag-based profiles, and weighted tag-based profiles. The track-based user profiles are generated using users, tracks which the users listened to, and number of playcounts. Therefore, the profiles do not use any ontology, and are directly generated from the folksonomy databases. The tag-based user profiles are generated using users, tracks which they added tags to, and tag frequency. Since the profiles do not consider semantics of tags and but consider tag frequency, only UniTag ontology is used. Finally, the weighted tag-based user profiles are generated using users, tracks which they added tags to, and the weighted values of the tag. The profiles use both UniTag ontology to preprocess the tags and UniEmotion ontology to assign the weighted values to emotional tags. After three types of user profiles are created, music recommendation is executed. The recommendation engine shown in the top right side in Figure 1 recommends n number of music items. The recommendation algorithm will be explained in detail in Section 4.2

4 UniEmotion Ontology and Music Recommendation

In the case of last.fm, 85% of tags are factual tags related to genre, region, instrumentation, while 10% of tags are emotional tags related to opinion such as excellent, sentiment such as favorite, or mood such as chill. The remaining 5% of tags are personal tags, such as seen it live or organizational tags, such as check out. Although the emotional tags comprise 10% of tags, we emphasize that the role of the emotional tags in tag-based recommendation is crucial.

Since the emotional tags represent users' musical preference explicitly, using only frequency of the tags is not enough to create user profiles in the tag-based recommendation. For example, a factual tag, e.g., progressive rock and an emotional tag, e.g., perfect, should not be considered with the same weight. Therefore, to classify the emotional tags and to assign the weight to them, we developed

UniEmotion ontology. In Section 4.1, we describe the UniEmotion ontology in detail, and In Section 4.2, we explain the music recommendation algorithm.

4.1 UniEmotion Ontology

The UniEmotion ontology is composed of four negative emotions, i.e., *fear, sadness, disgust, anger* and four positive emotions i.e., *joy, trust, anticipation, surprise,* as chosen by Plutchik's model [14]. Therefore, the emotion class has eight subclasses mentioned above. The emotion class has the *hasIntensity* property which describes intensity of an emotion. The intensity of an emotion is classified as strong, middle, and weak.

Let us take a closer look at the tags belonging to the category, *happiness*. For example, the tag, *beautiful*, represents a strong intensity of the emotion, while the tag, *joyful*, represents a weak intensity of the emotion. Also, the tags, *favorite* and *favorites*, are considered as a middle intensity of the emotion. Sometimes, tags have synonym, and for the definition of synonym, the emotion class has the *hasSynonym* property. In the case of several emotional tags, users prefer to use their native language. Therefore, for the definition of the foreign language of a tag, the emotion class has the *ForeignLanguage* property, and the ForeignLanguage property has subproperties, such as *French*, *Deutsch*, *Korean*, etc. The intensity of emotion is assigned manually based on the SentiWordNet online dictionary [15].

4.2 Music Recommendation

The first step of a recommendation process is generating user profiles. In our recommendation algorithm, a user profile is represented by a set of (*userID*, *itemID*, *preference*). The *userID* is assigned to each user sequentially, and the *itemID* is assigned to each music artist listened and tagged by the users. Finally, the *preference* represents a user's preference about the specific item. According to the types of the *preference*, three types of user profiles are generated in this study: track-based profiles, tag-based profiles, and weighted tagcount-based profiles. In the case of track-based profiles, play counts of the item is calculated, which are used for a conventional recommender system. This is used for comparison with the proposed approach.

In the case of the tag-based profiles, the number of tags used for the item is calculated. This is based on the simple assumption that if a user adds tags to an item and uses the tags often, then he seems to be interested in the item. However, in the profiles, the meaning of a tag is not considered. Finally, in the case of the weighted tag-based profiles, semantics of tags are considered. After classifying tags using the UniEmotion ontology, the weighted value of a tag is assigned to each tag, and then the value of preference is calculated.

For both the tag-based profiles and the weighted tag-based profiles, the factual tags are calculated as 1. Also, negative words are calculated as -1. The main difference between the tag-based profiles and the weighted tag-based profiles lies in handling the emotional tags. In the tag-based profiles, the emotional tags are also calculated as 1,

but in the weighted tag-based profiles, the emotional tags have the weighted values based on the classification using the UniEmotion ontology. Positive emotional tags have the value 1.5, 2, and, 2.5 according to intensity of emotion, weak, middle, and strong, respectively, while negative emotional tags have the value, -1.5, -2, and, -2.5 according to the emotional intensity.

Let us assume that user 1 adds a tag, *fantastic*, to item 10. In the case of the tag-based profiles, it is represented by (1, 10, 1), while in the case of the weighted tag-based profiles, it is represented by (1, 10, 2.5). Since the tag, *fantastic*, is classified as category of happiness with the strong intensity using the UniEmotion ontology, the preference value is weighted as 2.5. Based on the user profiles, a user-based collaborative filtering algorithm is executed. Algorithm 1 shows the music recommendation process.

First of all, once a target user is selected among 1,100 users, n nearest neighbors are chosen by calculating the Pearson Correlation similarity, which finds the ratio between the covariance and the standard deviation of both objects. This is because calculating every item is too much overhead. Then, the estimated preferences of every item which the target user does not have any preference for yet are calculated by computing the estimated average preference. The estimated average preference is calculated on average of preference for the target item, and by weighting similarity between the target user and a neighbor user belonging to the top n neighborhood. As a result, items with the estimated highest average preference are recommended.

Algorithm 1. Music Recommendation

Input: set of User Profiles {$\mathbf{U_n}$ X $\mathbf{I_n}$ X $\mathbf{P_n}$}
Output: set of recommended items
  For every user $\mathbf{U_n}$
    compute user similarity between $\mathbf{U_n}$ and $\mathbf{U_{n+1}}$
    lists the top n users ranked by similarity as a neighborhood **N**
  For every item $\mathbf{I_n}$ which a user $\mathbf{U_n}$ in **N** has a preference for, but a target user $\mathbf{U_t}$ does not have a preference yet
    For every user $\mathbf{U_v}$ in **N** which has a preference for $\mathbf{I_n}$
      compute a similarity **s** between $\mathbf{U_t}$ and $\mathbf{U_v}$
      calculate $\mathbf{U_v}$'s preference for $\mathbf{I_n}$ weighted by **s** on average
  lists the top 10 items ranked by estimated preferences

5 Experimental Results

Diverse evaluation metrics of recommender systems such as precision, recall, F1 measure and ROC curve are used. Among them, we choose the precision metric, which relates the number of hits to the total number of recommended items. According to McLaughlin and Herlocker [16], the precision metric reflects the real

use experience better than other evaluation metrics, because users receive ranked lists from a recommender instead of predictions for ratings of specific items.

The precision metric is calculated by (1)

$$\text{Precision} = \frac{\text{number of correct recommendation}}{\text{number of all recommended items}} \quad (1)$$

We randomly select 1,100 users from last.fm. There are about 18,700 artists, which the users have listened to, and 12,600 tags which the users added. Among the tags, 500 emotional tags are processed using UniEmotion ontology, and the weighted values are assigned to them according to the emotion's intensity. We chose 70% of data as training set and 30% of data as test set.

Figure 2 shows precisions of the three approaches when 10 similar users are chosen. We evaluated precisions for the number of recommended items at 5, 10, 15, 20, and 25. A shows the track-based recommendation, and B shows the tag-based recommendation. Finally, C is the weighted tag-based recommendation. As shown in Figure 2, the weighted tag-based approach outperforms other two approaches. While in case of B and C, the precisions are increasing with increasing the number of recommended items, in case of A, the precisions are decreasing. The average values of precisions are 1.99%, 2.22%, and 2.62% for A, B, and C, respectively, when 10 neighbors are chosen.

Also, we evaluated the precision according to the number of similar users at 10 recommended items shown in Figure 3. Different from Figure 2, precisions are decreasing when the number of similar users is increasing for all the methods. This is because the large number of similar users does not contain a target user's preference. The average values of the precision are 1.653%, 1.541%, and 2.128%, for A, B, and C, respectively. In this case, the experimental results also show that the weighted tag-based recommendation outperforms other two approaches.

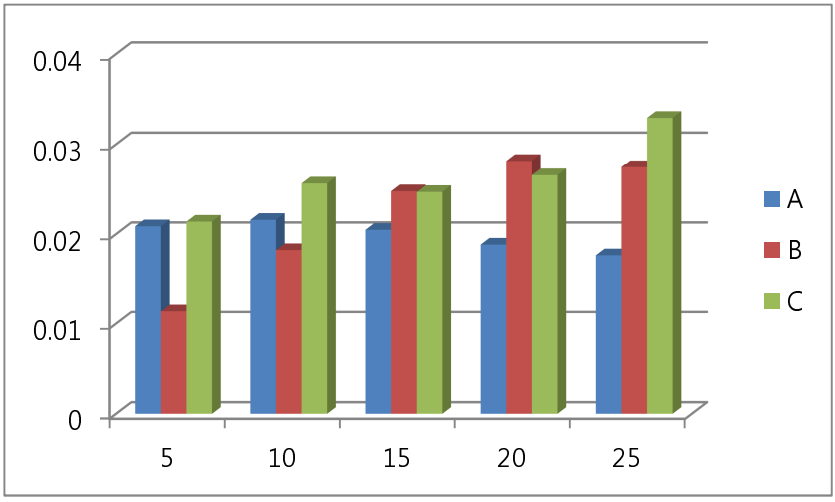


Fig. 2. Precisions with the number of recommended items at 10 similar users

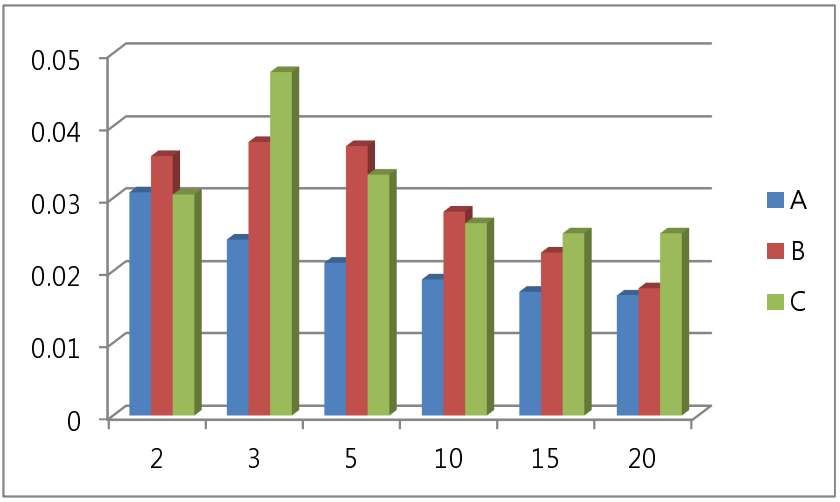


Fig. 3. Precisions with the number of similar users at 10 recommended items

6 Conclusions and Future Work

In this paper, we present a semantically enhanced tag-based music recommendation. To capture semantics of tags, the UniEmotion ontology is developed. Using the UniEmotion Ontology, every emotional tag is classified into four positive emotions and four negative emotions. Also, the emotional tags have three intensity of emotion: strong, middle, and weak. According to classification of emotional tags, weighted values are assigned to the emotional tags. The music recommendation algorithm is executed based on the weighted tag-based user profiles. Our experiments show that the weighted tag-based recommendation outperforms both the conventional track-based recommendation and the unweighted tag-based recommendation.

Two points are considered in this study. One is elaborating the UniEmotion ontology and the other is evaluation of the recommender systems. In general, context plays a central role in representing intensity of emotion. However, since tags are used without context, only the tag itself is considered in definition of the intensity of the emotional tags. Therefore, we have some difficulties in classifying several tags into 8 categories defined in the UniEmotion ontology. Also, emerging internet slangs should be handled with care. Definition of emotional tags is based on the SentiWordNet dictionary, but in the case of internet slangs, negative words can be used to represent positive emotion. Therefore, it needs further study.

Presently, we are doing diverse types of evaluation considering recalls, F1 measure, and ROC curve. Precision is one of the main evaluation techniques in the recommender systems, but other evaluation matrices should be also considered.

References

1. Firan, C.S., Nejdl, W., Paiu, R.: The Benefit of Using Tag-Based Profiles. In: LA-Web 2007. Snatiago de Chile (2007)
2. Sen, S., Vig, J., Riedl, J.: Tagommenders: Connecting Users to Items through Tags. In: Proceedings of World Wide Web Conference, pp. 671–680 (2009)
3. Adomavicius, A., Tuzhilin, A.: Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Trans. on Knowledge and Data Engineering 17(8), 743–749 (2005)
4. Celma, O., Ramirez, M., Herrera, P.: Foafing the Music: a Music Recommendation System Based on RSS Feeds and User Preferences. In: Proc. of International Conference on Music Information Retrieval (ISMIR 2005), Londun, UK (2005)
5. Kim, H.L., et al.: The State of the Art in Tag Ontologies: A Semantic Model for Tagging and Folksonomies. In: Proceedings of the International Conference on Dublin Core and Metadata Applications (2008)
6. Cantador, I., Konstas, I., Joemon, M.J.: Categorising Social Tags to Improve folksonomy-based Recommendations. Journal of Web Semantics 7(1), 1–15 (2011)
7. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: A Large Ontology from Wikipedia and WordNet. Journal of Web Semantics 6(3), 203–217 (2008)
8. Lamere, P., Pampalk, E.: Social Tagging and Music Information Retrieval. In: Proceedings of Inernational Conference on Music Information Retrieval (2008)
9. Nanopoulos, A., et al.: MusicBox: Personalized Music Recommendation based on cubic Analysis of Social Tags. IEEE Trans. on Audio, Speech and Language Proceeding 18(2), 1–7 (2010)
10. Kim, H.H., Jo, J., Kim, D.: Generation of Tag-Based User Profiles for Clustering Users in a Social Music Site. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ACIIDS 2012, Part II. LNCS, vol. 7197, pp. 51–61. Springer, Heidelberg (2012)
11. Gruber, T.: Ontology of Folksonomy: A Mash-up of Apples and Oranges. Int. J. on Semantic Web & Information Systems 3(2), 1–11 (2007)
12. Baldoni, M., et al.: From Tags to Emotions: Ontology-driven Sentiment Analysis in the Social Semantic Web. Intelligenza Artificiale 6(1), 41–54 (2012)
13. Han, B., et al.: Music Emotion Classification and Context-based Music Recommendation. Multimedia Tools and Applications 47(3), 433–460 (2010)
14. Plutchik, R.: The Nature of Emotions. American Scientist 89, 344–350 (1997)
15. Esuli, A., Sebastiani, F.: SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In: Proc. of the 5th Conference on Language Resources and Evaluation (LREC 2006), Genova, Italy, pp. 417–422 (2006)
16. McLaughlin, M.R., Herlocker, J.L.: A collaborative filtering algorithm and evaluation metric that accurately model the user experience. In: Proc. of the 27th ACM International Conference on Research and Development in Information Retrieval (SIGIR 2004), Sheffield, UK, pp. 329–336 (2004)

A Method for Determination of an Opening Learning Scenario in Intelligent Tutoring Systems

Adrianna Kozierkiewicz-Hetmańska and Dariusz Zyśk

Institute of Informatics, Wroclaw University of Technology, Poland
adrianna.kozierkiewicz@pwr.wroc.pl, dariusz.zysk@gmail.com

Abstract. The intelligent tutoring systems should guarantee an effective learning. Students who use those systems should achieve better learning results in a shorter time. Our previous research pointed out that the personalization of the learning scenario allows to satisfy the mentioned postulates. In this paper the method for determination of an opening learning scenario is presented. Before a student begins to learn an opening scenario is determined based on information provided during a registration process. User is offered the optimal learning path suitable for his learning styles and a current knowledge level. Worked out method applied the ant colony optimization technique. The effectiveness of the proposed solution was tested in a specially implemented environment. The researches demonstrate that the algorithm gives quite good results, because 66% of the learning material in the determined learning scenario were adapted to student's learning styles.

1 Introduction

Intelligent tutoring systems (in this work also called distance education systems, intelligent e-learning systems or simply e-learning systems) are a solution for the problem of maintaining a high quality education together with its low cost. The traditional in-class learning is often ineffective because a teacher lectures passive students. In e-learning systems students can decide when and how often they want to learn and a learning material is often well suited for students' needs, preferences, knowledge level, learning styles etc. This advantages cause that researchers are interested in the distance education systems.

An e-learning system has to take over from a teacher and support students during the learning process. The main task of intelligent tutoring systems is offering a proper order and form of the education material which is called a learning scenario. By an opening learning scenario we mean a first learning scenario proposed to a student after the registration process. Determining the opening learning scenario is a very important task because it influences student's motivation. If the opening learning scenario is completely unsuitable for a student he or she can give up learning. Moreover, it is a difficult and a complicated problem because it requires finding the opening learning scenario which is the best adapted to student's needs and preferences but at the beginning of the learning process we do not have a lot of information about students. So far, there has been no solution for the problem of determining an opening

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 129–138, 2013.

learning scenario based on a student's characteristic. In our previous works [8,11] we worked out methods based on assumption that similar students will learn in the same or a very similar way [12]. Before starting the learning process student provides information about himself which are stored in a student's profile. Next, the student is classified to a group of similar learners. Based on successfully finished scenarios of students who belong to the same class as the new learner the opening learning scenario is chosen. The consensus theory was applied in the algorithm to generate the opening learning scenario. In such solution it is expected that if students described by similar features finish a similar learning scenario with positive notes, the new learner completes the offered learning scenario successfully. The described procedure has some limitations which occur in case of starting up the e-learning system. If a system has no information about finished scenarios he has to choose the opening learning scenario in a random way.

The solution to the problem described above is presented in this paper. This work is devoted to presenting a method for determination of an opening learning scenario based only on information collected during the registration process. It is original, innovative and presented for the first time. The proposed method uses the ant colony techniques which are tested in the specially implemented environment.

This paper is organized as follows. In the next Section methods used for determining a suitable learning material in e-learning systems are presented. Section 3 contains a description of our approach for determining an opening learning scenario based on the ant colony algorithm. Next, the results of experiments which tested the effectiveness of the proposed solution are showed. Finally, conclusion and future works are described.

2 Related Works

In many intelligent tutoring systems different methods for determination of a learning scenario were applied such as: Bayesian Network, Neural Network, genetic algorithms, or consensus theory.

The very popular methodology is Bayesian Networks. In [6] author designed the Pedagogical Module which is responsible for determining a suitable learning material. The role of this module is to choose between the following actions: show a new topic, deepen a current topic, review of a previous topic and present the next page in the index. The Bayesian network for the pedagogical action is built and two variables: time spent on the corresponding topic as well as the answered question are considered during planning of the learning material. The Naïve Bayesian algorithm is applied also in EDUCE [7]. The analysis of the information about time of learning, how many times a user looked at the type of material and attempted to answer a question and after which resource he/she gets the question right, allow to personalize the learning material.

In [2] and [16] for providing the personalized learning the genetic algorithm is used. This task required to describe an abstract representation of a solution named chromosome, fitness function, selection operation, crossover and mutation operation

and stop criterion. Those parameters of genetic algorithms are defining, based on assumed knowledge, representation and goals of the system which should be achieved.

Neural Networks are applied in Learning Assistant [13] which infer using metadata describing pupils and didactic material. An SOM neural network is used for grouping similar pupils based on student's level of expertise, learner preferences and a learning pace. The learning path for a new user is generated based on the learning path that is appropriate for the cluster to which the pupil has been classified by the trained SOM.

The different approaches are presented in [8], [12] and [14] where algorithms for determination of an opening learning scenario are based on the consensus theory. In the first paper the learning scenario consists of presentations and corresponding tests to presentations. In [14] the algorithm of determination of an opening learning scenario is based on the choice of concepts' order, presentations and presentations' orders. This method is improved by adding the third step: the choice of suitable versions of lessons in [8].

In [1] author uses a fuzzy decision making process to update the learner model and specifies his learning level to provide an appropriate teaching material to each learner. The 18 rules were worked out which allow choosing the best fitted learning material. System could propose to a learner a new unit or a current unit on three different difficulty levels: for beginners, normal students and high quality pupils.

In recent years the ontologies have been used for solving different tasks and problems. In [4] ontology is applied to describe learning scenarios and a process of personalization of a learning scenario. In the proposed ontology the operations of the Guilford model are associated to the levels of the Blooms taxonomy. A student achieves goals defined according to the Bloom's taxonomy.

The new approaches for determination of the learning scenario are multi-agent systems. The examples of this solution are presented in [3] and [15].

Despite many methods for generation of learning scenario were worked out, very often, in real systems, only an adaptive navigation and presentation are implemented. Such solution requires students to make decision about the order of the learning material. Many described methods need information about the finished learning path, time spent on learning and the number of correct answers given by other students. The method presented in this paper solves the problems mentioned above.

3 The Ant Colony Algorithm

In this work we propose a method for determining an opening learning scenario based on the ant colony optimization technique. The algorithm tries to choose the learning material the best fitted to student's learning styles and current knowledge level stored in the learner profile. Method for determination of an opening learning scenario required defining a learner profile and a knowledge representation.

The learner profile p is created during the registration process. Student fills in a questionnaire to provide demographic data such as: login, password, name, e-mail which allow only for identification of the user. The learning styles are assessed using

the proper test which is used in the personalization process. ILS questionnaire is applied to assess student's learning styles [5]. The learner's behaviour is considered in four dimensions: perception (sensitive or intuitive), receiving (verbal or visual), processing (active or reflective) and understanding (sequential or global). The original results obtained from the ILS test are presented as a pair, where the first element refers to learner's preferred direction and the second is a score on a scale 1-11. In this work we transform the results from the ILS questionnaire into scale 0-1. We assume that the value of the following attributes: perception, receiving, processing, understanding greater than 0.5 refers to the strength of learner's preferred direction into: intuitive, verbal, active and sequential, respectively. Otherwise, student is characterized by sensitive of information perception, visual of information receiving, reflective of information processing and global of information understanding. The learner profile *p* is presented as a tuple:

$$t : A \rightarrow V$$

where: *A*- a finite set of attributes, *V*- a set of attributes' values, $V = \bigcup_{a \in A} V_a$, V_a- a set of attribute's value for $a \in A : \forall_{a \in A} (t(a) \in V_a)$.

The table below shows the content of the learner profile.

Table 1. The content of learner profile

Attribute name	Attribute domain
login	sequence of symbols
password	sequence of symbols
first name	sequence of letters
second name	sequence of letters
perception	[0,1]
receiving	[0,1]
processing	[0,1]
understanding	[0,1]

In our approach we assume that the knowledge structure consists of learning materials and relations between them. By the learning material we understand elementary, indivisible units. The knowledge structure is defined in the following way [11]:

Definition 1. *Knowledge structure is a directed and a weighted graph Gr:*

$$Gr = (S, R, X)$$

where: $S = \{s_1, s_2, ..., s_q\}$ *- set of nodes representing a learning material, R - set of edges representing relationships among learning materials,* $X = [x_{ij}]_{i,j=1,...,q}$, $x_{ij} \in [0,1]$.

By X we denote a matrix which reflects an expert's/teacher's opinion that material s_i should be learnt before material s_j, $i, j \in \{1,2,3,...q\}$. The greater value of x_{ij} refers that the expert recommends such connection of a learning material.

Definition 2. *By the learning scenario ls we mean a Hamiltonian path in a graph Gr.*

Figure 1 presents an example of the defined knowledge structure.

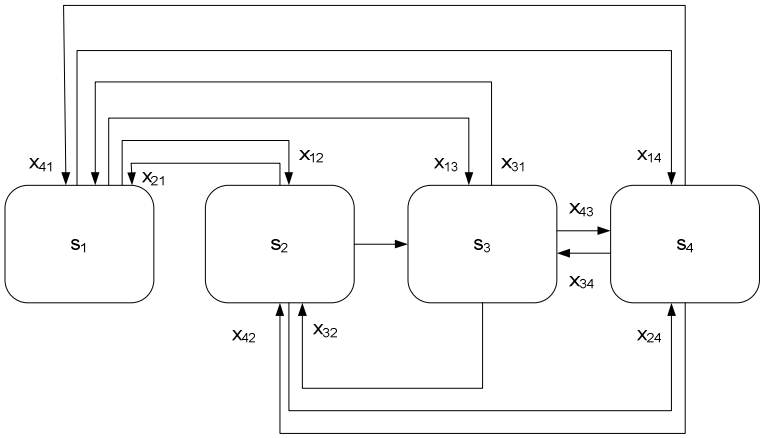


Fig. 1. The sample of graph-based knowledge structure, where S- set of nodes, X-weight matrix

Each learning material $s_j \in S$, $j \in \{1,2,3,...q\}$ is described by six attributes: perception, receiving, processing and understanding, required knowledge level and difficulty level. The attributes: perception, receiving, processing and understanding are interpreted similarly like in the learner profile. For the needs of this paper we assume the easy way to determine the value of those attributes. The domain of those attributes are intervals [0,1]. If the following attributes: perception, receiving, processing, understanding, describing a learning material have value greater than 0.5, it means that this learning material is dedicated to intuitive, verbal, active and sequential students, respectively. Otherwise, learning material is prepared for students characterized by the opposite direction of information perception, receiving, processing and understanding. Let us suppose that the attribute receiving describing learning material has value equal to 0.3. That learning material contains 70% of pictures, graphs, diagrams, films etc. which are preferred by visual students and only 30% of text. Additionally each learning material is described by a level of difficulty $d(s_j) \in D$ where D-set of difficulty level, $D = \{1,2,3,4,5\}$ and 5 means that learning material is dedicated to an expert and 1 to a beginner, and required knowledge level, which is a set of learning materials and their levels of difficulty, which should be learnt before: $rkl(s_j) = \{(s, dl(s)) : s \in S, dl \in D\}$. The presented knowledge representation allows

to offer to student a learning scenario which is adapted to his learning styles and current knowledge level.

As it has already been mentioned, our approach for determining the opening learning scenario is based on the ant colony optimization algorithm. The general idea of the ant colony optimization could be described in the following steps:

```
                ANT COLONY OPTIMIZATION PROCEDURE

BEGIN
Set parameters, initialize pheromone trails

WHILE termination conditions not met DO
{
1. Construct solutions();
2. Update pheromone trails();
}
END
```

Our ant colony optimization approach is similar to an ant algorithm for the travelling salesperson problem. We assumed that algorithm is finish after the assumed number of cycles *nc*. In every cycle *na* ants construct solutions which are the Hamiltonian path in graph Gr.

1. Construct Solutions()

A solution constructed by an ant $k \in \{1,...,na\}$ is dependent on chosen nodes s_j after nodes s_i $i, j \in \{1,2,3,...q\}$ with the probability estimated in the following way:

$$p_{ij} = \frac{[\tau_{ij}]^{\alpha}[\eta_{ij}]^{\beta}}{\sum_{s_k \notin tabu}[\tau_{ik}]^{\alpha}[\eta_{ik}]^{\beta}}, \; if \; s_j \notin tabu$$

where: α and β are constants that determine the relative influence of the pheromone values τ_{ij} and the heuristic values η_{ij} on the decision of the ant, $i, j \in \{1,2,3,...q\}$, *tabu*- the vector which contains the list of nodes visited by an ant *k*.

The greater heuristic values for learning material s_i and s_j increases the probability of choosing the material s_j after s_i in a learning scenario *ls* . The heuristic function represents how much the learning material s_i is connected with the learning material s_j. After the learning material s_i the most probable learning material is chosen which is the best suited for student's learning styles, where connection between lessons are the most recommended by experts and student has required knowledge level by learning a proper lessons before. The heuristic functions is defined as:

$$\eta_{ij} = w_1(4 - RS_j) + w_2 x_{ij} + w_3(4 - RU_{ij})$$

where: w_1, w_2, $w_3 \in [0,1]$ - weights of particular elements in the heuristic function, $RS_j = |pp - pm| + |rp - rm| + |prp - prm| + |up - um|$, *pp*- the value of attribute perception describing the learner, *pm*- the value of attribute perception describing the learning material s_j, *rp*- the value of attribute receiving describing the learner, *rm*- the value of attribute receiving describing the learning material s_j, *prp*- the value of attribute processing describing the learner, *prm*- the value of attribute processing describing the learning material s_j, *yp*- the value of attribute understanding describing the learner, *um*- the value of attribute understanding describing the learning material s_j, $RU_{ij} = \begin{cases} d(s_i) - dl(s_i) \;\; if \;\; rkl(s_j) \;\; contains \;\; s_i \;\; and \;\; d(s_i) \geq dl(s_i) \\ 4 \;\; otherwise \end{cases}$

2. Update Pheromone Trails()

The best solution found so far and the best solution found in the current generation are then used to update the pheromone matrix which are done in two stages. In the first step some of the old pheromone is evaporated on all the edges according to formula:

$$\tau_{ij}^{new} = (1 - \rho)\tau_{ij}^{old}$$

Then an ant $k \in \{1,...,na\}$ lay pheromone on edges which belong to ant's tour:

$$\tau_{ij}^{new} = \tau_{ij}^{old} + \frac{1}{\sum_{s \in ls} \eta_{ij}}$$

4 The Results of Experiment

For evaluating the proposed method the special environment was implemented [17]. The 50 learners' profiles were chosen in a random way from a normal distribution and 25 learning materials were prepared to conduct the research. The analysis is made for the significance level $\alpha = 0.05$.

In this work we assume that proposed solution for determining the opening learning scenario gives effective results if the learning scenario is the closest to the learner's profile. We take into account only those attributes which describe the learning styles because our previous work pointed out that students achieved better results if they were offered the learning scenario suitable for their learning styles [9],[10]. Taking into account student's learning styles improves the learning results by more than 7.182% and less than 7.976%. For assessing the distance between the learning scenario *ls* and the learner's profile *p* the following metric is proposed:

$$eff(ls, p) = \sum_{s \in ls} d(s, p)$$

$$where: \; d(s, p) = \sqrt{\frac{(pp - pm)^2 + (rp - rm)^2 + (prp - prm)^2 + (up - um)^2}{4}}.$$

Before testing the effectiveness of our approach for determining the opening learning scenario we set the best parameters as follows: $w_1 = 0.7$, $w_2 = 1.0$, $w_3 = 0.5$ $\rho = 0.5, \alpha = 0.8, \beta = 1.0$, nc=10, na=10. The value of algorithm's parameters such as α, β, evaporation rate, number of cycle, number of ants, were chosen based on experimental simulations. The parameters of heuristic function w_1, w_2 i w_3 were set based on researcher's knowledge. We decided that the algorithm should choose a learning scenario where the form of presentation of the learning material is the most preferable by students. The order of learning materials should have a smaller influence on the construction of the solution.

Next, we analyze the distribution of obtained data using Shapiro-Wilk test. We obtain the value of statistical test equal to 0.9827 and p-value equal 0.6699 therefore we cannot reject the null hypothesis that sample come from a normal distribution. For the further analysis we use t-Student test (parametric). We test a null hypothesis that the mean of *eff* is equal to 8.5 against an alternative hypothesis that the mean of *eff* is less than 8.5. The statistical value t=-2.4337 and p-value equal to 0.00932 suggests that the null hypothesis should be rejected and the alternative hypothesis accepted.

The conclusion of conducted statistical analysis is that the mean of the distance between the learner profile and the determined learning scenario is less than 8.5. Each learning scenario consists of 25 learning materials. Therefore we can interpret the obtained results that determined by the ant colony optimization algorithm learning scenario contains less than 34% of learning materials which are not fitted to student's learning styles. It is quite a good result particularly that the determined learning scenario could be modified and adapted to student's characteristic during the learning process.

5 Conclusion and Further Work

The personalization of the learning scenario is a very important task in designing of the intelligent tutoring systems because researches pointed out that students achieve better learning results if the learning material is suitable for their learning styles.

In this paper we present method which solves the problem of choosing the opening learning scenario adapted to a student's learning styles and a current knowledge level. This problem is an NP-hard therefore the heuristic algorithm based on the ant colony optimization technique is proposed. The novelty of this approach relies on a determination of the opening learning scenario in case of starting up the intelligent tutoring system and based only on information provided by a student during the registration process. Moreover, the ant colony optimization technique has not been used to choose the opening learning scenario before.

The preliminary research shows us that our approach for determination of the opening learning scenario gives good results which we can interpret that less than 34% of learning materials in a learning scenario were not well fitted to student's learning styles. The determined learning scenario is quite good for a beginning of the learning process (better than learning scenario chosen in a random way), especially it

could be modified during the learning process [11] and adapted to current user's characteristic.

In our further work we are planning to apply the proposed method in the intelligent tutoring system and conduct researches with real students. Furthermore we want to test our algorithm by taking into account other attributes (not only learning styles) describing the learner and learning materials e.g. the level of difficulty. We would like to change the knowledge structure because we suppose that more elaborates on the knowledge structure and possibility of choosing some learning materials could improve the effectiveness of the proposed algorithm.

References

[1] Al-Aubidy, K.M.: Development of a web-based distance learning system using fuzzy decision making. In: Intr. Conf. On Smart Systems & Devices, Tunisia (March 2003)

[2] Bhaskar, M., Das, M.M., Chithralekha, T., Sivasatya, S.: Genetic Algorithm Based Adaptive Learning Scheme Generation For Context Aware E-Learning. International Journal on Computer Science and Engineering (IJCSE) 2(4), 1271–1279 (2010)

[3] Chen, P., Meng, A., Zhao, C.: Constructing Adaptive Individual Learning Environment Based on Multiagent System. In: Proc. IEEE International Conference on Computational Intelligence and Security Workshops (2007)

[4] Essalmi, F., Jemni, L., Ayed, L.: A process for the generation of personalized learning scenarios based on ontologies. In: Proc. of ICTA 2007, pp. 173–175 (2007)

[5] Felder, R.M., Soloman, B.: Index of Learning Styles (1998), http://www2.ncsu.edu/unity/lockers/users/f/felder/public/ILSdir/ILS-a.htm

[6] Gamboa, H., Fred, A.: Designing intelligent tutoring systems: a bayesian approach. In: ICEIS 2001 – Artificial Intelligence and Decision Support Systems, pp. 452–458 (2001)

[7] Kelly, D., Tangney, B.: Predicting Learning Characteristics in a Multiple Intelligence Based Tutoring System. In: Lester, J.C., Vicari, R.M., Paraguaçu, F. (eds.) ITS 2004. LNCS, vol. 3220, pp. 678–688. Springer, Heidelberg (2004)

[8] Kozierkiewicz, A.: Determination of opening learning scenarios in intelligent tutoring systems. In: Zgrzywa, A., Choroś, K., Siemiński, A. (eds.) New Trend in Multimedia and Network Information Systems, pp. 204–213. IOS Press, Amsterdam (2008)

[9] Kozierkiewicz-Hetmanska, A.: Evaluation of an Intelligent Tutoring System Incorporating Learning Profile to Determine Learning Scenario. In: Jezic, G., Kusek, M., Nguyen, N.-T., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2012. LNCS (LNAI), vol. 7327, pp. 44–53. Springer, Heidelberg (2012)

[10] Kozierkiewicz-Hetmańska, A.: Evaluating the Effectiveness of Intelligent Tutoring System Offering Personalized Learning Scenario. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ACIIDS 2012, Part I. LNCS (LNAI), vol. 7196, pp. 310–319. Springer, Heidelberg (2012)

[11] Kozierkiewicz-Hetmanska, A., Nguyen, N.T.: A method for learning scenario determination and modification in intelligent tutoring system. International Journal Applied of Mathematics and Computer Science 21(1) (2011)

[12] Kukla, E., Nguyen, N.T., Danilowicz, C., Sobecki, J., Lenar, M.: A model conception for optimal scenario determination in an intelligent learning system. ITSE— International Journal of Interactive Technology and Smart Education 1(3), 171–184 (2004)

[13] Kwasnicka, H., Szul, D., Markowska-Kaczmar, U., Myszkowski, P.: Learning Assistant – Personalizing learning paths in elearning environments. In: 7th Computer Information Systems and Industrial Management Applications. IEEE (2008)
[14] Nguyen, N.T.: Advanced Methods for Inconsistent Knowledge Management. Springer, New York (2008)
[15] Rishi, O.P., Govil, R., Sinha, M.: Distributed Case Based Reasoning for Intelligent Tutoring System: An Agent Based Student Modeling Paradigm. World Academy of Science, Engineering and Technology 29 (2007)
[16] Tan, X., Shen, R., Wang, Y.: Personalized course generation and evolution based on genetic algorithms. Journal of Zhejiang University Science Computers and Electronic (2012) (in Press)
[17] Zyśk, D.: A Method for Determination of an Opening Learning Scenario in Distance Education Systems, master thesis (2012)

Recommending QA Documents for Communities of Question-Answering Websites

Duen-Ren Liu*, Chun-Kai Huang, and Yu-Hsuan Chen

Institute of Information Management
National Chiao Tung University, Hsinchu 300, Taiwan
dliu@mail.nctu.edu.tw

Abstract. Question & Answering (Q&A) websites have become an essential knowledge-sharing platform. This platform provides knowledge-community services where users with common interests or expertise can form a knowledge community to collect and share QA documents. However, due to the massive amount of QAs, information overload can become a major problem. Consequently, a recommendation mechanism is needed to recommend QAs for communities of Q&A websites. Existing studies did not investigate the recommendation mechanisms for knowledge collections in communities of Q&A Websites. In this work, we propose a novel recommendation method to recommend related QAs for communities of Q&A websites. The proposed method recommends QAs by considering the community members' reputations, the push scores and collection time of QAs, the complementary relationships between QAs and their relevance to the communities. Experimental results show that the proposed method outperforms other conventional methods, providing a more effective manner to recommend QA documents to knowledge communities.

Keywords: Knowledge Community, Group Recommendation, Knowledge Complementation, Question-Answering Websites, Link Analysis.

1 Introduction

Question & Answering (QA) websites become an important knowledge sharing platform, where question answering knowledge is formed through the mechanism of question posting and answering. The Yahoo! Answers Taiwan website (http://tw.knowledge.yahoo.com/) is a community-driven knowledge website which provides a knowledge community service, so that users with common interests or expertise can form a knowledge community to collect and share question answering knowledge regarding their interests. As the number of posting questions and answers increases rapidly through time, the massive amount of question answering knowledge creates a problem of information overload. Consequently, a recommendation mechanism is needed to recommend QAs to communities of Q&A websites and enhance the effectiveness of knowledge sharing.

* Corresponding author.

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 139–147, 2013.

Currently, related research in Question Answering Websites focuses on finding appropriate experts for answering target questions [1]. Previous researches did not investigate the recommendation mechanisms for knowledge collection in question answering websites. Moreover, previous studies on recommender systems focus on recommending items of interest to individual users via collaborative filtering or content-based approaches [2, 3]. Traditional group-based recommendation methods mainly include two kinds of approaches [4]. The first one aggregates interest profiles for each member in a group to form the group's interest profile. The group's interest profile is then used to filter recommended items. The second kind of approach generates a group recommendation list via aggregating the recommendation list of each member derived from personalized recommendations. To the best of our knowledge, there is no study on the recommendation mechanisms for knowledge collections in communities of question answering websites. Traditional recommendation mechanisms have not considered certain factors, such as knowledge complementation and the reputation of the community member in terms of his/her collected QAs.

In this work, a novel group recommendation method is proposed to recommend QA documents to communities of QA websites. The proposed recommendation method generates community profiles from previously collected QAs by considering community members' reputations in collecting and answering QAs, the push scores and collection time of QAs. Moreover, users are usually interested in browsing relevant QAs of related questions to get more complete and complementary information. The proposed approach generates recommendations of QA documents by considering the complementary knowledge of the documents and the relevance degree between the QA document and the community profile. Finally, we use the data collected from Yahoo! Answers Taiwan to conduct our experimental evaluation. Experimental results show that the proposed method outperforms other conventional methods, providing a more effective manner of recommending QAs to communities.

The rest of this paper is organized as follows. Section 2 describes related works. Section 3 describes the proposed methods for recommendation. Section 4 presents experiments and evaluation results. Finally, the conclusion is presented in Section 5.

2 Related Work

Existing group-based recommendation researches were divided into two aspects: the first kind of method aggregates the interest profile of each member in a group to form the group's interest profile. The group's interest profile is then used to filter recommended items. The second kind of approach generates a group recommendation list via aggregating the recommendation list of each member derived from personalized recommendations [4]. However, the second method does not take into account the importance of each member and the interaction between members. The current group-based recommendation systems are widely utilized in different fields, especially in the life and entertainment field. For example, in MusicFx [5], each member can give a rating to the music based on their preference. Group-based recommendations are also used for movies or TV programs such as PolyLens [6] and TV4M [7]. These recommendation systems combine individual preference of movies or programs and then

generate a common recommendation list for the group. In addition, group recommendation is generally used to recommend tourism schedules or scenic spots [8].

The identification of knowledge complementation is unclear due to the definition of complementary knowledge depending on the users themselves. Ma and Tanaka [9] use the concept of topic-structure to measure the complementary degree between two documents. Liu, Chen and Lu [10] define two types of knowledge complementation in a QA website: partial complementation and extended complementation, and propose a method to predict complementation relationships between QA documents by building a classification model based on three measures: question similarity, answer novelty and answer correlation.

3 Proposed QA Recommendation Approach

3.1 Overview of Recommendation for Community Knowledge Collection

The framework of our proposed recommendation method for a knowledge community contains three stages. In the first stage, the content of each QA document is preprocessed into a document profile vector. The term vector of a QA d is denoted as KP_d. The content of each QA is analyzed using the TF-IDF approach [11] to calculate the weight of term i in a profile of QA d, KP_d. In the second stage, the collected QAs are grouped into several topics according to their tags. A community topic profile is derived from a weighted aggregation of document profiles of a topic's collected documents by considering members' reputations in collecting QAs and answering questions, push scores of QAs as well as the time factors of collected QAs. QAs with higher push scores more clearly represent the community's interests. The most recent QAs collected can better reflect the current interest of the community. In the third stage, each target QA is compared with each collected QA of the community to determine a complementary score based on question similarities, answer novelty and answer correlation. Finally, the approach combines the community preference score and complementary score of each target QA to generate a recommendation list.

3.2 Preference Analysis of Knowledge Community

The topic relevance score of target QA q to a topic of community G can be derived by calculating the cosine similarity between q's profile and the community topic profile. A community G's preference score on the target QA q can then be derived as the maximal topic relevance score over all topics of G. The diversity of QAs exists in each topic of a community. Accordingly, we derive a community G's preference score on a target QA q by considering the top-k QAs in each topic collected by G that have highest weighted relevance scores to the target QA q, as shown in Eq. (1). Let $D_{G,q}^{z,topk}$ be the set of top-k QAs in topic z collected by community G that have highest weighted relevance scores to the target QA q. The weighted relevance score of a QA d to the target QA q is derived from their cosine similarity multiplied with the QA d's collection weights, including the collection member's reputation, push score of QA d, and the collection time of QA d. The community G's topic-based preference score on

target QA q in topic z, denoted as $TPR^{z}_{G,q}$, is an aggregation of the weighted relevance scores between the target QA and the QAs in $D^{z,topk}_{G,q}$.

$$TPR^{z}_{G,q} = \frac{\sum_{d \in D^{z,topk}_{G,q}} sim\left(KP_d, KP_q\right) \times MI^{z:d}_{u_c,G} \times WRec^{z}_{d,G} \times WT_{d,G}}{\left|D^{z,topk}_{G,q}\right|}, \quad (1)$$

$$GTPR_{G,q} = Max_z(TPR^{z}_{G,q})$$

where KP_d is the document profile of QA d; $MI^{z:d}_{u_c,G}$ is the importance of member u_c that collected QA d for topic z; $WRec^{z}_{d,G}$ is the push score of d within topic z; $WT_{d,G}$ is the weight of d's collection time. The community's preference score on the target QA q, denoted as $GTPR_{G,q}$, is the maximal topic-based preference score over all topics of community G.

Important community members usually play an important role in collecting QAs. A community member u's importance in topic z, $MI^{z}_{u,G}$ consists of two parts: the reputation of member u for collecting/pushing QA documents in community G, $MCR_{u,G}$. and the reputation of member u for answering questions on topic z on behalf of community G, $MAR^{z}_{u,G}$. The importance of community members is defined, as shown in Eq. (2), which adjusts the relative importance between the member's reputation for collecting QA ($MCR_{u,G}$) and for answering questions ($MAR^{z}_{u,G}$) by parameter α:

$$MI^{z}_{u,G} = \alpha \times MCR_{u,G} + (1-\alpha) MAR^{z}_{u,G} \quad (2)$$

$MCR_{u,G}$ is derived from the link analysis of the knowledge collection and push interactions between community members, while $MAR^{z}_{u,G}$ is derived based on the number of best answers obtained by member u. We adopt a link analysis algorithm, PageRank [12] to calculate members' reputations according to the collect/push relationships among community members. $MAR^{z}_{u,G}$ is a normalized number of best answers obtained by member u on topic z for knowledge community G.

QAs pushed by members with greater importance hold more importance and should generally have higher push scores for the community. In addition, a QA with a higher number of recommendations will be given a higher push score. The push score of a QA d in topic z of community G, $WRec^{z}_{d,G}$ is shown in Eq. (3). $MI^{z}_{u_r,G}$ is the importance of recommender u_r in topic z of community G; $UR^{z}_{d,G}$ is the set of members who recommend the collected QA d in topic z of community G.

$$WRec^{z}_{d,G} = 1 + \left[\frac{\sum_{u_r \in UR_{d,G}} MI^{z}_{u_r,G}}{\left|UR^{z}_{d,G}\right|} \times \left(1 - \frac{1}{\left|UR^{z}_{d,G}\right| + 1}\right) \right] \quad (3)$$

The more recent QA documents collected by a community can better reflect the current interest of the community. The time weight of a QA d collected by community G, $WT_{d,G}$ is adopted from the formula given in [13] to compute time factor.

3.3 Complementary Analysis and Recommendations of Complementary QAs

The complementary relationships among QAs include partial complementation and extended complementation. The information provided in the answer part of a collected QA may be partial and incomplete, so the community may wish to search for related QAs to get complete information. However, the information in some related QAs may be redundant to the collected QA and of no interest to the community. QAs that provide related information that is not redundant are called *partially complementary QA*s of a collected QA. Moreover, some information in the collected QA's answer may not be clear, so the community may wish to search for related QAs that contain extended complementary information. Such QAs are called *extended complementary QAs* of a collected QA. Given two QAs, suppose one is called the Collected QA and the other is called a Target QA. We use the cosine similarity measure to determine the degree of similarity between the question of a collected QA and the question of a target QA. If the question similarity is high, the questions of the two QAs are related, so we analyze their answers to derive each answer's novelty. Let A_d and A_k denote the answers of the Collected QA d and the Target QA q, respectively. We measure the novelty of the two answers, A_d and A_q by Eq. (4), which refers to [10]. We use the term vectors generated by TF-IDF to measure the cosine similarity between the answers of the two QAs. If the similarity is high, this means that the answers contain a lot of common information, so their novelty is low:

$$Nov\left(A_q, A_d\right) = 1 - sim\left(A_q, A_d\right) \tag{4}$$

If the question similarity score is high, this implies that the two questions are related; and if the answers are not redundant, i.e. the answer novelty score is high, partial complementation is inferred. If the question similarity is low, the two questions are different; thus, we have to check to see if any term appears in both the answer of the collected QA and the question of the target QA. If such a term exists, we consider that the target QA may contain some information that can explain the unknown subject in the collected QA's answer. However, the answers of the two QAs may be redundant or unrelated, so we have to check the answer novelty and correlation between the collected QA and the target QA. The answer correlation is measured by the correlation of terms in the answers of the two QAs. Extended complementation generally can be inferred if the answer novelty and answer correlation are high. We use the all-confidence metric [14], which measures the mutual dependence of two variables, to derive the answer correlation, as shown in Eq. (5). The correlation between the two answers, A_d and A_q, denoted by $AC(A_d, A_q)$, is derived by summing the all-confidence (x,y) scores for $x \in S_d^A$ and $y \in S_q^A$. Note, that S_d^A / S_q^A is the term set of A_d / A_q:

$$AC\left(A_q, A_d\right) = \sum_{x \in A_q} \sum_{y \in A_d} \frac{P(x \wedge y)}{MAX\left(P(x), P(y)\right)} \tag{5}$$

where *x/y* is the term contained in the answer for document *q/d*; *P(x)* is the probability of documents containing term *x* and $P(x \wedge y)$ is the probability of documents containing both term *x* and term *y*. The dependence of two terms (probability) is measured by the number of documents which contain the two terms returned by the Google search engine. We use a decision tree classification approach to build a classification model and predict the complementary relationships among QAs based on three input variables: question similarity, answer novelty and answer correlation. Specifically, we use Weka's Classification and Regression Tree (CART) model [15] to build a classification model. In the prediction of a target QA, the decision process reaches a leaf node of the classification tree based on question similarities, answer novelties and answer correlations between two QAs. The complementary score of target QA, *q,* to the collected QA, *d*, $CPS_{q,d}$, is the partial or extended probability which can be calculated as the ratio of the number of training cases in the leaf node with a positive label to the total number of training cases in the leaf node.

A target QA may be complementary to very few QAs collected in a community. Therefore, in order to enhance the effect of complementary QA recommendations, we derive the complementary score of target QA *q* to a topic *z* of a community *G*, $CPS^{z}_{G,q}$ by aggregating the complementary scores of a target QA to the QAs collected in *z*, as shown in Eq. (6):

$$CPS^{z}_{G,q} = f_{d \in D^{z}_{G}}\left(CPS_{q,d}\right) \tag{6}$$

where D^{z}_{G} is the set of QAs in topic *z* collected by community *G*; $CPS_{q,d}$ is the complementary score of target QA *q* to collected QA *d*; and *f*() is an aggregation function, such as the average, max or sum of the complementary scores of target QA to the QAs collected in *z*, that can be used to determine the complementary score of target QA to a topic. In the experiment, the max function is applied for measuring the complementary score. Once the complementary score of target QA *q* to a topic *z* $CPS^{z}_{G,q}$ is derived, we consider $CPS^{z}_{G,q}$ to enhance the effect of recommending complementary QAs in deriving community *G*'s preference score on target QA *q*, $GPRC^{topic}_{G,q}$ by the topic-based complementary approach, as shown in Eq. (7):

$$TPRC^{z,topic}_{G,q} = \frac{\sum_{d \in D^{z,topk}_{G,q}} sim\left(KP_d, KP_q\right) \times MI^{z:d}_{u_c,G} \times WRec^{z}_{d,G} \times WT_{d,G} \times \left(1 + CPS^{z}_{G,q}\right)}{\left|D^{z,topk}_{G,q}\right|}, \tag{7}$$

$$GTPRC^{topic}_{G,q} = Max_{z}\left(TPRC^{z,topic}_{G,q}\right)$$

where $CPS^{z}_{G,q}$ is the complementary score of target QA *q* to a topic *z* collected by community *G*. Community *G*'s preference score on target QA *q* of topic *z* is obtained by multiplying the two factors, including the weighted relevance scores of target QA *q* to top-*k* QAs in topic *z* and the complementary score of target QA to topic *z*. Finally, community *G*'s preference score on the target QA document *q*, $GTPRC^{topic}_{G,q}$, is the maximal topic-based preference score over all topics of community *G*. We enhance the effect of the complementary QA recommendation by using the *Max* function to

derive complementary scores of target QA to topics. QAs with high preference scores are used to compile a recommendation list, from which the top-*N* QAs are chosen and recommended to the target user.

4 Experimental Evaluations

4.1 Experiment Design

We evaluate the performance of the proposed approach by using the QA documents collected in knowledge communities at Yahoo! Answers Taiwan. We choose 15 knowledge communities from three domains: computer, medicine and finance. The F1 performance metric [3, 16] is used to evaluate the performance of the proposed approach. F1-measure is the harmonic means of precision and recall. We divide the data set into training data and testing data. The data from each community is separated into two parts, 80% for training data and 20% for testing data. Our proposed methods are compared with the traditional content-based group recommendation method. The content-based group recommendation method consolidates individual profiles to generate group profiles, which in turn are used to filter out items of recommendation. The top-*N* QAs are recommended to the target user.

The traditional ***GP-CB*** method mainly considers the content similarity between the recommended document and the community profile in order to recommend related QAs to the community without considering community topics and QA collection weights. The ***GPT*** method recommends QAs to the community based on the relevance of target QA to the community-topic profiles without considering QA collection weights. A community *G*'s preference score on a target QA *q* is derived by considering the top-*k* relevant QAs in each topic. The ***GTPR*** method uses QA collection weights to derive weighted relevance scores and derive a community *G*'s preference score from top-*k* QAs in each topic. The topic-based complementary method (*GTPRC-T*) recommends QAs to the community not only considering the relevance of QAs and QA collection weights, but also the complementary scores of QAs.

4.2 Experimental Results

A community *G*'s preference is derived by considering the top-*k* QAs rather than all QAs of each topic. Our experiment result shows the recommendation quality is the best when *k* equals 10. Based on the result, we choose top-10 QAs of each topic to derive community preferences for *GPT* method and our proposed methods. The *GTRPC-T* performs better than the *GTPR*. The results imply that considering complementary QAs helps to improve the recommendation quality. Fig. 1 shows the performance comparison (F1 measures) among various recommendation methods. The *GPT* recommends QAs based on top-*k* (top-10) QAs in topics that are most relevant to target QA without considering QA collection weights. The *GTPR* uses QA collection weights to derive weighted relevance scores and derive a community *G*'s preference score from top-*k* QAs in each topic. The *GP-CB* method does not consider the topics and the three QA collection factors. The result shows that the recommendation quality of *GPT* is better than that of the traditional *GP-CB* method. The *GTPR* method

performs better than the *GPT* and the *GP-CB* method. Considering the topic profiles and the QA collection factors can achieve better recommendation performance than the traditional content-based group profiling method. Moreover, the results show that the *GTPRC-T* performs the best among all the methods. The recommendation quality is improved when we consider the complementary scores of the target QAs. In summary, our proposed approach is effective in recommending complementary QA documents to knowledge communities.

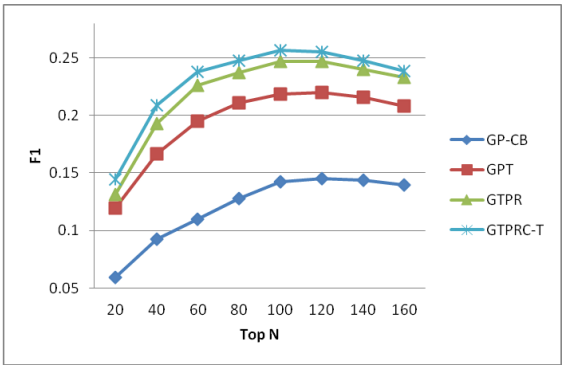


Fig. 1. F1 measures of various recommendation methods

5 Conclusion

In this research, a novel recommendation approach is proposed on recommending relevant and complementary QA documents to knowledge communities of Q&A websites. Recommending complementary QAs is important to increase the effectiveness of knowledge collections. The novel ideas of our proposed approach are as follows: 1) It generates community topic profiles by considering QA collection factors such as community members' reputations in collecting and answering QAs, push scores of QAs and the collection time of QAs from the historically collected QA documents on specific topics. 2) It predicts the complementary scores of QAs based on question similarity, answer novelty and answer correlation. 3) It proposes a QA-based complementary approach and topic-based complementary approach to recommend complementary QA documents. Experimental results show that consider partial or extended complementary QAs help improve the recommendation quality. Moreover, our proposed approach, that considers community topic profiles with QA collection factors and complementary scores of QAs, performs better than traditional recommendation methods. Our proposed approach is effective in recommending complementary QA documents to knowledge communities.

Acknowledgments. This research was supported by the National Science Council of Taiwan under grant NSC 100-2410-H-009-016 and NSC 99-2410-H-009-034-MY3.

References

1. Liu, D.-R., Chen, Y.-H., Kao, W.-C., Wang, H.-W.: Integrating Expert Profile, Reputation and Link Analysis for Expert Finding in Question-Answering Websites. Information Processing and Management 49, 312–329 (2013)
2. Balabanovic, M., Shoham, Y.: Fab: content-based, collaborative recommendation. Communication of the ACM 40(3), 66–72 (1997)
3. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Analysis of recommendation algorithms for e-commerce. In: Proceedings of ACM Conference on Electronic Commerce, pp. 158–167 (2000)
4. Kim, J.K., Kim, H.K., Oh, H.Y., Ryu, Y.U.: A group recommendation system for online communities. International Journal of Information Management 30, 212–219 (2010)
5. McCarthy, J.F., Anagnost, T.D.: MusicFX: an arbiter of group preferences for computer supported collaborative workouts. In: Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work (CSCW), pp. 363–372. ACM, Seattle (1998)
6. O'Connor, M., Cosley, D., Konstan, J.A., Riedl, J.: PolyLens: a recommender system for groups of users. In: Proceedings of the Seventh Conference on European Conference on Computer Supported Cooperative Work, pp. 199–218. Kluwer Academic Publishers, Bonn (2001)
7. Yu, Z., Zhou, X., Hao, Y., Gu, J.: TV Program Recommendation for Multiple Viewers Based on user Profile Merging. User Modeling and User-Adapted Interaction 16, 63–82 (2006)
8. Jameson, A.: More than the sum of its members: challenges for group recommender systems. In: Proceedings of the Working Conference on Advanced Visual Interfaces, pp. 48–54. ACM, Gallipoli (2004)
9. Ma, Q., Tanaka, K.: Topic-structure-based complementary information retrieval and its application. ACM Transactions on Asian Language Information Processing 4, 475–503 (2005)
10. Liu, D.-R., Chen, Y.-H., Lu, P.-J.: Complementary QA-Network Analysis for Q&A Retrieval in Question-Answering Websites (2012) (submitted manuscript)
11. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing and Management 24, 513–523 (1988)
12. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project (1998)
13. Zhang, J., Ackerman, M., Adamic, L., Nam, K.: QuME: a mechanism to support expertise finding in online help-seeking communities. In: Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology, pp. 111–114. ACM (2007)
14. Omiecinski, E.R.: Alternative interest measures for mining associations in databases. IEEE Transactions on Knowledge and Data Engineering 15, 57–69 (2003)
15. Weka: Data Mining Software at URL: `http://www.cs.waikato.ac.nz/ml/weka/`
16. Rijsbergen, C.J.V.: Information retrieval. Butterworth-Heinemann, London (1979)

Using Subtree Agreement for Complex Tree Integration Tasks

Marcin Maleszka and Ngoc Thanh Nguyen

Wroclaw University of Technology, Wyb. Wyspianskiego 27, 50-370 Wroclaw
Marcin.Maleszka@pwr.wroc.pl,
Ngoc-Thanh.Nguyen@pwr.edu.pl

Abstract. Hierarchical structures are common in modern applications. Tree integration is one of the tools for them that is not fully researched. We define a complex tree to model other common hierarchical structures. Complex tree integration is parametrized by specific integration criteria. Sub-tree agreement is a group of criteria that describes the relation of sub-tree number and structure between input trees and the integrated tree. This paper provides several definitions of sub-tree agreement, the most important properties of these criteria, and examples of algorithms based on sub-tree agreement.

Keywords: subtree agreement, tree integration, integration criteria, integration algorithms.

1 Introduction

Hierarchical data formats have become a common occurence in theoretical and practical applications. Even documents are nowadays stored in the XML format. Consequently, there is now a need for tools operating with hierarchical structures.

In our previous research ([10], [11] and others) we have created tools for hierarchical structures integration. By defining a generalized structured called complex tree, we are able to work with most existing structures by translating them to the new one. We have proposed the integration task for complex trees with specific integration criteria as properties of the process. These integration criteria allow easy determining of the aims in each task. In previous papers we proposed multiple such criteria and expanded on some of them, including completeness (all elements from input should remain in output), mininality (the output should not be much larger than the inputs) and optimality (the output should be a median of the inputs). In this paper we focus on the last undescribed group of integration criteria - sub-tree agreement and its variants.

The sub-tree agreement may be understood as a form of completeness - its aim may be for all sub-trees from the input to remain in the output (or similar, depending on the specific criterion used). Unfortunatelly, the methods developed for standard forms of completeness do not work with sub-trees. Consequently, it is necessary to analyze this group of criteria anew. In this paper we present

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 148–157, 2013.

some most important properties of sub-tree agreement, as well as some basic integration algorithms using the criterion.

In our research we are aiming to create a collaborative recommendation system with hierarchically represented profiles. In such a system creating a centroid representing a group of profiles may be done by integrating profiles of this group. Classical approach to this is done by selecting the most average solution – in our research this is called Optimality criterion. Using Sub-Trees as an alternative allows transfering whole areas of user interest to the centroid, which may be preferable in many cases.

The rest of this paper is organized as follows: Section 2 provides a survey of related works from information retrieval and knowledge integration areas; Section 3 defines the complex tree used in this paper as well as the integration task; in Section 4 we define the sub-tree related criteria. Section 5 contains a short list of properties of these criteria, and section 6 provides an example of an algorithm maximizing the criteria. The paper is concluded with some summarizing remarks in Section 7.

2 Related Works

First research on integration of hierarchically structured data may be found in papers such as [1], [4], [6]. In those works a problem of determining a median tree was defined for structures called n-trees. At that time a single n-tree was desired an aggregation of results from multiple biochemical experiments giving different elementary trees. The inconsistencies between the input data had to be eliminated. The proposed solution was finding a so-called median tree that minimized the sum of distances to all other structures. Several approximate solutions for the problem were defined, like clusters [4] and their variations [1], [15], so-called Maximum Agreement Sub-Trees [8] or triads [6].

The Maximum Agreement Sub-Trees [8] were the main inspiration for the set of criteria presented in this paper. As stated before, the paper operates on a simple structure of n-trees. [8] does not provide an integration algorithm, but instead defines means to calculate the "distance" between two trees by calculating the number of common sub-trees. The trees are more similar, if more sub-trees are identical. In this research we use criteria based on this measure.

The domain literature also provides research that defines some basic integration criteria, similarly to the approach used by authors of this paper. The criterion of optimality first appeared in works on n-trees (but it was not always explicitly stated). There are also some works done on classification of schema integration (including hierarchical XML schemas). A survey by Rahm and Bernstein [14] provides an interesting classification of matching approaches. Schema matching in general is a much wider area than just tree integration, but with widespread of hierarchical structures in practical applications, it is also used in the area.

The research done by Do [7] describes some criteria for XML schema integration, divided into four general areas: input criteria, output criteria, quality

measures, and effort criteria. The most relevant criteria for tree integration are named by the authors as: schema information (a criterion based on the size of input schemas), schema similarity (the closer the inputs are, the smaller the space to find the output in will be), element representation (if the elements are found in the output), cardinality (cardinality of relationships in the output), precision and recall (as defined in information retrieval).

Passi [13] provides definitions for the following three basic criteria for integrating XML schemas: completeness (all elements from the initial schemas are in the merged schema), minimality (each unique element is defined only once in the schema) and understandability (in this case, a proper formatting of the output). Although those criteria are based on the criteria created for schema integration, authors modify them for integrating a constructed hierarchical structure. Further work in the area [9] modifies those criteria to postulates known in the literature: completeness and correctness (the integrated schema must contain all concepts presented in any component schema correctly; the integrated schema must be a representative of the union of the application domains associated with the schemas), minimality (if the same concept appears in more than one component schema, it must occur only once in the integrated schema) and understandability (the integrated schema should be easy to be understood for the designer and the end user; this implies that among the several possible representations of results of integration allowed by a data model, the most understandable one should be chosen). The same definitions may be found in other papers, e.q. in [2] and [3]. A thorough analysis of the minimality criterion (although not specifically for the tree structures) was done by Batista and Salgado [3] and Comyn-Wiattiau and Bouzeghoub [5].

For ontologies, integration criteria are gathered in [16], where the authors describe legibility (comprising of minimality - every element appears only once - and clarity - it is easily readable), simplicity (a minimal possible number of elements occur), correctness (all elements are properly defined), completeness (all relevant features are represented) and understandability (the ease of navigation by the user). For ontologies the scope of transformation during the integration process is much larger than for simple data structures. This is based on the fact that not only the amount of knowledge included in the integrated ontology is often greater than the sum of knowledge represented in input ontologies, but also the structure of the output might be very different form each other. The criteria are constructed to describe more what the user would gain after the transformation, less how mathematically correct the effect would be.

3 Complex Tree Integration

The research described in this paper is based on authors' previous work in [10] and [11]. These papers proposed a criteria-based approach to integration, with specific normalized criteria measures. Due to parameterizing the integration process with different criteria these papers shown that it is possible to attain different goals. For example, the completeness criterion was used to measure how

much of initial data (knowledge) was retained after integration; 0 meaning that all data was lost and 1 that all data remained. In this paper the same approach is used for Sub-Tree Agreement criteria.

This research is conducted on a specific structure, the complex tree:

Definition 1. *Complex Tree*
A complex tree is a four $t = (Y, S, V, E)$*, where:*

- Y *is a set of allowed node types in the tree*
- S *is a function determining required attributes for types*
- *the pair* (V, E) *is a a labeled tree, with nodes defined as a triple* (l, y, A)*, where:*
  - l *is the label of the node*
  - y *is the type of the node*
  - A *is the set of attributes of the node*

Additionally, the set of all complex trees will be denoted as $\mathbf{T}$.

This definition of complex tree is a basic extension of the known labeled tree. In fact, most of the criteria researched by the authors work correctly with labeled trees. The complex tree structure was adopted to allow common mathematical description for all practical structures (i.e. n-trees, XML, ontologies) modelled by complex trees.

For the complex tree the integration task may be defined as follows:

Definition 2. *Criteria-based Integration Task*
Given a multiset of N *complex trees*

$$T = \{t_1, t_2, \ldots, t_N\}$$

one should determine a complex tree $t^* \in \boldsymbol{T}$ *which best represents the trees from* T*.*

The use of words „best represents" in the definition means that t^* maximizes some defined criteria measures. In previous works we have proposed several such criteria,

In our research we use a specific description of the criteria, using normalized functions to measure their values. The arguments of these functions are the integrated tree and the input tree (for ease of readability, we use | instead of a comma to distinguish different types of arguments). A criterion is thus defined as follows:

$$C(t^*|t_1, t_2, \ldots, t_N) \geq \alpha$$

This notation represents the requirement that the criterion measure is equal or greater than the given threshold value. Thus, the integration aim is clearly stated.

4 Sub-Tree Agreement

In our previous work [cybernetics] we have defined two main criteria for Sub-Tree Agreement, based on a common definition of Sub-Tree. The Sub-Tree Agreement has several practical applications. For example, one may observe the structure of company employees before and after a corporate merger. It may be necessary to keep large sub-structures as close to the input as necessary as the cost of detail reorganization may be high, thus keeping entire divisions unchanged. This is directly translated to Initial Sub-tree Agreement, which should be maximum in that case. More detailed examples are provided at the end of this section

Definition 3. *Sub-Tree*
A Complex Tree $t_s = (Y_s, S_s, V_s, E_s)$ *is called a sub-tree of a complex tree* $t = (Y, S, V, E)$ *if* $Y_s = Y$, $S_s = S$, *and* (V_s, E_s) *is a connected sub-graph of* (V, E).

Accordingly $ST(t)$ is a set of all sub-trees of t. We will measure the size of a sub-tree or a complex tree $d(t)$ as the number of nodes in it.

Definition 4. *Initial Sub-Tree Agreement*
Initial Sub-Tree Agreement is a measure for a criterion comparing the size of the largest sub-tree from the input trees to be found in the integrated tree.

$$A_I(t^*|t_1, \ldots, t_N) = \frac{max_{t_s \in ST(t_1) \cup \ldots \cup ST(t_N)}\{d(t_s)\}}{d(t^*)} \tag{1}$$

Definition 5. *Final Sub-Tree Agreement*
Final Sub-tree Agreement is a measure for a criterion comparing the size of the largest sub-tree from the integrated tree to be found in the integrated trees.

$$A_F(t^*|t_1, \ldots, t_N) = \frac{max_{t_s \in ST(t^*)}\{d(t_s)\}}{max\{d(t_1), \ldots, d(t_N)\}} \tag{2}$$

Initial Sub-Tree Agreement and Final Sub-Tree Agreement attain the maximum value of 1 if t^* is identical to the largest of the set $\{t_1, \ldots, t_N\}$. They attain the minimum value of 0 if there is no common sub-tree in t^* and any of $\{t_1, \ldots, t_N\}$.

Alternately, the following are also proper sub-tree criteria:

Definition 6. *Input Sub-Tree Agreement*
Input Sub-Tree Agreement is a measure for a criterion comparing the number of unique subtrees in the input complex trees with the number of unique subtrees in the integrated tree.

$$A_{In}(t^*|t_1, \ldots, t_N) = min\{1, \frac{card\{ST(t_1) \cup \ldots \cup ST(t_N)\}}{card\{ST(t^*)\}}\} \tag{3}$$

Definition 7. *Output Sub-Tree Agreement*
Output Sub-Tree Agreement is a measure for a criterion comparing the number of unique subtrees in the integrated complex tree with the number of unique subtrees in the input trees.

$$A_{Out}(t^*|t_1, \ldots, t_N) = min\{1, \frac{card\{ST(t^*)\}}{card\{ST(t_1) \cup \ldots \cup ST(t_N)\}}\} \tag{4}$$

4.1 Sub-Tree Agreement in Practical Applications

Sub-Tree Agreement may be applied in several different practical applications. The most natural application is the case of employee hierarchy. Each node in the complex tree represents an employee (or, in some cases, only a work position). If a node is a parent to another, this represents a person being a direct supervisor of another. Consequently, all descendants of a node are direct and indirect underlings of some employee.

Sub-Tree Agreement becomes necessary in case of reorganizations in such structure. This may occur e.q. during a company merger. Such situation directly translates to an integration task for the employee hierarchies. Different aims are possible during this process, but here we only focus on sub-trees. A sub-tree is a representation of some estabilished department or team in one of the companies. If it is desirable for the departments and teams to be kept intact after the reorganization, the Sub-Tree Agreement should be used. The simplest approach would be using the Output Sub-Tree Agreement.

In such case, if Output Sub-Tree Agreement were to be 1, all sub-trees from the input would remain intact. Consequently, all departments from the source companies would be kept. This may lead to additional connections (or even nodes) to be created in the integrated structure - a situation that is not desirable. Thus, Output Sub-Tree Agreement of $0.8 - 0.9$ may be a better alternative. In this case, while most departments from the source companies remain unchanged, some may be removed to provide a clearer result. The application of the Minimality criterion for the addtional aim of integration may be desirable.

5 Properties of Sub-Tree Agreement

Our research determined that the various Sub-Tree Agreement criteria from the previous section have the following properties. Due to space constraints we only provide proof outlines instead of full proofs. These will be provided in other publications.

Theorem 1. *High values of Sub-Tree Agreement are possible only in cases where the Relationship Completeness and Path Completeness criteria have high values.*

Proof Outline. The Sub-Tree Agreement requires that some sub-structures of the complex tree remain the same in input and output trees. These structures have the same edges (and in lower number - paths), which means that the calculated value of Relationship Completenes (and Path Completeness) has to be high. The opposite does not hold, as the same edges in the tree may not mean the same sub-trees.

Theorem 2. *Output Sub-Tree Agreement is always higher than Input Sub-Tree Agreement.*

$$\forall_{t_1,\ldots,t_N \in \boldsymbol{T}} A_{Out}(t^*|t_1,\ldots,t_N) \geq A_{In}(t^*|t_1,\ldots,t_N)$$

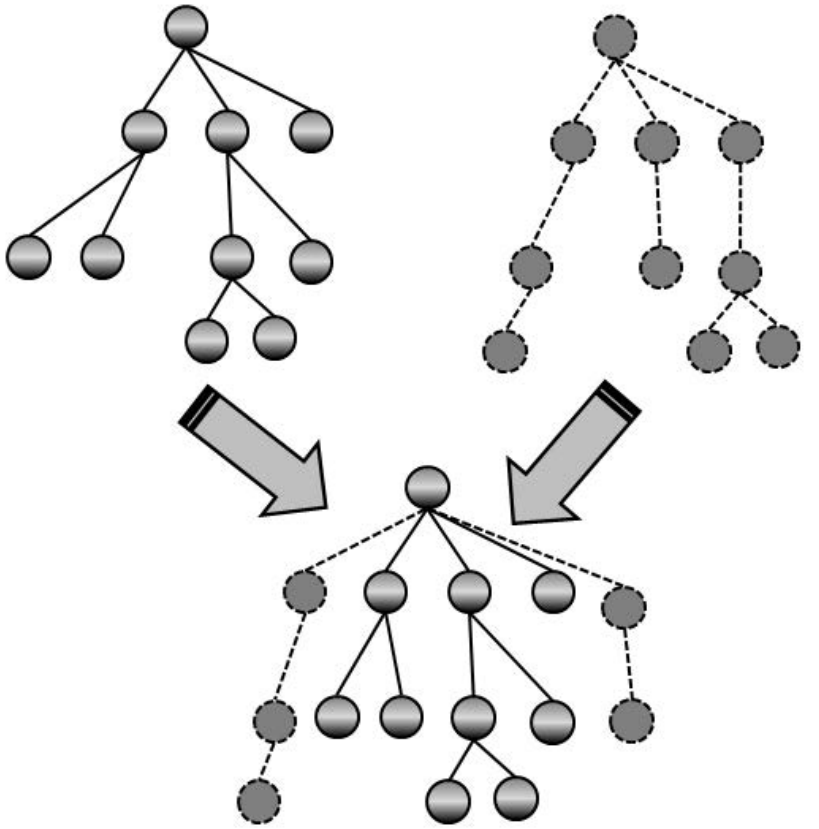


Fig. 1. Example of Sub-Tree based Integration. The left-side tree is included as a whole in the integrated tree. Two out of three main sub-trees from the right-side tree are included. The specific values of sub-tree agreement in this case are: $A_I = 1$, $A_F = 1$, $A_{In} = 1$, $A_{Out} = \frac{5}{6}$.

Proof outline. This is a natural consequence of using set cardinality in calculation. One of the cardinalities is always higher than the other.

Theorem 3. *For specific types of trees, Sub-Tree Agreement is inversly propostional to Precision criterion.*

Proof Outline. The Precision criterion is used to minimize redundancy in the integrated tree. If the same node appears in multiple sub-trees in different input trees, then high Sub-Tree Agreement requires that this node is also present in multiple locations after integration. This leads to low Precision.

Theorem 4. *For specific types of trees, Sub-Tree Agreement is inversly propostional to various Minimality criteria.*

Proof outline. Minimality criteria are used to minimize the size of integrated tree, with various size measures. The proof shows that high Minimality in some variants prevents high Sub-Tree Agreement by reducing the number of allowed sub-trees.

6 Sub-Tree Agreement Based Algorithms

Preliminary research by the authors indicates, that it is not possible to archieve maximum input or output sub-tree agreement in practical applications. Algorithms created must then provide a high value of the criterion, for example above a given threshold. Below we provide an algorithm that guarantees Final Sub-Tree Agreement equal to 1 and Input Sub-Tree Agreement above $\frac{max\{card\{ST(t_1)\},\ldots,card\{ST(t_N)\}\}}{card\{ST(t_1)\cup\ldots\cup ST(t_N)\}}$, as Algorithm 1.

The algorithm works in three simple steps:

1. Create the basic output tree by selecting the input tree with the largest sub-tree.
2. Divide other input trees into sub-trees
3. Attach selected sub-trees to the output tree

Algorithm 1. Basic STA Algorithm

Input: A set $T = \{t_1, \ldots, t_N\}$ of input trees
Output: A single output tree t_{STA}
BEGIN
Set $t_{STA} = t_1$ and $int_{max} = card\{V_1\}$;
foreach *Tree* t_i *in* T **do**
  if $card\{V_i\} > int_{max}$ **then**
    $t_{STA} = t_i$
    $int_{max} = card\{V_i\}$
Create a set of all sub-trees $ST = (ST(t_1) \cup \ldots \cup ST(t_N)) - ST(t_{STA})$
foreach *sub-tree* $st \in ST$ **do**
  if *root of st is a child of the* t_{STA}*'s root* **then**
    Add st to t_{STA}
END

Another simple algorithm is a modification of work done in [4]. In that work the authors use structures defines as clusters, that are sets of tree leafs with

a common ancestor. Such structures are very similar to sub-trees used in this paper. An algorithm modified to maximize some Sub-Tree Agreement consists of following steps:

1. For each input tree create the set of all clusters.
2. Select clusters that are to occur in the integrated tree.
3. Build the integrated tree out of the created clusters.

It may be noted that some new sub-trees may be created by using this approach, so this is not an universal solution. For example, using all clusters from step 1 in step 2 will maximize Input Sub-Tree Agreement and Final Sub-Tree Agreement, but Initial Sub-Tree Agreement and Output Sub-Tree Agreement may in some cases be smaller.

7 Conclusions

In this paper the various sub-tree agreement criteria were described. These criteria may be useful in multiple applications, with the simplest example being the case of reorganizing companies.

Multiple properties were presented for the defined criteria, with short outlines of the proofs. Relations between different criteria are the most important properties, as common applications require the use of multiple criteria – this represents multiple parallel aims in a single integration task.

Examples of basic integration algorithms were provided to show the applicability of the approach used.

In our future research we aim to use Sub-Tree Agreement criteria in a collaborative recommendation system, as described in the introduction. Initial and Input Sub-Tree Agreement types may be used to define the minimal number of user interest hierarchies that we want to transfer unchanged from the input to the integrated centroid of the group. The earlier approach of finding an average profile as the centroid is mostly satisfying, with the result representing a group of users. By slightly diverging from that solution, through the introduction of Sub-Tree Agreement criteria, we may be also able to represent more heterogeneous groups in a situation where splitting them is impossible.

Acknowledgment. This research was co-financed by Ministry of Science and Higher Education grant no. B20073/I32 and by the Fellowship co-financed by European Union within European Social Fund.

References

1. Adams, E.N.: N-Trees as Nestings: Complexity, Similarity, and Consensus. Journal of Classification 3, 299–317 (1986)
2. Batini, C., Lenzerini, M., Navathe, S.B.: A comparative analysis of methodologies for database schema integration. ACM Computing Surveys (CSUR)

3. Batista, M.D.C.M., Salgado, A.C.: Minimality Quality Criterion Evaluation for Integrated Schemas. In: Proceedings of 2nd International Conference on Digital Information Management, ICDIM 2007, pp. 436–441 (2007)
4. Barthelemy, J.P., McMorris, F.R.: The Median Procedure for n-Trees. Journal of Classification 3, 329–334 (1986)
5. Comyn-Wattiau, I., Bouzeghoub, M.: Constraint Confrontation: An Important Step in View Integration. In: Loucopoulos, P. (ed.) CAiSE 1992. LNCS, vol. 593, pp. 507–523. Springer, Heidelberg (1992)
6. Day, W.H.E.: Optimal Algorithms for Comparing Trees with Labeled Leaves. Journal of Classification 2, 7–28 (1985)
7. 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)
8. Farach, M., Przytycka, T.M., Thorup, M.: On the agreement of many trees. Information Processing Letters 55, 297–301 (1995)
9. Madria, S., Passi, K., Bhowmick, S.: An XML Schema integration and query mechanism system. Data & Knowledge Engineering 65, 266–303 (2008)
10. Maleszka, M., Nguyen, N.T.: Path-Oriented Integration Method for Complex Trees. In: Jezic, G., Kusek, M., Nguyen, N.-T., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2012. LNCS (LNAI), vol. 7327, pp. 84–93. Springer, Heidelberg (2012)
11. Maleszka, M., Nguyen, N.T.: A Method for Complex Hierarchical Data Integration. Cybernetics and Systems 42(5), 358–378 (2011)
12. Nguyen, N.T.: Inconsistency of Knowledge and Collective Intelligence. Cybernetics and Systems 39(6), 542–562 (2008)
13. Passi, K., Lane, L., Madria, S., Sakamuri, B.C., Mohania, M., Bhowmick, S.: A Model for XML Schema Integration. In: Bauknecht, K., Tjoa, A.M., Quirchmayr, G. (eds.) EC-Web 2002. LNCS, vol. 2455, pp. 193–202. Springer, Heidelberg (2002)
14. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. The VLDB Journal 10, 334–350 (2001)
15. Stinebrickner, R.: s-Consensus Trees and Indices. Bulletin of Mathematical Biology 46, 923–935 (1984)
16. Trinkunas, J., Vasilecas, O.: Ontology Transformation: from Requirements to Conceptual Model. Scientific Papers, University of Latvia, Computer Science and Information Technologies 751, 52–64 (2009)

Data Sets for Offline Evaluation of Scholar's Recommender System

Bahram Amini, Roliana Ibrahim, and Mohd Shahizan Othman

Dept. of Information Systems, Faculty of Computer Science and Information Systems,
Universiti Teknologi Malaysia (UTM), Malaysia
avbahram2@live.utm.my, {roliana,shahizan}@utm.my

Abstract. In an offline evaluation of recommender systems, data sets have been extensively used to measure the performance of recommender systems through statistical analysis. However, many data sets are domain and application dependent and cannot be engaged in different domains. This paper presents the construction of data sets for the offline evaluation of a scholar's recommender system that suggests papers to scholars based on their background knowledge. We design a cross-validation approach to reduce the risk of false interpretations by relying on multiple independent sources of information. Our approach addresses four important issues including the privacy and diversity of knowledge resources, the quality of knowledge, and the timely knowledge. The resulting data sets represent the instance of scholar's background knowledge in clusters of learning themes, which can be used to measure the performance of the scholar's recommender system.

Keywords: Recommender System, Evaluation, Data Set, Scholar Domain.

1 Introduction

A recommender system customizes its outputs to individual user requests. It implies that the objective or the goal of recommender systems is to guide the users to interesting items looking for. The evaluation of recommender systems should assess how much of this goal has been achieved [1]. Most of the time, recommender systems are evaluated in an offline setting with appropriate pre-collected data sets before doing an online performance test in real environments [2].

Each recommender algorithm models the user's behavior and focuses on a specific recommendation problem. In fact, data sets simulate the user behavior in real mode and are used to measure the algorithm's performance. Good examples of such data sets are MovieLens, Jester, Book-Crossing, or the EachMovie data set [3]. These data sets are employed as a common standard to evaluate the recommendation algorithms based on the known machine-learning measures such as accuracy, coverage, novelty, diversity, and so on [4].

However, the data sets are domain and application dependent in which one data set cannot be applied for evaluation to different recommender systems. As an example, a data set of measuring the performance of a recommender system in e-commerce

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 158–167, 2013.

differs from the scholar's domain in several ways: In e-commerce, the used data are often simple web server access logs or ratings of users on items (behavioral information) while in the scholar's domain very sophisticated information such as research objective, scholar's background knowledge, and learning style (conception information) is used. Even recommender systems in the same domain expose different properties that require algorithm-aware data sets for effective evaluation.

Regarding our previous study, a recommender system for researchers that re-ranks the articles based on the scholar's background knowledge, the instances of candidate data sets should contain information such as scholars ID, publication body, research interests, homepage address, and studied articles. An exhaustive investigation of publicly available data sets for the learning domain on the web indicated little match between the existing data set instances and our evaluation requirements. In the best cases, relevant data sets for researcher domain contain links between authors and their bibliographic information such as publications, citation indices, and publisher names that are not useful for our purpose [5].

These issues motivate the development of special purpose and application centric data sets which fit with the type of our evaluation. Moreover, to develop the domain-dependent data sets, four issues must be addressed including:

- **Public vs. private:** Many digital libraries such as Springer and ACM do not allow unauthenticated users to access the main parts or the whole publications, making such data unavailable for processing.
- **Diversity of data sources:** The obscurant property of scholar knowledge is the diversity and homogeneity of the knowledge items, residing on different locations such as digital libraries, homepages, and so forth.
- **Quality of knowledge:** In many circumstances, the topic of articles does not match exactly with the main interest or prior knowledge of scholars because they are co-authors who are supervising the main author. Moreover, many papers contain so much formula, algorithm codes, figures, and similar non-textual knowledge that burden extra text processing which inherently bias the topic detection.
- **Timely knowledge:** The scholar's knowledge is continually growing and perennial maturity. Scholars who were interested in a topic few years ago may be interested in different topics now. Therefore, modeling the outdated scholar's knowledge produces false interpretations of scholars' interests, and in turn, lessens the recommendation accuracy.

This paper develops a practical framework for offline evaluation of a content-based recommender system for digital libraries in which scholar's profiles represent their background knowledge. We also describe the construction of different data sets for offline evaluation in terms of accuracy and precision. We identify and collect the knowledge items comprising three collections of scholar's background knowledge.

The rest of the paper has been organized as follows: Section 2 describes the related works and highlights the lack of appropriate data sets for scholar recommenders. Section 3 represents the proposed framework and the methodology in detail. Section 4 discusses the implementation for construction three data sets. In Section 5 and 6 the discussion and conclusion are explained.

2 Related Work

Mendeley's data set [6] contains three collections of information about the scholar libraries. The first one includes a set of articles that appear in scholarly libraries featured by UserId and ArticleID. The second one provides readership information, showing which article users have already read using Mendeley Desktop, featured by UserId, ArticleID, and ReadingStatus. The last one shows which articles the user has marked with "stars" sign, featured by UserId, ArticleID, and StarStatus.

The PSLC (Pittsburgh Science of Learning Center) DataShop [7] is a data repository that provides access to a large number of educational data sets derived from the intelligent tutoring systems. It contains 270 data sets which record 58 million learner actions. Similarly, LinkedEducation.org [8] provides an open platform of data for educational purposes including the structure of organizations, institutions, courses, learning resources and interrelationships between academicians.

The Organic.Edunet data set [9] collects data from the Organic.Edunet Web portal which is a learning portal for organic agriculture educators. It provides more than 10,000 learning resources collected from eleven federal institutional repositories. The data set starting from January 2010 and includes information about 345 tags, 250 ratings and 325 textual reviews upon 3 different dimensions: the relevance to the organic theme, the usefulness of a resource, and the quality of metadata.

Also, a few relevant data sets for learning domain are described in dataTEL data collection site [10]: CAM for MACE which contains a data set for advanced graphical metadata access to learning resources in architecture, stored in different repositories all over Europe; SidWeb which is private and contains the data from a learning management system, posts on forums, creation or editing of resources, and practices of students on quizzes. The data set is a complete reflection of the data obtained for a period of 4 years and around 4 million events are recorded on 35 thousands resources.

However, existing data sets lack information content; full documentation, or irrelevant to the scholar domain, which particularly impede the task of content analysis and theme extraction. Indeed, our approach rather focuses on the domain-relevant data sets, which provide scholar's conception in terms of knowledge items over a time interval.

3 The Measurement Framework

Figure 1 represents the framework of offline experimental analysis using data sets. As shown, we firstly develop three different data sets and then extract the key terminologies from the content of those data instances. Next, the instances are disambiguated by means of Wikipedia to transform to homogeneous and discriminant key terms. Since the extracted terms are in large volume, the clustering method is successively applied. After extracting the key terms, it is crucial to organize the scholar's knowledge into relevant research topics. After that, key terms are sampled based on the features of the scholar's domain. Finally, the measurement of recommender system takes place using the training and testing parts of data sets.

3.1 Methodology

The creation of the data set is motivated by several factors including the desire of having real data, reasonable size in term of adequacy and processing time, availability and open access, expected quality, and application tailored.

To neutralize any bias inherent in particular data sources and validate the empirical results, we employ a cross-validation or so-called triangulation approach which reduces the risk of false interpretations by drawing upon multiple independent sources of information [11]. It relies on the idea that no single source of data can fully explain a phenomenon. Triangulation, as an analytical approach, integrates multiple data sources to improve the understanding of a problem. It is particularly useful when the data are rare, abundant but not similar, rapid response is needed, or the best single data source is not available.

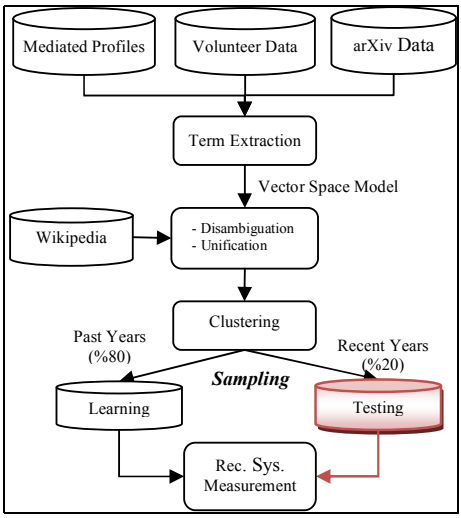


Fig. 1. The framework of experimental analysis of recommender system for scholar's domain

Thus, we collect three independent collections of data from different information resources. We also concern about the four issues mentioned in Section 1 dealing with the collecting and organizing the data sets. We engaged the Information Retrieval techniques [12] and adapted the common guidelines for constructing data sets for such quantitative evaluation.

To address the issues, we employ different strategies by focusing on three kinds of knowledge resources in the field of Computer Science as follows:

- **Volunteer Researchers:** It collects scholarly knowledge of 25 volunteer researchers at FSKSM, UTM who provide relevant papers in line with their research topics. From the scholar's name and affiliation, we would be able to locate their academic homepages, research interests and published papers, which provide complementary knowledge about the individual scholar.
- **Digital Library (DL):** It accesses to the free domain relevant digital libraries such as arXiv and extracts a collection of 65 scholar's publications who were the first or second author of the articles and published at least 10 publications during the years 2003 and 2012.

- **Mediated Profiles:** It locates recently published papers of 35 international full researchers whose authors were the first or second one. It collects knowledge from the mediated profiles provided by advanced digital libraries such as Google Scholar, ACM Profile Page, and Microsoft Academic Research.

3.2 Data Sampling

The "data sampling" task partitions each primary data set into the so-called "training" and "testing" parts. The training part is used to learn the user model or set up the algorithms in the analysis phase, while the testing part is used to evaluate the user model to check whether the proposed model performs well and predicts unseen data accurately.

Furthermore, it is a common approach to apply "standard" random sampling, i.e., the 80/20 partitioning, which is a random sampling without the replacement. It selects 20 percent of the instances for testing and the remaining 80 percent for training phase. Since sampling can lead us to an over-specialization to the data set parts (training and testing), the training process is repeated several times and the average performance of the learned models is calculated as the final performance. This process is called as cross-validation [13].

Although the standard random sampling is sufficiently applicable in most cases, we apply the sampling method for the scholar's domain in a different way: we sample only from the most recent publications because the recommender system would be predicting in a real scenario where the portion of user interests is measured in recent years [14]. For instance, if a scholar has 20 publications in years 2006 to 2011, we consider 16 publication of years 2006-2010 as the training and the remaining in years 2010-2011 as the testing part.

3.3 Term Clustering

Since there are many extracted key terms, the clustering method is required. The issue with the scholar's knowledge is that the scholar's focus has been changing over the time from one topic to another. Thus, after extracting the key terms, it is crucial to organize the scholar's knowledge into relevant research topics. Another issue is that a researcher may do search on several topics simultaneously. So, the key question is that what collection of key terms constitutes the proper knowledge corresponding to different context of the scholar's research. To address these issues and organize the research areas, we partition the key terms into categories of semantically related terms using k-bisecting clustering algorithm [15] which is a top-down hierarchical method.

To measure the distance between two terms in the clustering algorithm, the semantic similarity (semantic relatedness) method is employed [16]. It is based on the simplified Jensen-Shannon method [15] in which all key terms appearing in the same Wikipage are similar. For example, if term A appears very frequently in a Wiki page entitled B, then A and B are semantically related and B is the superior topic of A. Finally, such clusters represent the scholar's knowledge in several topics whereas underlying key terms referring to the same topic are grouped together.

4 Implementation

This section describes the details and directions of implementing the proposed approach. We explain a stepwise approach to collect and organize the three data sets, deciding on the parameters, and report the statistics in each step. The main objectives are simplicity, non-intrusive gathering data, and cross-checking of several knowledge resources.

4.1 Volunteer Researchers

For this data set, we invited volunteer scholars, both master and PhD students, who were studying in any sub-fields of Computer Science. They were asked to provide a collection of 20 to 25 relevant articles which have been studied in their recent research. From the scholar names, we located their home page, their publications and identified complementary keywords as their research interests. Then, we did preprocessing on the articles to build several corpora of plain text and then removed knowledge-free parts including publishing information on header and footer, images, formulas, tables, acknowledgments, and references parts. Next, important concepts or key phrases- so called "theme", are extracted from the articles. The "key phrase" refers to the multi-word terms that best describe the main topics of the articles [17].

We employed statistical approach which considers all important terms in the domain as potential concepts and took into account the quantitative and heuristic metrics to measure the importance of the terms. For this purpose, we employed Provalis' WordStat content analyzer module under QDAMiner 4 tool. This module supports domain independent, statistical corpus-based, and multi-word extraction method. The reliability as well as the validity of this module for recent versions are approved in many experiments [18].

Having employed the content analyzer, we automatically extracted about 220 multi-word key phrases from the text corpora based on the popular TF*IDF measurement with associated term's weights. This measure relies not only on term frequencies but also on the internal structure of the candidate terms. In facts, the weights represent the importance degree of key phrases in the field of respected scholar's research. Each term is represented as a vector of ⟨T, TF, %Cases, Score⟩, where T is a term, TF is frequency of T in the corpus, %Cases is the percentage of articles containing T, and Score is the TF*IDF measure.

There are many outliers or impure terms- the terms that refer to no domain concepts and describe no theme of the articles in a corpus [19]. To address this problem, the following steps are applied to the candidate terms to remove the less important terms and sieve irrelevant concepts:

1. The terms that are general, i.e. non discriminant, neutralized, and uninformative, which give no sense of conception in the domain are removed. It includes phrases that can be changed without affecting the main topics.
2. Variant terms in each list are aggregated together based on either syntactic or semantic similarities [20]. The variant terms which appear in different forms are

either inflectional variants (singular-plural variants), compounding variants, orthographic variants, misspellings, and abbreviations [21]. For example, the terms "BP Algorithm" and "Back Propagation Algorithm" are aggregated together because the first one is an abbreviation of the other. The corresponding score is equal to the maximum value of their scores.

4.2 Free Digital Library (arXiv)

The arXiv digital library provides an open access to 784,152 articles in Computer Science among all branches of Mathematics, Physics, Quantitative Biology, and Statistics. Although, there is a bulk downloading service for arXiv dumps, the performance of downloading terabytes of data and processing the articles in the Computer Science was a matter of concern. Thus, we directly searched for papers in the respected field, published between the years 2003 and 2011, and retrieved both article titles and author names in a page by page manner into Excel sheets, labeled with the publication year. For instance, the year 2011 encompassed 40,907 items including article's title and author's name. We put in order and selected the authors who have more than 10 publications by identifying the author's property in each entry and extracted 302 authors' names in total.

In the next step, we further filter out the authors whose research areas were non Computer Science, non-English, and involved long articles (more than 30 pages). Since arXiv restricts running crawler agents, due to the performance issue, we searched the library using the individual author name and retrieved all articles of each author using DownThemAll[1] tool. It is an add-ins for the popular internet browsers, which locates and downloads all visible pdf links on an HTML page in parallel and stores the associated files to the local storage. Fortunately, no tuning and setting up parameters was required.

We finally stored the articles into the distinctive corpora labeled by the author ID. We then randomly sampled 65 corpora, which contained the standard length of papers, having 8 to 30 normal pages. The challenge with arXiv library was that a great number of so-called spam articles were erroneously tagged with the Computer Science category. Thus, we discard such articles by masking article's keywords with the key terms of Computer Science ontology. For example, the main topics of such articles were Telecommunication, Probabilistic, and Signal Processing which do not fit our application.

4.3 Mediated Profiles

To effectively extract the mediated profiles from digital libraries, we employed ArnetMiner[2] web service [22] as an integrated interface. It automatically identifies, locates, and extracts the researcher profiles from the web by using social network analysis (SNA) [23] and Information Integration techniques. The scholar's profile

[1] http://www.downthemall.net/
[2] http://www.arnetminer.org

page integrates various information including personal information, citation statistics, dated research interest, educational history, expertise field, and publication records from the major federated digital libraries. We effectively engaged the scholar information provided by ArnetMiner as an index to the ACM Institutional Profile Pages for retrieving scholars' knowledge including published papers, supplied keywords for individual publications, and author profiles between the years 2003 and 2012.

For this data set, we collected the research interests and the key terms from their profiles for the publications dated in recent years (2003-2012). We directly collected key terms supplied by the researchers to the ACM digital library. Since the resulting terms were in large volume, the clustering method is applied. Moreover, the key terms had different significance to the scholar's knowledge. Thus, we simulated the concept of TF*IDF as a new weighting scheme over the collection of terms as follows: 1) The value 10 is assigned to the key terms of the year 2012, value 9 to the key terms of the year 2011, and former values to other years down to 2003 iteratively. 2) The key terms which are repeated as well as semantically similar are grouped together and the summation of corresponding weights is assigned. 3) The values are normalized to the base value 100 by dividing each value to the total sum. The resulting data set contained 3090 key terms in total and 88 key terms in average.

5 Discussion

We constructed three data sets by drawing upon multiple independent sources of information. We employed the triangulation approach that reduces the risk of false interpretations and applied different strategies by focusing on three different kinds of knowledge resources in the field of Computer Science: Volunteer Researchers, Digital Library, and Mediated Profiles. Table 1 shows the strategies that successfully addressed the issues of construction data set, as mentioned in Section 1. As depicted, volunteer strategy supports all four features but is intrusive and ineffective, while the other two strategies address three issues sufficiently. The overlapping parts among the strategies confirms the exerted triangulation approach and validates the functionality of our data sets.

Table 1. Three construction strategies which address four types of data sets issues

Strategy /Issues	Public vs. Private	Diversity of data sources	Quality of knowledge	Timely knowledge
Volunteers	√	√	√	√
Free DL (arXiv)	√	√		√
Mediated Profiles	√	√	√	

Additionally, our approach advances the body of knowledge in developing the data set for scholar's recommender system and contributes to the following aspects:

1. Semantic heterogeneity: Since the terms are extracted from the independent and autonomous knowledge resources, heterogeneity exists. Heterogeneity among different knowledge items (themes) has been resolved by using Wikipage contents.

2. Multiple focus of the scholar's research: Since full researchers contribute to different concurrent research, likely with different topics, we employed clustering method to distinguish either overlapping or non-overlapping research topics. In our approach, shared terms are assigned with different weight tags to discriminate the importance of terms in each cluster/topic.

Moreover, we examined the idea of using mediated scholar profiles supplied by some digital libraries as well as the use of a weighting schema which discriminated the key terms in each research topic. The comparison of our approach with the related works indicates similarities as well as differences: The most significant improvement is the engagement of real scholar's knowledge, implicit and as well as non-intrusive theme extraction, and the use of academic social network (ArnetMiner).

6 Conclusion

In this paper, we presented the construction of three application-tailored data sets for evaluation of a scholar's recommender system. It captured the sufficient and reliable key terms (theme) from qualified scholar's knowledge resources in the context of Computer Science. We employed a cross-validation or triangulation approach which reduces the risk of false interpretations by relying on multiple source of information. The resulting data sets were in reasonable size in terms of adequacy and processing time, which support the offline evaluation of the recommender system. We also engaged an academic social network, ArnetMiner, for reliable integrating various scholar's profiles on the Web.

Acknowledgement. This work is supported by the Ministry of Higher Education (MOHE) and Research Management Centre (RMC) at the Universiti Teknologi Malaysia (UTM) under Exploratory Research Grant Scheme Category (VOT R.J130000.7828.4L051).

References

1. Burke, R.: Knowledge-based recommender systems. In: Encyclopedia of Library and Information Systems, pp. 180–200. Marcel Dekker (2000)
2. Shani, G., Gunawardana, A.: Evaluating Recommendation Systems. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.) Recommender Systems Handbook, pp. 257–297. Springer, Heidelberg (2011)
3. Schafer, J.B., Frankowski, D., Herlocker, J., Sen, S.: Collaborative Filtering Recommender Systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) Adaptive Web 2007. LNCS, vol. 4321, pp. 291–324. Springer, Heidelberg (2007)
4. Drachsler, H., Hummel, H., Berg, B., Eshuis, J.: Evaluating the Effectiveness of Personalised Recommender Systems in Learning Networks. In: Learning Network Services for Professional Development, pp. 95–113. Springer, Heidelberg (2009)
5. Yao, L., Tang, J., Li, J.: A Unified Approach to Researcher Profiling. In: IEEE/WIC/ACM International Conference on Web Intelligence, pp. 359–365 (2007)

6. Jack, K., Hammerton, J., Harvey, D., Hoyt, J.J., Reichelt, J., Henning, V.: Mendeley's Reply to the DataTEL Challenge, pp. 1–3. Elsevier, Procedia Computer Science (2010)
7. Stamper, J., Koedinger, K., Baker, R.S.J.d., Skogsholm, A., Leber, B., Rankin, J., Demi, S.: PSLC DataShop: A Data Analysis Service for the Learning Science Community. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010, Part II. LNCS, vol. 6095, pp. 455–455. Springer, Heidelberg (2010)
8. Manouselis, N., Drachsler, H., Verbert, K., Duval, E.: TEL as a Recommendation Context. In: Manouselis, N. (ed.) Recommender Systems for Learning, pp. 21–37. Springer, New York (2010)
9. Manouselis, K., Kosmopoulos, N., Kastrantas, T.: Developing a Recommendation Web Service for a Federation of Learning Repositories. In: International Conference on Intelligent Networking and Collaborative Systems, INCOS 2009, pp. 208–211 (2009)
10. Verbert, K., Drachsler, H., Manouselis, N., Wolpers, M.: Dataset-driven Research for Improving Recommender Systems for Learning. In: Proceedings of the 1st International Conference on Learning Analytics and Knowledge, LAK 2011, pp. 44–53 (2011)
11. Karsten, J., Karen, A.J.: Using triangulation to validate themes in qualitative studies. Qualitative Research in Organizations and Management: An International Journal 4(2), 123–150 (2009)
12. Weiss, S.M., Indurkhya, N., Zhang, T., Damerau, F.J.: Information Retrieval and Text Mining. In: Fundamentals of Predictive Text Mining, pp. 75–90. Springer, Heidelberg (2010)
13. Ricci, F., Rokach, L., Shapira, B., Kantor, P.B.: Recommender Systems Handbook. In: Recommender Systems Handbook, pp. 63–95. Springer, Heidelberg (2011)
14. Amatriain, X., Jaimes, A., Oliver, N., Pujol, J.M.: Data Mining Methods for Recommender Systems, pp. 39–72. Springer Sience+Business Media (2011)
15. Wartena, C., Brussee, R.: Topic Detection by Clustering Keywords. In: Proceedings of the 19th International Conference on Database and Expert Systems Application, DEXA 2008, pp. 2–6 (2008)
16. Zhiqiang, L., Werimin, S., Zhenhua, Y.: Measuring Semantic Similarity between Words Using Wikipedia. In: International Conference on Web Information Systems and Mining, pp. 251–255 (2009)
17. Medelyan, O., Witten, I.H., Milne, D.: Topic Indexing with Wikipedia. In: Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, pp. 19–24 (2008)
18. Provalis Research, QDA Miner version 4.0 [Computer Software]. Provalis Research, Montreal, Canada (2011)
19. Zhang, K., Xu, H., Tang, J., Li, J.: Keyword Extraction Using Support Vector Machine. In: Yu, J.X., Kitsuregawa, M., Leong, H.-V. (eds.) WAIM 2006. LNCS, vol. 4016, pp. 85–96. Springer, Heidelberg (2006)
20. Budanitsky, A., Hirst, G.: Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In: Workshop on WordNet and Other Lexical Resources, NAACL 2001, pp. 29–34 (2001)
21. Kozakov, L., Park, Y., Fin, T., Drissi, Y.: Glossary extraction and utilization in the information search and delivery system for IBM Technical Support. IBM Systems Journal 43(3), 546–563 (2004)
22. Tang, J., Zhang, J.: ArnetMiner: Extraction and Mining of Academic Social Networks. In: KDD 2008, pp. 990–998. ACM, Las Vegas (2008)
23. Butts, C.T.: Social network analysis: A methodological introduction. Asian Journal Of Social Psychology 11(1), 13–41 (2008)

A Method for Collaborative Recommendation in Document Retrieval Systems

Bernadetta Mianowska and Ngoc Thanh Nguyen

Institute of Informatics, Wroclaw University of Technology,
Wybrzeze Wyspianskiego 27, 50-370, Wroclaw, Poland
Bernadetta.Mianowska@pwr.wroc.pl,
Ngoc-Thanh.Nguyen@pwr.edu.pl

Abstract. The most common problem in the context of recommendation systems is "cold start" problem which occurs when new product is recommended or a new user becomes to the system. A great part of systems do not personalize a user until they gather sufficient information. In this paper a novel method for recommending a profile for a new user based only on knowledge about a few demographic data is proposed. The method merges a content-based approach with collaborative recommendation. The main objective was to show that based on knowledge about other similar users, the system can classify a new user based on subset of demographic data and recommend him a non-empty profile. Using the proposed profile, the user will obtain personalized documents. A methodology of experimental evaluation was presented and simulations were performed. The preliminary experiments have shown that the most important demographic attributes are gender, age, favorite browser and level of education.

Keywords: user profile, collaborative recommendation, demographic data-based classifier.

1 Introduction

Recommender systems have been developed in variety of domains: documents, books, movies, products, etc. Due to the information overload in the Internet, it is difficult to obtain the relevant information. Therefore, recommender systems play an important role in filtering and customizing the desired information [4]. The system tries to guess what would be interesting for a user. Recommender systems are proposed to discover the implicit interests in user's mind based on the usage logs [11].

In this paper a novel method for collaborative recommendation is presented. In classical approach objects to classify are described by the set of attributes. Information about objects is divided into training and testing set. The first – training set is used to determine a classifier that in the next step will be used for classification of objects in the second – testing set. On both levels of the clustering procedure the whole set of attributes is used. Users are divided

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 168–177, 2013.

into groups of common values of attributes. The system recommends documents which are personalized for each group. All users in a group are proposed the same documents.

Our approach is based on the following idea. The objects are described by two independent set of attributes – users are described by demographic data and usage data is represented in the form of a profile. In the first step, the users are clustered based on usage data. The partition of users is used to determine the minimal set of attributes coming from the demographic data that the partition obtained using select demographic data is as close as possible to partition obtained in the first step.

Using proposed method it is enough to ask user only about actually important demographic information that user needs to introduce when starting interaction with the system. Each group of users is described by minimal set of attributes determined in the previous step. A new user should be classified into the group of users which are the most similar in terms of demographic information. The system does not need to wait until user has enough usage data to recommend him some propositions. In our approach a non-empty profile will be recommended for the new user. Based on the knowledge about other users in the same group the first profile will be determined. While user is interacting with the system and his usage data are gathered, his profile will be adapted to his preferences.

To verify recommendation methods one should check if proposed ranking of documents are useful for the user. In our approach it is not obvious how to check if the proposed profile reflects actual user preferences. To manage the problem the authors prepare the methodology of result verification.

The rest of the paper is organized as follows. In the Section 2 we present the overview of classical approaches to clustering. The model of user is presented in Section 3. Section 4 describes in detail a method for determining a minimal set of attributes. In Section 5 the way of simulating user activity, experimental evaluations are presented and obtained results are discussed. In the last Section 6 we gather the main conclusions and future works.

2 Related Works

Collaborative recommendation is a very popular area of the information retrieval domain. Due to the information overload in the Internet, it is very useful for a user when the system proposes him interesting information. The main purpose of recommender systems is to predict users' future likes and interests.

In classic approach presented in [2] collaborative recommendation is understood in two ways. The first one – given a new item to be recommended, predict the rating that the user would give. The second one – given a new user, find the best items and their ratings for being recommended, showing the results ordered by predicted rating.

Recommender systems are classified into following categories [5]:

1. Content-based recommendations: Recommended objects are those with content similar to the content of previously preferred objects of a target user.

2. Collaborative recommendations: Recommended objects are selected on the basis of past evaluations of a large group of users. They can be divided into:
   - Memory-based collaborative filtering: Recommended objects are those that were preferred by users who share similar preferences as the target user, or, those that are similar to the other objects preferred by the target user.
   - Model-based collaborative filtering: Recommended objects are selected on models that are trained to identify patterns in the input data.
3. Hybrid approaches: These methods combine collaborative with content-based methods or with different variants of other collaborative methods.

When new users enter the system, there is usually insufficient information to produce recommendation for them [5]. The usual solutions of this problem are based on using hybrid recommender techniques combining content and collaborative data and sometimes they are accompanied by asking for some base information (such as age, location and preferred genres) from the users.

Different approach is presented by the authors of [3] and [12] where demographic filtering methods are developed. They propose some attributes to describe the user: age, gender, occupation. The creation and management of personalized recommendations require mainly three distinct and important components: a user profile, an algorithm to update the profile given usage/input information, and an adaptive tool that exploits the profile in order to provide personalization.

Su et al. [11] has noted that the most important problems with recommender systems are: cold-start, first-rater, sparsity and scalability. They have proposed a method that integrates multiple contents and collaborative information to predict users' preferences based on the fusion of Rough-Set and Average-category-rating.

The next aspect of recommendation systems is verification of its quality. The result of recommendation is e.g. list of documents or products ranked according to user preferences. The quality in such systems are usually understood in terms of accuracy metrics. Most of recommendation algorithms are tested using existing benchmarks.

Ahn [1] presents the review of similarity measures often used in collaborative filtering that are limited to be used in new user cold-start situations where only a small number of ratings are available for similarity calculation. The problem is even more amplified by the sparsity of available data. A new measure proposed by him takes into account proximity, impact and popularity. The measure is tested using MovieLens, NetFlix and Jester benchmarks. Cold-start problem is simulated by generating 1-20 ratings for movies by a new user.

3 User Model

The most important problem in our approach is to recommend a profile for a new user instead of document rating prediction. The proposed approach is

justified while user interests are changing with time. User profile is a model of his preferences and stores information about his current interests.

In our model the user profile contains demographic data and usage data. The first part is information that can be provided by the user at the beginning of his interaction with the system and the system uses it in the classification procedure. The second part of the profile is connected with user activities in the system. Information about documents that are relevant for the user is the most important for the system, as it shows real user preferences. The user can declare fields of interests but the chosen domains might not reflect real user preferences. These two part of user profile are described in details in the next subsections.

3.1 Demographic Attributes

The user is described by a set of attributes that are provided to the system. The following list of user attributes are the most popular in the domain of information retrieval systems.

Based on few systems [3], [5], [12] connected with user modeling the following attributes were chosen as significant for classification procedure: gender, age, kind of living town, interest, level of education, employment (domain/job), family, language/foreign language, religion, favorite browser, favorite colour.

The main objective of this paper is to present the method for determining a minimal set of attributes needed in classification procedure. The partition of users' set obtained using whole list of attributes and the partition obtained using only the most significant attributes should be almost the same.

3.2 User Profile Based on Usage Data

In information retrieval system, usage data can be treated as a set of documents that were relevant to the users' queries. The most popular form of the usage data is a list of web pages that the user has visited. More sophisticated forms process this list and save the most frequent terms from user queries to build the profile.

In our approach, user profile will store knowledge about users' interests that user expressed in the form of users' queries. User activities are divided into sessions. In each session systems saves information about user queries and documents that were relevant for him. We assume that documents are described by the set of weighted terms (keywords). After each session the average weight for each term from users' queries is calculated and based on the weights differences between sessions, the user profile is built first, and subsequently updated. The procedure of user profiling is described in details in our previous works [6], [7].

User profile P is presented in the following form:

$$Profile = \{(t_j, w_j) : t_j \in T \wedge w_j \in [0.5, 1), j = 1, 2, \ldots, n_u\} \tag{1}$$

where t_j is index term, w_j is appropriate weight of user interests in particular term t_j and n_u is a number of user interests at the moment.

4 A Method for Collective Recommendation Using Usage-Based Classifier

An original idea of proposed profile recommendation method is presented in Figure 1. It contains three steps. At the beginning the system has a set of users that have both demographic data and profiles. To create a set of similar users (users with similar interests), a clustering method (K-means) is performed based on the data in profiles (step 1). We assume that each profile reflects user preferences. If two users are in the same group, it means that their preferences are similar.

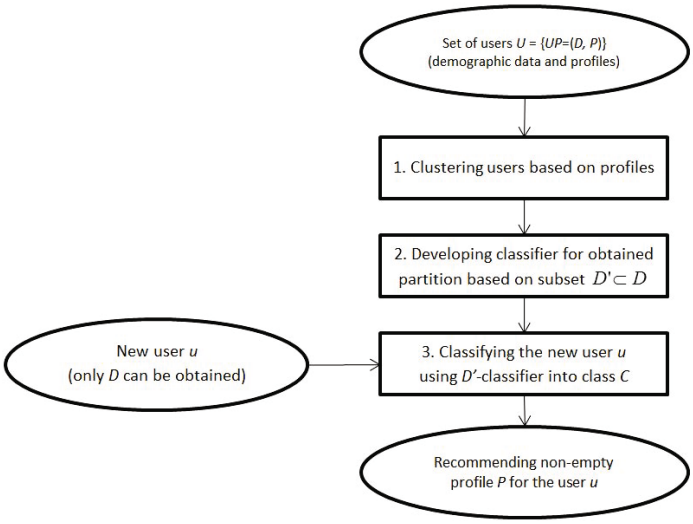


Fig. 1. Schema of profile recommendation based on a usage data classifier

The most important part is to describe obtained users' groups by the attributes coming from demographic data (step 2). The set of users can be clustered according to demographic attributes using associate rules algorithm. The obtained partition should reflect the partition that was a result of profile-based clustering. The method for determining the minimal set of needed attributes to obtain the most similar partitions is presented in Section 4.1.

When a new user is coming to the system, he is asked only about a few demographic data. Based on given information, system classifies him to appropriate class (step 3). The centroid profile for each group of similar user is determined using method described in Section 4.2. To verify if the user was classified into the proper group the system needs to observe his activity. Based on usage data, the system updates his profile. After few series of adaptation the system compares obtained profile with other profiles in the group and checks if the user should not be assign to another group.

4.1 Developing Demographic Data-Based Classifier

The main problem to solve in this task is: for a given classification of users set U a minimal subset D' of attributes from D should be determined, such that the distance between the classification generated by attributes from D' and the given classification is minimal ([8] and [10]). This task is is a NP-complete problem. Discovering the most significant attributes (the minimal attributes' set) that can describe the users, we can reduce the set of required data that user is asked to fulfill in the registration process. Based on this knowledge and using consensus method [9] a new non-empty profile will be generated for the new user.

Let $U = \{U_1, U_2, \ldots, U_N\}$ be a set of users profiles and each user is described by a set of attributes $D = \{d_1, d_2, \ldots, d_M\}$. While using the system, history of user activity is gathered and the profile is updated. Based on these data another grouping process can be carried out to obtain partition Q of user profiles U using a smaller set of attributes. By a partition of set U we call a finite set of nonempty and disjoint with each other classes which are subsets of U, such that their union is equal to U.

The task of a significant feature selection method is to select a set D' of attributes as a subset of the set of all attributes A, which would give the partition as close as possible to previously obtained partition Q, i.e. the distance between the partition Q and partition obtained by the significant feature is the smallest according to the assumed distance measure between partitions.

Designed method for feature selection determines which elements of a set of demographic data D have significant impact on further user activity during the search process. Knowledge obtained in this way will be used for profile recommendation for a new user in the system. To construct a method of recommendation, definition of the function of the distance between user profiles and a method for determining the representative profile of a class are needed.

Using attributes from set D' the set of users can be clustered. Two users $u_1, u_2 \in U$ belongs to the same group if and only if values of each attribute $d \in D'$ are the same. Let's assume that the clustering process on the basis of user profiles gives partition $P_{D'}$ of set U. The previous partition was based on users' profiles (partition Q). To compare these two partitions of the same set of objects a definition of distance measure $d(P, Q)$ between the partitions P and Q is proposed.

Distance measure between partitions is calculated as number of users that are assign to different groups.

$$d(P_D, Q) = \frac{1}{2\dot{c}ard(U)} \dot{\sum_{i,j=1}^{N}} |p_{ij} - q_{ij}|$$

where

$$p_{ij} = \begin{cases} 1, \text{ if } u_i \text{ and } u_j \text{ are in different class of } P_D \\ 0, \text{ if } u_i \text{ and } u_j \text{ are in the same class of } P_D \end{cases}$$

Algorithm 1: Algorithm for Determining Centroid Profile

Input: A set U of N input profiles; set of terms T
Output: A single output profile that is centroid of group
Create the output profile with weights 0
foreach *term* $t \in T$ **do**
  Calculate average weight of term t in user profiles U;
  Add considered term t with calculated weight.
Recommend obtained profile for a new user.

The problem of significant feature selection for classification method using usage–based classifier is as follows:

Definition 1. *Problem of determining demographic-based classifier*
For a given partition Q of set U one should determine a minimal set $D' \subseteq D$ such that $d(P_{D'}, Q)$ is minimal.

4.2 Recommending Non-empty Profile for a New User

Profile for a new user will be recommended based on a set of similar users that were determined on the previous level. When user registers to the system, he is asked to fulfill the questionnaire about his demographic data. Based on them, he is classified into group that has similar values of those data. Using knowledge about other users in this group, the centroid is determined and proposed to the new user.

5 Experimental Evaluation

Designed collective recommendation methods will be experimentally evaluated in a simulation environment using the prepared ways of modeling user behavior. Due to the free availability of Java the developed algorithms are implemented and a simulation is performed in it. The main objective of the experimental evaluations is to check if a profile recommended for a new user based on his demographic data is appropriate for this user. In other words, we would like to check if the user profile after many adaptation series (user is interacting with the system and his profile is updated according to his activities) would be classified to the same group as it was classified at the beginning (based on demographic data).

Due to the time-consuming problem of gathering real user data, instead of engaging actual users a method for modeling user behavior in information retrieval situations is developed. The way of user profile building and adapting was presented in our previous works [6] and [7].

The idea of simulating a user is as follows. The user declares a few terms as his interests. Based on the set of interests, the set of preferences is determined (using documents that user remarks as relevant, the weights of appropriate terms

are calculated). User preferences are not visible for the system. The system can observe only queries and documents relevant to those queries. Based on user activities, the system builds and updates the user profile. The assumption is that using the method for adapting the user profile, this profile converges to user preferences. The method was presented and evaluated in our previous work [7].

5.1 Simulation of User Activity Based on His Demographic Data

This part of the experiment is connected with methodology of simulating user preferences based on his demographic data. Based on statistical data from Polish Central Statistical Office and European Commission Eurostat [13] and [14] for the people from Poland the following values of demographic attributes have been assumed. Ranking of favorite browsers was developed by Gemius [15].

- gender – $V_1 = \{'M','W'\}$ – $51,7\%$ of population are women;
- age (age interval) – $V_2 = \{15-19; 20-24; 25-29; 30-34; 35-39; 40-45\}$;
- kind of the living town – $V_3 = \{'City','Village'\}$;
- interests (choose categories) – V_4 – list of hobbies;
- education – V_5 – 7 levels of education;
- employment (domain/job) – V_6 – list of jobs;
- family (no. of children) – V_7 – if married or not; how many children;
- language/foreign languages – V_8 – list of languages;
- religion – V_9 – list of religions;
- favorite browser – V_{10} = {Chrome, Firefox, IE, Opera, Safari, Other};
- favorite colour – V_{11} ={white, black, gray, red, blue, brown, green, yellow, gold}.

To generate population of users that would reflect a real population, the joint distribution were calculated for available attributes. An exemplary distribution of data for 3 dimensions: gender, age and kind of town are presented in Table 1.

Table 1. An exemplary data about distribution of gender, age and size of town for Poles

Age	City – Men	City – Women	Village – Men	Village – Women
15–19	5.6%	5.2%	7.5%	7.2%
20–24	7.2%	6.8%	7.5%	7.1%
25–29	8.6%	8.1%	7.6%	7.3%
30–34	8.2%	7.8%	7.3%	7.1%
35–39	7.1%	6.8%	7.1%	6.9%
40–44	5.9%	5.7%	6.5%	6.3%

5.2 Plan of Experiments

The simulations are performed according to following steps:

1. Generating set of users (demographic data and profiles).
2. Clustering users based on profiles (using existing method – K-means algorithm). The partition Q is obtained.
3. Developing classifier based on user demographic data
   - Using association rules algorithm to cluster users.
   - Determining a minimal set $D' \subseteq D$ such that $d(P_{D'}, Q)$ is minimal.
   - For each obtained group determining equivalent group in partition Q.
4. Collecting demographic data from a new user and classifying him to the group C_1 based on D'-classifier.
5. Observing user activities (building and adapting his profile).
6. Classifying the user to the group based on his profile C_2.
7. Checking if user was classified to the proper group (if the C_1 is the same as C_2).

5.3 Results Analysis

The training set of users has 100 users. The testing set contains 20 users. The users are searching for documents that contain terms from users' queries. Based on the set of relevant documents to each query, theirs profiles are built and adapted. The users were clustered into 2 – 4 groups using theirs profiles. The main contribution of the researches was to describe each group by demographic attributes. Different combinations of 3 out of 10 demographic attributes were evaluated. The following subsets of demographic data have given the best results:

1. $D'_1 = \{gender, color, education\}$
2. $D'_2 = \{age, gender, colour\}$

As the preliminary experiments has shown the most important demographic attributes are gender, age, favorite colour and level of education. It means that it is enough to ask user only for these attributes and the user should be classified to the group that would be the most similar in terms of his preferences.

6 Summary and Future Works

In this paper the authors propose a novel method for recommending profile for a new user based only on knowledge about partial demographic data. A methodology of experimental evaluation was presented and simulations were performed. The main objective was to show that based on knowledge about other similar users, the system can classify a new user based on subset of demographic data and recommend him a non-empty profile. Using proposed profile, the user will obtain personalized documents. As the preliminary experiments has shown the most important demographic attributes are gender, age, favorite color and level of education.

Acknowledgments. This research was partially supported by Polish Ministry of Science and Higher Education.

References

1. Ahn, H.J.: A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. Information Sciences 178, 37–51 (2008)
2. Campos, L.M., Fernández-Luna, J.M., Huete, J.F., Rueda-Morales, M.A.: Combining content-based and collaborative recommendations: A hybrid approach based on Bayesian networks. International Journal of Approximate Reasoning 51, 785–799 (2010)
3. Gemmis, M., Iaquinta, L., Lops, P., Musto, C., Narducci, F., Semeraro, G.: Preference Learning in Recommender Systems. In: Preference Learning (PL 2009) ECML/PKDD 2009 Workshop (2009)
4. Kardan, A.A., Ebrahimi, M.: A novel approach to hybrid recommendation systems based on association rules mining for content recommendation in asynchronous discussion groups. Information Sciences 219, 93–110 (2013)
5. Lu, L., Medo, M., Yeung, C.H., Zhang, Y.-C., Zhang, Z.-K., Zhou, T.: Recommender systems. Physics Reports 519, 1–49 (2012)
6. Mianowska, B., Nguyen, N.T.: Using Knowledge Integration Techniques for User Profile Adaptation Method in Document Retrieval Systems. In: Nguyen, N.T. (ed.) Transactions on CCIV. LNCS, vol. 6910, pp. 140–156. Springer, Heidelberg (2011)
7. Mianowska, B., Nguyen, N.T.: Tuning User Profiles Based on Analyzing Dynamic Preference in Document Retrieval Systems. Multimedia Tools and Applications (2012), doi:10.1007/s11042-012-1145-6
8. Nguyen, N.T.: Rough Classification – New Approach and Applications. Journal of Universal Computer Science 15(13), 2622–2628 (2009)
9. Nguyen, N.T.: Using Consensus Methodology in Processing Inconsistency of Knowledge. In: Last, M., et al. (eds.) Advances in Web Intelligence and Data Mining, vol. 23, pp. 161–170. Springer, Heidelberg (2006)
10. Pawlak, Z.: Rough classification. Int. J. Human-Computer Studies 51, 369–383 (1999)
11. Su, J.-H., Wang, B.-W., Hsiao, C.-Y., Tseng, V.S.: Personalized rough-set-based recommendation by integrating multiple contents and collaborative information. Information Sciences 180, 113–131 (2010)
12. Vozalis, M., Margaritis, K.G.: Collaborative filtering enhanced by demographic correlation. In: Proceedings of the AIAI Symposium on Professional Practice in AI, part of the 18th World Computer Congress, pp. 293–402 (2004)
13. Central Statistical Office, `http://www.stat.gov.pl/gus` (available on October 25, 2012)
14. European Commission Eurostat, `http://epp.eurostat.ec.europa.eu` (available on October 25, 2012)
15. Gemius Ranking, `http://www.ranking.pl/pl` (available on October 25, 2012)

Combining Multiple Clusterings of Chemical Structures Using Cumulative Voting-Based Aggregation Algorithm

Faisal Saeed[1,2,*], Naomie Salim[1], Ammar Abdo[3,4], and Hamza Hentabli[1]

[1] Faculty of Computing, Universiti Teknologi Malaysia, Malaysia
[2] Information Technology Department, Sanhan Community College, Sana'a, Yemen
[3] Computer Science Department, Hodeidah University, Hodeidah, Yemen
[4] LIFL UMR CNRS 8022 Universite´ Lille 1 and INRIA Lille Nord Europe, 59655 Villeneuve d'Ascq cedex, France
alsamet.faisal@gmail.com

Abstract. The use of consensus clustering methods in chemoinformatics is motivated because of the success of consensus scoring (data fusion) in virtual screening and also because of the ability of consensus clustering to improve the robustness, novelty, consistency and stability of individual clusterings in other areas. In this paper, Cumulative Voting-based Aggregation Algorithm (CVAA) was examined for combining multiple clusterings of chemical structures. The effectiveness of clusterings was evaluated based on the extent to which they clustered compounds, which belong to the same activity class, together. Then, the results were compared to other consensus clustering and Ward's methods. The MDL Drug Data Report (MDDR) database was used for experiments and the results were obtained by combining multiple clusterings that were applied using different distance measures. The experiments show that the voting-based consensus method can efficiently improve the effectiveness of chemical structures clusterings.

Keywords: Consensus clustering, Cumulative voting, Data fusion, Molecular datasets, Ward's clustering.

1 Introduction

Many clustering techniques have been used in the literature for chemical structures clustering [1-9] and were used to reduce the high costs and lengthy time needed to discover new drugs. The clustering helps the pharmaceutical industries to find faster and more effective ways of discovering and producing chemical compounds that can effectively react to the examined disease. Furthermore, there is a strong need for a rational and effective selection of a subset in the combinatorial chemical library so that the maximum amount of information can be obtained by testing minimal numbers of chemical compounds. Therefore, many approaches for compound selection have been used which are cluster-based compound selection, partition-based compound selection,

* Corresponding author.

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 178–185, 2013.

dissimilarity-based compound selection, and optimization-based compound selection. Among these approaches, the cluster-based compound selection, which is known as clustering, has become the most commonly used in compound selection [10].

Brown and Martin [9] considered the Ward's clustering method to be the most effective clustering method in compound clustering. However, as it is known, no clustering method is capable of correctly finding the best clustering results for all datasets and applications. So, the idea of combining different clustering results (which is known as consensus clustering) is considered as an alternative approach for improving the quality of the individual clustering algorithms [11].

Consensus clustering involves two main steps: (i) partitions generation and (ii) consensus function. In the first step, as many as possible individual partitions will be generated. Different generation mechanisms can be applied including the using of: (i) different object representations; (ii) different individual clustering methods; (iii) different parameters initialisation for clustering methods; and (iv) data resampling. In the second step, there are two main approaches which are the objects co-occurrence-based and the median partition-based approaches. Relabeling and voting-based consensus clustering is a method of the first approach.

In voting-based consensus clustering methods [12-18], the consensus partition is derived by seeking an optimal relabeling of the ensemble partitions. In general, the optimal relabeling of the ensemble partitions is addressed through a pairwise relabeling of each ensemble partition with respect to a representative partition, which is known as the voting problem, so that each cluster of a given ensemble partition is viewed as a "voter" that votes for the representative clusters according to a defined voting [18] . Then, the voting-based aggregation algorithm is used to obtain the consensus (aggregated) partition [17-18].

In chemoinformatics, Chu *et al*. [19] used consensus similarity matrix methods on sets of chemical structures and concluded that the consensus clustering methods can outperform the Ward's method, the current standard clustering method for chemoinformatics applications. However, based on the implemented methods, it was not the case if the clustering is restricted to a single consensus method. In addition, Saeed *et al*. [20] examined the use of the Cluster-based Similarity Partitioning Algorithm CSPA [21], for combining multiple clusterings of MDDR dataset and concluded that it can improve the effectiveness of individual clusterings and provide robust and stable clustering. However, the high cost of computations using similarity matrix consensus clustering methods lead to find efficient algorithm to improve the effectiveness of combining multiple clusterings of chemical structures. In this paper, cumulative voting-based aggregation algorithm is used. The computational complexity for similarity matrix based methods is O (n^2), whereas for CVAA is O (*n)*, where *n* is the number of objects [11].

2 Materials and Methods

2.1 Dataset

The MDL Drug Data Report (MDDR) database [22] was used for experiments. This database consists of 102516 molecules. The MDDR subset (DS1) was chosen from the MDDR database which has been used for many virtual screening experiments

[23-25]. The DS1 dataset contains eleven activity classes (8294 molecules), which involves homogeneous and heterogeneous active molecules. Details of this dataset are listed in Table 1. Each row in the table contains an activity class, the number of molecules belonging to the class, and the diversity of the class, which was computed as the mean pairwise Tanimoto similarity calculated across all pairs of molecules in the class. For the clustering experiments, two 2D fingerprint descriptors were used which were developed by Scitegic's Pipeline Pilot [26]. These were 120-bit ALOGP and 1024-bit ECFP_4 fingerprints.

Table 1. MDDR Activity Classes for DS1 Dataset

Activity Index	Activity class	Active molecules	Pairwise similarity
			Mean
31420	Renin Inhibitors	1130	0.290
71523	HIV Protease Inhibitors	750	0.198
37110	Thrombin Inhibitors	803	0.180
31432	Angiotensin II AT1 Antagonists	943	0.229
42731	Substance P Antagonists	1246	0.149
06233	Substance P Antagonists	752	0.140
06245	5HT Reuptake Inhibitors	359	0.122
07701	D2 Antagonists	395	0.138
06235	5HT1A Agonists	827	0.133
78374	Protein Kinase C Inhibitors	453	0.120
78331	Cyclooxygenase Inhibitors	636	0.108

2.2 Partitions Generation

Every consensus clustering method is made up of two steps: partitions generation and consensus functions. In this paper, the partitions (also called ensembles) were generated by using six individual clustering algorithms on each 2D fingerprint. These algorithms were single linkage, complete linkage, average linkage, weighted average distance, Ward and K-means clustering methods. The thresholds of 500, 600, 700, 800, 900 and 1000 were used to generate partitions with different sizes (number of clusters). Every individual clustering method was applied by using six distance measures in order to generate six ensembles for each 2D fingerprint (each ensemble includes 6 partitions). The distance measures, which were used with each clustering technique, were Correlation, Cosine, Euclidean, Hamming, Jaccard and Manhattan.

2.3 Cumulative Voting-Based Aggregation Algorithm

The cumulative voting based aggregation algorithm consists of two steps; the first one is to obtain the optimal relabeling for all partitions, which is known as the voting problem. Then, the voting-based aggregation algorithm is used to obtain the

aggregated (consensus) partition. The cumulative voting-based aggregation algorithm described by Ayed and Kamel [17-18] is shown in Figures 1-2.

Let χ denote a set of n data objects, and let a partition of χ into k clusters be represented by an $n \times k$ matrix $\mathbf{U}$ such that $\sum_{q=1}^{k} ujq = 1$, for $\forall$ j. Let $u = \{U_i\}_{i=1}^{b}$ denote an ensemble of partitions. The voting-based aggregation problem is concerned with searching for an optimal relabeling for each partition $\mathbf{V}^i$ with respect to representative partition $\mathbf{U}^0$ (with k^0 clusters) and for a central aggregated partition denoted as $\bar{\mathbf{U}}$ that summarises the ensemble partitions. The matrix of coefficients $\mathbf{W}^i$, which is a $k^i \times k^0$ matrix of $w_{lq}{}^i$ coefficients, is used to obtain the optimal relabeling for ensemble partitions.

In this paper, the fixed-reference approach is used, whereby an initial reference partition is used as a common representative partition for all the ensemble partitions and remains unchanged throughout the aggregation process. Instead of selecting random partition, the partition that is generated by the method, which showed high ability to separate active from inactive molecules in our experiments, is suggested to be the reference partition U_0; and this method is the Ward's clustering (the current standard clustering method for Chemoinformatics applications). The cumulative voting based aggregation algorithm is described in Figure 2.

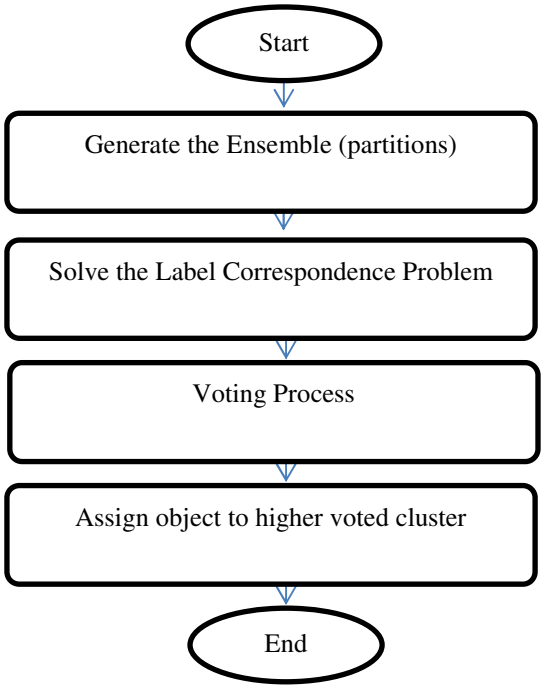


Fig. 1. Voting-based consensus clustering method

1: select a partition $\mathbf{U}^i \in \mathcal{U}$ which is generated by the Ward's method and assign to $\mathbf{U}^0$
2: for $i = 1$ to b do
3: $\mathbf{W}^i = (\mathbf{U}i^{\mathrm{T}}\ \mathbf{U}i)^{-1}\ \mathbf{U}i^{\mathrm{T}}\mathbf{U}^{\mathbf{0}}$
4: $\mathbf{V}^i = \mathbf{U}i\,\mathbf{W}^i$
5: $\mathbf{U}^0 = \frac{i-1}{i}\ \mathbf{U}^0 + \frac{1}{i}\ \mathbf{V}^i$
6: end for
7: $\bar{\mathbf{U}} = \mathbf{U}^0$.

Fig. 2. Cumulative Voting-based Aggregation Algorithm

2.4 Performance Evaluation

The results were evaluated based on the effectiveness of the methods to separate active from inactive molecules using Quality Partition Index (QPI) measure [27]. As defined by [8], an active cluster is a non-singleton cluster for which the percentage of active molecules in the cluster is greater than the percentage of active molecules in the dataset as a whole. Let p be the number of actives in active clusters, q be the number of inactives in active clusters, r be the number of actives in inactive clusters (i.e., clusters that are not active clusters) and s be the number of singleton actives. The high value occurs when the actives are clustered tightly together and separated from the inactive molecules. The QPI is defined to be:

$$QPI = \frac{p}{p+q+r+s} \tag{1}$$

3 Results and Discussion

The ensembles were generated by using the six individual clusterings, each with the six distance measures. Six ensembles (each with size of six partitions) were combined using graph-based consensus clustering (CSPA) and voting-based consensus clustering (CVAA). This process was done for each fingerprint (ALOGP and ECFP_4).

The mean of QPI values were averaged over the eleven activity classes of the dataset. Tables 2-3 show the effectiveness of MDDR dataset clustering for ALOGP and ECFP_4 fingerprints. The best values for QPI of consensus clustering methods for each column in all tables were bold-faced for ease of reference.

Visual inspection of QPI values in Tables 2-3 enables comparisons to be made between the effectiveness of two consensus clustering methods and the Ward's method. In addition, two fingerprints were used for the experiments in order to study the effectiveness of consensus clustering on different representations of molecular dataset.

For clustering of MDDR dataset which was represented by ALOGP fingerprint, as shown in Table 2, the performance of voting-based consensus method (CVAA) outperformed the Ward's method and the graph-based consensus method (CSPA). The highest QPI values were obtained by using Jaccard distance measure.

Table 2. Effectivenss of clustering of MDDR dataset using QPI: ALOGP Fingerprint

Clustering Method			No. of clusters					
			500	600	700	800	900	1000
Consensus Clustering	CVAA	Correlation	43.84	47.38	48.72	50.70	53.41	54.06
		Cosine	45.60	46.08	47.56	50.46	53.79	54.50
		Euclidean	44.43	45.54	47.95	48.65	52.68	54.86
		Hamming	53.13	56.08	59.07	60.58	64.02	67.76
		Jaccard	**57.86**	**60.62**	**64.07**	**66.49**	**70.68**	**73.53**
		Manhattan	56.01	58.10	60.99	61.86	64.56	65.97
	CSPA	Correlation	46.81	50.04	51.72	51.78	54.23	56.36
		Cosine	46.04	49.49	51.42	52.11	54.48	55.92
		Euclidean	46.20	49.86	51.05	51.88	54.36	56.33
		Hamming	54.67	58.50	60.27	61.78	62.33	65.66
		Jaccard	55.03	59.13	60.84	61.03	63.73	67.44
		Manhattan	55.08	59.00	59.10	60.84	61.78	64.61
Individual Clustering	Ward's method		52.33	54.86	56.90	59.00	61.33	63.17

Table 3. Effectivenss of clustering of MDDR dataset using QPI: ECFP_4 Fingerprint

Clustering Method			No. of clusters					
			500	600	700	800	900	1000
Consensus Clustering	CVAA	Correlation	74.86	78.02	82.39	84.16	85.71	87.04
		Cosine	74.79	78.12	81.85	84.78	85.91	87.18
		Euclidean	71.04	74.92	78.41	81.91	84.47	86.80
		Hamming	70.99	74.36	78.47	81.68	84.24	86.28
		Jaccard	**83.48**	**87.01**	**88.72**	**90.98**	**90.67**	**92.05**
		Manhattan	70.74	74.26	78.52	81.74	84.12	86.09
	CSPA	Correlation	70.58	73.29	74.86	76.86	79.17	82.03
		Cosine	71.23	71.85	76.43	76.55	78.06	81.21
		Euclidean	65.33	67.09	72.49	72.73	74.50	78.75
		Hamming	64.68	66.82	69.88	71.25	74.17	76.64
		Jaccard	69.91	71.73	74.20	76.01	77.72	79.26
		Manhattan	63.07	65.77	68.83	71.50	74.06	77.33
Individual Clustering	Ward's method		75.83	79.88	83.34	84.25	86.49	88.25

Similarly, the results in Table 3 show that, when ECFP_4 fingerprint is used, the CVAA consensus clustering performed very well and outperformed the Ward's and CSPA methods. The Jaccard distance measure was the method of choice to generate the ensembles which gives the highest QPI values.

4 Conclusion and Future Work

The results of the experiments show that the cumulative voting-based aggregation algorithm (CVAA) can efficiently improve the effectiveness of chemical structures clusterings. The performance of CVAA consensus clustering outperforms CSPA and Ward's methods for both ALOGP and ECFP_4 fingerprints. The experiments suggest that when using CVAA, the Jaccard distance measures is the method of choice for generating ensembles using different individual clusterings. In the future work, more voting-based consensus clustering methods will be examined and compared with Ward and other consensus clustering methods.

Acknowledgment. This work is supported by the Ministry of Higher Education (MOHE) and Research Management Centre (RMC) at the Universiti Teknologi Malaysia (UTM) under the Research University Grant Category (VOT Q.J130000.7826.4F011). We also would like to thank MIS-MOHE for sponsoring the first author.

References

1. Adamson, G.W., Bush, J.A.: A method for the automatic classification of chemical structures. Information Storage and Retrieval 9, 561–568 (1973)
2. Downs, G.M., Barnard, J.M.: Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures. Journal of Chemical Information and Computer Science 32, 644–649 (1992)
3. Willett, P.: Similarity and Clustering in Chemical Information Systems. Research Studies Press, Letchworth (1987)
4. Downs, G.M., Willett, P., Fisanick, W.: Similarity searching and clustering of chemical-structure databases using molecular property data. J. Chem. Inf. Comput. Sci. 34, 1094–1102 (1994)
5. Brown, R.D., Martin, Y.C.: The information content of 2D and 3D structural descriptors relevant to ligand–receptor binding. J. Chem. Inf. Comput. Sci. 37, 1–9 (1997)
6. Downs, G.M., Barnard, J.M.: Clustering of chemical structures on the basis of two-dimensional similarity measures. J. Chem. Inf. Comput. Sci. 32, 644–649 (1992)
7. Holliday, J.D., Rodgers, S.L., Willet, P.: Clustering Files of chemical Structures Using the Fuzzy k-means Clustering Method. Journal of Chemical Information and Computer Science 44, 894–902 (2004)
8. Varin, T., Bureau, R., Mueller, C., Willett, P.: Clustering files of chemical structures using the Székely–Rizzo generalization of Ward's method. Journal of Molecular Graphics and Modeling 28(2), 187–195 (2009)

9. Brown, R.D., Martin, Y.C.: Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. J. Chem. Inf. Compute. Sci. 36, 572–584 (1996)
10. Salim, N.: Analysis and Comparison of Molecular Similarity Measures. University of Sheffield. PhD Thesis (2003)
11. Vega-Pons, S., Ruiz-Schulcloper, J.: A survey of clustering ensemble algorithms. International Journal of Pattern Recognition and Artificial Intelligence 25(3), 337–372 (2011)
12. Fischer, B., Buhmann, J.M.: Bagging for path-based clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(11), 1411–1415 (2003)
13. Dudoit, S., Fridlyand, J.: Bagging to improve the accuracy of a clustering procedure. Bioinformatics 19(9), 1090–1099 (2003)
14. Evgenia, D., Andreas, W., Kurt, H.: A combination scheme for fuzzy clustering. International Journal of Pattern Recognition and Artificial Intelligence 16(7), 901–912 (2002)
15. Gordon, A.D., Vichi, M.: Fuzzy partition models for fitting a set of partitions. Psychometrika 66(2), 229–248 (2001)
16. Topchy, A., Law, M., Jain, A.K., Fred, A.: Analysis of consensus partition in clustering ensemble. In: Proceedings of the IEEE Intl. Conf. on Data Mining 2004, Brighton, UK, pp. 225–232 (2004)
17. Ayad, H.G., Kamel, M.S.: Cumulative voting consensus method for partitions with a variable number of clusters. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(1), 160–173 (2008)
18. Ayad, H.G., Kamel, M.S.: On voting-based consensus of cluster ensembles. Patt. Recogn. 43, 1943–1953 (2010)
19. Chu, C.-W., Holliday, J., Willett, P.: Combining multiple classifications of chemical structures using consensus clustering. Bioorgan. Med. Chem. 20(18), 5366–5371 (2012)
20. Saeed, F., Salim, N., Abdo, A., Hentabli, H.: Combining Multiple Individual Clusterings of Chemical Structures Using Cluster-Based Similarity Partitioning Algorithm. In: Hassanien, A.E., Salem, A.-B.M., Ramadan, R., Kim, T.-h. (eds.) AMLTA 2012. CCIS, vol. 322, pp. 276–284. Springer, Heidelberg (2012)
21. Strehl, A., Ghosh, J.: Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions. J. Machine Learning Research 3, 583–617 (2002)
22. Sci Tegic Accelrys Inc., the MDL Drug Data Report (MDDR) database is available from at `http://www.accelrys.com/` (accessed November 1, 2012)
23. Abdo, A., Chen, B., Mueller, C., Salim, N., Willett, P.: Ligand-Based Virtual Screening Using Bayesian Networks. J. Chem. Inf. Model. 50, 1012–1020 (2010)
24. Abdo, A., Salim, N.: New Fragment Weighting Scheme for the Bayesian Inference Network in Ligand-Based Virtual Screening. J. Chem. Inf. Model. 51, 25–32 (2011)
25. Abdo, A., Saeed, F., Hentabli, H., Ali, A., Salim, N., Ahmed, A.: Ligand expansion in ligand-based virtual screening using relevance feedback. Journal of Computer-Aided Molecular Design 26, 279–287 (2012)
26. Pipeline Pilot, Accelrys Software Inc., San Diego (2008)
27. Varin, T., Saettel, N., Villain, J., Lesnard, A., Dauphin, F., Bureau, R., Rault, S.J.: 3D Pharmacophore, hierarchical methods, and 5-HT4 receptor binding data. Enzyme Inhib.Med. Chem. 23, 593–603 (2008)

Investigation of Incremental Support Vector Regression Applied to Real Estate Appraisal

Tadeusz Lasota[1], Petru Patrascu[2], Bogdan Trawiński[2], and Zbigniew Telec[2]

[1] Wrocław University of Environmental and Life Sciences, Dept. of Spatial Management
ul. Norwida 25/27, 50-375 Wrocław, Poland
[2] Wrocław University of Technology, Institute of Informatics,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
tadeusz.lasota@up.wroc.pl,
{zbigniew.telec,bogdan.trawinski}@pwr.wroc.pl

Abstract. Incremental support vector regression algorithms (SVR) and sequential minimal optimization algorithms (SMO) for regression were implemented. Intensive experiments to compare predictive accuracy of the algorithms with different kernel functions over several datasets taken from a cadastral system were conducted in offline scenario. The statistical analysis of experimental output was made employing the nonparametric methodology designed especially for multiple $N \times N$ comparisons of N algorithms over N datasets including Friedman tests followed by Nemenyi's, Holm's, Shaffer's, and Bergmann-Hommel's post-hoc procedures. The results of experiments showed that most of SVR algorithms outperformed significantly a pairwise comparison method used by the experts to estimate the values of residential premises over all datasets. Moreover, no statistically significant differences between incremental SVR and non-incremental SMO algorithms were observed using our stationary cadastral datasets. The results open the opportunity of further research into the application of incremental SVR algorithms to predict from a data stream of real estate sales/purchase transactions.

Keywords: support vector regression, incremental SVR, SMO for regression, property valuation.

1 Introduction

Several computational intelligence methods to create data driven models to assist with real estate appraisals were devised during the last two decades. The main attention was focused on neural networks [1], [2], less researchers have been involved in the application of fuzzy systems [3], [4] and support vector machines [5], [6]. We explored several approaches to build predictive models for property valuation including evolutionary fuzzy systems, neural networks, and decision trees using MATLAB, KEEL, RapidMiner, and WEKA data mining systems [7], [8], [9]. We also constructed and evaluated models employing evolving fuzzy systems eTS [10] and FLEXFIS [11] which treated cadastral data on property sales/purchase transactions as a data stream which in turn could reflect the changes of real estate

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 186–195, 2013.

market in the course of time. In this paper we report our recent study into the application of incremental support vector regression to this purpose.

Online machine learning algorithms are especially useful when dealing with very large or non-stationary data. In online learning scenario subsequent instances of training data come in one by one whereas in classical batch mode all instances are available at once. Batch algorithms are inefficient when applied to an online setting because they have to be retrained from scratch as new data come in and this is expensive and time consuming. Support vector machine was shown to be valuable technique to solve many classification and regression problems revealing a very good generalization performance. Majority of support vector regression algorithms assume batch processing of all available training instances [12], [13], [14]. Several researchers tackled the problem of adapting support vector machine to online learning; early proposed algorithms were approximate. Accurate incremental and decremental support vector regression algorithms were devised by Ma [15] and Gâlmeanu [16] based on the approach of Cauwenberghs and Poggio for incremental SVM classification [17]. Incremental and decremental training consists in the migration of vectors in and out of the support set along with modifying the associated thresholds and preserving the Karush-Kuhn-Tucker (KKT) conditions.

We implemented a SVR algorithm for online settings based on Ma's and Gâlmeanu's proposals. In order to able to conduct a comparative evaluation we implemented also non-incremental Smola and Schölkopf's sequential minimal optimization (SMO) algorithm for support vector machine for regression [18], [19]. Both algorithms were applied to real-world regression problem of predicting the prices of residential premises based on historical data of sales/purchase transactions obtained from a cadastral system. The experiments were carried out using a cross-validation approach for offline processing. Moreover, we compared SVR algorithms with a property valuating method employed by professional appraisers in reality.

The analysis of the results was performed using statistical methodology including nonparametric tests followed by post-hoc procedures designed especially for multiple $N \times N$ comparisons [20], [21], [22]. The idea behind statistical methods applied to analyse the results of experiments was as follows. The commonly used paired tests i.e. parametric t-test and its nonparametric alternative Wilcoxon signed rank tests are not appropriate for multiple comparisons due to the so called family-wise error. The proposed routine starts with the nonparametric Friedman test, which detect the presence of differences among all algorithms compared. After the null-hypotheses have been rejected the post-hoc procedures should be applied in order to point out the particular pairs of algorithms which produce differences. For $N \times N$ comparisons nonparametric Nemenyi's, Holm's, Shaffer's, and Bergamnn-Hommel's procedures are recommended.

2 Methods Used and Experimental Setup

The investigation was conducted with our experimental system developed in C# in .NET environment [23] designed to carry out research into support vector regression employed to create data driven property valuation models. We implemented in the system an incremental / decremental SVR following Ma's and Gâlmeanu's proposals

[15], [16] as well as the Smola and Schölkopf's sequential minimal optimization algorithm for regression [18], [19]. Different kernel functions were employed in both algorithms. Following denotation will be used in the rest of the paper:

- *INC* – incremental /decremental SVR algorithm,
- *SMO* – sequential minimal optimization algorithm for regression,
- *R* – Gaussian radial basis function of the form $k(x,y)=\exp(-\gamma\|x-y\|^2)$,
- *RE* – exponential kernel in the form of $k(x,y)=\exp(-\gamma\|x-y\|)$,
- *RG* – Gaussian kernel in the form of $k(x,y)=\exp(-0.5\|x-y\|^2/\sigma^2)$,
- *Expert* – pairwise comparison approach developed by professional appraisers.

The predictive accuracy of seven property valuation models, i.e. *INC-R, INC-RE, INC-RG, SMO-R, SMO-RE, SMO-RG*, and *Expert* was compared over several cadastral datasets. As performance measure the root mean square error (RMSE) was used; data was normalized using min-max approach; 10-fold cross-validation was applied in offline processing. Much emphasis was laid on the parameter selection for our SVR algorithms. In the system we implemented two methods for determining the optimal parameters, namely grid search based on principles from design of experiments (DOE) elaborated by Staelin [24] and pattern search (PS) developed by Momma and Bennett [25]. So, we were able to carry out a few thousand preliminary runs to obtain parameters providing the models with the best predictive accuracy.

2.1 Dataset Used in Experiments

Real-world dataset used in experiments was drawn from an unrefined dataset containing above 50,000 records referring to residential premises transactions accomplished in one Polish city with the population of 640,000 within eleven years from 1998 to 2008. In this period most transactions were made with non-market prices when the council was selling flats to their current tenants on preferential terms. First of all, transactional records referring to residential premises sold at market prices were selected. Then the dataset was confined to sales transaction data of apartments built before 1997 and where the land was leased on terms of perpetual usufruct. Hence, the final dataset counted 5303 records and comprised all premises which values could be estimated by the experts.

The prices of premises change substantially in the course of time what justifies the application of incremental learning methods. However, in our experiment we assumed the batch scenario to create data-driven models. Therefore, whole dataset could not be used in scenario. It should have been spilt into subsets covering smaller periods and we might assume that within one year the prices of premises with similar attributes were roughly comparable. We prepared two series of subsets one-year and half-year ones in the following way. Starting from the beginning of 1998 the prices were updated for the last day of subsequent one-years or half-years. The trends were modelled by polynomials of degree three. The chart illustrating the change trend of average transactional prices per square metre is given in Fig. 1. We might assume that one-year and half-year datasets differed from each-other and might constitute different observation points to compare the accuracy of ensemble models in our study and carry out statistical tests. The sizes of one-year and half-year datasets are given in Table 1 and Table 2, respectively.

Table 1. Number of instances in one-year datasets

1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
269	477	329	463	530	653	546	580	677	575	204

Table 2. Number of instances in half-year datasets

199802	199901	199902	200001	200002	200101	200102	200201	200202	200301
202	213	264	162	167	228	235	267	263	267
200302	**200401**	**200402**	**200501**	**200502**	**200601**	**200602**	**200701**	**200702**	**200801**
386	278	268	244	336	300	377	289	286	181

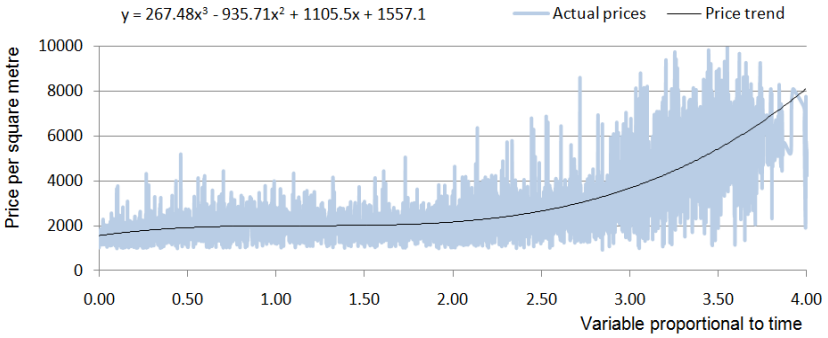


Fig. 1. Change trend of average transactional prices per square metre over time

2.2 Expert's Valuation Method

In order to compare evolutionary machine learning algorithms with techniques applied to property valuation we asked professional appraisers to evaluate premises using historical data of sales/purchase transactions obtained from a cadastral system. In this section the pairwise comparison method used by the experts to estimate the values of premises comprised in our dataset is described. The experts simulated professional appraisers' work in the way it is done in reality.

First of all the whole area of the city was divided into 6 quality zones: 1 - the central one, 2 - near-medium, 3 - eastern-medium, and 4 – western-medium zones, and finally 5 - south-western-distant and 6 - north-eastern-distant zones. Next, the premises located in each zone were classified into 243 groups determined by 5 following quantitative features selected as the main price drivers: *Area, Year, Storeys, Rooms,* and *Centre.* Domains of each feature were split into three brackets as follows.

Area denotes the usable area of premises and comprises small flats up to 40 m^2, medium flats in the bracket 40 to 60 m^2, and big flats above 60 m^2.

Year (*Age*) means the year of a building construction and consists of old buildings constructed before 1945, medium age ones built in the period 1945 to 1960, and new buildings constructed between 1960 and 1996, the buildings falling into individual ranges are treated as in bad, medium, and good physical condition respectively.

Storeys are intended for the height of a building and are composed of low houses up to three storeys, multi-family houses from 4 to 5 storeys, and tower blocks above 5 storeys.

Rooms are designated for the number of rooms in a flat including a kitchen. The data contain small flats up to 2 rooms, medium flats in the bracket 3 to 4, and big flats above 4 rooms.

Centre stands for the distance from the city centre and includes buildings located near the centre i.e. up to 1.5 km, in a medium distance from the centre - in the brackets 1.5 to 5 km, and far from the centre - above 5 km.

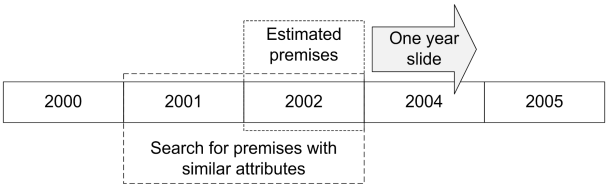


Fig. 2. Time windows used in the pairwise comparison method of experts' estimation

Premises estimation procedure employed a two-year time window to take transaction data of similar premises into consideration (Fig. 2).

1. Take next premises to estimate.
2. Check the completeness of values of all five features and note a transaction date.
3. Select all premises sold earlier than the one being appraised, within current and one preceding year and assigned to the same group.
4. If there are at least three such premises calculate the average price taking the prices updated for the last day of a given transaction year (or half year).
5. Return this average as the estimated value of the premises.
6. Repeat steps 1 to 5 for all premises to be appraised.
7. For all premises not satisfying the condition determined in step 4 extend the quality zones by merging 1 & 2, 3 & 4, and 5 & 6 zones. Moreover, extend the time window to include current and two preceding years.
8. Repeat steps 1 to 5 for all remaining premises.

3 Results of Experiments

The performance of seven models, i.e. *INC-R, INC-RE, INC-RG, SMO-R, SMO-RE, SMO-RG*, and *Expert* built over one-year and half-year datasets was presented in Tables 3 and 4 and compared graphically in Figures 3 and 4, respectively. The best results for each dataset are stressed in boldface in Tables 3 and 4. It can be easily seen that the performance of the experts' method fluctuates strongly achieving for some datasets, i.e. 2001, and 2008, excessively high RMSE values. However, differences among incremental and non-incremental SVR are not apparent therefore we should refer to statistical tests of significance.

Table 3. Performance of models in terms of RMSE generated using SVR for one-year datasets

Dataset	INC-R	INC-RE	INC-RG	SMO-R	SMO-RE	SMO-RG	Expert
1998	**0.09863**	0.09920	0.10689	0.10010	0.09875	0.10589	0.11232
1999	0.09075	**0.08270**	0.09028	0.08977	0.08354	0.09022	0.11556
2000	**0.08042**	0.08076	0.08269	0.08322	0.08213	0.08243	0.10340
2001	0.06381	**0.06181**	0.06667	0.06604	0.06303	0.06718	0.13262
2002	0.06700	**0.06514**	0.06964	0.06806	0.06573	0.07156	0.08717
2003	**0.07387**	0.07699	0.07625	0.07711	0.07771	0.07683	0.08680
2004	0.09175	**0.08669**	0.09249	0.09066	0.08751	0.09420	0.10242
2005	0.07464	**0.07053**	0.07666	0.07509	0.07132	0.07671	0.09672
2006	0.08932	0.08582	0.08998	0.09202	**0.08560**	0.09210	0.09515
2007	**0.06113**	0.06555	0.06408	0.06533	0.06960	0.06436	0.08115
2008	**0.07342**	0.07518	0.07475	0.07342	0.07615	0.07477	0.11832

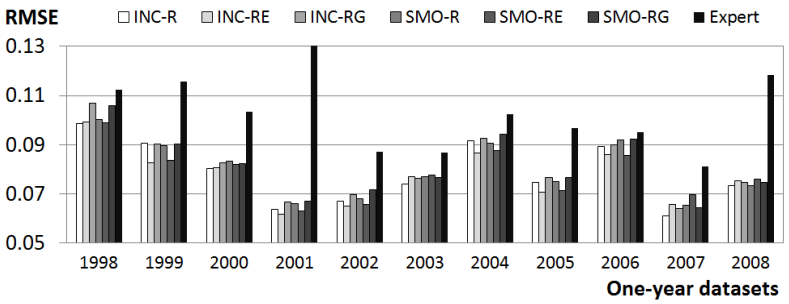


Fig. 3. Comparison of models built using SVR for one-year datasets

Table 4. Performance of models in terms of RMSE generated using SVR for half-year datasets

Dataset	INC-R	INC-RE	INC-RG	SMO-R	SMO-RE	SMO-RG	Expert
199802	0.11187	0.11141	0.11418	**0.10266**	0.10512	0.11223	0.12275
199901	0.09919	**0.09065**	0.09997	0.10018	0.09696	0.10024	0.09719
199902	0.09705	**0.08882**	0.09948	0.09688	0.09138	0.09863	0.11985
200001	0.09362	0.09148	0.09194	0.09656	0.09129	0.09249	**0.09122**
200002	0.10779	0.10262	**0.09259**	0.10775	0.10262	0.09444	0.11501
200101	0.07355	**0.07262**	0.07422	0.07355	0.07460	0.07529	0.14852
200102	0.09398	0.09672	**0.09240**	0.09398	0.09672	0.09352	0.10104
200201	0.09754	0.08654	0.10315	0.09279	**0.08470**	0.10353	0.10192
200202	0.07822	0.07875	**0.07641**	0.07822	0.07881	0.07691	0.15045
200301	0.09608	0.09298	0.09547	0.09295	0.09635	**0.09266**	0.09361
200302	**0.08030**	0.08365	0.08050	0.08100	0.08336	0.08083	0.09333
200401	0.10070	**0.09434**	0.10128	0.10070	0.09436	0.10261	0.11183
200402	0.09939	**0.08977**	0.09986	0.09939	0.09305	0.10139	0.11284
200501	0.09650	**0.07681**	0.09644	0.09185	0.07693	0.09272	0.12704
200502	**0.08684**	0.08962	0.08695	0.09073	0.08947	0.09067	0.10525
200601	0.10363	0.09467	0.10283	0.10425	**0.09445**	0.10768	0.10115
200602	0.09729	**0.09106**	0.09850	0.09921	0.09461	0.09918	0.10430
200701	0.07329	0.07769	**0.07235**	0.07733	0.08409	0.07346	0.08581
200702	0.08773	0.09040	0.09335	**0.08609**	0.09192	0.09158	0.09682
200801	0.08339	0.08111	**0.08102**	0.08339	0.08411	0.08694	0.14577

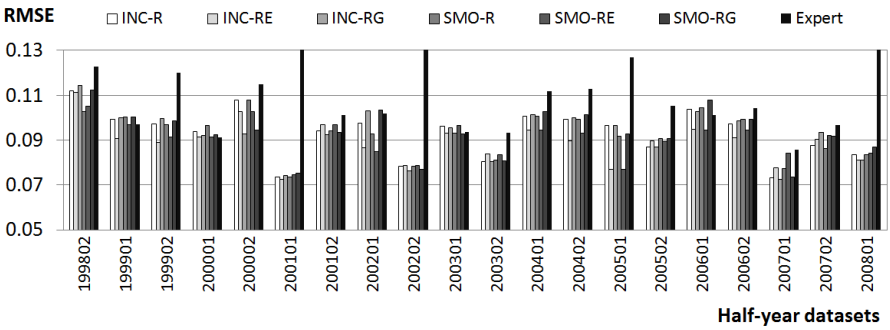


Fig. 4. Comparison of models built using SVR for half-year datasets

The Friedman test performed in respect of RMSE values of all models built over 11 one-year and 20 half-year datasets showed that there were significant differences between some models. Average ranks of individual models are shown in Table 5, where the lower rank value the better model. The incremental algorithm *INC-RE* was the first in both lists. Adjusted p-values for Nemenyi's, Holm's, Shaffer's, and Bergmann-Hommel's post-hoc procedures for *N*×*N* comparisons for all possible pairs of algorithms are shown in Tables 6 and 7 for one-year and half-year datasets, respectively. For comparison, the results of paired Wilcoxon tests are placed in both tables. The p-values indicating the statistically significant differences between given pairs of algorithms are marked with italics. The significance level considered for the null hypothesis rejection was 0.05. Following main observations could be done based on the most powerful Shaffer's, and Bergmann-Hommel's post-hoc procedures: *INC-R, INC-RE, SMO-R*, and *SMO-RE* algorithms revealed significantly better performance than Expert method. The same could be observed in the case of *INC-RG* but for only half-year datasets. There are not significant differences among the incremental and non-incremental SVR algorithms. The paired non-parametric Wilcoxon test can lead to over-optimistic decisions because it allows for rejection of a few more null hypotheses.

Table 5. Average rank positions of SVR models determined during Friedman test

	1st	2nd	3rd	4th	5th	6th	7th
One-year	INC-RE (2.36)	INC-R (2.50)	SMO-RE (3.09)	SMO-R (3.95)	INC-RG (4.27)	SMO-RG (4.82)	Expert (7.00)
Half-year	INC-RE (2.70)	SMO-RE (3.45)	INC-RG (3.70)	SMO-R (3.70)	INC-R (3.85)	SMO-RG (4.55)	Expert (6.05)

Table 6. Adjusted p-values for *N*×*N* comparisons of SVR models over 11 one-year datasets

Model vs Model	**pWilcox**	**pNeme**	**pHolm**	**pShaf**	**pBerg**
INC-RE vs Expert	*0.003346*	*1.01E-05*	*1.01E-05*	*1.01E-05*	*1.01E-05*
INC-R vs Expert	*0.003346*	*2.17E-05*	*2.07E-05*	*1.55E-05*	*1.55E-05*
SMO-RE vs Expert	*0.003346*	*4.62E-04*	*4.18E-04*	*3.30E-04*	*2.42E-04*
SMO-R vs Expert	*0.003346*	*0.019859*	*0.017022*	*0.014185*	*0.010402*
INC-RG vs Expert	*0.003346*	0.064440	0.052166	*0.046029*	*0.033754*

Table 6. (*continued*)

INC-RE vs SMO-RG	*0.016369*	0.161818	0.123290	0.115584	0.115584
INC-R vs SMO-RG	*0.004439*	0.248790	0.177707	0.177707	0.118471
SMO-RG vs Expert	*0.003346*	0.374940	0.249960	0.196397	0.196397
INC-RE vs INC-RG	*0.016369*	0.802502	0.496787	0.420358	0.382144
INC-R vs INC-RG	*0.004439*	1.000000	0.651490	0.597199	0.382144
SMO-RE vs SMO-RG	*0.040861*	1.000000	0.668474	0.668474	0.425393
INC-RE vs SMO-R	*0.016369*	1.000000	0.841455	0.841455	0.589019
INC-R vs SMO-R	*0.028418*	1.000000	1.000000	1.000000	1.000000
INC-RG vs SMO-RE	0.091162	1.000000	1.000000	1.000000	1.000000
SMO-R vs SMO-RE	0.109512	1.000000	1.000000	1.000000	1.000000
SMO-R vs SMO-RG	0.050461	1.000000	1.000000	1.000000	1.000000
INC-RE vs SMO-RE	*0.007646*	1.000000	1.000000	1.000000	1.000000
INC-R vs SMO-RE	0.722108	1.000000	1.000000	1.000000	1.000000
INC-RG vs SMO-RG	0.109512	1.000000	1.000000	1.000000	1.000000
INC-RG vs SMO-R	0.286004	1.000000	1.000000	1.000000	1.000000
INC-R vs INC-RE	0.286004	1.000000	1.000000	1.000000	1.000000

Table 7. Adjusted p-values for *N*×*N* comparisons of SVR models over 20 half-year datasets

Model vs Model	**pWilcox**	**pNeme**	**pHolm**	**pShaf**	**pBerg**
INC-RE vs Expert	*0.000103*	*1.97E-05*	*1.97E-05*	*1.97E-05*	*1.97E-05*
SMO-RE vs Expert	*0.000189*	*0.002966*	*0.002825*	*0.002118*	*0.002118*
INC-R vs Expert	*0.000390*	*0.012214*	*0.011051*	*0.008724*	*0.006398*
INC-RG vs Expert	*0.000780*	*0.012214*	*0.011051*	*0.008724*	*0.006398*
SMO-R vs Expert	*0.000390*	*0.026876*	*0.021757*	*0.019197*	*0.014078*
INC-RE vs SMO-RG	*0.030366*	0.142097	0.108264	0.101498	0.101498
SMO-RG vs Expert	*0.000892*	0.590269	0.421621	0.421621	0.309188
INC-RE vs SMO-R	*0.027622*	1.000000	1.000000	1.000000	0.922923
SMO-RE vs SMO-RG	0.100459	1.000000	1.000000	1.000000	1.000000
INC-RE vs INC-RG	0.073139	1.000000	1.000000	1.000000	1.000000
INC-R vs INC-RE	*0.022769*	1.000000	1.000000	1.000000	1.000000
INC-R vs SMO-RG	0.125860	1.000000	1.000000	1.000000	1.000000
INC-RG vs SMO-RG	0.172996	1.000000	1.000000	1.000000	1.000000
INC-RE vs SMO-RE	0.093604	1.000000	1.000000	1.000000	1.000000
SMO-R vs SMO-RG	0.313464	1.000000	1.000000	1.000000	1.000000
SMO-R vs SMO-RE	0.125860	1.000000	1.000000	1.000000	1.000000
INC-RG vs SMO-RE	0.295878	1.000000	1.000000	1.000000	1.000000
INC-R vs SMO-RE	0.135358	1.000000	1.000000	1.000000	1.000000
INC-R vs SMO-R	0.777565	1.000000	1.000000	1.000000	1.000000
INC-RG vs SMO-R	0.910825	1.000000	1.000000	1.000000	1.000000
INC-R vs INC-RG	0.881293	1.000000	1.000000	1.000000	1.000000

4 Conclusions and Future Work

The experiments aimed to compare the performance of three versions of incremental SVR algorithms and three versions of sequential minimal optimization algorithms for regression were conducted. The algorithms differed in types of kernel function used. Moreover, the predictive accuracy of a pairwise comparison method employed by professional appraisers in reality was compared with the SVR models applied for residential premises valuation.

The overall results of our investigation were as follows. Four of six SVR algorithms revealed prediction accuracy significantly better than the experts' method employed in reality. It confirms that automated valuation models based on SVR can be successfully utilized to support appraisers' work. Moreover, no statistically significant differences among both incremental SVR and non-incremental SMO algorithms were observed using our stationary cadastral datasets. The results open the opportunity of further research into the application of incremental SVR algorithms to predict from data stream of real estate sales/purchase transactions.

Non-parametric statistical procedures especially designed for the comparison of multiple algorithms over multiple datasets were applied to the analysis of the experimental results. They consisted of the Friedman test followed by Nemenyi's, Holm's, Shaffer's, and Bergmann-Hommel's post-hoc procedures. Compared to pairwise Wilcoxon test they discard smaller number of null hypotheses. Therefore, they are able to detect only stronger differences among algorithms.

Acknowledgments. This paper was partially supported by the Polish National Science Centre under grant no. N N516 483840.

References

1. Peterson, S., Flangan, A.B.: Neural Network Hedonic Pricing Models in Mass Real Estate Appraisal. Journal of Real Estate Research 31(2), 147–164 (2009)
2. Pi-ying, L.: Analysis of the Mass Appraisal Model by Using Artificial Neural Network in Kaohsiung City. Journal of Modern Accounting and Auditing 7(10), 1080–1089 (2011)
3. González, M.A.S., Formoso, C.T.: Mass appraisal with genetic fuzzy rule-based systems. Property Management 24(1), 20–30 (2006)
4. Kusan, H., Aytekin, O., Özdemir, I.: The use of fuzzy logic in predicting house selling price. Expert Systems with Applications 37(3), 1808–1813 (2010)
5. Kontrimas, V., Verikas, A.: The mass appraisal of the real estate by computational intelligence. Applied Soft Computing 11(1), 443–448 (2011)
6. Zurada, J., Levitan, A.S., Guan, J.: A Comparison of Regression and Artificial Intelligence Methods in a Mass Appraisal Context. Journal of Real Estate Research 33(3), 349–388 (2011)
7. Graczyk, M., Lasota, T., Trawiński, B.: Comparative Analysis of Premises Valuation Models Using KEEL, RapidMiner, and WEKA. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) ICCCI 2009. LNCS (LNAI), vol. 5796, pp. 800–812. Springer, Heidelberg (2009)

8. Król, D., Lasota, T., Trawiński, B., Trawiński, K.: Investigation of Evolutionary Optimization Methods of TSK Fuzzy Model for Real Estate Appraisal. International Journal of Hybrid Intelligent Systems 5(3), 111–128 (2008)
9. Lasota, T., Mazurkiewicz, J., Trawiński, B., Trawiński, K.: Comparison of Data Driven Models for the Validation of Residential Premises using KEEL. International Journal of Hybrid Intelligent Systems 7(1), 3–16 (2010)
10. Lasota, T., Telec, Z., Trawiński, B., Trawiński, K.: Investigation of the eTS Evolving Fuzzy Systems Applied to Real Estate Appraisal. Journal of Multiple-Valued Logic and Soft Computing 17(2-3), 229–253 (2011)
11. Lughofer, E., Trawiński, B., Trawiński, K., Kempa, O., Lasota, T.: On Employing Fuzzy Modeling Algorithms for the Valuation of Residential Premises. Information Sciences 181, 5123–5142 (2011)
12. Basak, D., Pal, S., Patranabis, D.C.: Support Vector Regression. Neural Information Processing – Letters and Reviews 11(10), 203–224 (2007)
13. Smola, A.J., Schölkopf, B.: A Tutorial on Support Vector Regression. Statistics and Computing 14, 199–222 (2004)
14. Chang, C.C., Lin, C.J.: Training *v*-support vector regression: Theory and algorithms. Neural Computation 14, 1959–1976 (2002)
15. Ma, J., Thelier, J., Perkins, S.: Accurate on-line Support Vector Regression modeling. Neural Computation 15(11), 2683–2703 (2003)
16. Gâlmeanu, H., Andonie, A.: Incremental / decremental SVM for function approximation. In: Proc. of the 11th International Conference on Optimization of Electrical and Electronic Equipment, OPTIM 2008 (2008), doi:10.1109/OPTIM.2008.4602473
17. Cauwenberghs, G., Poggio, T.: Incremental and Decremental Support Vector Machine Learning. In: Leen, T.K., et al. (eds.) Advances in Neural Information Processing Systems, vol. 13, pp. 409–415. MIT Press, Cambridge (2001)
18. Scholkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2001)
19. Shevade, S.K., Keerthi, S.S., Bhattacharyya, C., Murthy, K.R.K.: Improvements to SMO Algorithm for SVM Regression. IEEE Transactions on Neural Networks 11(5), 1188–1193 (2000)
20. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30 (2006)
21. García, S., Herrera, F.: An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons. Journal of Machine Learning Research 9, 2677–2694 (2008)
22. Trawiński, B., Smętek, M., Telec, Z., Lasota, T.: Nonparametric Statistical Analysis for Multiple Comparison of Machine Learning Regression Algorithms. International Journal of Applied Mathematics and Computer Science 22(4) (2012) (in print)
23. Patrascu, P.: Implementation and testing of incremental support vector regression system to assist with real estate appraisals. Master's Thesis, Wrocław University of Technology, Wrocław, Poland (2011) (in Polish)
24. Staelin, C.: Parameter selection for support vector machines, HP Laboratories Israel. Tech. Rep. HPL-2002-354, R.1 (2002)
25. Momma, M., Bennett, K.P.: A Pattern Search Method for Model Selection of Support Vector Regression. In: SIAM Conference on Data Mining (2002)

A Descriptive Method for Generating siRNA Design Rules

Bui Thang Ngoc, Tu Bao Ho, and Kawasaki Saori

Japan Advanced Institute of Science and Technology 1-1 Asahidai, Nomi City, Ishikawa, 923-1292 Japan

Abstract. Short-interfering RNAs (siRNAs) suppress gene expression through a process called RNA interference (RNAi). Current research focuses on finding design principles or rules for siRNAs and using them to artificially generate siRNAs with high efficiency of gene knockdown ability. Design rules have been reported by analyzing biology experiments and applying learning methods. However, possible good design rules or hidden characteristics remain undetected. In contribution to computational methods for finding design rules which are mostly employed by discriminative learning techniques, in this paper we propose a novel descriptive method to discover two design rules for effective siRNA sequences with 19 nucleotides (nt) and 21 nt in length that have important characteristics of previous design rules and contain new characteristics of highly effective siRNA. The key idea of the method is first to transform siRNAs to transactions then apply an Apriori adaptation with automatic *min_support* values to detect descriptive rules for effective and ineffective siRNAs. Rational design rules are created by analyzing graphical representations of descriptive rules. Experimental evaluation on the two siRNA data sets including 5737 siRNA sequences shown that our design rules are promising to design siRNAs effectively.

Keywords: RNAi, siRNA, Apriori algorithm, discriminative learning techniques, descriptive method, design rules.

1 Introduction

In 2006, Fire and Mello received Nobel prize for their contributions to RNA interference (RNAi). Their contributions as well as other research groups' to discovery of RNAi have already had an immense impact on biomedical research and will most likely lead to novel medical applications in the future. RNAi is a powerful technique for post–transcriptional silencing of messenger RNA (mRNA). In RNAi, double strand RNA (dsRNA) sequences are introduced into cells and cleaved into short interfering RNA sequences (siRNAs). After that, each siRNA binds to its complementary target mRNA and induces its degradation. Therefore, the translation process of the mRNA into protein will be prevented and infection by RNA viruses can be blocked. On RNAi research, designing of effective siRNAs, which can silence mRNA sequences efficiently, is one of the most

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 196–205, 2013.

important challenges. Numerous biological works have been carried out in order to clarify rational design rules to generate effective siRNAs.

In the view point of biological research, the first rational design rules for siRNAs were proposed by Elibalshir [1],[2],[3]. They suggested that effective siRNAs having 19–21 nt in length with 2 nt overhangs at 3' end can silence mRNA efficiently. LiSa J Scherer et al. [4] reported that the thermodynamic properties (G/C content of siRNA) to target specific mRNA are important characteristics. Soon after these early works, many rational rules [5],[6],[7],[8],[9] for effective siRNAs have been reported. Characteristics of these rules relate to thermodynamic properties, point-specific nucleotides and specific motif sequences.

Although the positional nucleotide characteristics for siRNA design rules are considered as the most important factor to determine effective siRNAs, there exist inconsistencies among proposed design rules. Most of previous design rules have the same statements at position 1 and 19 on siRNAs but have some inconsistencies at other positions. This also implies that these rules might result in the generation of many candidate siRNAs and thus make it dificult to extract a few of them for synthesizing effective siRNAs. Furthermore, previous emprical analysises only based on small data sets and focused on specific genes. Therefore, these rules may be not enough information to design effective siRNAs.

In computational approach, some discriminative methods have been applied to find design rules and select effective siRNAs. Chalk et al [10] reported thermodynamic properties by using the regression tree tool in BioJava software. According to them, the score of a siRNA candidate is incremented by one for each rule fulfilled giving a score range of (0,7). Teramoto et al. [11] and Ladunga [12] used Support Vector Machine (SVM) to select effective siRNAs. Teramoto adapted the string kernel with Libsvm program to classify siRNAs into effective and ineffective classes by representing each siRNA as k–mer subsequences. Ladunga used SVMLight package with polynominal kernels to train over 2200 siRNA sequences. Huesken et al. [13] discovered motifs for effective and ineffective siRNA sequences based on the significance of nucleotides by applying the artificial neural network to train more than 2,400 siRNAs targeting human as well as rodent genes. Shigeru Takasaki [15] proposed prediction methods based on neural networks and decision trees for selecting effective siRNA from many possible candidates. In the first method, the author used K-means algorithm to calculate variances and centers of each Radial Basis Function corresponding to K nodes on the hidden layer. The similarity of two sequences used is Euclidean distance. In the second one, a decision tree is divided into growing and pruning steps. The testing data were used to check the increasing missclassification error in the tree pruning step. Morever, he combined two methods to increase efficiency of the prediction.

However, these discriminative techniques are potentially unsuitable to detect hidden characteristics of data. The relationships of charateristics are not explicit and visualiable. Neural networks can not guarantee the solution and produce diffirent results when training again with the same data. In Takasaki work, the meaning of clusters were not mentioned and Euclidean distance is also not good

to assess similarity of each pair of siRNAs. Thus, K-means algorithm in this case can get bad results. Moreover, the decision tree method can not generalise the data well because of overfitting and it can be unstable because small variations in data may result different trees or the different design rules.

To overcome those above drawbacks, descriptive approach is the promising way to find important characteristics from data and describe data explicitly. It also clarifies the relationships between characteristics of data. Therefore, a new descriptive method will be proposed to detect rational design rules for effective siRNAs that mostly have characteristics of previous design rules and contain new characteristics of effective siRNAs. The method will detect descriptive rules by apdating Apriori algorithm with automatic $min\_support$ values after transforming siRNAs to transactions. The detected descriptive rules are filtered, graphically represented and analyzed to generate design rules for effective siRNAs.

2 Related Work

In this section, we describe previous design rules for effective siRNA sequences as follows.

Reynolds et al.[5] analyzed 180 siRNAs systematically and reported the following eight criteria for improving siRNA selection: (1) G/C content 30−52%, (2) at least 3 As or Us at positions 15−19, (3) absence of internal repeats, (4) an A at position 19, (5) an A at position 3, (6) a U at position 10, (7) a base other than G or C at position 19, (8) a base other than G at position 13.

Ui-Tei et al. [6] examined 72 siRNAs targeting six genes and reported four rules for effective siRNA designs. They summarized the following charateristics: (1) A or U at position 19, (2) G or C effective at position 1, (3) at least five U or A residues from positions 13–19, (4) no GC stretch more than 9 nt long.

Amarzguioui and Prydz [7] analyzed 46 siRNAs targeting single genes and reported the following six rules for effective siRNA designs based on their literature: (1) $\Delta T_3 = T_3 - T_5$, the difference between the number of A/U residues in the three terminal positions at $3'$ and $5'$ ends of sense strand. $\Delta T_3 > 1$ is positively correlated with functional siRNA; (2) G or C residue at position 1, positively correlated; (3) an U residue at position 1, negatively correlated; (4) an A residue at position 6, positively correlated; (5) A or U at position 19, positively correlated; (6) G at position 19, negatively correlated.

Hsieh et al.[8] did experiment with 138 siRNAs targeting 22 genes and reported the following characteristics: (1) Nucleotide C is negative at position 6, (2) nucleotide C or G is positive and A or U is negative at position 11, (3) Nucleotide A is positive at position 13, (4) Nucleotide G is positive at position 16, (4) Nucleotide U is positive and nucleotide G is negative at position 19.

Jagla et al.[9] tested 601 siRNAs targeting one exogenous and three endogenous genes and reported four rules in the following way: (1) A or U positive at position 19, (2) A or U positive at position 10, (3) G or C positive at position 1, (4) more than three A/Us between positions 13 and 19.

3 Method

To generate design rules for effective siRNAs, our framework is described as following steps:

1. Transform siRNA sequences in original data set to transactions.
2. Apply the adaptive Apriori algorithm with automatic $min\_support$ values to detect decriptive rules for effective and ineffective siRNAs.
3. Filter descriptive rules and generate design rules for effective siRNAs.

3.1 Transforming siRNA Sequences to Transactions

Let n be length of siRNA sequences. In order to transform each siRNA sequence $v_1v_2 \dots v_n$ to a transaction, it is considered as a set of pairs $\{(1, v_1), (2, v_2), \dots, (n, v_n)\}$ where each pair (p, v) indicates the nucleotide v at position p $(1 \leq p \leq n)$. A function is built to map each pair (p, v) to a positive integer number. Each function value is considered as an item in the transaction data set. The nucleotides 'A', 'C' , 'G' and 'U' are respectively mapped to 1, 2, 3 and 4 values to define this function as following:

$$f : \{1 \dots n\} \times \{1, 2, 3, 4\} \to \mathbb{N}$$

$$f(p, v) = 4(p - 1) + v$$

Hence, each siRNA sequence $v_1v_2 \dots v_n$ is transformed to a set of items $\{f(1, v_1), f(2, v_2), \dots, f(n, v_n)\}$. It is easy to see that the function f is a bijective map with the determination region of this function ranging from 1 to $4n$. If function f receives value x, p and v correspond to $x \ mod \ 4$ and $(x - v) \ div \ 4 + 1$. A k-itemset $(1 \leq k \leq n)$ is a set of items $\{f(p_1, v_1), f(p_2, v_2), \dots, f(p_k, v_k)\}$ on this new feature space. The problem is to find frequent itemsets whose frequencies are not less than $min\_support$ value.

3.2 The Adaptive Apriori Algorithm to Detect Descriptive Rules

In this section, $min\_support$ value is defined to determine $(k+1)$-frequent itemset joined by two k-frequent itemsets. Let P be a set of transactions having items in k-frequent itemset $\{f(p_1, v_1), f(p_2, v_2), \dots, f(p_{k-1}, v_{k-1}), f(p_k, v_k)\}$ and Q be a set of transactions having items in k-frequent itemset $\{f(p_1, v_1), f(p_2, v_2), \dots, f(p_{k-1}, v_{k-1}), f(p'_k, v'_k)\}$.

In set P, we consider items at position p'_k on transactions. There are four items $f(p'_k, 1)$, $f(p'_k, 2)$, $f(p'_k, 3)$ and $f(p'_k, 4)$ at this position. Thus, by applying Dirichlet principle, when the set P is didived into four subsets based on the above position, there is at least one subset A of P that its cadinality satifies the following inequation:

$$|A| \geq \lceil \frac{|P|}{4} \rceil + 1 \quad (1)$$

The subset A of P that its transactions have $f(p'_k, v'_k)$ item is considered. It is clear that these transactions have $(k + 1)$ items $f(p_1, v_1), f(p_2, v_2), \dots, f(p_k, v_k)$

Algorithm 1. The Adaptive Apriori algorithm to detect descriptive rules for effective siRNA

Input: Effective siRNA sequences in C_1 class.
Output: Descriptive rules for effective siRNAs S_1.
for $s = 1 \rightarrow |C_1|$ **do**
  Transform siRNA sequence s to transaction using f function.
end for
k=1;
L_k= {2-itemset}, count frequence of 2-itemsets.
repeat
  k=k+1
  $S_1 = S_1 \cup L_k$
  for each pair of k-itemsets in L_k **do**
    Generate $(k+1)$-itemset t.
    Compute $min\_support$
    if frequence of $t \geq min\_support$ **then**
      $L_{k+1} = L_{k+1} \cup \{t\}$
    end if
  end for
until ((k$\geq$ n) $\vee$ ($L_{k+1} = \emptyset$))

and $f(p'_k, v'_k)$ and frequence of this $(k+1)$-itemset is cadinality of A. In case the cadinality of A satifies the above inequation, we call this $(k+1)$-itemset to be frequent.

We analyse the same way for Q set when considering items at the position p_k on transactions. $(k+1)$-itemset is frequent if cardinality of subset A satifies the following inequation:

$$|A| \geq \lceil \frac{|Q|}{4} \rceil + 1 \quad (2)$$

From (1) and (2): $(k+1)$-itemset joined two above k-frequent itemsets is frequent if its frequence satifies the equation (1) or (2). Thus, frequence of $(k+1)$-itemset satifies the following inequation:

$$|A| \geq \min\{\lceil \frac{|P|}{4} \rceil + 1, \lceil \frac{|Q|}{4} \rceil + 1\}$$

$$\Leftrightarrow |A| \geq \lceil \frac{\min\{|P|, |Q|\}}{4} \rceil + 1$$

Therefore, $min\_support$ is defined as right formula of inequation.

$$min\_support = \lceil \frac{\min\{|P|, |Q|\}}{4} \rceil + 1 \quad (3)$$

Let C_1 and C_2 denote effective and ineffective siRNA classes, respectively. Let S_1 and S_2 denote sets of frequent itemsets in class C_1 and C_2. Let L_k is a set of k-frequent itemsets

Descriptive rules are detected by applying the apdative Apriori algorithm. Unlike traditional Apriori algorithm, the candidate gereration and frequent itemset

mining steps are combined into one step. In this step, $min\_support$ value is computed using equation 3, $(k+1)$-itemset joined by the two k-frequent itemsets will be checked whether it is frequent or not. The adaptive Apriori algorithm is described at the Algorithm 1. It is also applied to detect descriptive rules for ineffective siRNA sequences in class C_2.

3.3 Filtering Descriptive Rules and Generating Design Rules

The adaptive Apriori algorithm can result many redundant frequent itemsets. It means that there exist rules that generalize these rules. Therefore, redundant rules have to be eliminated from S_1 and S_2. On the other hand, frequent itemsets in S_1 and S_2 have to be accurate rules. Thus, the confidences of itemsets in S_1 or in S_2 have to equal to 1. The Algorithm 2 shows the filtering descriptive rules in S_1. The filtering descriptive rules S_2 is done the same way. After filtering descriptive rules in S_1 and S_2, filtered descriptive rules are graphically represented as sequence logos by using Weblogo tool. On sequence logos, the height of a nucleotide at each position represents its contribution to design rule for siRNAs. Therefore, design rules are generated by choosing nucleotides in descreasing order of their height at each position on sequence logos.

Algorithm 2. Filtering descriptive rules for effective siRNAs

```
Input: Descriptive rules in S1, siRNAs in C1 and C2.
Output: Filtered descriptive rules.
// eliminate inconfident rules in S1
for each descriptive rule t in S1 do
  for each siRNA s in C2 do
    if ( s contains t) then
      S1 = S1\ {t}
    end if
  end for
end for
// eliminate redundant rules in S1
for each descriptive rule t in S1 do
  for each descriptive rule r in S1 do
    if (r! = t ) & (r contains t) then
      S1 = S1 \ {t}
    end if
  end for
end for
```

4 Experimental Evaluation

In this section, our method is applied to generate two design rules for effective siRNAs with 19 nt and 21 nt in length. Our rules are also assessed as previous rules.

In our experiment, two data sets are used. These data sets are collected from siRecord data set which is the biggest siRNA data set. In siRecord, very high and low siRNA sequences with 19 nt in length are collected for our first data set. This data set consists of 2470 effective and 1261 ineffective siRNA sequences corresponding to very high and low siRNA sequences, respectively. The second one contains 1461 effective and 538 ineffective siRNA sequences with 21 nt in length. $min\_support$ values are automatically defined, however, it can be descreased to zero. Therefore, low bound of $min\_support$ is set 10 in our experiment. The programs are coded in C++ on Dev-cpp environment. The processor speed of computer is 2.52 GHz and memory is 4 GB.

In process to generate design rule for effective siRNAs with 19 nt in length by using the first data set, 153 and 5 filtered descriptive rules for effective and ineffective siRNA sequences are detected. Figure 1 and Figure 2 represent the two sequence logos for two above types of filtered rules. The above two sequence logos are analysed to generate a rational design rule for highly effective siRNAs with 19 nt in length. Our rule shows that effective siRNA sequences with 19 nt in length have the sixteen following characteristics: A 'G/C' and absence of 'A/U' at position 1 (1), An 'A' and absence of 'U' at position 2 (2), an 'A' at position 3 (3), absence of 'A' at position 4 (4), An 'A' and absence of 'C' at position 6(5), an 'A/G' at position 7 (6), a 'C' at position 9 (7), an 'U' at position 10 (8), an 'A/G/U' at position 11 (9), an 'A/C/U' at position 13 (10), an 'A/G' at position 14 (11), an 'A/U' and absence of 'C' at position 15 (12), an 'A/G/U' at position 16 (13), An 'A/U' and absence of 'G/C' at position 17 (14), An 'A/U' and absence of 'G/C' at position 18 (15), An 'A' and absence of 'G' at position 19 (16). This rule is called DR19 and represented on Table 1.

When applying our method on the second data set, 332 filtered descriptive rules for effective siRNAs with 21 nt in length are detected. However, the set of filtered descriptive rule for ineffective siRNAs is empty. it may be caused by the imbalance of the data set. Sequence logo of filtered descriptive rules for effective siRNAs is shown on Figure 3. The design rule for effective siRNA with 21 nt in length has twenty following characteristics: an 'A/G' at position 1 (1), an 'A' at position 2 (2), a 'G/C' at position 3 (3), an 'A/C' and absence of 'U' at position 4 (4), an 'U/A' and absence of 'C' at position 5 (5), an 'G/C/U' at position 6 (6), absence of 'A' at position 7 (2), an 'A/U' at position 8 (8), a 'G/U' and absence of 'A' at position 9 (9), an 'U' at position 10 (10), an 'U' at position 11 (11), absence of 'U' at position 13 (12), absent 'C' at position 14 (13), absence of 'C' at position 15 (14), an 'A' and absence of 'U' at position 16 (15), an 'A/G' and absence of 'C' at position 17 (16), an 'U' and absence of 'C' at postion 18 (17), an 'A' at position 19 (18), 'A/U' and absence of 'G/C' at position 20 (19), 'A/U' and absence of 'G' at position 21 (20). The rule is called DR21 and represented on Table 2.

The DR19 is in a good agreement with previous design rules at some characteristics as follows.

Fig. 1. Sequence logo of design rules for effective siRNA with 19 nt in length

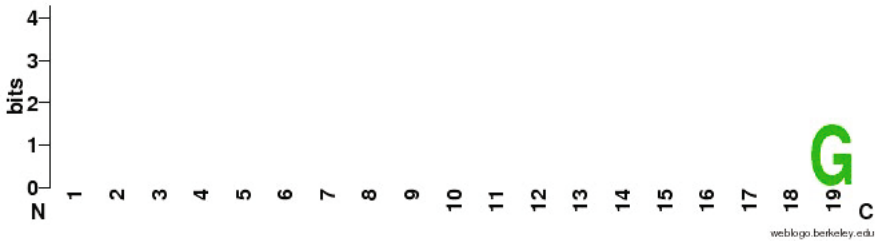


Fig. 2. Sequence logo of design rules for ineffective siRNA with 19 nt in length

Table 1. Rational rules for effective siRNA with 19 nt in length

	Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Reynolds	effective			A							U			A/C/U						A/U
Ui-Tei	effective	G/C																		A/U
	ineffective	A/U																		G/C
Amarzguioui	effective	G/C					A							U						A/U
	ineffective	U									U									G
Jagla	effective	G/C									A/U									A/U
Hsieh	effective											G/C		A			G			U
	ineffective						C					A/U								G
Our rule	effective	G/C	A	A			A	A/G		C	U	A/G/U		A/C/U	A/G	A/U	G/A/U	A/U	U/A	A
	ineffective				A		G													G

– A ‘G/C’ at position 1 but nucleotide ‘G’ is more important than ‘C’ at this position
– An ‘A’ at position 3 as Reynolds’ rule
– An ‘A’ at postion 6 as Amarzguioui’s rule. Absence of ‘C’ at this position as Hsieh’s rule
– An ‘U’ at position 10 as Reynolds’ rule and Jagla’s rule
– Absence of ‘G’ at position 13 as Reynolds’ rule
– Absence of ‘G’ at position 13 as Reynolds’ rule
– A ‘G’ at position 16 as Hsieh’s rule
– An ‘A’ at position 19
– Absence of ‘G’ at position 19 as Reynolds’ rule, Amarzguioui’s rule and Hsieh’s rule

Interestingly, DR19 contains new characteristics that makes it satisfy other important characteristics of previous rules such as thermodynamic propertities or GC content ranging from 30% to 52% (In our case, GC content ranges from 36% to 52%); at least three ’A/U’ at positions from 15 to 19; at least five ‘A/U’ at

Fig. 3. Sequence logo of design rules for effective siRNA with 21 nt in length

Table 2. Rational rules for effective siRNA with 21 nt in length

	Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Huesken	effective	A/G		C								U	A			A/U	A			A/U	A/U	A/U
	ineffective	no G		A								C				C						
Our rule	effective	A/G	A	G/C	C/A, no U	U/A, no C	G	G/C, no A	A/G	G/U, no A	U	U		G, no U	A, no C	U/A, no C	A, no U	G/A, no C	U	A	U/A	A/U
	ineffective																					

positions from 13 to 19 (characterizing for effective silencing, efficience siRNAs entry into RISC and ability of siRNA duplex to unwind); no GC stretch more than 9 nucleotides. In addition, new characteristics in the seed region ranging from position 2 to position 7 may play an important role to avoid off-target effects of siRNA that is also one of challenging proplems in RNAi [14]. Moreover, characteristics of DR19 in the region (9-11) can make siRNAs recognite and cleave target mRNA sequences [5]. Therefore, DR19 not only integrates characteristics of previous design rules but also provides new characteristics for effective siRNAs.

The DR21 is compared to Huesken's motifs which is generated by using neural network. These two rules have the same conclusion at following points:

- An 'A/G' at position 1
- A 'G/C' at position 3
- An 'A' at position 8
- An 'U' at position 11
- An 'A/U' at position 15, no 'C' at this position
- An 'A' at position 16
- An 'A' at position 19
- An 'A/U' at position 20
- An 'A/U' at position 21

The DR21 rule does not give any conclusion at positions 12 as Huesken's rule because in DR21, contributions of different nucleotides at this position are similar to together. Thus, no nucleotide has more significant than the other. Another different point between these two rules is that DR21 rule contains new characteristics in the seed region (4-9) as the DR21 to avoid off-target effects of siRNAs. Moreover, Huesken's rule does also not include characteristics of two nucleotides overhang at 3' end although these nucleotides can improve effective silencing.

5 Conclusion

We have a descriptive approach to find two design rules for effective siRNAs with 19 nt and 21 nt in length. The design rules not only contain the important characteristics of previous design rules but also have new characteristics to design siRNA effectively. In addition, we also define automatic *min_support* values for adatptive Apriori algorithm to detect descriptive rule efficiently.

References

1. Elbashir, S.M., Lendeckel, W., Tuschl, T.: RNA interference is mediated by 21– and 22–nucleotide RNAs. Genes Dev. 15, 188–200 (2001)
2. Elbashir, S.M., Martinez, J., Patkaniowska, A., Lendeckel, W., Tuschl, T.: Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate. EMBO J. 20, 6877–6888 (2001)
3. Elbashir, S.M., Harborth, J., Weber, K., Tuschl, T.: Analysis of gene function in somatic mammalian cells using small interfering RNAs. Methods 26, 199–213 (2002)
4. Scherer, L.J., Rossi, J.J.: Approaches for the sequence-specific knockdown of mRNA. Nat. Biotechnol. 21, 1457–1465 (2003)
5. Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W.S., Khvorova, A.: Rational siRNA design for RNA interference. Nat. Biotechnol. 22(3), 326–330 (2004)
6. Ui-Tei, K., Naito, Y., Takahashi, F., Haraguchi, T., Ohki-Hamazaki, H., Juni, A., Ueda, R., Saigo, K.: Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. Nucleic Acids Res. 32, 936–948 (2004)
7. Amarzguioui, M., Prydz, H.: An algorithm for selection of functional siRNA sequences. Biochem. Biophys. Res. Commun. 316(4), 1050–1058 (2004)
8. Hsieh, A.C., Bo, R., Manola, J., Vazquez, F., Bare, O., Khvorova, A., Scaringe, S., Sellers, W.R.: A library of siRNA duplexes targeting the phosphoinositide 3-kinase pathway: determinants of gene silencing for use in cell-based screens. Nucleic Acids Res. 32(3), 893–901 (2004)
9. Jagla, B., Aulner, N., Kelly, P.D., Song, D., Volchuk, A., Zatorski, A., Shum, D., Mayer, T., De Angelis, D.A., Ouerfelli, O., Rutishauser, U., Rothman, J.E.: Sequence characteristics of functional siRNAs. Rna 2005 11(6), 864–872 (2005)
10. Chalk, A.M., Wahlestedt, C., Sonnhammer, E.L.L.: Improved and automated prediction of effective siRNA. Biochem. Biophys. Res. Commun. 319, 264–274 (2004)
11. Teramoto, R., Aoki, M., Kimura, T., Kanaoka, M.: Prediction of siRNA functionality using generalized string kernel and support vector machine. FEBS Lett. 579, 2878–2882 (2005)
12. Ladunga, I.: More complete gene silencing by fewer siRNAs: Transparent optimized design and biophysical signature. Nucleic Acids Res. 35, 433–440 (2007)
13. Huesken, D., Lange, J., Mickanin, C., Weiler, J., Asselbergs, F., Warner, J., Mellon, B., Engel, S., Rosenberg, A., Cohen, D., Labow, M., Reinhardt, M., Natt, F., Hall, J.: Design of a Genome-Wide siRNA Library Using an Artificial Neural Network. Nature Biotechnology 23(8), 955–1001 (2005)
14. Pei, Y., Tuschl, T.: On the art of identifying effective and specific siRNAs. Nat. Methods 3, 670–676 (2006)
15. Takasaki, S.: Efficient prediction methods for selecting effective siRNA sequences. Comput. Biol. Med. 40, 149–158 (2010)

A Space-Time Trade Off for FUFP-trees Maintenance

Bac Le[1], Chanh-Truc Tran[1], Tzung-Pei Hong[2], and Bay Vo[3]

[1] University of Science, Ho Chi Minh City, Viet Nam
lhbac@fit.hcmus.edu.vn, tructc@gemadept.com.vn
[2] Department of CSIE, National University of Kaohsiung, Taiwan, R.O.C.
tphong@nuk.edu.tw
[3] Information Technology College, Ho Chi Minh City, Viet Nam
vdbay@itc.edu.vn

Abstract. In the past, Hong *et al.* proposed an algorithm to maintain the fast updated frequent pattern tree (FUFP-tree), which was an efficient data structure for association-rule mining. However in the maintenance process, the counts of infrequent items and the IDs of transactions with those items were determined by rescanning all the transactions in the original database. This step might be quite time-consuming depending on the number of transactions in the original database and the number of rescanned items. This study improves that approach by storing 1-items during the maintenance process and based on the properties of FUFP-trees, such that the rescanned items and inserted items are processed more efficiently to reduce execution time. Experimental results show that the improved algorithm needs some more memory to store infrequent 1-items but the performance is better than the original one.

Keywords: data mining, frequent itemset, FUFP-tree, infrequent itemset, incremental mining.

1 Introduction

Data mining is one of the most interesting subjects with many techniques and algorithms developed [1]. Among the research topics of data mining, improving the efficiency of mining association rules from transaction databases has attracted much attention [1-11]. The first several algorithms for mining association rules were based on the *Apiori* algorithm [2], which repeatedly scanned a database to generate and process candidate itemsets level by level and thus needed a high computational cost. In 2000, the frequent pattern-tree (FP-tree) structure was proposed by Han *et al.* [6] for efficiently mining association rules without the generation of candidate itemsets. In real-world applications, a transaction database keeps being updated, and insertion is a very common operation. Efficient maintenance algorithms are thus needed when transactions are inserted [8-9]. In 2008, the incremental fast updated frequent pattern-tree (FUFP-tree) maintenance algorithm for handling transaction insertion was proposed [8]. In that approach, the FUFP-tree is incrementally handled without reconstructing the FUFP-tree from the beginning. However, the original database needs to be rescanned to determine the occurrence of infrequent items, which are not stored during

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 206–214, 2013.

the maintenance process, and to determine the transaction IDs in which the rescanned items appear. This paper improves the above approach for transaction insertion by storing 1-items during the maintenance process and using the properties of FUFP-trees, such that the rescanned items and inserted items are processed more efficiently to reduce execution time.

2 Review of FUFP-trees

An FUFP-tree [8] is similar to an FP-tree except that it has bi-directional links between parent nodes and their child nodes. When new transactions are inserted to the original database, Hong *et al.*'s algorithm processes them to maintain the FUFP-tree without reconstructing it from the updated database. Depending on whether items are frequent (large) in the original database and in the new transactions, there are 4 cases to consider, which are shown in **Table 1**. Each case is processed separately. The Header-Table and the FUFP-tree are then appropriately updated if necessary.

Table 1. Four cases for transaction insertion [8]

Case	*Org. DB*	*New Trans*	*Results*
1	Frequent	Frequent	Always Frequent
2	Frequent	Infrequent	Determined from existing info.
3	Infrequent	Frequent	Determined by rescanning DB
4	Infrequent	Infrequent	Always infrequent

There are some points which can be improved in the original approach. When the original approach processes the items in case 3, the transactions in the original database need to be rescanned for determining the occurrences of infrequent items, which are not stored during the maintenance process. This step is thus the most time-consuming step. The computation time of this step is positively related to the number of transactions in the original database, the number of items in each transaction (the length of each transaction) and the number of items in the set of rescanned items.

3 Improved Algorithm

3.1 Notations

D, T, U: the original database, new transactions, updated database, respectively;
Sup: the minimum support threshold for frequent itemsets;
minSup_Org, *minSup_New*, *minSup*: *the* minimum support count of *D*, *T*, *U*, respectively;
$Count_{Org}(I)$, $Count_{New}(I)$, $Count_{Upd}(I)$: frequency of *I* in *D*, *T*, *U,* respectively;
Flist, *IFlist*: the list of large and small items in *D,* respectively*;*
Flist_New, *IFlist_New*: the list of large and small items in *T,* respectively*;*
Item_Case1, *Item_Case2*, *Item_Case3*, *Item_Case4:* list of items of the four cases, respectively;

Items: a temporary list to store items;
Htable: the *Header-Table* of FUFP-tree;
FUFP_tree: the current FUFP-tree;
Rescan_Items: the list of items to update the FUFP-tree based on the original database;
Insert_Items: the list of items to update the FUFP-tree based on new transactions;
Corresponding branch: the branch generated from the frequent items in a transaction according to the order of items appearing in *Header-Table.*

3.2 Proposed Algorithm

The details of the improved algorithm are shown below.

INPUT: Original database (*D*), Header-Table (*Htable*), FUFP-tree (*FUFP_tree*), support threshold (*Sup*), set of *t* new transactions (*T*).

OUTPUT: A new FUFP-tree for the updated database (*U*).

STEP 1: Scan the new transactions *T* to find their items and counts, and store large items into *Flist_New* and small items into *IFlist_New*.
STEP 2: Based on *Flist*, *IFlist, Flist_New* and *IFlist_New*, find and store items into *Items_Case1*, *Items_Case2*, *Items_Case3 and Items_Case4*, respectively.
STEP 3: For each item *I* in *Items_Case1*, do the following substeps:
Substep 3-1: The new count of *I* in *U*: $Count_{Upd}(I) = Count_{Org}(I) + Count_{New}(I)$.
Substep 3-2: Set the count of *I* in *Htable* = $\text{Count}_{Upd}(I)$.
Substep 3-3: Set the count of *I* in *Flist* = $\text{Count}_{Upd}(I)$.
Substep 3-4: Add *I* to the set of *Insert_Items*.
STEP 4: For each item *I* in *Items_Case2*, do the following substeps:
Substep 4-1: The new count of *I* in *U*: $Count_{Upd}(I) = Count_{Org}(I) + Count_{New}(I)$.
Substep 4-2: Set the count of *I* in *Flist* = $\text{Count}_{Upd}(I)$.
Substep 4-3: If ($Count_{Upd}(I) \geq minSup$), item *I* will still be large in updated DB; update the count of *I* in *Htable* as $\text{Count}_{Upd}(I)$ and add *I* to the set of *Insert_Items*.
Substep 4-4: If ($Count_{Upd}(I) < minSup$), item *I* will become small in updated DB; move *I* from *Flist* to *IFlist,* and remove *I* from the *Htable* and the *FUFP-tree*.
STEP 5: For each item *I* in *Items_Case3*, do the following substeps:
Substep 5-1: The new count of *I* in *U*: $Count_{Upd}(I) = Count_{Org}(I) + Count_{New}(I)$.
Substep 5-2: Set the count of *I* in *IFlist* = $\text{Count}_{Upd}(I)$.
Substep 5-3: If ($Count_{Upd}(I) \geq minSup$), add *I* both to *Insert_Items* and *Rescan_Items*.
STEP 6: Sort the items in *Rescan_Items* in descending order of their updated counts.
STEP 7: Insert the items in the *Rescan_Items* to the end of the *Htable* according to the descending order of their counts and move *I* from *IFlist* to *Flist.*
STEP 8: Update the *FUFP-tree* according to the set of *Rescan_Items*. For each transaction *J* in the original database, do the following substeps:
Substep 8-1: Determine which items of *Rescan_Items* appear in *J*, and store the results to a temporary list *Items*. If the list *Items* has no items, it means that there is no items of *Rescan_Items* appearing in *J*, and redo substep 8-1 with next transaction *J*.

Substep 8-2: Find the corresponding branch *B* of *J* in *FUFP-tree*, and store *B* to the temporary branch, *Branch*.
Substep 8-3: For each item *I* in *Items*, if *I* appears in the corresponding branch *Branch*, add 1 to the count of the node *I* and remove node *I* from *Branch* (from the properties of FUFP-trees, if a node in a specific branch is different from the others, it should not be considered in the next run after being processed. This will speed up the algorithm); otherwise, insert *I* at the end of the branch, set its count as 1, then re-find the new corresponding branch *B*, and store *B* to *Branch*.
STEP 9: Update the FUFP-tree according to the set of *Insert_Items*. For each transaction *J* in the new transactions, do the following substeps:
Substep 9-1: Determine which items of *Insert_Items* appear in *J*, and store the results to a temporary list *Items*. If the list *Items* has no items, it means that there is no items of *Insert_Items* appearing in *J*, and redo substep 9-1 with the next transaction *J*.
Substep 9-2: Find the corresponding branch *B* of *J* in *FUFP-tree* and store *B* to the temporary branch, *Branch*.
Substep 9-3: For each item *I* in *Items*, if *I* appears in the corresponding branch *Branch*, add 1 to the count of the node *I* and remove node *I* from *Branch*, (like substep 8-3); otherwise, insert *I* at the end of the branch, set its count as 1, re-find the new corresponding branch *B*, and store *B* to *Branch*.
STEP 10: For each item *I* in *Items_Case4*, do the following substeps:
Substep 10-1: The new count of *I* in *U*: $Count_{Upd}(I) = Count_{Org}(I) + Count_{New}(I)$.
Substep 10-2: Set the count of *I* in *IFlist* = $Count_{Upd}(I)$.

4 An Example

This section illustrates the proposed algorithm for maintaining an FUFP-tree after transactions are inserted. An original database with 10 transactions and 8 items, from *a* to *h*, is used in this example, which shown in **Table 2**.

Table 2. Original database used for the example

No	Items	No	Items
1	*a, b, c, d, e*	6	*a, c, d, e, g*
2	*a, b, c, f, h*	7	*a, b, h*
3	*b, c, d, e, g*	8	*b, c, d, g*
4	*a, b, f, h*	9	*a, b, d, f*
5	*a, b, f*	10	*a, b, d, h*

Assume the support threshold was set at 50%. For the original database, *minSup_Org* is 5, and the frequent 1-itemsets are *b*, *a*, *d*, and *c*, which are used to construct the *Header-Table.* The FUFP-tree is then built from the original database and *Header-Table*. **Fig.1** shows the results. Assume there are five transactions inserted to the original database as in **Table 3**.The proposed algorithm proceeds as follow.

STEP 1: The five new transactions are first scanned to get the items and their counts. Large items are stored in *Flist_New = {b:4, f:4, a:3, e:3}* and small items are stored in

IFlist_New = {c:2, d:2, g:1} based on *minSup_New* = 5 × 50% = 2.5 (3 by integer). The large items and small items of the original database are stored in *Flist = {b:9, a:8, d:6, c:5}* and *IFlist = {e:3, f:4, h:4, g:4}*, respectively, during the FUFP-tree construction.

Table 3. New inserted transactions

No	Items
1	*a, b, e, f*
2	*c, e, f*
3	*a, b, f*
4	*a, b, d, f, g*
5	*b, c, d, e*

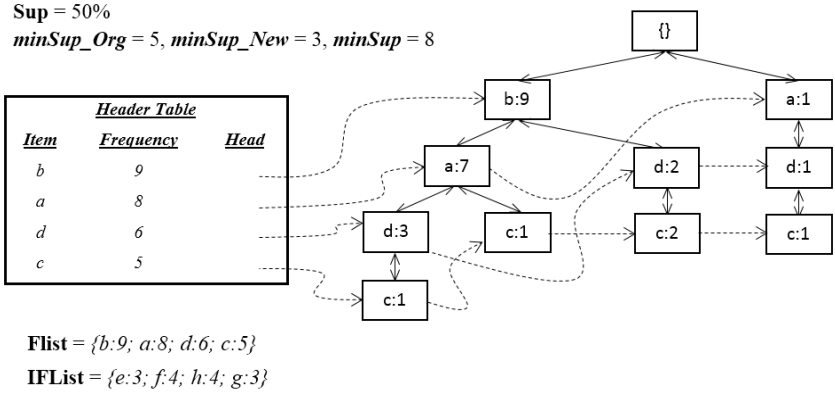


Fig. 1. FUFP-tree and *Header-Table* for the example

STEP 2: From *Flist*, *IFlist*, *Flist_New*, *IFlist_New*, the items of the 4 cases are calculated. In case 1, the items which appear both in *Flist* and *Flist_New* are stored in *Items_Case1* (= {*b, a*}). In case 2, the items which appear in *Flist* but don't exist in *Flist_New* are stored in *Items_Case2* (= {*d, c*}). In case 3, the items which appear in *Flist_New* but do not exist in *Flist* are stored in *Items_Case3* (= {*f, e*}). In case 4, the items which appear in *IFlist* but do not exist in *Flist_New* are stored in *Items_Case4* (= {*h, g*}).

STEP 3 to STEP 5: Each item in *Items_Case1, Items_Case2* and *Items_Case3* are processed by its individual step. After STEP 5, *Insert_Items = {b, a, d, f}* and *Rescan_Items = {f}*. *FUFP-tree*, *Header-Table*, *Flist* and *IFlist* are also updated correspondingly.

STEP 6: The items in the set of *Rescan_Items* are sorted in descending order of their updated counts. In this example, there is only *f*, thus no sorting is needed.

STEP 7: The items in the *Rescan_Items* are inserted to the end of the *Header-Table* according to the descending order of their counts. Thus, *f* is added to the end of *Header-Table*, and then *f* is moved from *IFlist* to *Flist.* The results after STEP 7 are shown in **Fig. 2**.

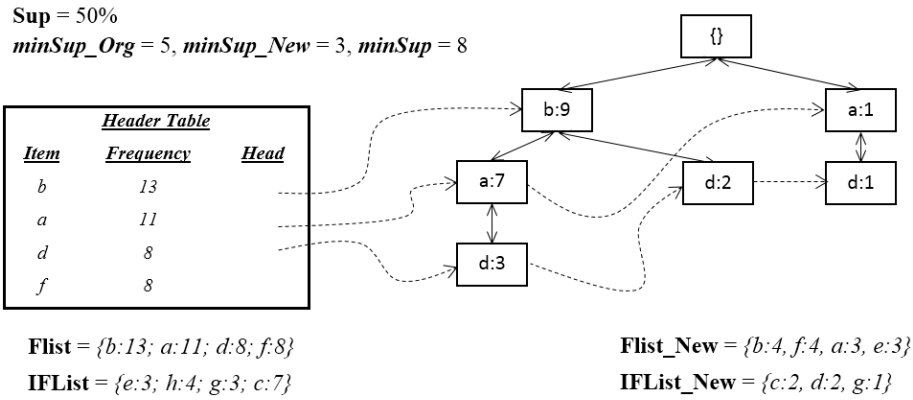


Fig. 2. FUFP-tree, *Header-Table*, *Flist* and *IFlist* after step 7 has been processed

STEP 8: The FUFP-tree is updated according to the transactions in the original database and the *Rescan_Items* (= {*f*}). **Table 4** shows the corresponding branches of the original database with items in *Rescan_Items*.

Table 4. Original transactions and items appear in *Rescan-Items*

No	Original DB	Items	Cor. branch	No	Original DB	Items	Cor. branch
1	*a, b, c, d, e*	-	-	6	*a, c, d, e, g*	-	-
2	*a, b, c, f, h*	*f*	*b* → *a*	7	*a, b, h*	-	-
3	*b, c, d, e, g*	-	-	8	*b, c, d, g*	-	-
4	*a, b, f, h*	*f*	*b* → *a* → *f*	9	*a, b, d, f*	*f*	*b* → *a* → *d*
5	*a, b, f*	*f*	*b* → *a* → *f*	10	*a, b, d, h*	-	-

In this example, each transaction in the original database is processed. Transactions 2, 4, 5 and 9 are processed because they include an item appearing in *Rescan_Items.* The results are shown in **Fig. 3**.

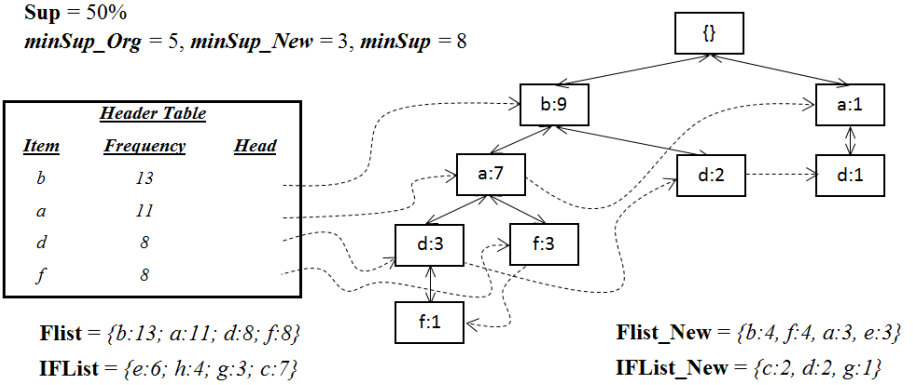


Fig. 3. FUFP-tree, *Header-Table*, *Flist* and *IFlist* after STEP 8

STEP 9: The FUFP-tree is updated according to the transactions in the new transactions and the *Insert_Items* (= {*b, a, d, f*}). **Table 5** shows the corresponding branches of the new transactions with items in *Insert_Items*. Each transaction with its corresponding branch in the new transactions is then processed.

Table 5. New transactions and items appear in *Insert-Items*

No	New trans.	Items	Cor. branch
1	*a, b, e, f*	*b, a, f*	$b \rightarrow a \rightarrow f$
2	*c, e, f*	*f*	-
3	*a, b, f*	*b, a, f*	$b \rightarrow a \rightarrow f$
4	*a, b, d, f, g*	*b, a, d, f*	$b \rightarrow a \rightarrow d \rightarrow f$
5	*b, c, d, e*	*b, d*	$B \rightarrow d$

STEP 10: The counts in *IFlist* of items in case 4 are then updated. Each item in *Items_Case4* is processed. After STEP 10, the final results are shown in **Fig. 4**.

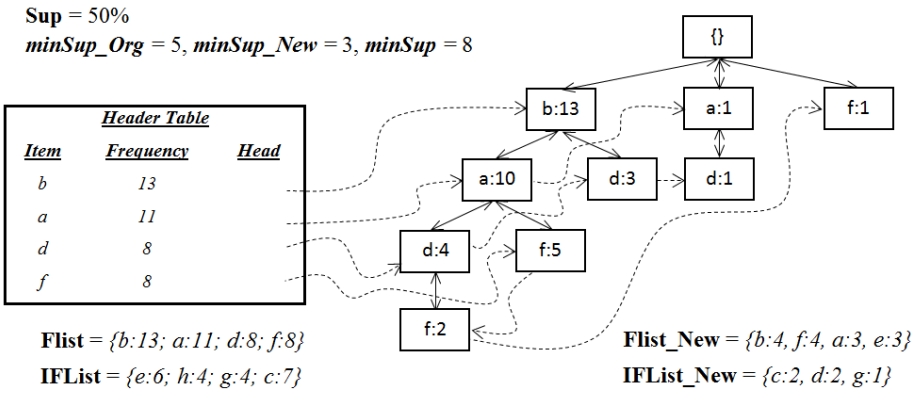


Fig. 4. FUFP-tree, *Header-Table*, *Flist* and *IFlist* after STEP 10 has been processed

5 Experimental Results

Experiments were programmed in C# on a laptop with an Intel 1.73 GHz quad-core CPU and 8GBs of RAM, running Windows 7 Ultimate 64 bits. Two real databases were used in the experiments. One is the BMS-POS and the other is MUSHROOM. The BMS-POS contained several years of point-of-sale data from a large electronics retailer with 515,597 transactions and 1,657 items. The maximal length of a transaction was 164 and the average length of the transactions was 6.5. There are 8,124 transactions with 22 items in the MUSHROOM. The parameters were set the same as Hong et al.'s. For the BMS-POS, the first 400,000 transactions were used to build the initial FUFP-tree and the next 5,000 transactions were sequentially used as new transactions; while for the MUSROOM, the first 5,000 transactions were used initially and the next 500 transactions were inserted each time. The *minSup* was set to 4%, 6%, and 8%. **Table 6** shows the execution time of the two algorithms with three different minimum support thresholds. Each value is the average execution time over 5 runs.

Table 6. Execution time of the two algorithms with different thresholds

%	Algorithms	Run time(s) of each 5,000 trans. inserted					BMS-POS
		5,000	10,000	15,000	20,000	25,000	
4	Hong et al.'s alg.	12.703	9.184	9.355	9.189	9.145	
	Proposed alg.	0.104	0.055	0.054	0.052	0.059	
6	Hong et al.'s alg.	10.861	9.157	9.270	9.173	9.176	
	Proposed alg.	0.128	0.054	0.055	0.056	0.055	
8	Hong et al.'s alg.	11.802	9.224	9.176	9.210	9.143	
	Proposed alg.	0.164	0.055	0.054	0.055	0.054	
%	Algorithms	Run time(s) of each 500 trans. inserted					MUSHROOM
		500	1,000	1,500	2,000	2,500	
4	Hong et al.'s alg.	0.367	0.278	0.291	0.304	0.314	
	Proposed alg.	0.031	0.024	0.021	0.019	0.017	
6	Hong et al.'s alg.	0.353	0.301	0.292	0.253	0.135	
	Proposed alg.	0.028	0.019	0.020	0.065	0.017	
8	Hong et al.'s alg.	0.363	0.382	0.288	0.241	0.139	
	Proposed alg.	0.031	0.141	0.018	0.019	0.019	

The results indicated that the proposed algorithm ran faster than the original approach. The main reasons are that Hong et al.'s approach has to rescan the transactions in the original database to determine the counts of infrequent items and the IDs of transactions in which the infrequent items appear, while the new approach gets the counts of infrequent items directly from *IFlist*, which is stored during FUFP-tree construction. Additionally, the proposed algorithm processes the *Recan_Items* and *Insert_Items* more efficiently based on the properties of the FUFP-tree. The number of nodes and the structure of the result trees generated are the same.

6 Conclusion and Future Work

An improved FUFP-tree maintenance approach for transaction insertion has been proposed. The proposed algorithm does not need to rescan the original database by storing the 1-items during the maintenance process. Moreover, based on the properties of the FUFP-tree, the item of a node in a specific branch is different from the others, thus the steps of updating the FUFP-tree according to *Rescan_Items* and *Insert_Items* are processed more efficiently by pruning out the processed item steps by steps. The execution time of the proposed algorithm is much lower than that of the original algorithm. The numbers of nodes of the FUFP-tree constructed by the two algorithms are the same. The proposed approach, however, requires some more memory to store 1-items. There is a trade-off between memory and execution time. The proposed approach is more efficient for large databases. For small databases with a few thousand of records, such as MUSHROOM, the difference is not very clear.

Lattice-based approaches for efficient mining association rules have been proposed in recent years [12-13]. In the future, we will study how to build frequent itemsets lattice when the database is changed. Besides, we will consider expanding the work in [14] to mine high utility itemsets.

Acknowledgement. This work was supported by Vietnam's National Foundation for Science and Technology Development (NAFOSTED).

References

1. Agrawal, R., Imielinski, T., Swami, A.: Database mining: A performance perspective. IEEE Transactions on Knowledge and Data Engineering 5(6), 914–925 (1993)
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: The 20th International Conference on Very Large Databases, pp. 487–499 (1994)
3. Agrawal, R., Srikant, R., Vu, Q.: Mining association rules with item constraints. In: The Third International Conference on Knowledge Discovery in Databases and Data Mining, pp. 67–73 (1997)
4. Fukuda, T., Morimoto, Y., Morishita, S., Tokuyama, T.: Mining optimized association rules for numeric attributes. In: The ACM Sigact-Sigmod Symposium on Principles of Database Systems, pp. 182–191 (1996)
5. Han, J., Fu, Y.: Discovery of multiple-level association rules from large database. In: The Twenty-first International Conference on Very Large Data Bases, pp. 420–431 (1995)
6. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: SIGMOD Conference, pp. 1–12 (2000)
7. Hong, T.P., Lin, C.W., Wu, Y.L.: Maintenance of fast updated frequent pattern trees for record deletion. Computational Statistics & Data Analysis 53(7), 2485–2499 (2009)
8. Hong, T.P., Lin, C.W., Wu, Y.L.: Incrementally fast updated frequent pattern trees. Expert Systems with Applications 34(4), 2424–2435 (2008)
9. Lin, C.W., Hong, T.P., Wu, Y.L.: The Pre-FUFP algorithm for incremental mining. Expert Systems with Applications 36(5), 9498–9505 (2009)
10. Mannila, H., Toivonen, H., Verkamo, A.I.: Efficient algorithm for discovering association rules. In: The AAAI Workshop on Knowledge Discovery in Databases, pp. 181–192 (1994)
11. Park, J.S., Chen, M.S., Yu, P.S.: Using a hash-based method with transaction trimming for mining association rules. IEEE Transactions on Knowledge and Data Engineering 9(5), 812–825 (1997)
12. Vo, B., Le, B.: Mining minimal non-redundant association rules using frequent itemsets lattice. Journal of Intelligent Systems Technology and Applications 10(1), 92–106 (2011)
13. Vo, B., Le, B.: Interestingness for association rules: Combination between lattice and hash tables. Expert Systems with Applications 38(9), 11630–11640 (2011)
14. Le, B., Nguyen, H., Vo, B.: An efficient strategy for mining high utility itemsets. International Journal of Intelligent Information and Database Systems 5(2), 164–176 (2011)

Adaptive Splitting and Selection Method for Noninvasive Recognition of Liver Fibrosis Stage

Bartosz Krawczyk[1], Michał Woźniak[1], Tomasz Orczyk[2], and Piotr Porwik[2]

[1] Department of Systems and Computer Networks, Wroclaw University of Technology, Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
{bartosz.krawczyk,michal.wozniak}@pwr.wroc.pl
[2] University of Silesia, Institute of Computer Science, Bedzinska 39, 41-200 Sosnowiec, Poland
{tomasz.orczyk,piotr.porwik}@us.edu.pl

Abstract. Therapy of patients suffer form liver diseases strongly depends on the liver fibrosis progression. Unfortunately, to asses it the liver biopsy has been usually used which is an invasive and raging medical procedure which could lead to serious health complications. Additionally even when experienced medical experts perform liver biopsy and read the findings, up to a 20% error rate in liver fibrosis staging has been reported. Nowadays a few noninvasive commercial tests based on the blood examinations are available for the mentioned above problem. Unfortunately they are quite expensive and usually they are not refundable by the health insurance in Poland. Thus, the cross-disciplinary team, which includes researches form the Polish medical and technical universities has started work on new noninvasive method of liver fibrosis stage classification. This paper presents a starting point of the project where several traditional classification methods are compared with the originally developed classifier ensembles based on local specialization of the classifiers in given feature space partitions. The experiment was carried out on the basis of originally acquired database about patients with the different stages of liver fibrosis. The preliminary results are very promising, because they confirmed the possibility of outperforming the noninvasive commercial tests.

Keywords: machine learning,multiple classifier system, clustering and selection, evolutionary algorithm, feature selection, medical informatics, liver fibrosis.

1 Introduction

Medical diagnosis is important area of computer aided diagnosis. Liebowitz [16] reports that 11% of expert systems are dedicated to the medical aided diagnosis and 21% of articles devoted intelligent method applications are connected with the medical cases. Such software is widely used especially when we do not

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 215–224, 2013.

have enough accurate diagnostic tools to make a reliable diagnosis. Our research is focusing on evaluating liver fibrosis stage. Diseases causing liver damages are serious thread for patient life as they may progress without any significant symptoms until the very final stage. Such a disease may be caused by the liver hepatitis virus which may lead to liver fibrosis and, in the terminal stage, to liver cirrhosis and death. Early detection of liver fibrosis is very important because, despite there is no cure for the virus itself, there is a therapy which slows down or even stops the progression of fibrosis. Unfortunately, in most cases the condition stays in so called compensated state, so no visible changes nor dysfunctions might be observed. Although most medical examination results are within their normal results, some slight discrepancies may be observed and used to evaluate the liver fibrosis stage [19]. We propose to use easy accessible noninvasive biomedical examinations as a blood test gathered from patients infected with (liver hepatitis type B virus and type C respectively). Our aim is to create an accurate medical decision support tool that will allow for an automatic classification of patients under the observation. For the problem under consideration we use the modified Adaptive Splitting and Selection method (AdaSS), previously developed by our team [9] and it is compared with several machine learning methods. This work is a starting point of the interdisciplinary research which the main objective is to design the reliable decision support system which could outperform the expensive commercial tools for the task of liver fibrosis stage.

The outline of the work is as follows. Firstly the medical background is presented, then the propose algorithm and its method of training are described. Next section focuses on the experimental evaluation of the pool of available classification methods which results are compared with proposed approach. The last part of the paper includes some conclusions and future research directions.

2 Medical and Clinical Aspects

As mentioned neither the chronic liver hepatitis nor the early stages of liver fibrosis give noticeable symptoms. During this early stage the remaining healthy regions of liver compensate the dysfunction of degraded ones. As the condensation of scare tissue within the liver may vary in different regions of the organ the only method of liver fibrosis stage recognition which gives the confidence is an autopsy and histopathological examination of the liver tissues. For the same reason the most common examination method - liver biopsy does not guarantee the correct diagnosis. This method is unfortunately not only inaccurate, but also may lead to serious health complications including risk of patient's death. Despite of that it is still a so called "gold standard" in the liver examination and is used as a reference method of alternate diagnostic methods.

There are three common description methods for liver biopsy samples. One used in the article is METAVIR[2] (4 stages of fibrosis) and the other are Histological Activity Index (HAI Score) also known as Knodell Score[11] (3 stages of fibrosis) and it's modified version called Ishak Score [7] (6 stages of fibrosis). METAVIR has been specifically designed and validated for patients with

hepatitis C. All these systems rely on a histological image of the liver and thus their quality depends on a sample size and doctor's experience.

Also some noninvasive examination methods have been developed. Most common and simplest one is the APRI-test [22], but also the ELF-Test [17] and the FIBRO-Test [4] have been developed by medical companies. All these methods are blood test based, but the APRI is very general and can be used only to detect advanced fibrosis or cirrhosis. Two other are more specific, but were developed and are held by commercial organizations, so they are expensive for patients and usually they are not refundable by the health insurance in Poland. All blood test based methods aim to detect some dependencies between liver functionality and blood test results, so they may be classified as indirect and noninvasive methods. This is very important, because in opposite to liver biopsy, they may be repeated in regular periods of time without a harm for a patient.

3 Classification Algorithm

Let's present shortly the classifier which we propose to use for the problem under consideration.Let's assume that we have n individual classifiers $\Psi^{(1)}, ..., \Psi^{(n)}$ at our disposal. They are able to classify a given object $x = [x^{(1)}, ..., x^{(d)}]^T \in X$ to the one of the predefined class $i \in \mathcal{M} = \{1, ..., M\}$. The l-th classifier makes the decision according to the following formula:

$$\Psi^{(l)}(x) = i \Leftrightarrow F^{(l)}(i, x) = \max_{k \in \mathcal{M}} F^{(l)}(k, x), \tag{1}$$

where $F^{(l)}(k, x)$ denotes the support function of the l-th classifier for the k-th class and a given object x. In this research we consider homogeneous classifiers i.e., all individual classifiers use the same model.

The feature space is partitioned into competence areas [10]:

$$X = \bigcup_{h=1}^{H} \hat{\mathcal{X}}_h, \quad \forall k, l \in \{1, ..., H\}, \quad k \neq l, \quad \hat{\mathcal{X}}_k \cap \hat{\mathcal{X}}_l = \emptyset. \tag{2}$$

As the representation method of them we propose to use:

$$\mathcal{C}_h = [c_h^{(1)}, c_h^{(2)}, ..., c_h^{(d)}]^T \in \mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_H\}. \tag{3}$$

$$x \in \hat{\mathcal{X}}_h \Leftrightarrow d(x, C_h) = \min_{k=1}^{H} d(x, C_k), \tag{4}$$

where $d(a, b)$ is a distance metric embedded in the system, as the Euclidean metric which we use in this research. The proposed classifier makes the decision according to the following formula:

$$\Psi(x) = i \Leftrightarrow \Psi_{member(c,x)}(x) = i. \tag{5}$$

where

$$member((C), x) = h \Leftrightarrow x \in \hat{\mathcal{X}}_h, \tag{6}$$

and

$$\Psi_h(x) = i \Leftrightarrow F_h^{(l)}(i,x) = \max_{k \in \mathcal{M}} F_h^{(l)}(k,x), \tag{7}$$

where

$$F_h^{(l)}(k,x) = \sum_{l \in \prod_h} w_h^{(l)}(k) F^{(l)}(k,x). \tag{8}$$

$\prod_h$ denotes a pool of individual classifiers used by the h-th area classifier Ψ_h and $w_h^{(l)}(k)$ is the weight assigned to the discriminant function of the l-th individual classifier for class k. The proposed method of decision making was described in the previous works [24,23] where we showed that it could be implemented as a single-layer perceptron.

4 Training Algorithm

4.1 Assuring the Diversity of the Classifier Pool

It is well-known that Multiple Classifier Systems (MCS) in order to work properly should consist of high quality and diverse individual classifiers. Diversity in classification can be understood as covering different aspects of the classified problem - this way predictors can be mutually supplementary to each other. Main works in the field of machine learning diversity concentrates on the problem of how to assure the diversity [13] or how to measure it [3].

In this work we propose two methods for creating a diverse pool of classifiers for the AdaSS:

- Utilizing different predictors in the pool.
- Utilizing different subset of features for each of the predictors.

The first approach can be achieved by manipulating the training algorithm of the base classifier. As in this work we use neural networks we propose to initialize each of them with random starting points and then stop the training procedure prematurely. This will lead to a set of different models.

The second approach concentrates on training the individual classifiers on the basis of different subsets of available features. Let us propose the following representation of the classifier $\Psi_h(x)$:

$$\mathcal{A}_h = \begin{pmatrix} a_h^{(1)}(x^{(1)}) & \cdots & a_h^{(1)}(x^{(d)}) \\ \vdots & \ddots & \vdots \\ a_h^{(n)}(x^{(1)}) & \cdots & a_h^{(n)}(x^{(d)}) \end{pmatrix}, \tag{9}$$

where $a_h^{(p)}(x^{(q)}) = 1$ if the q-th feature is used by the p-th individual classifier used by $\Psi_h(x)$, otherwise $a_h^{(p)}(x^{(q)}) = 0$ means that mentioned above attribute is not used by it. This allows for the feature selection step to be embedded in the AdaSS training procedure.

4.2 Training Goal

The objective defined for the training algorithm is the maximization of the accuracy of the classification. It is based on the frequency of correct classification of object from a learning set $\mathcal{LS}$

$$\mathcal{LS} = \{(x_1, j_1), (x_2, j_2), ..., (x_K, j_K)\}, \tag{10}$$

where x_i denotes observations described in the i-th object and j_i denotes its correct class label. It leads to the commonly known criterion:

$$Q(\Psi) = \frac{1}{K}\sum_{n=1}^{K}(\delta(\Psi_{member(C,x_n)}(x_n), j_n)), \tag{11}$$

where δ denotes Kronecker's delta. This criterion was defined for so-called 0-1 loss function [5] but let us note that it can be easily changed to incorporate a case where the misclassification costs between pairs of classes are different [12,8].

4.3 Optimization Algorithm

The process of searching for maximum value of target criterion was treated as a compound optimization problem solved by an evolutionary algorithm (EA) [18]. Our solutions represent by chromosomes differ according to the chosen diversity assurance method.

In case of the first solution it consists of two components:

$$Chromosome = [\mathcal{C}, \mathcal{W}], \tag{12}$$

The first component $\mathcal{C} = \{C_1, C_2, ..., C_H\}$ represents centroids according to the Eq 3. The second one $\mathcal{W} = \{\mathcal{W}_1, \mathcal{W}_2, ..., \mathcal{W}_H\}$ consists of weights assigned to each individual classifier and each class i.e.:

$$\mathcal{W}_h = \begin{pmatrix} w_h^{(1)}(1) & \cdots & w_h^{(1)}(M) \\ \vdots & \ddots & \vdots \\ w_h^{(n)}(1) & \cdots & w_h^{(n)}(M) \end{pmatrix}. \tag{13}$$

In case of the second solution the chromosome includes the additional component:

$$Chromosome = [\mathcal{A}, \mathcal{C}, \mathcal{W}], \tag{14}$$

where features are represented by:

$$\mathcal{A} = [\mathcal{A}_1, ..., \mathcal{A}_h, ..., \mathcal{A}_H], \tag{15}$$

For the purpose of the EA computer implementation $\mathcal{A}$, $\mathcal{C}$ and $\mathcal{W}$ matrices are transformed into vectors. One must have in mind that the two (or three) parts of the chromosome have entirely different meaning and nature. Therefore we ensure that no data exchange can be done between them while processing the population.

Training procedure steps can be summarized as follows:

- Initialization - the random generation of the initial population. It also sets the parameters for the EA approach: N_c - upper limit of algorithm cycles, N_p - size of population, β - mutation probability, γ - crossover probability, Δ_m - mutation range factor and V - upper limit of algorithm iterations with falling quality ($V < N_c$).
- Selection and reproduction process according to the roulette selection with elitism for further operations.
- Mutation genetic operator alter chromosomes of selected individuals by adding some random noise allowing for exploration of the search space. It is generated according to the Gaussian Normal distribution with mean value equal to 0 and standard deviation set to Δ_m. Each constituents of the chromosome is treated separately and can be affected with the following probabilities:

$$P_a(t) = 2\beta\frac{t}{N_c}, \quad P_c(t) = \beta\frac{t}{N_c}, \quad P_w(t) = \beta - P_c(t), \tag{16}$$

  where t is the iteration index of the algorithm, $P_a(t)$ is the mutation probability of the feature vector, $P_c(t)$ is the mutation probability of the centroid vector and $P_w(t)$ is the mutation probability of the weight vector.
- Crossover procedure according to the two-point rule.
- Protecting against overfitting procedure which cancels training process when the accuracy of classification (controlled at each generation over a validation data set) deteriorates.

5 Experimental Results

To evaluate the quality of the previously described methods for the problem of liver fibrosis stage diagnosis we propose carried out the computer experiments. We have acquired medical data records from 103 real patients of the Branch of the Gastroenterology and Hepatology of the Independent Public Central Hospital of the Silesian Medical University, Poland. From the practical point of view it is very useful for the future therapy to distinguish among five fibrosis stages. The numbers of examined patients for each fibrosis stage (F0..F4) are presented in Table 1.

Table 1. Number of patients with given fibrosis stage [n (%)]

$F0$	$F1$	$F2$	$F3$	$F4$
2	34	5	16	46
(2%)	(33%)	(5%)	(15%)	(45%)

As the input data we used the the traditional blood test includes the following characteristics: HB, RBC, ,WBC, PLT, PT, PTP, APTT, INR, ASPT, ALAT, ALP, BIL, GGTP, KREA, GLU, Na, K, Fe, CRP, TG, CHO, Ur. acid, TP, TIBC, Neutr, Lymph, Mono, Eos, Baso, Albu, Glb. 1, Glb. 2, Glb. , Glb. .

5.1 Set-Up

As reference methods we have selected most popular ensembles - Bagging, Boosting, Random Forest and Random Subspace - as they were used in our previous work [14] - in there one may also find the details of used parameters for these ensemble classifiers.Additionally we have compared our method with the single best classifier from the pool (i.e. built on the basis of the most effective feature selection algorithm), all classifiers from the pool (i.e. without the pruning procedure) and with simple majority voting (i.e. without the trained fuser). By this we can establish the influence of three steps in our proposed ensemble on the final accuracy.

The combined 5x2 cv F test [1] was carried out to asses the statistical significance of obtained results.

All experiments were carried out in the R environment [21] and computer implementations of the classification methods used were taken from dedicated packages built into the above mentioned software. This ensured that results achieved the best possible efficiency and that performance was not diminished by a bad implementation.

5.2 Results

As a base classifier we have used a Neural Network, trained with the quickprop procedure. The number of neurons in the input layer was equal to the number of features, in final layer equal to the number of classes and in hidden layer equal to the half of the sum of total neurons in mentioned layers. Each committee consisted of five classifiers. Tested approach used neural fuser described in [24,23].

For the training phase following parameters have been set: $H = 5$, $N_c = 200$, $N_p = 100$, $\beta = \{0.7;0.3\}$, $\gamma = \{\ 0.3;0.7\}$, $\Delta_m = 0.2$ and $V = 15$.

As reference methods we have selected most popular ensembles - Bagging, Boosting [20], Random Forest and Random Subspace [6] - as they were used in our previous work [14] - in there one may also find the details of used parameters for these ensemble classifiers.Additionally we have compared our method with the single best classifier from the pool (built with and without feature selection step) and with standard Clustering and Selection method [15]. By this we can establish the influence of three steps in our proposed ensemble on the final accuracy.

Results are presented in the table 2, where *AdaSS* stands for the Adaptive Splitting and Selection, $AdaSS + FS$ for the same method with incorporated feature selection, *Bagg* for Bagging, *Boost* for Boosting, *RandS*for Random Subspace, *RandF* for Random Forest, *CS* for Clustering and Selection, *SB* for the Single Best model form the AdaSS pool and $SB + FS$ for the Single Best model with feature selection.

5.3 Results Discussion

The AdaSS algorithm delivered the best results, outperforming the reference methods regardless of how the diversity for the pool of classifiers were assured.

Interestingly one may clearly see how the AdaSS improves the quality of the final prediction - we spotted more than 12% accuracy gain in comparison to a single best model from the pool of classifiers. This proves that assigning classifiers to areas in which they are most competent can greatly boost the quality of the model.

In comparison with a standard Clustering and Selection method AdaSS achieved better results by more than 9%. This highlights the strength of our proposal - embedding the clustering and selection step into EA allowed for better exploitation of the available pool of individual classifiers. As for the diversity assurance method the incorporation of feature selection step lead to a small but statistically significant improvement.

Table 2. Results of the experiment

id	Classifier	accuracy [%]	ids of outperformed classifiers
1	$Bagg$	80.50	6
2	$Boost$	84.92	1,6,7
3	$RandS$	88.54	1,2,4,5,6,7
4	$RandF$	87.02	1,2,5,6,7
5	CS	83.31	1,6,7
6	SB	80.11	-
7	$SB + FS$	81.85	1
8	$AdaSS$	90.12	1,2,3,4,5,6,7
9	$AdaSS + FS$	92.74	1,2,3,4,5,6,7

6 Conclusions

Despite problems like getting the real medical records of patients with diagnosed chronic hepatitis C, infected with Hepatitis type B and C Virus from the hospital or the fact that blood test results which were available for the research were incomplete (not all patients had a complete set of blood tests done). Together with the medical experts we attested that it is possible to reach similar or even lower error level than commercial tests what encourages us to continue work on this topic to develop reliable decision support system which may be used in practice by several leading Polish hospitals.

Our future works will concentrate on the extension the database with the patient infected with HBV or HCV, additional we are going to implement new method for the the problem of imbalanced class distribution among biopsy patients and possible presence of class label noise in the data.

References

1. Alpaydin, E.: Combined 5 x 2 cv f test for comparing supervised classification learning algorithms. Neural Computation 11(8), 1885–1892 (1999)

2. Bedossa, P., Poynard, T.: An algorithm for the grading of activity in chronic hepatitis c. the metavir cooperative study group. Hepatology 24, 289–293 (1996)
3. Bi, Y.: The impact of diversity on the accuracy of evidential classifier ensembles. International Journal of Approximate Reasoning 53(4), 584–607 (2012)
4. BioPredictive. Website, http://www.biopredictive.com/intl/physician/fibrotest-for-hcv/view?set_language=en
5. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley, New York (2001)
6. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20, 832–844 (1998)
7. Ishak, K., Baptista, A., Bianchi, L., Callea, F., De Groote, J., Gudat, F., Denk, H., Desmet, V., Korb, G., MacSween, R.N., et al.: Histological grading and staging of chronic hepatitis. Hepatology 22, 696–699 (1995)
8. Jackowski, K., Krawczyk, B., Woźniak, M.: Cost-Sensitive Splitting and Selection Method for Medical Decision Support System. In: Yin, H., Costa, J.A.F., Barreto, G. (eds.) IDEAL 2012. LNCS, vol. 7435, pp. 850–857. Springer, Heidelberg (2012)
9. Jackowski, K., Woźniak, M.: Algorithm of designing compound recognition system on the basis of combining classifiers with simultaneous splitting feature space into competence areas. Pattern Analysis and Applications 12(4), 415–425 (2009)
10. Jackowski, K., Woźniak, M.: Method of classifier selection using the genetic approach. Expert Systems 27(2), 114–128 (2010)
11. Knodell, R.G., Ishak, K.G., Black, W.C., Chen, T.S., Craig, R., Kaplowitz, N., Kiernan, T.W., Wollman, J.: Formulation and application of a numerical scoring system for assessing histological activity in asymptomatic chronic active hepatitis. Hepatology 1, 431–435 (1981)
12. Krawczyk, B., Woźniak, M.: Designing Cost-Sensitive Ensemble – Genetic Approach. In: Choraś, R.S. (ed.) Image Processing and Communications Challenges 3. AISC, vol. 102, pp. 227–234. Springer, Heidelberg (2011)
13. Krawczyk, B., Woźniak, M.: Analysis of Diversity Assurance Methods for Combined Classifiers. In: Choraś, R.S. (ed.) Image Processing and Communications Challenges 4. AISC, vol. 184, pp. 177–184. Springer, Heidelberg (2013)
14. Krawczyk, B., Woźniak, M., Orczyk, T., Porwik, P., Musialik, J., Błońska-Fajfrowska, B.: Classification techniques for non-invasive recognition of liver fibrosis stage. Journal of Medical Informatics & Technologies 20, 121–127 (2012)
15. Kuncheva, L.I.: Clustering-and-selection model for classifier combination. In: Proceedings of the Fourth International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies, vol. 1, pp. 185–188 (2000)
16. Liebowitz, J.: The handbook of applied expert systems. CRC Press, Boca Raton (1998)
17. Siemens Medical. Website, http://www.medical.siemens.com/webapp/wcs/stores/PSGenericDisplay~q_catalogId~e_-111~a_langId~e_-111~a_pageId~e_103713~a_storeId~e_10001.html
18. Michalewicz, Z.: Genetic algorithms + data structures = evolution programs, 3rd edn. Springer, London (1996)
19. Orczyk, T., Pałys, M., Porwik, P., Musialik, J., Błońska-Fajfrowska, B.: Simple and non-invasive liver fibrosis stage prediction method. Journal of Medical Informatics & Technologies 17, 227–232 (2011)
20. Polikar, R.: Ensemble based systems in decision making. IEEE Circuits and Systems Magazine 6(3), 21–45 (2006)
21. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2008)

22. Wai, C.T., Greenson, J.K., Fontana, R.J., Kalbfleisch, J.D., Marrero, J.A., Conjeevaram, H.S., Lok, A.S.: A simple noninvasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis c. Hepatology 38, 518–526 (2003)
23. Woźniak, M., Krawczyk, B.: Combined classifier based on feature space partitioning. Journal of Applied Mathematics and Computer Science 22(4) (2012) (in press) (to appear)
24. Woźniak, M., Zmyslony, M.: Designing combining classifier with trained fuser - analytical and experimental evaluation. Neural Network World 20(7), 925–934 (2010)

Investigation of Mixture of Experts Applied to Residential Premises Valuation

Tadeusz Lasota[1], Bartosz Londzin[2], Bogdan Trawiński[2], and Zbigniew Telec[2]

[1] Wroclaw University of Environmental and Life Sciences, Dept. of Spatial Management
ul. Norwida 25/27, 50-375 Wroclaw, Poland
[2] Wrocław University of Technology, Institute of Informatics,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
tadeusz.lasota@wp.pl, bartek.londzin@gmail.com,
{zbigniew.telec,bogdan.trawinski}@pwr.wroc.pl

Abstract. Several experiments were conducted in order to investigate the usefulness of mixture of experts approach to an online internet system assisting in real estate appraisal. All experiments were performed using real-world datasets taken from a cadastral system. The analysis of the results was performed using statistical methodology including nonparametric tests followed by post-hoc procedures designed especially for multiple *1×N* and *N×N* comparisons. The mixture of experts architectures studied in the paper comprised: four algorithms used as expert networks: *glm* – general linear model, *mlp* – multilayer perceptron and two support vector regression *ε-SVR* and *ν-SVR* as well as and three algorithms *glm, mlp*, and *gmm* – gaussian mixture model employed as gating networks.

Keywords: mixture of experts, SVR, statistical tests, real estate appraisal, MATLAB.

1 Introduction

One of the most popular and interesting method of machine learning, which can be widely applied to resolve problems related to classification and regression, is mixture of expert model. The main idea of mixture of experts (MoE) is based on "divide and conquer" principle common in computer science. In terms of mixture of experts it consists of divide hard to solve problem to number of simpler sub-problems, whose solution can be in easy way combine to obtain resolution of original problem.

Research in this area has been initiated by Jacob et al. [1]. They proposed the mixture of experts model, in which the set of neural networks and a gating network have been trained together. The idea behind the system is to learn the model how cases should be allocated to experts. This is realized by a gating network which is responsible for allocating individual cases to one or several experts and when the output of the expert is incorrect then the associated with the expert weights are changed. The overall output of the model is produced from combining outputs of experts and weights generated by gating network. The most efficient method for

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 225–235, 2013.

calculating model parameters is maximum likelihood approach, for which Jordan and Jacobs [2] proposed expectation-maximization algorithm. The MoE approach has been then developed and extended by Avnimelech [3], Srivastava [4], Lima [5], and others. The biggest disadvantage of the method is its tendency for over-fitting due to the complexity of the model.

Numerous works related to MoE model have been published to date; recent comprehensive surveys can be found in [6], [7]. The latest studies have been mainly focused on using the model in regression and classification applications in finance, bioinformatics, healthcare and recognition. MoE techniques have been applied to speech [8], face [9], handwriting [10], and hand gesture [11] recognition, in molecular biology [12], and medicine [13], [14]. Another line of research is focused on regression approach. MoE methods are widely used in financial analysis for risk estimation of asset returns [15], forecasting of daily S&P500 returns [16] and time series forecasting [17].

The main goal of our study was to investigate the usefulness of the MoE models to real world application such as an internet system to assist in property valuation. The second goal was to carry out the comparative analysis of different machine learning algorithms composing the MoE architecture.

2 Mixture of Expert Approach Used

The mixture of experts (MoE) architecture divides the covariate space, i.e. the space of all possible values of the explanatory variables, into regions, and fit simple surfaces to the data that fall in each region. The architecture consists of M modules referred to as expert networks and module referred to as gating network. MoE is a network architecture for supervised learning, which comprises a number of expert networks and a gating network (see Fig. 1). Expert networks approximate the data within each region of the input space: expert network i maps its input, the input vector x, to an output vector y_i. It is assumed that different expert networks are appropriate in different regions of the input space. The architecture contains a module, referred to as a gating network, that identifies for any input x, the expert or mixture of experts whose output is most likely to approximate the corresponding target output y. The task of a gating network is to combine the various experts by assigning weights to individual networks, which are not constant but are functions of the input instances. Both expert and the gating networks are fed with the input vector $\boldsymbol{x}$ and the gating network produces one output per expert. For each input x, these probability of selection are constrained to be nonnegative and to sum to one. The total output of the architecture, given by

$$y(\boldsymbol{x}) = \sum_{j=1}^{M} w_j(\boldsymbol{x}) y_j(\boldsymbol{x})$$

is a convex combination of the expert outputs for each x. The output of MoE is the weighed sum of the expert outputs. The expectation-maximization (EM) algorithm is usually employed to learn the parameters of the MoE architecture.

So far the authors of the present paper have investigated the basic MoE architecture to construct regression models to assist with real estate appraisal [18]. In the current contribution we extended the MoE architecture to include two support vector regression algorithms [19], namely *ε-SVR* and *ν-SVR*, as expert networks. Moreover, we ensured that different machine learning algorithms could be used as expert and gating networks. Thus, the MoE architecture investigated in the research reported in the paper encompassed: four algorithms playing the role of expert networks, namely general linear model – *glm* [20], neural network of multilayer perceptron type – *mlp*, two support vector regression *ε-SVR* [21] and *ν-SVR* [22] and three algorithms serving as gating networks, namely *glm, mlp*, and gaussian mixture model – *gmm* [23]. The location of individual algorithms within the MoE architecture is shown in Fig. 1.

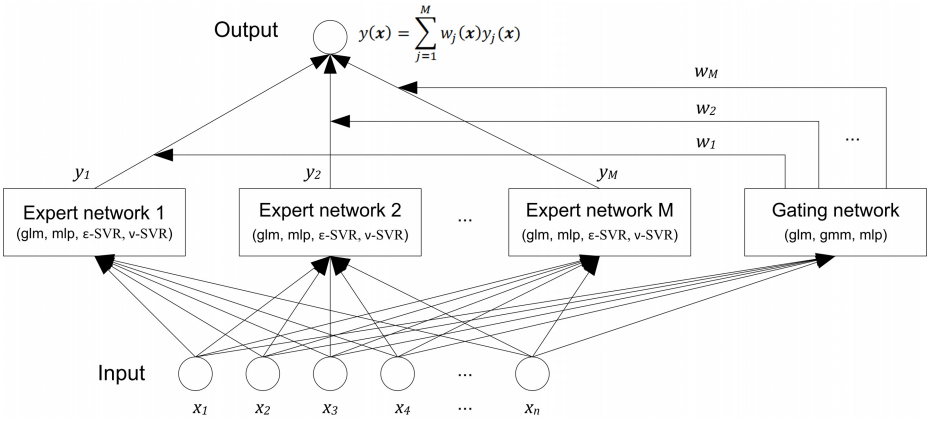


Fig. 1. The architecture of mixture of experts

3 Plan of Experiments

The experiments were conducted using *Mixlab* – a matlab tool developed by Perry Moerland [24] based on the *netlab* library. We extended the *Mixlab* package to include two versions of support vector machine for regression, namely *ε-SVR* and *ν-SVR*, using the *libsvm* library.

Real-world dataset used in experiments was drawn from an unrefined dataset containing above 50 000 records referring to residential premises transactions accomplished in one Polish big city with the population of 640 000 within 11 years from 1998 to 2008. The final dataset counted the 5213 samples. Five following attributes were pointed out as main price drivers by professional appraisers: usable area of a flat (*Area*), age of a building construction (*Age*), number of storeys in the building (Storeys), number of rooms in the flat including a kitchen (*Rooms*), the distance of the building from the city centre (*Centre*), in turn, price of premises (*Price*) was the output variable.

Due to the fact that the prices of premises change substantially in the course of time, the whole 11-year dataset cannot be used to create data-driven models. In order to obtain comparable prices it was split into 20 subsets covering individual half-years. Then the prices of premises were updated according to the trends of the value changes over 11 years. Starting from the beginning of 1998 the prices were updated for the last day of subsequent half-years. The trends were modelled by polynomials of degree three. We might assume that half-year datasets differed from each-other and might constitute different observation points to compare the accuracy of ensemble models in our study and carry out statistical tests. The sizes of half-year datasets are given in Table 1. The descriptive statistics of input and output attributes of residential premises are shown in Table 2.

Table 1. Number of instances in half-year datasets

1998-2	**1999-1**	**1999-2**	**2000-1**	**2000-2**	**2001-1**	**2001-2**	**2002-1**	**2002-2**	**2003-1**
202	213	264	162	167	228	235	267	263	267
2003-2	**2004-1**	**2004-2**	**2005-1**	**2005-2**	**2006-1**	**2006-2**	**2007-1**	**2007-2**	**2008-1**
386	278	268	244	336	300	377	289	286	181

Table 2. Descriptive statistics of input and output attributes of residential premises

Name	Max	Min	Avg	Median	StdDev	Description
Area	199.83	13.70	51.61	47.76	22.09	usable area of premises
Age	214	8	66.5	64.0	29.3	age of building construction
Storeys	12	1	5.2	5.0	2.6	no. of storeys in a building
Rooms	8	1	3.0	3.0	0.9	no. of rooms in a flat
Centre	13263	49	2434	2192	1475	distance from the centre of a city
Price	1075000	17000	149628	120000	102504	price of premises

The experiments were conducted in two phases. First phase consisted in the selection of optimal parameters of learning algorithms employed to expert and gating networks using the trial and error approach. For *glm* five optimisation routines of EM algorithms were examined: *irls* – iteratively reweighted least-squares algorithm, *heuristic* – weighted least-squares, *scg* – scaled conjugate gradient, *quasinew* – quasi-Newton, and *graddesc* – gradient descent. For *gmm* three types of covariance matrices were tested: *diagonal, spherical*, and *full* ones. For *mlp* three optimisation routines of EM algorithms were considered: *scg* – scaled conjugate gradient, *graddesc* – gradient descent, and *quasinew* – quasi-Newton as well as the number of neurons in a hidden layer. For ε-SVR and ν-SVR four kernel functions: *linear*, *polynomial*, *radial basis*, and *sigmoid* as well as the values of a cost function and ε and ν values, respectively, were explored. The selected parameters of MoE architectures for glm and mlp expert networks and ε-SVR and ν-SVR ones are placed in Table 3 and 4, respectively.

In the second phase the MoE models were built for 12 architectures and parameters described in Table 3 and 4. The models were generated by running Matlab software individually for each of 20 half-year datasets using 10-fold cross validation (10cv). The 10cv was repeated five times for each dataset and the median of the mean square errors (MSE) was applied as performance measure. The data was normalised using the min-max approach.

Table 3. Selected parameters of ME architectures with *glm* and *mlp* expert networks

Expert network	EM routines	No. of neurons	Gating network	No. of experts	EM routines or cov. matrix	No. of neurons	No. of trials
Glm	scg	-	glm	4	heuristic	-	
Glm	scg	-	gmm	2	spherical	-	540
Glm	scg	-	mlp	3	quasinew	3	
Mlp	scg	4	glm	3	heuristic	-	
Mlp	scg	4	gmm	3	spherical	-	1660
Mlp	scg	4	mlp	5	scg	2	

Table 4. Selected parameters of ME architectures with ε-SVR and ν-SVR expert networks

Expert network	Kernel function.	Cost fun. values	ε or ν	Gating network	No. of experts	EM routines or cov. matrix	No. of neurons	No. of trials
ε-SVR	radial basis	4	0.0039	glm	3	irls	-	
ε-SVR	radial basis	1	0.0039	gmm	3	spherical	-	5000
ε-SVR	linear	16	0.0039	mlp	3	quasinew	3	
ν-SVR	radial basis	4	0.8	glm	3	irls	-	
ν-SVR	radial basis	2	0.8	gmm	3	full	-	4500
ν-SVR	linear	11	0.4	mlp	3	scg	3	

The rationale behind statistical methods employed to analyse the output of experiments was as follows. Several articles on the use of statistical tests in machine learning for comparisons of many algorithms over multiple datasets have been published recently [25], [26], [27], [28]. Their authors argue that the commonly used paired tests i.e. parametric t-test and its nonparametric alternative Wilcoxon signed rank tests are not adequate when conducting multiple comparisons due to the so called multiplicity effect resulting in the family-wise error. They recommend following methodology. First of all the Friedman test or its more powerful derivative the Iman-Davenport test should be performed. Both tests can only inform the researcher about the presence of differences among all samples of results compared. After the null-hypotheses have been rejected he can proceed with the post-hoc procedures in order to find the particular pairs of algorithms which produce differences. The latter comprise Bonferroni-Dunn's, Holm's, Hochberg's, Hommel's, Holland's, Rom's, Finner's, and Li's post hoc procedures for $1 \times N$ comparisons, where one algorithm of the best accuracy is used as the control algorithm and Nemenyi's, Holm's, Shaffer's, and Bergamnn-Hommel's procedures in the case of $N \times N$ comparisons.

4 Statistical Analysis of Experimental Results

The performance, in terms of MSE, of the MoE models comprising expert networks built by *glm, mlp, ε-SVR*, and *v-SVR* is presented in Figures 2-5, respectively. For each type of expert network the impact of gating networks in the form of *glm, gmm*, and *mlp* is compared. Taking into account the fact that each half-year dataset was normalized individually and the scale of the vertical axis in each chart is the same we can make following observations. The best results are provided by the MoE models with expert networks created by *glm* as well as the ones with gating networks constructed by *glm*, too. In turn, the application of *mlp* gating networks for MoE with *ε-SVR* and *v-SVR* expert networks results in the worst performance.

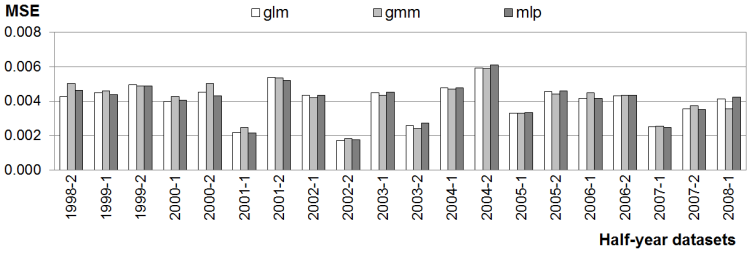


Fig. 2. Performance of MoE models with *glm* as expert network

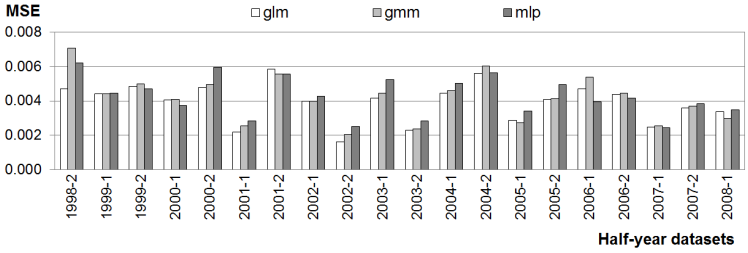


Fig. 3. Performance of MoE models with *mlp* as expert network

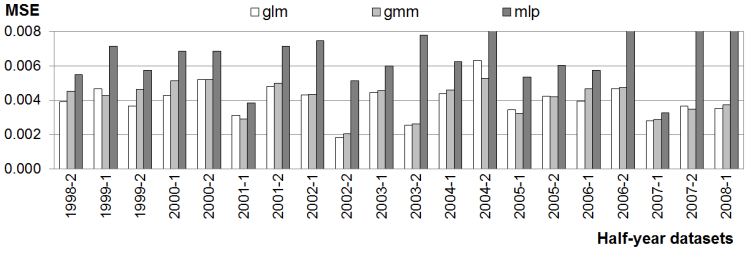


Fig. 4. Performance of MoE models with *ε-SVR* as Expert network

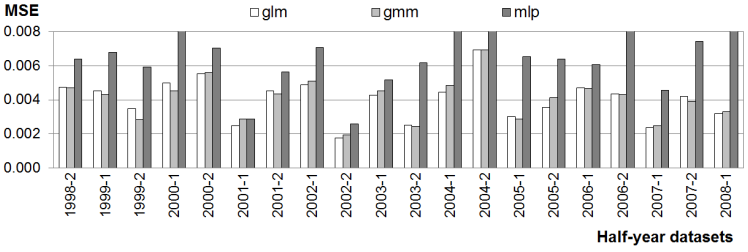


Fig. 5. Performance of MoE models with *v-SVR* as expert network

Statistical analysis of the results of experiments was performed using a software available on the web page of Research Group "Soft Computing and Intelligent Information Systems" at the University of Granada (http://sci2s.ugr.es/sicidm). In the paper we present the selected output produced by this JAVA software package comprising Friedman tests as well as post-hoc multiple comparison procedures.

The average ranks of the MoE models produced by Friedman tests applied to individual expert networks are shown in Table 5, where the lower rank value the better model. For each type of expert network the ranks are statistically significant except for *glm*. The *glm* expert networks reveal the best performances for each type but one.

Table 5. Average rank positions of MoE models produced by Friedman tests

Expert: **glm** p-value=0.8187		Expert: **mlp** p-value=*0.0193*		Expert: **ε-SVR** p-value=*0.0000*		Expert: **ν-SVR** p-value=*0.0000*	
Rank	Gating	Rank	Gating	Rank	Gating	Rank	Gating
1.90	glm	1.50	glm	1.35	glm	1.45	gmm
2.00	mlp	2.15	gmm	1.65	gmm	1.55	glm
2.10	gmm	2.35	mlp	3.00	mlp	3.00	mlp

Adjusted p-values for post hoc procedures for *1×N* comparisons, where the top gating methods in Friedman's ranks were the control algorithms, are placed in Tables 6a and 6b for respective expert networks. In all tests the p-value less than 0.05 means that the control method significantly outperforms the compared algorithm. The p-values indicating the statistically significant differences between given pairs of algorithms are marked with italics.

Table 6a. Adjusted p-values for *1×N* comparisons of MoE with glm and mlp as experts

Expert	**glm**		**mlp**	
Control	**glm**		**glm**	
Gating	**gmm**	**mlp**	**gmm**	**mlp**
pBonf	1.00000	1.00000	0.07967	*0.01438*
pHolm	1.00000	1.00000	*0.03983*	*0.01438*
pHoch	0.75183	0.75183	*0.03983*	*0.01438*
pHomm	0.75183	0.75183	*0.03983*	*0.01438*
pHoll	0.77636	0.77636	*0.03983*	*0.01433*
pRom	0.75183	0.75183	*0.03983*	*0.01438*
pFinn	0.77636	0.77636	*0.03983*	*0.01433*
pLi	0.67989	0.75183	*0.03983*	*0.00743*

Table 6b. Adjusted p-values for *1×N* comparisons of MoE with ε-SVR and ν-SVR as experts

Expert	**ε-SVR**		**ν-SVR**	
Control	**glm**		**gmm**	
Gating	**gmm**	**mlp**	**glm**	**mlp**
pBonf	0.68556	*3.62E-07*	1. 00000	*1.90E-06*
pHolm	0.34278	*3.62E-07*	0.75183	*1.90E-06*
pHoch	0.34278	*3.62E-07*	0.75183	*1.90E-06*
pHomm	0.34278	*3.62E-07*	0.75183	*1.90E-06*
pHoll	0.34278	*3.62E-07*	0.75183	*1.90E-06*
pRom	0.34278	*3.62E-07*	0.75183	*1.90E-06*
pFinn	0.34278	*3.62E-07*	0.75183	*1.90E-06*
pLi	0.34278	*2.76E-07*	0.75183	*3.83E-06*

In all cases the Bonferroni-Dunn tests provided the highest adjusted p-values, because it is the simplest but also the least powerful procedure so that we omit its results in further considerations. For *mlp* expert network the *glm* gating network revealed significantly better performance than *gmm* and *mlp* ones. In turn, for both SVR expert's networks the control methods outperformed significantly only the *mlp* gating networks. Similarly to the results of Friedman tests no significant differences between gating algorithms for *glm* expert network were observed.

Table 7. Average rank positions of all 12 MoE models produced by Friedman test with p-value=0.0000 (the models are denoted by Expert/Gating)

Rank	Model	Rank	Model	Rank	Model	Rank	Model
4.00	mlp/glm	5.40	glm/mlp	5.75	glm/gmm	6.65	mlp/mlp
5.15	ν-SVR/glm	5.55	ν-SVR/gmm	5.90	ε-SVR/glm	11.30	ε-SVR /mlp
5.15	glm/glm	5.70	mlp/gmm	6.10	ε-SVR/gmm	11.35	ε-SVR /mlp

Table 8. Adjusted p-values for $N \times N$ comparisons of MoE models over 20 half-year datasets showing 20 hypotheses rejected out of 66

Model vs Model	pNeme	pHolm	pShaf
mlp/glm vs ε-SVR/mlp	*7.56E-09*	*7.56E-09*	*7.56E-09*
mlp/glm vs ν-SVR/mlp	*1.01E-08*	*9.93E-09*	*8.41E-09*
ν-SVR/glm vs ε-SVR/mlp	*3.56E-06*	*3.45E-06*	*2.97E-06*
glm/glm vs ε-SVR/mlp	*3.56E-06*	*3.45E-06*	*2.97E-06*
ν-SVR/glm vs ν-SVR/mlp	*4.55E-06*	*4.27E-06*	*3.79E-06*
glm/glm vs ν-SVR/mlp	*4.55E-06*	*4.27E-06*	*3.79E-06*
glm/mlp vs ε-SVR/mlp	*1.19E-05*	*1.08E-05*	*9.92E-06*
glm/mlp vs ν-SVR/mlp	*1.51E-05*	*1.35E-05*	*1.26E-05*
ν-SVR/gmm vs ε-SVR/mlp	*2.40E-05*	*2.11E-05*	*2.00E-05*
ν-SVR/gmm vs ν-SVR/mlp	*3.02E-05*	*2.61E-05*	*2.52E-05*
mlp/gmm vs ε-SVR/mlp	*4.76E-05*	*4.04E-05*	*3.97E-05*
glm/gmm vs ε-SVR/mlp	*5.96E-05*	*4.97E-05*	*4.97E-05*
mlp/gmm vs ν-SVR/mlp	*5.96E-05*	*4.97E-05*	*4.97E-05*
glm/gmm vs ν-SVR/mlp	*7.45E-05*	*5.98E-05*	*5.19E-05*
ε-SVR/glm vs ε-SVR/mlp	*1.16E-04*	*9.12E-05*	*8.06E-05*
ε-SVR/glm vs ν-SVR/mlp	*1.44E-04*	*1.11E-04*	*1.00E-04*
ε-SVR/gmm vs ε-SVR/mlp	*2.73E-04*	*2.07E-04*	*1.90E-04*
ε-SVR/gmm vs ν-SVR/mlp	*3.36E-04*	*2.50E-04*	*2.35E-04*
mlp/mlp vs ε-SVR/mlp	*0.002477*	*0.001802*	*0.001726*
mlp/mlp vs ν-SVR/mlp	*0.002994*	*0.002132*	*0.002087*

Statistical tests adequate to $N \times N$ comparisons were conducted for all 12 MoE models considered. The Friedman and Iman-Davenport tests were performed and the calculated values of their statistics were 92.79 and 13.86, respectively, whereas the critical values at $\alpha=0.05$ are $\chi^2(11)=21.92$ and $F(11,297)=1.82$, what means that there are significant differences between some models. Average ranks of the models are shown in Table 7, where the lower rank value the better model. The three top ranks

gained the models with *mlp, v-SVR*, and *glm* as expert networks; all of them comprised *glm* as gating networks. In Table 8 adjusted p-values for Nemenyi, Holm, and Shaffer post-hoc procedures for $N \times N$ comparisons are shown for 20 pairs of models out of 66, where significant differences were noticed. Following main observations could be done: *ε-SVR* and *v-SVR* with *mlp* gating networks revealed significantly worse performance than any other MoE architecture. There are not significant differences among the MoE architectures composed of *glm* and *mlp* expert networks with any gating networks as well as embracing *ε-SVR* and *v-SVR* expert networks with *glm* and *gmm* gating networks

5 Conclusions and Future Work

Several experiments were conducted in order to investigate the usability of mixture of experts approach to an online internet system assisting with real estate appraisal. The mixture of experts architectures studied in the paper comprised: four algorithms used as expert networks: glm – general linear model, mlp – multilayer perceptron and two support vector regression ε-SVR and ν-SVR as well as and three algorithms glm, mlp, and gmm – gaussian mixture model employed as gating networks. In the tests 20 real-world datasets taken from a cadastral system, were employed. The analysis of the results was performed using recently proposed statistical methodology including nonparametric tests followed by post-hoc procedures designed especially for multiple comparisons. Following general conclusions could be drawn on the basis of the experiments: three algorithms *mlp, v-SVR*, and *glm* used as expert networks in the MoE architectures can lead to low values of prediction error provided *glm* is applied to a gating network. In turn, using *mlp* to gating networks results in significantly worse performance. The investigation proved the usefulness and strength of multiple comparison statistical procedures to analyse and select machine learning algorithms for real estate appraisal.

Acknowledgments. This paper was partially supported by the Polish National Science Centre under grant no. N N516 483840.

References

1. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural Computation 3, 79–87 (1991)
2. Jordan, M.I., Jacobs, R.A.: Hierachical mixtures of experts and the EM algorithm. Neural Computation 6, 181–214 (1994)
3. Avnimelech, R., Intrator, N.: Boosted mixture of experts: An ensemble learning scheme. Neural Computation 11(2), 483–497 (1999)
4. Srivastava, A.N., Su, R., Weigend, A.S.: Data mining for features using scale-sensitive gated experts. IEEE Transactions on Pattern Analysis and Machine Intelligence 21, 1268–1279 (1999)

5. Lima, C.A.M., Coelho, A.L.V., Von Zuben, F.J.: Hybridizing mixtures of experts with support vector machines: Investigation into nonlinear dynamic systems identification. Information Sciences 177(10), 2049–2074 (2007)
6. Masoudnia, S., Ebrahimpour, R.: Mixture of experts: a literature survey. Artificial Intelligence Review (2012), doi:10.1007/s10462-012-9338-y
7. Yuksel, S.E., Wilson, J.N., Gader, P.D.: Twenty Years of Mixture of Experts. IEEE Transactions on Neural Networks and Learning Systems 23(8), 1177–1193 (2012)
8. Jianping, D., Bouchard, M., Yeap, T.H.: Linear Dynamic Models With Mixture of Experts Architecture for Recognition of Speech Under Additive Noise Conditions. IEEE Signal Processing Letters 13(9), 573–576 (2006)
9. Ebrahimpour, R., Kabir, E., Esteky, H., Yousefi, M.R.: View-independent face recognition with Mixture of Experts. Neurocomputing 71, 1103–1107 (2008)
10. Ebrahimpour, R., Sarhangi, S., Sharifizadeh, F.: Mixture of Experts for Persian Handwritten Word Recognition. Iranian Journal of Electrical & Electronic Engineering 7(4), 217–224 (2011)
11. Yoon, J.-W., Yang, S.-I., Cho, S.-B.: Adaptive mixture-of-experts models for data glove interface with multiple users. Expert Systems with Applications 39(5), 4898–4907 (2012)
12. Caragea, C., Sinapov, J., Dobbs, D., Honavar, V.: Mixture of experts models to exploit global sequence similarity on biomolecular sequence labeling. BMC Bioinformatics 10(suppl. 4), S4 (2009)
13. Goodband, J.H., Haas, O.C.L., Mills, J.A.: A mixture of experts committee machine to design compensators for intensity modulated radiation therapy. Pattern Recognition 39, 1704–1714 (2006)
14. Güler, I., Übeyli, E.D.: A modified mixture of experts network structure for ECG beats classification with diverse features. Engineering Applications of Artificial Intelligence 18, 845–856 (2005)
15. Yumlu, M.S., Gurgen, F.S., Okay, N.: Financial time series prediction using mixture of experts. In: Proc. 18th Int. Symp. Comput. Inf. Sci., pp. 553–560 (2003)
16. Weigend, A.S., Shi, S.: Predicting daily probability distributions of S&P500 returns. J. Forecast. 19(4), 375–392 (2000)
17. Cheung, Y.M., Leung, W.M., Xu, L.: Application of mixture of experts model to financial time series forecasting. In: Proc. Int. Conf. Neural Netw. Signal Process., pp. 1–4 (1995)
18. Graczyk, M., Lasota, T., Telec, Z., Trawiński, B.: Application of Mixture of Experts to Construct Real Estate Appraisal Models. In: Graña Romay, M., Corchado, E., Garcia Sebastian, M.T. (eds.) HAIS 2010, Part I. LNCS, vol. 6076, pp. 581–589. Springer, Heidelberg (2010)
19. Basak, D., Pal, S., Patranabis, D.C.: Support Vector Regression. Neural Information Processing – Letters and Reviews 11(10), 203–224 (2007)
20. Makhoul, J.: Linear prediction. A Tutorial Review. Proceedings of the IEEE 63(4), 561–580 (1975)
21. Smola, A.J., Schölkopf, B.: A Tutorial on Support Vector Regression. Statistics and Computing 14, 199–222 (2004)
22. Chang, C.C., Lin, C.J.: Training *v*-support vector regression: Theory and algorithms. Neural Computation 14, 1959–1976 (2002)
23. Yuan, C., Neubauer, C.: Variational mixture of Gaussian process experts. In: Koller, D., et al. (eds.) Advances in Neural Information Processing Systems, vol. 21, pp. 1897–1904. MIT Press, Cambridge (2009)
24. Moerland, P.: Some methods for training mixtures of experts, Technical Report IDIAP-Com 97-05, IDIAP Research Institute (1997)

25. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30 (2006)
26. García, S., Herrera, F.: An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons. Journal of Machine Learning Research 9, 2677–2694 (2008)
27. Luengo, J., García, S., Herrera, F.: A Study on the Use of Statistical Tests for Experimentation with Neural Networks: Analysis of Parametric Test Conditions and Non-Parametric Tests. Expert Systems with Applications 36, 7798–7808 (2009)
28. Trawiński, B., Smętek, M., Telec, Z., Lasota, T.: Nonparametric Statistical Analysis for Multiple Comparison of Machine Learning Regression Algorithms. International Journal of Applied Mathematics and Computer Science 22(4) (2012) (in print)

Competence Region Modelling in Relational Classification

Tomasz Kajdanowicz[1], Tomasz Filipowski[1,2], Przemysław Kazienko[1], and Piotr Bródka[1]

[1] Wroclaw University of Technology, Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland
[2] Research Engineering Center Sp. z o.o., ul. Strzegomska 46B, 53-611 Wrocław, Poland
tomasz.kajdanowicz@pwr.wroc.pl,tomasz.filipowski@pwr.wroc.pl, piotr.brodka@pwr.wroc.pl, kazienko@pwr.wroc.pl

Abstract. Relational classification is a promising branch of machine learning techniques for classification in networked environments which does not fulfil the iid assumption (independent and identically distributed). During the past few years, researchers have proposed many relational classification methods. However, almost none of them was able to work efficiently with large amounts of data or sparsely labelled networks. It is introduced in this paper a new approach to relational classification based on competence region modelling. The approach aims at solving large relational data classification problems, as well as seems to be a reasonable solution for classification of sparsely labelled networks by decomposing the initial problem to subproblems (competence regions) and solve them independently. According to preliminary results obtained from experiments performed on real world datasets competence region modelling approach to relational classification results with more accurate classification than standard approach.

Keywords: competence region modelling, relational classification, collective classification, social network analysis.

1 Introduction

In the past decade the Internet has evolved from the web of static content to a complex social medium were each piece of information is created or influenced by users. Social web services has become the most frequently used parts of the network and almost all web pages had to integrate with them or introduce own user and community wise interfaces. From the computer science perspective, the nature of most online data change from independent to relational – that is a fact that data mining techniques researchers have to face.

One of the currently growing needs is classification of relational data in different areas. A wide range of various relational classifiers and their applications were introduced so far. However, there is still a big space for improvements. Existing classifiers, mostly because of correlated error [12], perform poorly with sparsely labelled networks, which is common in relational datasets. Secondly, relational

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 236–244, 2013.

measures used for classification have very high computational complexity. The main reason behind it is a necessity of recursive algorithm usage on a whole network structure. Moreover, most of valuable and interesting social data repositories contain big amounts of data, which makes calculations even more time and resource consuming.

In order to address these issues the paper presents how the general idea of Competence Region Modelling (CRM) [22], [23], known already in traditional classification techniques which fulfil the iid assumption (independent and identically distributed), can be applied to relational classification, namely to collective classification problem. CRM uses a clustering method on order to decompose data set into the competence regions. Then a number of classifiers are trained using data instances from each competence region. In the inference phase each instance from testing data is classified respectively to its region membership.

In this paper, Competence Region Modelling was adjusted and applied for collective classification problem using appropriate clustering and classification methods – the k-means algorithm has been used in order to group network nodes according to their network structural feature – eigenvalues and the independent random forest compound classifier was run separately for each group found.

2 Related Work

The concept of relational classification has more than a decade. It was extensively used in areas like labelling science papers and web content [1-3], segmentation and labelling of text [4], molecules and compounds classification in genetics and chemistry [5-6], fraud detection [7-8] or directed marketing [9-10]. A lot of the conceptual and experimental work was done in this subject over the years [24-25]. Different types of approaches and classifiers were proposed and examined; different features of relational data were utilized [11]. Yet, there is still a broad field for improvement in the field.

One of the most important issues that still have to be faced is computational complexity of the relational classification methods. Entities within the network are connected to each other and there is a strict dependence between node network positions and its graph based measures that are used for classification [28]. As a natural consequence of these two facts most of the existing algorithms are based on recurrence [27]. Because of this recurrence and the structural complexity of the real world networks, relational classifiers are highly expensive in terms of calculation. The bigger and denser network is, the more expensive necessary computations are. While we cannot simplify the social measures without redefining them, a lot can be done during the data pre-processing and sampling stage. Decomposing a social graph may be one of the possible approaches [18]. Learning on selected subgraphs for instance may decrease calculation time without classification accuracy lost. Structural decomposition by excluding some nodes and edges from the network may potentially increase performance. Grouping and clustering methods can be used to compress training examples sets, reduce calculation time and the cost of relational classification.

The accuracy of relational classification is a second important problem. There are many factors that can affect the accuracy of classification based on relational data [26]. One of the most important ones is network labelling density. Real world social networks are sparsely labelled and this leads to high correlated errors [18]. Restricting learning and classification areas by structural decomposition is a way to prevent correlated error spread [12].

Based on the previous work on the social network homophily [13] and social groups [20] it seems natural to limit single classifier competence to a single network cluster instead of using whole network to learn. Especially homophily seems to be promising in this case. It was one of the first phenomenon observed by the social network researchers [13-16] and is a tendency of positive bonds existence between object that are similar to each other. In other words, it is a positive correlation between features and relations of network nodes. The more similar object are the stronger they are connected. The reverse process is called autocorrelation. Homophily and autocorrelation are common in all types of social data [17]. Autocorrelation is the basis of collective inference approach – the foundation of most of the relational classification methods introduced so far. Homophily, social groups, relational measures (clustering) or other social network statistical features can be utilized in terms of performing relational decomposition and increasing final classification performance.

The subject of structural network decomposition in relational classification itself was barely scratched in a few publications like [18] but the results clearly show a potential that lays in hidden information that can be obtained from social network structure and used to cluster network graph to improve classification efficiency.

In this paper the original Competence Regions Modelling (CRM) [22-23] concept was adapted and used as a realization of structural decomposition in relational classification.

3 Competence Regions Modelling in Relational Classification

The general idea of Competence Regions Modelling (CRM) [22-23] consists in splitting training data into separate regions with an independent profile. The region identification is achieved by means of some clustering algorithms applied to regular input features describing learning cases (observations). A separate classifier (or an ensemble of classifiers) is trained for each such cluster – competence region. In this way classifiers may be more specialized – they are limited only to their smaller regions. During the testing phase, each testing case is assigned to the nearest competence region (cluster) and appropriate region's model is applied to it.

The proposal of Competence Regions Modelling (CRM) in Relational Classification is quite similar to traditional CRM in its main idea but it has to be adapted to relational, collective classification, see Fig. 1. The differences between traditional CRM and CRM in relational modelling are presented in Sec. 4.

It also aims at decomposing the initial space of observations (a network) into smaller clusters and performs more specialized generalization within them. Assuming

that a single relational classifier has limited generalization abilities for whole network it is proposed the idea to decompose it into smaller parts and generalize them individually.

In order to perform such generalization CRM initially assumes to calculate structural measures for each node using the whole network (Step 1, see Fig. 1 and Algorithm 1). Then, disregarding the class labels, it clusters the whole relational dataset (training and testing nodes) into k clusters using the clustering algorithm and structural measures – Step 2.

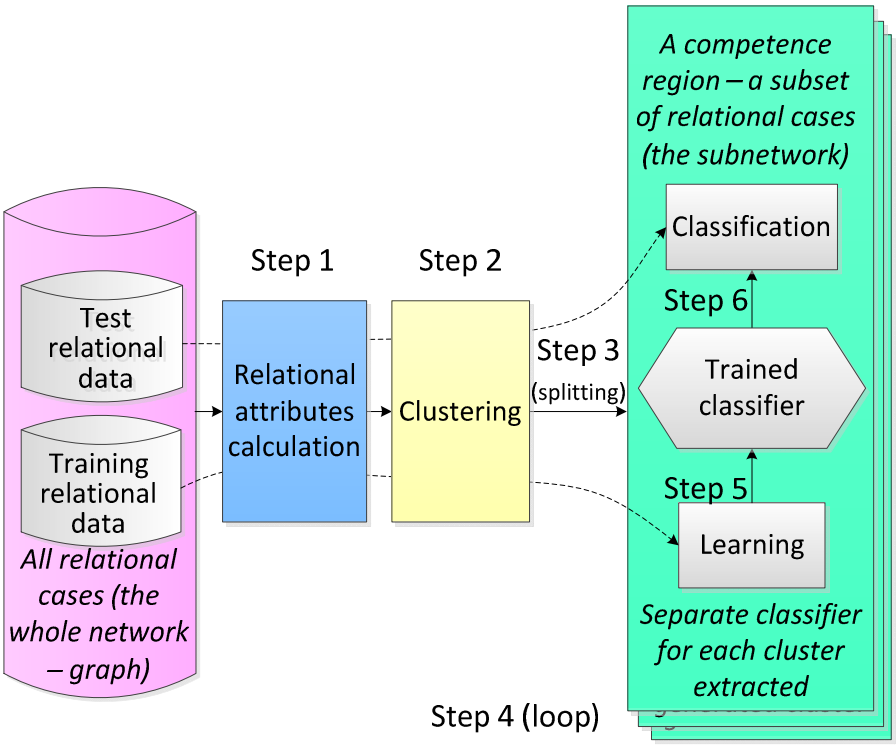


Fig. 1. The block diagram of Competence Regions Modelling in Relational Classification. Steps correspond to those from Algorithm 1.

Afterwards for each cluster that actually defines a competence region, training set and testing set is separated – Step 3. There is no special split technique defined for the CRM relational method, thus any can be used in this place. Once the data is split, an individual classifier is trained using calculated structural attributes of training nodes, separately within each competence region (cluster) – Step 4 (loop) and Step 5 (learning). There are k trained classifiers, which are competent in their clusters, after the learning phase. Having the trained classifiers obtained, we can generalize, i.e. apply knowledge extracted and stored within the individual models to unknown observations (cases). For each testing node in the network (relational observation) the

appropriate trained classifier is selected and applied – the one, which was prepared for the region the node belongs to (Step 6). Subsequently, a selected classifier returns classification results – classes for unknown cases in the region.

Algorithm 1. Algorithm for Competence Regions Modelling approach to relational classification:

```
(1)For each node n that belongs to the network N
     For each structural measure m from defined set M
         Calculate m for n using whole N
(2)Cluster whole N into k groups (clusters) using
   nodes' eigenvalues and the clustering/grouping
   algorithm
(3)Split N into training N_L and testing N_T sets
(4)For each competence region (cluster) k
     Train classifier c_k on nodes N_L that belong to
     cluster k
     Test classifier c_k on nodes N_T that belong to
     cluster k
(5)Validate results
```

4 Traditional CRM vs. CRM in Relational Modelling

There are some differences between traditional Competence Regions Modelling (CRM) and CRM in relational classification. Particularly, we have the following unique activities within CRM in relational classification:

1. For each case (observation) some structural, relational measures need to be calculated. It refers both training and testing data N_L and N_T - the whole network N (Step 1).
2. Clustering algorithms use relational features instead of typical ones. As a result, some graph-based clustering algorithms can be applied for the purpose of competence region extraction (Step 2).
3. After clustering, the training and testing data need to be separated (Step 3). In traditional CRM they are separated at the beginning, however in case of relational classification, relational measures need to be calculated before splitting.
4. There is no need to assign testing cases to regions found in the training data (in opposite to traditional CRM). In case of relational CRM both sets are split into regions during the clustering step.

5 Experiments

In order to validate the utility of proposed method preliminary experiments were recording its predictive accuracy in comparison to the results of reference method. The reference method was not using competence region modelling, but was similar to the proposed one in all other settings.

5.1 Setup of Competence Regions Modelling Method

In order to evaluate the predictive accuracy of proposed Competence Region Modelling (CRM) in relational classification a new experimental environment has been developed in Matlab. To perform the experiments the following setup was applied to each steps of the CRM algorithm (see Algorithm 1):

Step 1: As an input for the classifier the following relational measures calculated in the networks were used: degree centrality, betweenness centrality, clustering coefficient, hubs and authority, and page rank [21].

Step 2: The competence regions for specialized classifiers were obtained from graph clustering that used k-means algorithm to group nodes according to their eigenvalues calculated on the whole network.

Step 3: The experiment was evaluated using 10-fold cross-validation with splits of nodes into training and testing cases. The split was accomplished by node sampling using uniform distribution.

Step 5: As the proposed approach required classification algorithm, the random forest compound classifier has been used as a base classifier.

Step 7: In order to assess proposed relational classification approach a standard measure of classification mean error was recorded.

5.2 Datasets

The experiments were carried out on two relational datasets. The genealogy dataset CS_PHD was the network that contained the ties between Ph.D. students and their advisers in theoretical computer science where arcs pointed from advisers to students [19]. The dataset consisted of 1,061 nodes interconnected with 924 arcs. Each node was assigned with one of 16 classes (research areas). Average node degree was equal to 0.636.

The dataset NET_SCIENCE contained a co-authorship network of scientists working on network theory and experiment [20]. It was extracted from the bibliographies of two review articles on networks. The network consisted of 1,588 nodes and 2,742 edges. There were 26 classes and the average node degree was equal to 1.726.

5.3 Results

The mean classification error for various numbers of groups representing regions of competence in proposed algorithm is presented in Fig. 2 and Fig. 3. As we can see the mean classification error is different for various number of competence regions. For the CS_PHD dataset, Fig. 2, results obtained from CRM method exceeds up to 8% results obtained from relational classification with no regions modelling, irrespectively to the number of modelled competence regions. This does not hold for NET_SCIENCE dataset. Only using 2,5,7 and 10 competence regions allowed to obtain smaller classification error, see Fig. 3.

What is worth emphasizing is quite high classification error obtained by relational classification for both examined datasets. According to high number of classes in both datasets (16 and 26) the modelled problem is complex and results as obtained in experiments may be considered as satisfactory. Additionally in order to obtain high accuracy it might be required further reformulation of the problem (e.g. subsampling) which was not considered in this work.

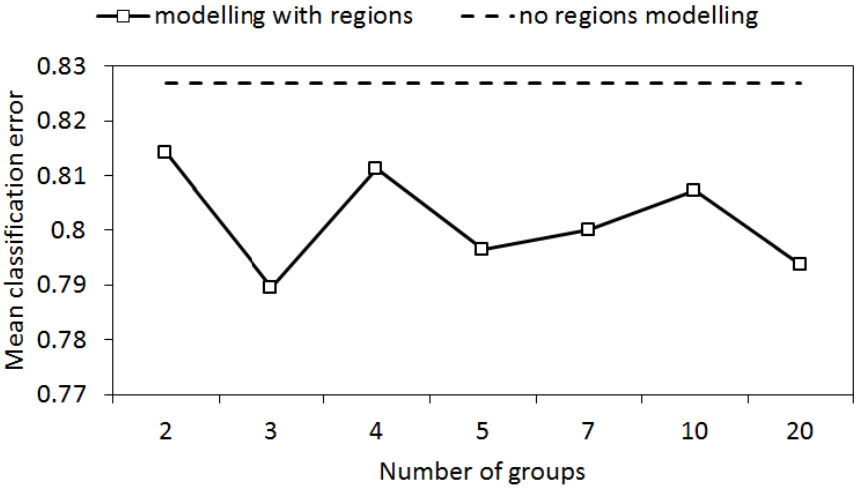


Fig. 2. Mean classification error for CS_PHD dataset using Competence Regions Modelling compared to results obtained from classification with no regions modelling

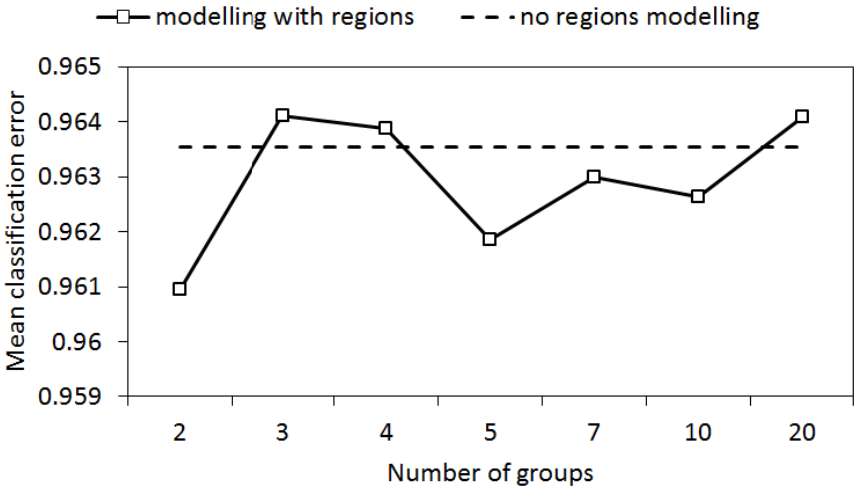


Fig. 3. Mean classification error for NET_SICENCE dataset using Competence Regions Modelling compared to results obtained from classification with no regions modelling

6 Conclusions and Future Work

The paper presented a new approach to relational classification based on competence region modelling. In general the approach aims at solving relational data classification problem by decomposing the initial problem to competence regions and solve them independently within them. It significantly differs from traditional competence regional modelling (CRM), see sec. 4. In particular, training and testing data are clustered together – in traditional CRM it refers only training data.

The results of the preliminary experiments showed interesting potential of the CRM approach to relational classification. Though, additional experiments with different settings have to be performed in order to gather deeper insight.

In further work there will be considered much more sophisticated clustering techniques, especially relational clustering, as well as other networks with various distributions of parameters. Moreover additional relational measures will be examined in order to discover their influence on CRM performance.

Acknowledgments. The work was partially supported by The Polish National Science Centre - the research projects, 2010-2013, 2011-2012 and fellowship co-financed by the European Union under the European Social Fund.

References

1. Chakrabarti, S., Dom, B., Indyk, P.: Enhanced hypertext categorization using hyperlinks. In: Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, pp. 307–319 (1998)
2. Taskar, B., Segal, E., Koller, D.: Probabilistic classification and clustering in relational data. In: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI), pp. 870–878 (2001)
3. Neville, J., Jensen, D.: Collective classification with relational dependency networks. In: Proceedings of the Multi-Relational Data Mining Workshop (MRDM) at the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2003)
4. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning (ICML), pp. 282–289 (2001)
5. Segal, E., Wang, H., Koller, D.: Discovering molecular pathways from protein interaction and gene expression data. Bioinformatics 19, 264–272 (2003)
6. Segal, E., Yelensky, R., Koller, D.: Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. Bioinformatics 19, 273–282 (2003)
7. Fawcett, T., Provost, F.: Adaptive fraud detection. Data Mining and Knowledge Discovery 3, 291–316 (1997)
8. Cortes, C., Pregibon, D., Volinsky, C.: Communities of Interest. In: Hoffmann, F., Adams, N., Fisher, D., Guimarães, G., Hand, D.J. (eds.) IDA 2001. LNCS, vol. 2189, p. 105. Springer, Heidelberg (2001)
9. Domingos, P., Richardson, M.: Mining the network value of customers. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 57–66 (2001)

10. Huang, Z., Chen, H., Zeng, D.: Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. ACM Transactions on Information Systems (TOIS) 22, 116–142 (2004)
11. Macskassy, S.A., Provost, F.: Classification in networked data: A toolkit and a univariate case study. J. Mach. Learn. Res. 8, 935–983 (2007), http://portal.acm.org/citation.cfm?id=1248693
12. Neville, J., Gallagher, B., Eliassi-rad, T.: Evaluating statistical tests for Within-Network classifiers of relational data, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.156.8070
13. Almack, J.C.: The influence of intelligence on the selection of associates. School and Society 16, 529–530 (1922)
14. Bott, H.: Observation of play activities in a nursery school. Genetic Psychology Monographs 4, 44–88 (1928)
15. Richardson, H.M.: Community of values as a factor in friendships of college and adult women. Journal of Social Psychology 11, 303–312 (1940)
16. Loomis, C.P.: Political and occupational cleavages in a Hanoverian village. Sociometry 9, 316–3333 (1946)
17. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Annual Review of Sociology 27, 415–444 (2001)
18. Neville, J., Jensen, D.: Leveraging relational autocorrelation with latent group models. In: Proceedings of the 4th International Workshop on Multi-Relational Mining, MRDM 2005, pp. 49–55. ACM, New York (2005), http://dx.doi.org/10.1145/1090193.1090201
19. Nooy, W., Mrvar, A., Batagelj, V.: Exploratory Social Network Analysis with Pajek, ch. 11. Cambridge University Press (2004)
20. Newman, M.: Finding community structure in networks using the eigenvectors of matrices. Physical Review E 74, 36–104 (2006)
21. Musiał, K., Kazienko, P., Bródka, P.: User position measures in social networks. In: Proceedings of the 3rd Workshop on Social Network Mining and Analysis, SNA-KDD 2009 (2009)
22. Kuncheva, L.I.: Clustering-and-selection model for classifier combination. In: Proceedings of KES 2000 Fouth International Conference on Knowledge-Based Intelligent Engineering Systems & Allied Technologies, vol. 1, pp. 185–188 (2000)
23. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley Interscience (2004)
24. Dzeroski, S., Lavrac, N.: Relational Data Mining. Springer, Berlin (2001)
25. Jensen, D., Neville, J.: Data mining in social networks. National Academy of Sciences workshop on Dynamic Social Network Modeling and Analysis (2002)
26. Jensen, D., Neville, J., Gallagher, B.: Why collective inference improves relational classification. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2004)
27. Neville, J., Jensen, D.: Iterative classification in relational data. In: AAAI Workshop on Learning Statistical Models from Relational Data, pp. 13–20 (2000)
28. Scott, J.: Social network analysis: A handbook, 2nd edn. Sage Publications Ltd., London (2000)

Approach to Practical Ontology Design for Supporting COTS Component Selection Processes

Agnieszka Konys, Jarosław Wątróbski, and Przemysław Różewski

West Pomeranian University of Technology in Szczecin,
Faculty of Computer Science and Information Systems,
ul. Żołnierska 49, 71-210 Szczecin, Poland
{akonys,jwatrobski,prozewski}@wi.zut.edu.pl

Abstract. The COTS (Commercial Off-The-Shelf components) selection process is difficult due to the huge number of existing COTS components. Moreover, the price of a mistake is great due to the complex nature of information systems. In this paper an analysis of different COTS component selection methodologies is presented. Based on this, the ontology for supporting COTS component selection processes is proposed. In order to achieve a high level of practicality on different levels of decision making, the ontology is implemented in Protégé software.

Keywords: COTS components, ontology, methodology supporting COTS component selection and evaluation.

1 Introduction

In literature, COTS (Commercial Off-The-Shelf components) products are defined as ready to sell products [12]. The increasing role of COTS components in the software marketplace influences the growth of popularity of that kind of software solutions and enables the construction of a whole system from COTS components [2]. COTS products can be part of a bigger and more complex COTS-Based System (CBS) [14]. The COTS market offers a wide range of software components supporting enterprise functions in different domains. The proper location of available components (and after then the choice of optimal solution) is one of the existing problems for enterprises. The COTS market is still developing, offering new COTS systems and reuse of existing solutions.

The adaptation of knowledge-engineering mechanisms should improve the process of knowledge acquisition on COTS components and related selection methodologies. Hence it is postulated that the ontology for COTS should provide mechanisms for updating information about particular components and methods, and extracting the information about these components according to inquiries posed by a decision-maker. It is premised that the ontology for methodologies supporting COTS component selection and evaluation enables a reduction in most research problems (e.g. knowledge systematization about COTS methodologies, the choice of a proper

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 245–255, 2013.

methodology for a given decision problem). Apart from the methodological aspect, the practical values allow determination of the technological implementation of the acquired research results.

The process of knowledge acquisition about COTS components is time consuming, limited by restricted access to component information and documentation [3]. The growing popularity of COTS components and the huge number of components available in the marketplace causes a data and information redundancy for decision makers. It is worth emphasising that knowledge about components is relatively low.

The possible alternatives to knowledge resource acquisition (such as COTS component repositories, semantic techniques, independent reports or expert knowledge etc.) are still in development phases. Some COTS repositories or semantic techniques are in prototype stages. It has been observed that COTS component repositories available on the market are excessively general solutions and it is difficult to use them for a specified domain. Owing to the huge number of COTS components on the market, a general repository application could not cope with the basic demands and requirements of decision makers, nor the functional conditions required by a particular domain either. Nowadays a small number of COTS repositories (e.g. COTSTrader, CLARiFi, CeBASE COTS Lessons-Learned Repository) supporting selection and evaluation processes exist in the market, but these solutions maybe inadequate through continuous development of the COTS marketplace. Another alternative for knowledge acquisition about COTS components is taxonomy (e.g. GOThIC). However, the information about software components included in a taxonomy system are scattered, only encompassing a short or even lacking a well-defined description of collected components. Moreover, both the precision and trustworthiness of collected information is not well documented [11].

The choice of appropriate software components from any number of available software solutions is one of the most important issues in the selection and development process of an enterprise's Information System. The existence of a huge amount of diffused information is one of the existing inconveniences related to the selection process. Moreover it could increment the risk of the decision making process.

In this paper a practical ontology supporting COTS component selection processes is presented. The ontology encompasses the set of selected COTS methodologies and was built using the Protégé application. The general aim of this solution is to provide and systematize knowledge about available COTS methodologies and to support the proper choice of the COTS component selection process as well.

The general statement of building an ontology supporting the COTS component selection process, is to provide a systematic and repeatable way to collect, organize and manage information about available COTS methodologies, and to support decision makers in the selection process [9]. Furthermore, the expected benefits are: (1) knowledge systematization about available COTS methodologies in the marketplace, (2) a time reduction in the selection process of a proper solution, (3) it will not be necessary for decision makers to have specific knowledge about available solutions on the marketplace, (4) the set of selected solutions fulfills the pre-defined requirements by a decision-maker, (5) the set of selected solutions complies the pre-defined criteria including varied level of details.

2 Analysis of COTS Component Selection Methodology

The are many differences between COTS software products available in the marketplace considering both the libraries and the applied components [1]. Hence, the problem of COTS component selection and evaluation is characterized by a high level of complexity. There are many approaches for COTS software selection that apply dissimilar methods in supporting the same software evaluation process.

The existence of repeatable and organized methodologies for software evaluation improve the whole decision-making process of software selection and decrease the eventual negative consequences. The analysis of literature identified a number of methodologies that support COTS software component selection and evaluation. It is worth noticing that the available methodologies present the evaluation process of COTS components in different ways. The differences encompass both our proposed approach, the applied method and its final effect as well. Moreover, the development and modification of pre-existing methodologies appears very often (e.g. OTSO methodology development) [1].

On the basis of literature research, a noticeable evolution of methodologies for COTS component evaluation has taken place [1]. The different types of specification, heterodox methodological approach and possible spectrum of practical applications conduce a comparative analysis of available COTS methodologies. The aim of this analysis is to provide a specification of their practical application areas in a COTS domain. The analysis encompasses 38 methodologies supporting COTS component evaluation (APCS, CAP, CARE, CBCPS, CDSEM, CEP, CSID, COSTUME, COTS-Agent Based System, CRE, Cil, Colombo and Francalanci, DBCS, Erol and Ferrel, FCS, GOThIC, IusWare, Jung and Choi, Lai, MAS, MRETS, Merad and Lemos, MiHOS, Morera, OTSO, PECA, PORE, RCPEP, SCARLET, SMI, STACE, Scenario-based technique, Sedigh Ali, StoryBoard, Teltumbde, Wang, Wei and Wang, WinWinSpiral Model). The specified characteristics of the selected methodologies are presented in publication [6]. The author's schema in Fig. 1 proposes an organizational schema of the selected methodologies depending on the applied technique and its further evolution.

The analyzed methodologies were organized in chronological order including the most important phases of COTS component evolution. Some of the analyzed methodologies can be grouped into several categories, but in the analysis the most dominant factor was considered. On the basis of this elaboration the general schema of a COTS software evaluation process was proposed. The relationships between available methodologies results from the application of the same, similar or modified approach proposed by the precursors of that method. The evolution of a COTS software evaluation process is driven by the development of the evaluation tools, new technologies and the enterprise requirements for its continuous development. Some of the analyzed methodologies have been elaborated for years. The latest solutions propose the application of hybrid methods or mixed techniques for COTS evaluation.

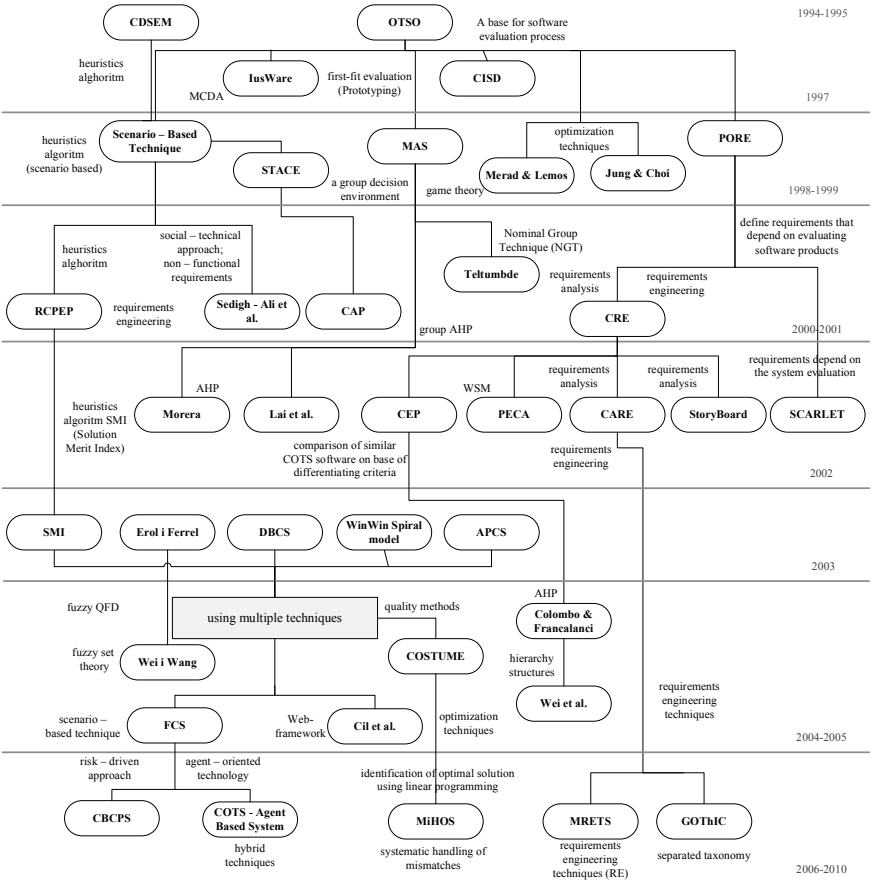


Fig. 1. Diagram of COTS component selection methodology evolution

3 The Approach to COTS Ontology Development

3.1 COTS Ontology Characteristics

In literature many approaches towards classification of the various types of ontology exist. For example Lassila and McGuinness [10] present a possibility of organizing an ontology based on the rising degree of formal semantics, whereas Oberle [13] proposes an idea of multi-dimensional fusion. On the basis of these approaches, Hepp [4] identifies six characteristics of variables of an ontology project (tab. 1).

Table 1. The general aspects of COTS ontology design

Aspects of ontology design	COTS ontology context
expressiveness	- The specified information about COTS methodologies are required; - The higher reasoning process provides the specified results; - The time-consuming process
size of relevant community	-Domain selection of projected ontology (COTS methodologies); -Defined set of final users of COTS ontology
conceptual dynamics in a given domain	-The strategy of updating the COTS ontology; -The process of making changes with relation to changes in COTS methodologies; -The limited level of specification of COTS ontology
number of conceptual elements in a particular domain	-Defining a dimension to the COTS ontology (the COTS ontology should be divided into the smaller parts if it is necessary); -A smaller size and higher specification level for COTS ontology provides more valuable results; -The process of COTS ontology adaptation and development is much easier
degree of subjectivity in conceptualization of the domain	-Defining the degree of notions considering particular concepts between actors in the COTS ontology
average size of the specification per element	-Defining the range of COTS ontology

3.2 The COTS Ontology Construction

The main advantage of the proposed COTS ontology is to provide systematic and repeatable knowledge about each of the presented methodologies. The ontology can then enable the selection of a proper solution for a given problematic area. Thus it provides the possibility to select the components in a simple and unique way including the set of criteria defined by a decision maker.

The basis for the ontology construction was a thorough analysis of considered solutions and then the experiment of identification of the set of criteria and sub-criteria that were used to create the taxonomy. For the ontology for methodologies supporting COTS component selection and evaluation, the set of criteria was created on the basis of available characteristics of these methodologies (see: Table 2).

The defined set of criteria was a basis for a taxonomy construction for methodologies supporting COTS component selection and evaluation. The aim of the taxonomy is to ensure systematization and classification for particular solutions. The taxonomy was a basis for an ontology construction as a next step.

The ontology was created including information about each considered COTS methodology and in many cases the names of criteria should be generalized. It helps in limitation of the total number of criteria in the ontology project. Furthermore it improves the speed of computing provided by a reasoner. The general structure of ontology for methodologies supporting COTS component selection and evaluation encompasses the following phases: (1) identification and selection of available COTS methodologies, (2) specified characteristics of selected COTS methodologies, (3) defining a set of criteria and sub-criteria, (4) taxonomy construction, (5) ontology, (6) process of defining classes.

Thereafter practical implementation of the constructed ontology is possible. That process is composed of the following steps: (1) defining preferences, (2) using of reasoning mechanisms, (3) selection of the solutions, (4) evaluation of the solutions.

Table 2. The relations between selected COTS methodologies and the defined set of criteria

Name of critera		Name of subcriteria	CEP	COSTUME	COTS-Agent Based System	CRE	Erol&Ferrel	IusWare	Jung&Choi	MRETS	Merad&Lemos	PORE	SCARLET	STACE
Software evaluation	1	Using of AHP method	-	-	-	-	-	-	-	-	-	-	-	yes
	2	Using of agents	-	-	yes	-	-	-	-	-	-	-	-	-
	3	Using of heuristics algorithm	-	-	-	-	-	-	-	-	-	-	-	-
	4	Using of MCDA	-	-	-	-	-	yes	-	-	-	yes	yes	-
	5	Using of experts methods	-	-	-	-	-	-	-	-	-	-	-	-
	6	Using of quality models	-	yes	-	-	yes	-	-	-	-	-	-	-
	7	Using of evaluation supporting tools	yes	yes	-	-	-	-	-	yes	-	-	-	-
	8	Using of risk-driven evaluation	-	-	-	-	-	-	-	-	-	-	-	-
	9	Using of software description from knowledge base or repository	-	-	-	-	-	-	-	-	-	-	-	-
	10	Using of first-fit evaluating (prototyping)	-	-	-	-	-	-	-	-	-	-	-	-
	11	Using of social-techniques	-	-	-	-	-	-	-	-	-	-	-	yes
	12	Using of taxonomy	-	-	-	-	-	-	-	-	-	-	-	-
	13	Using of software engineering techniques	-	-	-	-	-	-	-	yes	-	-	-	-
	14	Using of goal-oriented programming	-	-	-	-	-	-	yes	-	-	-	-	-
	15	Using of linear programming	-	-	-	-	-	-	-	-	-	-	-	-
	16	Using of scenario-based techniques	-	-	-	-	-	-	-	-	-	-	-	-
	17	Using of game theory	-	-	-	-	-	-	-	-	yes	-	-	-
	18	Using of fuzzy set theory	-	-	-	-	-	-	-	-	-	-	-	-
	19	Using of domain model utility	-	-	-	-	-	-	-	-	-	-	-	-
	20	Using of WSM	-	-	-	yes	-	-	-	-	-	-	-	-
	21	Using of confidence factors	yes	-	-	-	-	-	-	-	-	-	-	-
	22	Defining requirements to aggregate use-cases and screenshoots	-	-	-	-	-	-	-	-	-	-	-	-
Defining criteria importance	23	Group AHP	-	-	-	-	-	-	-	-	-	-	-	-
	24	Hierarchical decision model	-	-	-	-	-	-	yes	-	-	-	-	yes
	25	Grouping and DSS mechanisms	-	-	-	-	-	-	-	yes	-	-	-	-
	26	Global and local weights adding	yes	-	-	-	-	-	-	-	-	-	-	-
	27	Fuzzy set theory	-	-	-	-	-	-	-	-	-	-	-	-
	28	Fuzzy QFD	-	-	-	-	yes	-	-	-	-	-	-	-
	29	Specified software description	-	-	-	-	-	-	-	-	-	-	-	-
	30	Multi-criteria decision supporting method	-	-	-	-	-	yes	-	-	-	yes	yes	-
	31	Using of tools	yes	yes	-	-	-	-	-	-	-	-	-	-
COTS evaluation process	32	Process recurence	-	-	-	-	-	yes	-	-	-	-	-	-
	33	Using of non-functional attributes	-	yes	-	-	-	-	-	-	-	-	-	yes
	34	Using of technical attributes	-	-	-	-	-	-	-	-	-	-	-	-
	35	Goal-oriented approach	-	-	-	-	-	-	-	-	-	-	-	-
Criteria definition process	36	Model's spreadsheet	yes	-	-	-	-	-	-	-	-	-	-	-
	37	Brainstorm	-	-	-	-	yes	-	-	-	-	-	-	-
	38	Requirements identification	-	yes	-	-	-	-	-	-	-	-	-	-
	39	DSS mechanisms	-	-	-	-	-	-	-	yes	-	-	-	-
	40	Nominal Group Technique (NGT)	-	-	-	-	-	-	-	-	-	-	-	-
	41	Defined by a decision-maker	-	-	yes	yes	-	-	-	-	yes	-	-	-
	42	Defining criteria template	-	-	-	-	-	-	-	-	-	-	-	-
	43	Comparison of CBS elements	-	-	-	-	-	-	-	-	-	-	-	-
	44	Specified requirements analysis	-	-	-	-	-	-	-	-	-	yes	yes	yes
	45	Goals and specifications defining	-	-	-	-	-	-	-	-	-	-	-	-
	46	Interview or questionnaire	-	-	yes	yes	-	yes	-	-	-	-	-	-
Type of preferential information	47	Deterministic type of preferential information	-	-	-	-	-	-	yes	-	-	-	-	yes
	48	Cardinal type of preferential information	-	-	-	-	-	-	yes	-	-	-	-	yes
	49	Non-deterministic type of preferential	-	-	-	-	-	-	-	-	-	-	-	yes
	50	Ordinal type of preferential information	-	-	-	-	-	-	yes	-	-	-	-	-
	51	Fuzzy type of preferential information	-	-	-	-	-	-	-	-	-	-	-	yes

3.3 Practical Implementation of COTS Ontology

The ontology was built using the Protégé application (http://www.semanticweb.org/ontologies/2010/11/Ontology1292334387792.owl).

The language supporting building the ontology is OWL (Ontology Web Language). It provides both the possibility for description of concepts and new additional functions for describing possible relationships. Each group of criteria is referred to subclasses with a higher level of specification. The whole ontology is based on a tree structure. The developed ontologies with huge numbers of classes and complex inheritances almost always require a tree class hierarchy [5]. The process of COTS ontology construction was determined by a few statements. The correctness of activity of the ontology is ensured by unified notation of the names of the classes, properties, objects and data types without using national letters, space bars or symbols. Moreover, it is necessary for reduplications in applied nomenclature for criteria and sub-criteria in each ontology to be non-extant. Then each of the primitive classes should be disjointed from each other. Thereafter for each class both the slots should be defined and the proper values should be described.

The specified analysis of available COTS methodologies allows defining the set of classes and subclasses. Thus 5 classes and 51 subclasses were defined. On the basis of identified relations between 38 COTS methodologies and the pre-defined set of criteria (see: Table 2), the semantic relationships were created. Table 3 presents the selected set of criteria and sub-criteria of the COTS ontology.

Table 3. Selected criteria and sub-criteria of COTS ontology

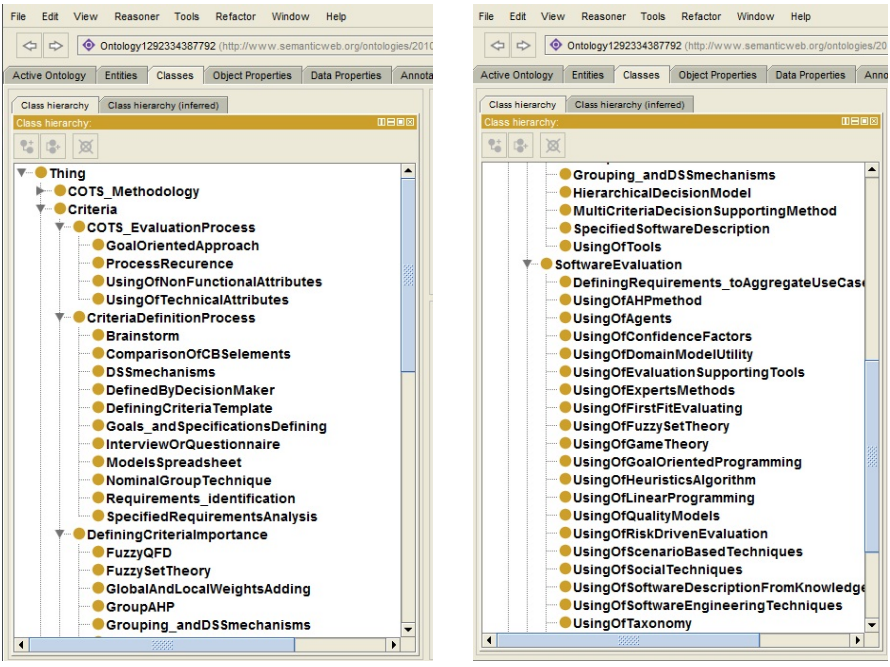


Figure 2 depicts 5 selected COTS methodologies (IusWare, PORE, GOThIC, OTSO and SCARLET) and the specified relations between the criteria for each of the selected methodologies.

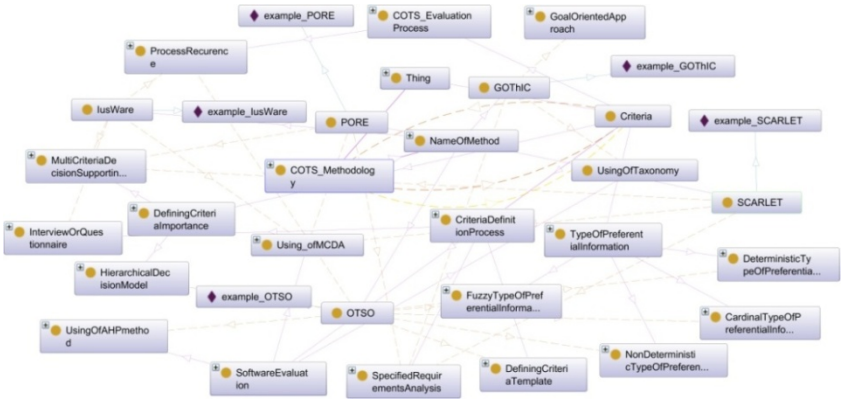


Fig. 2. The specified characteristics of COTS methodologies

The verification, consistency in evaluation and computable process was made using a HermiT reasoning mechanism. This process begins when the reasoning mechanism starts. The further steps encompass building the class hierarchy, initializing the class instance data structure and computing instances for all classes. The final result of computing encompasses the set of solutions for pre-defined queries. In the case of an identified inconsistency, the computable process is not workable.

4 Case Studies: Ontology Supporting COTS Component Selection Process Considering a Proper Methodology Choice for a Given Decision Problem

The case study presents a practical example of a COTS ontology application. It is supposed that a decision-maker is looking for the COTS methodologies that fulfill a set of pre-defined requirements. In this case the following requirements were identified by the decision-maker. The preferable methodology should satisfy the following criteria: (1) defining criteria importance: hierarchical decision model, (2) software evaluation: using of AHP method and software evaluation tools, (3) type of preferential information: deterministic type of preferential information.

The application of the reasoning mechanism provides a set of results with regard to the pre-defined requirements. In this case only two methodologies (STACE and CAP) fulfill these defined criteria (Fig.3). The same query was posed using a DL Query mechanism (Fig.4). The identified set of results is exactly the same as earlier.

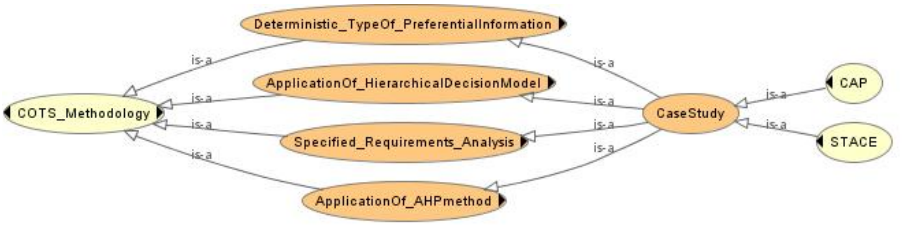


Fig. 3. A practical example of COTS ontology application – a limited set of results

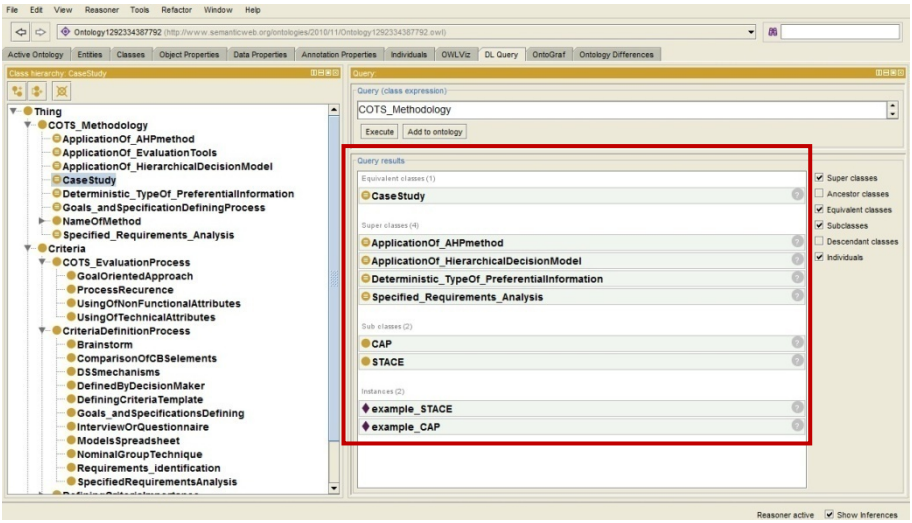


Fig. 4. A practical example of COTS ontology application – usage of DL query mechanism

This case study presented a practical example of ontology application taking into consideration the proper choice of a COTS methodology for a given decision problem. In this case, an exemplary set of requirements was defined, especially including the approach for software evaluation, criteria importance definition and type of preferential information offered by the particular methodologies. On the basis of the pre-defined criteria of the analyzed methodologies, it is possible to specify a non-limited set of queries for the COTS ontology.

5 Conclusion

In this paper an approach to design a practical ontology supporting COTS component selection process was presented. The general statements of COTS ontology construction were described. Furthermore the significant role of the proper selection of components was introduced. On the basis of specified characteristics of 38 COTS methodologies, the ontology was built using the Protégé application and OWL standard. Due to the limited space of publication, only the small portion of the

practical application of a COTS ontology was presented. Further results can be found in publications [7, 8].

The general aim of the COTS ontology construction was to provide a systematic and repeatable way for selection of a proper methodology for a given decision problem. The COTS ontology enabled the knowledge systematization about the available solutions on the marketplace. The specification level of the acquired results is defined by a decision-maker. The more valuable the requirements definition, the smaller the number of identified results. The decision-maker then does not have to have a broad knowledge of methodologies, but can still make a reasonable choice.

References

1. Ayala, C.: Systematic Construction of Goal-Oriented COTS Taxonomies. In: Proceedings of the 3rd Doctoral Consortium at the 18th Conference on Advanced Information Systems Engineering, CAISE 2006, June 5-9, Luxembourg (2006)
2. Carvallo, J.P., Franch, X., Grau, G., Quer, C.: COSTUME: A Method for Building Quality Models for Composite COTS-Based Software Systems, Technical Report LSI-04-12-R, Dept.LSI (2004)
3. Clark, J., Clarke, C., De Panfilis, S., Granatella, G., Predonzani, P., Silitti, A., Succi, G., Vernazza, T.: Selecting Components in Large COTS Repositories. Journal of Systems and Software - Special Issue: Applications of Statistics in Software Engineering 73, 323–331 (2004)
4. Hepp, M.: Ontologies: State of the art, business potential and grand challenges. In: Hepp, M., De Leenher, P., de Moor, A., Sure, Y. (eds.) Ontology Management: Semantic Web, Semantic Web Services, and Business Applications, pp. 3–22. Springer (2007)
5. Horridge, M. (ed.): A Practical Guide To Building OWL Ontologies Using Protege 4 and CO-ODE Tools Edition 1.2. The University of Manchester, Manchester (2009), `http://protege.stanford.edu/`
6. Konys, A., Wątróbski, J.: Methods and techniques of selecting scalable components for structure requirements of enterprise management systems. In: Łatuszyńska, M., Nermend, K. (eds.) Modern Enterprise Problems, pp. 145–170 (2009)
7. Konys, A., Wątróbski, J.: A model of ontology supporting COTS component selection process in management information system domain. In: Proceedings of Advanced Information Technologies for Management, AITM 2010, vol. 147, pp. 189–204. Wroclaw University of Economics Research Papers (2010)
8. Konys, A.: Ontologies supporting the process of selection and evaluation of COTS software components. In: Proceedings of Advanced Information Technologies for Management, AITM 2011. Wroclaw University of Economics Research Papers (2011)
9. Kushtina, E., Różewski, P., Zaikin, O.: Extended Ontological Model for Distance Learning Purpose. In: Reimer, U., Karagiannis, D. (eds.) PAKM 2006. LNCS (LNAI), vol. 4333, pp. 155–165. Springer, Heidelberg (2006)
10. Lassila, O., McGuinness, D.L.: The Role of Frame-based Representation on the Semantic Web, Knowledge Systems Laboratory Report KSL-01-02, vol. 6. Stanford University (2001)

11. Li, J., Conradi, R., Slyngstad, O., Bunse, C., Torchiano, M., Morisio, M.: An Empirical Study on Decision Making in Off-The-Shelf Component-Based Development. In: Proceedings of the 28th International Conference on Software Engineering, ICSE 2006, pp. 897–900 (2006)
12. Morisio, M., Torchiano, M.: Definition and Classification of COTS: A Proposal. In: Palazzi, B., Gravel, A. (eds.) ICCBSS 2002. LNCS, vol. 2255, pp. 165–175. Springer, Heidelberg (2002)
13. Oberle, D.: Semantic Management of Middleware, vol. 1. Springer, New York (2006)
14. Torchiano, M., Morisio, M.: Overlooked aspects of COTS-based development. IEEE Software 21(2), 88–93 (2004)

Planning of Relocation Staff Operations in Electric Vehicle Sharing Systems*

Junghoon Lee and Gyung-Leen Park

Dept. of Computer Science and Statistics,
Jeju National University, Republic of Korea
{jhlee,glpark}@jejunu.ac.kr

Abstract. This paper designs a computerized operation planner for relocation staffs in electric vehicle sharing systems, in which uneven vehicle distribution can lead to severe service quality degradation. After relocation pairs are created based on the target vehicle distribution and vehicle-to-station matching, our scheme finds an operation sequence for a relocation team. To overcome the time complexity of the ordering problem, a genetic algorithm is developed. It encodes a relocation schedule based on numbering of relocation pairs, defines a fitness function accounting for the inter-relocation move, and finally tailors genetic operators. The performance measurement result obtained from a prototype implementation shows that the proposed scheme finds an efficient schedule having a converged fitness value with just small-size population. The difference in relocation distance does not go beyond 24.8 % even in the case of extremely unbalanced distribution for the given parameters.

Keywords: Smart transportation, electric vehicle sharing system, vehicle relocation, genetic algorithm, relocation distance.

1 Introduction

According to the cohesive integration of high-end information and communication technologies, the modern power system, called the smart grid, is becoming more reliable and intelligent [1]. The smart grid pursues energy efficiency in a variety of grid objects belonging to power generation, transmission, distribution, and consumption. Not just restricted to efficiently managing such power supply domains, the energy system can regulate different type of consumer devices in homes, buildings, and the like [2]. In the mean time, electric vehicles, or EV in short, are one of the most important smart grid entities [3]. Many countries are making plans to replace gasoline-powered vehicles with EVs, as EVs much reduce air pollution and achieve significant energy efficiency. However, to say nothing of the short driving range, the cost is still much too high for personal

* Prof. Gyung-Leen Park is the corresponding author.
This research was financially supported by the Ministry of Knowledge Economy (MKE), Korea Institute for Advancement of Technology (KIAT) through the Inter-ER Cooperation Projects.

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 256–265, 2013.

ownership. Hence, carsharing is a reasonable business model for EVs for the time being and it contributes to the fast penetration of EVs into our everyday life.

In its service scenario, a customer rents out an EV at a sharing station and returns to either the same or a different station. The latter case is called one-way rental and it is most convenient from the customers' perspective [4]. However, due to ever-changing demand patterns, the number of EVs in each station will be uneven soon. Some stations cannot meet the balance between demand and supply. Then, customers can possibly fail to rent an EV in the station they want. Such a stock imbalance problem can be solved by explicit vehicle relocation, but how to relocate EVs is a very complex problem dependent on many variables. However, if the requirement is clearly specified, this relocation planning can be empowered by computational intelligence in obtaining an efficient relocation schedule [5]. Here, the communication between the relocation coordinator and respective EVs is indispensable, and networked vehicles allow many sophisticated services to be integrated in smart transportation.

[6] investigates existing relocation strategies in vehicle sharing systems where vehicles can be returned to any parking lot within the specific district. This survey classifies them into user-based and operator-based approaches. The user-based strategies give incentives to the trip which can compensate for the imbalance of supply and demand in the pick-up points. Even if these schemes do not need any cost for hiring service staffs, how many customers will be influenced by the reward is quite questionable. On the contrary, operator-based relocation works under the coordination of system managers and it can take advantage of sophisticated computer algorithms, especially on prediction and optimization. According to this work, operator intervention as well as subsequent manual relocation is indispensable for EV sharing systems to provide an acceptable service quality to customers, in spite of high cost for employment of service staffs and additional vehicle movement not taking any passengers.

The first thing to decide is when to execute the relocation procedure. A practical answer is during the non-operation time, as relocation staffs can redistribute EVs without being interrupted during this interval. However, it can be also triggered when demand-supply mismatch is forecasted. Next, the relocation coordinator determines how many EVs will be moved from overflow stations to underflow stations, considering the target EV distribution after relocation. Here, each EV in the overflow station is assigned to an underflow station. The final step is staff operation planning which searches an efficient route for the whole EV relocation. Multiple EVs can be simultaneously moved in a car transporter or each EV can be separately moved by relocation staffs. In the latter case, the basic operation is as follows: two staffs go to an overflow station driving a staff car. One drives EV to the assigned station while the other follows by the staff car. Then, two go to the next overflow station together in the staff car again.

In the path taken by the relocation staffs, underflow and overflow stations appear alternately. The relocation cost usually depends on the driving distance and time, and thus on the sequence of relocation operations. This is a variant of the TSP (Traveling Salesman Problem), where each relocation pair is the visiting

node, as every pair must appear in a visiting sequence just once. The cost of a sequence is calculated based on the distance between each station. As it belongs to the NP category, suboptimal methods are indispensable in spite of optimality loss for better responsiveness. Genetic algorithms are one of the most widely used suboptimal search techniques in many different areas, not restricted to just engineering problems [7]. In this regard, this paper designs a staff operation scheduler for EV relocation based on genetic algorithms. Our design embraces an encoding scheme for relocation routes, the definition of a fitness function, and customization of genetic operators.

This paper is organized as follows: After issuing the problem in Section 1, Section 2 introduces the background and related work. Section 3 explains the overall layout of the relocation coordinator and then designs a staff operation scheduler. After performance measurement results are demonstrated and discussed in Section 4, Section 5 summarizes and concludes the paper with a brief introduction of future work.

2 Background and Related Work

Jeju City in the Republic of Korea has established one of the world's largest smart grid testbed, aiming at verifying and developing advanced technologies in 5 major areas consisting of smart power grid, smart place, smart transportation, smart renewables, and smart electricity services. Here, EVs are the key element of smart transportation. Noticeably, due to the terrain effect, the vehicle speed can hardly exceed 100 kmh, while the average daily driving distance is below 100 km. Hence, this area is considered as the best place for the penetration of EVs and also for the development of EV-related business models. Currently, hundreds of EVs are deployed and also hundreds of EV chargers are installed over the whole island surrounded by 200 km coastline. Several consortiums participating in this enterprise are working on business models on EV sharing, EV rent-a-cars, and integration of wind energies to EV charging.

As for the management of carsharing systems, [8] designs a three-phase decision support system for vehicle relocation. It begins with *Optimizer*, which allocates staff resources for the sake of minimizing the relocation cost, considering customer pick-up and return patterns, number of parking stalls, and inter-station relocation cost. After formulating a mixed integer linear programming model and necessary constraints, the problem is solved by a branch-and-bound technique. Then, phase 2, or *Trend Filter*, calculates the optimized results by means of a series of heuristics to finalize a recommendation set of operating parameters such as shift hours as well as relocation techniques. Here, the relocation manager extracts upper and lower buffer thresholds, by which the relocation procedure is triggered. In phase 3, its simulator part evaluates the effectiveness of recommended parameters in terms of zero-vehicle-time, full-port-time, and number of relocations.

In addition, [6] designs a two-step relocation strategy. In the first step, *Offline module* captures demand patterns repeating periodically, scaling from daily-basis

to seasonal-basis. Moreover, demand patterns for unusual events such as sports games or fare trades are also parameterized. Demand patterns are clustered according to their spatio-temporal behaviors and then cluster-specific relocation strategies are pre-selected for respective clusters. In the second step, *Online optimization module* continuously monitors the current state of spatial vehicle distribution to decide whether to trigger the relocation procedure. Here, an ordinal number is estimated by the comparison of optimum and current state. The selected strategy relocates vehicles to the optimal distribution with the minimal cost. Here, the offline selection of appropriate strategies can be fulfilled without any restriction on execution time length.

[9] proposes a relocation method consisting of focus-forecasting, inventory replenishing, and microscopic simulation. Impressively, it handles the uneven vehicle distribution brought by one-way rentals on the existing carsharing services in Singapore. The focus-forecasting model traces and forecasts the total number of vehicles rented out and returned in each sharing station. It exploits several well-known time-series techniques for selective moving average and takes the best one that would have been most accurate for the most recent period. Next, if an inventory decision is made in favor of vehicle relocation, the understocked station will be replenished from the nearest over-stocked station so as to minimize the travel cost for relocation. The microscopic simulator estimates the current traffic status and runs a link-to-link shortest path algorithm to find an efficient relocation path. However, this work omits how to actually move or redistribute individual vehicles.

3 Relocation Scheduler

3.1 Preliminaries

Figure 1 depicts how our relocation scheme works. It consists of 3 steps covering relocation strategy, action planning, and staff operation planning, while our design makes them independent as much as possible. That is, any relocation strategy can work with our action and staff operation planning schemes and vice versa. To begin with, our previous work has built a performance analysis framework for EV sharing systems based on the actual trip data consisting of pick-up and drop-off points collected from a taxi telematics system in Jeju City [10]. Each point is mapped to a specific sharing station and considered as a renting out or returning activity at the mapped station. This framework can approximate the demand dynamics for EV sharing requests, which are rarely available due to immaturity of EV sharing business. For the given parameters including the number of EVs, it is possible to conduct experiments to measure service ratio, moving distance, and per-station EV distribution at relocation time.

Upon a practical assumption that the relocation procedure is carried out during non-operation hours, relocation strategies decide the relocation vector, which is the target EV distribution after relocation [11]. Intuitively, 3 intuitive strategies have been considered including even, utilization-based, and morning-focused schemes. The even relocation scheme makes all station have the same number

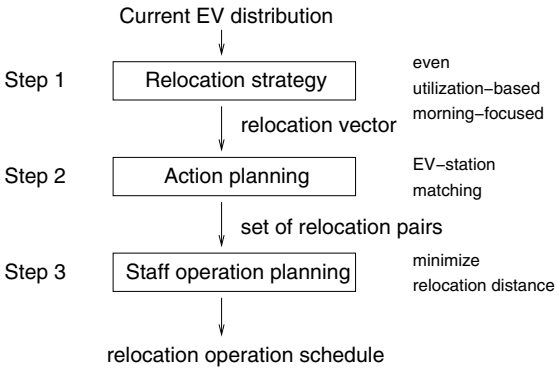


Fig. 1. Relocation coordination

of EVs. It doesn't consider demand patterns, but may have reasonable service ratio in the long term. Second, the utilization-based scheme relocates EVs according to the average demand ratio of each station. It assigns more EVs to the station having more pick-ups. Next, the morning-focused scheme redistributes EVs according to just the pick-up requests during the first few hours after the beginning of operation hours. [11] shows that the morning-focused scheme outperforms others in Jeju City data. Anyway, in overflow stations, the number of EVs is larger than the target distribution.

In the second step, the relocation action planner decides the set of (E_i, S_j) pairs ($0 \leq i < m$ and $0 \leq j < n$), where m and n are the number of EVs and stations, respectively. Each of them corresponds to relocating an EV, say E_i, to an underflow station, say S_j. The number of relocation pairs coincides with the number of EVs in overflow stations, while only underflow stations can be included. For the relocation vector, or target vehicle distribution given by a relocation strategy, the proposed planner builds two preference lists, one for EVs in overflow stations and the other for underflow stations [12]. Then, the matching procedure assigns each electric vehicle to a station in such a way to reduce the relocation cost by means of a modified stable marriage problem solver. The relocation distance is the sum of distances for all (S_i, S_j) pairs in the relocation plan, where S_i is the overflow station E_i is currently parked in. To calculate the relocation distance, it is necessary to know the distances of all station pairs.

Finally, individual EVs must be moved according to the relocation plan represented by a set of (S_i, S_j) pairs, and this is the main subject of this paper. Here, we consider the operation schedule of a single relocation team consisting of 2 service staffs, and the basic relocation activity is described previously. Actually, it's true that more than 2 EVs can be moved with more than 3 staffs. Moreover, two teams or more can run respective relocation vehicles, and some EV relocation can take place in parallel. However, it is hard to concretely model those activities and single team scheduling will be a preliminary building block for complex relocation planning. After all, each (S_i, S_j) pair is moving an EV

from an overflow station, S_i, where the EV is located to an underflow station, S_j. According to the sequence of each pair, the total relocation distance and time will be different. Apparently, another cost function can be defined to account for different criteria.

3.2 Staff Operation Planner

This section designs a staff operation planner exploiting genetic algorithms. The ordering problem usually needs $O(n!)$ complexity, where n is the number of elements, namely, the number of relocation pairs. Even though the genetic algorithm can possibly fail to find an optimal solution, it can generate a relocation plan within an acceptable time bound. Moreover, its execution time is controllable by adjusting the number of genetic iterations. It can also combine a variety of efficient heuristics in selecting crossover points, processing given constraints, and the like. To begin with, it is necessary to encode a schedule to a chromosome, which is represented by an integer-valued vector. Second, a fitness function must be defined to evaluate the quality of each schedule. It will affect which one will survive and be better selected for mating. Finally, genetic operators such as selection, crossover, and mutation, are tailored for the schedule generation based on the given system goal.

Figure 2 outlines our staff operation planning. Basically, after a relocation team moves an EV to the designated underflow station, the team goes to the next overflow station. To apply genetic algorithms, each sequence or feasible schedule is encoded to a integer-valued vector. Our design assigns a unique number to each relocation pair. There can be the same pairs in the relocation action plan, if more than one EV in an overflow station needs to be moved from the same underflow station. The example in Figure 2 contains 2 (A, B) pairs. Those pairs are assigned different numbers. The sequence of these numbers can represent a staff operation schedule, and we can calculate the relocation distance for this vector, or chromosome. Here, we assume that the distance between every pair of stations is known in priori, as it can be easily calculated by the A* algorithm provided that current traffic information is available.

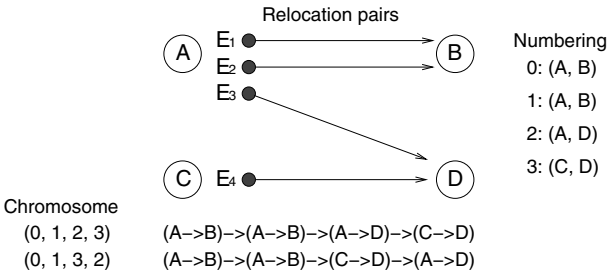


Fig. 2. Encoding of relocation schedules

Now, genetic operators are iteratively executed, continuously improving the quality of population for each generation. For initial population, a predefined number of chromosomes are generated randomly. The iteration mainly consists of selection and reproduction. Selection is a method that picks parents according to the fitness function. The Roulette wheel selection gives more chances to chromosomes having better fitness values for mating. Reproduction, or crossover, is the process taking two parents and producing offsprings. This operation randomly selects a pair of two crossover points and swaps the substrings from each parent. After crossover, some elements disappear while others come more than once in a single chromosome. Hence, duplicated elements will be replaced by disappearing ones to keep the validity of schedules. Moreover, reproduction may generate the same chromosome with already existing ones and they will be replaced by new random ones.

In the example shown in Figure 2, there are 4 sharing stations from A to D. Among them, A and C are overflow stations. A has 3 more EVs than its target value, while C has 1. On the contrary, B and D are underflow stations, each of which needs 2 EVs to meet the target distribution. The action planner creates the set of relocation pairs, $\{(E_1, B), (E_2, B), (E_3, D), (E_4, D)\}$. They are numbered from 0 to 3 and each E_i is replaced with an overflow station where it belong, resulting in the list of $\{(A, B), (A, B), (A, D), (C, D)\}$. There are 4! feasible operation sequences for them and the relocation distance can be calculated based on the route taken by them. Figure 2 also shows two staff operation plans encoded by (0, 1, 2, 3) and (0, 1, 3, 2). With a vector and numbering table, we can be aware of station-level visiting sequences. Namely, for the sequence of (0, 1, 2, 3), the cost will be the sum of every pair of consecutive stations, namely, $\overline{AB} + \overline{BA} + \overline{AB} + \overline{BA} + \overline{AD} + \overline{DC} + \overline{CD}$.

4 Experiment Result

We implement a prototype of the proposed staff operation planner using the C programming language for performance assessment. The performance parameters include genetic algorithm-related ones such as the number of iterations and population size as well as relocation-related ones such as the number of underflow stations and the number of moves, or relocation pairs. In the mean time, the relocation distance is the main performance metric. Here, the distance scale is not specified, as it depends on the vehicle type. For EVs, it can be implicitly assumed to be km. For each parameter setting, we generate 30 experiment sets of relocation pairs, measure the relocation distances by running our prototype implementation, finally and average them. In addition, we set the number of sharing stations to 10 and inter-station distance exponentially distributes with the average of 3.0, but ranging between the lower and upper bounds of 0.5 and 5.0, respectively. These values are chosen, mainly targeting at Jeju City area.

The first experiment measures the fitness value change according to the progress of genetic iterations and the result is shown in Figure 3. In this experiment, the population size is set to 96, the number of underflow stations to

4, and the number of moves to 15. After the first iteration, the fitness value is 59.88. It is cut down to 42.07 after 37 iterations and remains unchanged to the end. On the contrary, the average fitness value of all schedules in the population keeps changing according to the iteration, as new solutions are included and compete for survival. Other relocation pair sets show the same pattern and converges within 100 iterations in most cases. This result indicates that 1,000 iterations are sufficient to make the fitness value converge for operation planning. It takes much less than 1 second with an average-performance computer and our scheduler can practically work even in mobile devices.

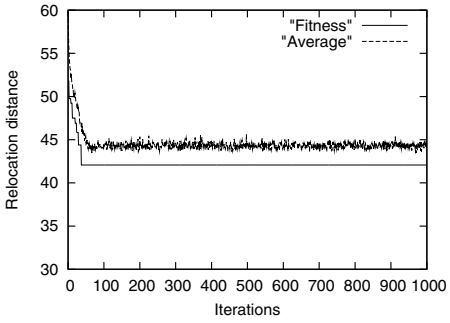


Fig. 3. Fitness change for iteration

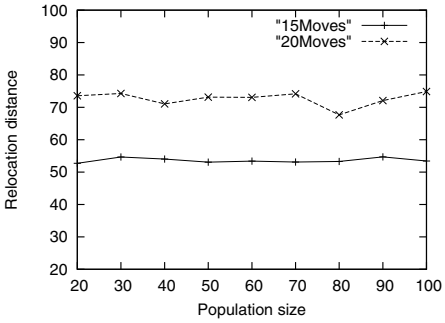


Fig. 4. Effect of population size

Next, Figure 4 plots the effect of population size to the relocation distance. As each generation has the step of sorting chromosomes in the population by their fitness values, the population size affects the execution time. We change the population size from 20 to 100. The number of underflow stations is set to 4 and that of iterations to 1,000. Figure 4 shows 2 curves, one for 15 moves and the other for 20 moves. For the case of 15 moves, the population size hardly affects the relocation distance for all of 30 sets. The relocation distance averaged from them ranges from 52.7 to 54.6, and the difference is less than 3.5 %. On the contrary, for the case of 20 moves, the relocation distance can be reduced to 67.6 with the population size of 80, while its maximum reaches 74.8, showing the difference of 9.7 %. Large population does not always lead to a better solution, as different chromosomes are selected for mating and evolved in different ways. This result indicates that our scheduler can find an operation of reasonable quality even with small-size population.

Figure 5 shows the effect of the number of underflow stations to the relocation distance. When generating a relocation pair, the source station is selected randomly out of overflow stations, while the destination is from underflow stations. Hence, if the number of underflow stations is small, the relocation pairs tend to be many-to-one mapping, and relocation pairs have a common destination. Otherwise, in the one-to-many mapping situation, the relocation staffs are more likely to return to the same overflow station for the next move. Our scheduler can find a better schedule when overflow and underflow stations are well-balanced.

In this case, there can be a circular loop which can cut down the inter-relocation distance. In this experiment, we change the number of stations from 1 to 9, and set the number of iterations and population size to 1,000 and 96, respectively. Figure 5 includes 3 curves for the cases of 15, 20, and 30 moves. The relocation distance differs by 22.2 %, 24.8 %, and 20.1 % for each case.

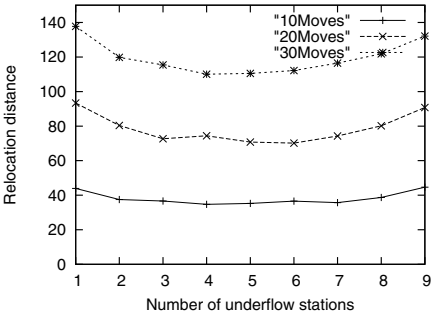


Fig. 5. Effect of the underflow stations

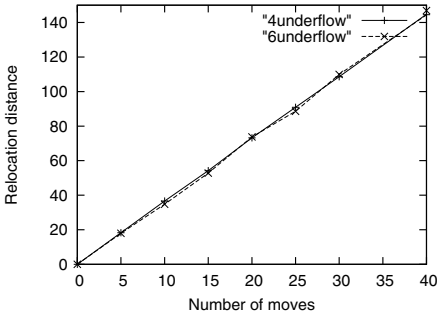


Fig. 6. Effect of the number of moves

Finally, our experiment measures the effect of the number of moves to the relocation distance, and the result is shown in Figure 6. We change the number of moves from 0 to 40 for the cases of 4 and 6 underflow stations, respectively. Here again, the number of iterations and population size to 1,000 and 96. Basically, the relocation distance linearly increases according to the number of moves, as the per-pair relocation distance will be the same as the average inter-station distance in the long run. Figure 6 plots two curves, one for the case of 4 underflow stations and the other for 6 underflow stations. In both cases, the relocation distance goes stably, and this experiment discovers that the proposed scheme always finds the converged solution.

5 Conclusions

Smart transportation is an important area in the smart grid, and electric vehicles are the key components, as they can reduce CO_2 emissions and achieve energy efficiency in transportation systems. EV sharing is one of the most practical business models for EVs due to the high cost for personal ownership. Without explicit vehicle relocations, a sharing system will suffer from severe mismatch between supply and demand. The relocation procedure consists of relocation strategies, relocation pair matching, and staff operation planning. In this paper, we have designed a relocation coordinator for EV sharing systems, focusing on staff operation scheduling. For the given relocation pairs created by our previous work, the proposed scheme generates an efficient schedule for a relocation team. Genetic algorithms are exploited to overcome the non-polynomial time complexity of the ordering problem, by encoding a schedule, defining a fitness

function, and customizing genetic operations. According to the experiment results, the proposed scheme finds an efficient schedule having a converged fitness value with just small-size population.

As future work, we are considering to extend our work to cover the case of multiple relocation teams to further improve the relocation overhead by parallel operations. With the efficiency in the relocation staff operation, it is also possible to design a proactive relocation procedure which is triggered even during the operation hours based on demand forecast.

References

1. Ipakchi, A., Albuyeh, F.: Grid of the Future. IEEE Power & Energy Magazine, 52–62 (2009)
2. Logenthiran, T., Srinivasan, D., Shun, T.: Demand Side Management in Smart Grid using Heuristic Optimization. IEEE Transactions on Smart Grid, 1244–1252 (2012)
3. He, Y., Venkatesh, B., Guan, L.: Optimal Scheduling for Charging and Discharging of Electric Vehicles. IEEE Transactions on Smart Grid, 1095–1105 (2012)
4. Barth, M., Todd, M., Xue, L.: User-based Vehicle Relocation Techniques for Multiple-Station Shared-Use Vehicle Systems. Transportation Research Record 1887, 137–144 (2004)
5. Cepolina, E., Farina, A.: A New Shared Vehicle System for Urban Areas. Transportation Research Part C, 230–243 (2012)
6. Weikl, S., Bogenberger, K.: Relocation Strategies and Algorithms for Free-Floating Car Sharing Systems. In: 15th International Conference on Intelligent Transportations Systems, pp. 355–360 (2012)
7. Sivanandam, S., Deepa, S.: Introduction to Genetic Algorithms. Springer (2008)
8. Kek, A., Cheu, R., Meng, Q., Fung, C.: A Decision Support System for Vehicle Relocation Operations in Carsharing Systems. Transportation Research Part E, 149–158 (2009)
9. Wang, H., Cheu, R., Lee, D.: Dynamic Relocating Vehicle Resources Using a Microscopic Traffic Simulation Model for Carsharing Services. In: 3-rd International Joint Conference on Computational Science and Optimizations, pp. 108–111 (2010)
10. Lee, J., Kim, H., Park, G., Kwak, H., Lee, M.: Analysis Framework for Electric Vehicle Sharing Systems Using Vehicle Movement Data Stream. In: Wang, H., Zou, L., Huang, G., He, J., Pang, C., Zhang, H.L., Zhao, D., Yi, Z. (eds.) APWeb Workshops 2012. LNCS, vol. 7234, pp. 89–94. Springer, Heidelberg (2012)
11. Lee, J., Park, G.-L., Kang, M.-J., Kim, J., Kim, H.-J., Kim, I.-K., Ko, Y.-I.: Design of an Efficient Matching-Based Relocation Scheme for Electric Vehicle Sharing Systems. In: Kim, T.-h., Ramos, C., Abawajy, J., Kang, B.-H., Ślęzak, D., Adeli, H. (eds.) MAS/ASNT 2012. CCIS, vol. 341, pp. 109–115. Springer, Heidelberg (2012)
12. Lee, J., Kim, H.-J., Park, G.-L.: Relocation Action Planning in Electric Vehicle Sharing Systems. In: Sombattheera, C., Loi, N.K., Wankar, R., Quan, T. (eds.) MIWAI 2012. LNCS (LNAI), vol. 7694, pp. 47–56. Springer, Heidelberg (2012)

Thematic Analysis by Discovering Diffusion Patterns in Social Media: An Exploratory Study with TweetScope

Duc Nguyen Trung[1], Jason J. Jung[1,⋆], Namhee Lee[2], and Jinhwa Kim[2]

[1] Department of Computer Engineering
Yeungnam University Gyeongsan, Korea 712-749
[2] Sogang Business School Sogang University
Seoul, Korea 121-742
{duc.nguyentrung,j2jung,namhee.lee80}@gmail.com,
jinhwakim@sogang.ac.kr

Abstract. The goal of this work is to capture diffusion patterns in social media, and to understand meaningful associations between the diffusion patterns and thematic features of the corresponding information. To do so, we have developed a Twitter-based diffusion monitoring system (called TweetScope) to efficiently collect the datasets from Twitter and conduct the proposed discovery process. Particularly, we expect that this work is feasible on establishing business strategies of various organizations.

Keywords: Sentiment analysis, Social media, Twitter, Information diffusion.

1 Introduction

Recently, many social networking services (e.g., Twitter, Facebook, and so on) have been regarded as an important online communication channel to exchange information among users. Such social media are efficiently providing public users with latest information. Hence, various organizations are trying to interact with users through the social media. Particularly, most of enterprises have been exploiting the social media for advertising their new products and promotions to customers.

However, it is very difficult for most of the organizations to take the publicity actions (e.g., advertisement and promotions) without understanding feedbacks and responses of public users. We can expect that most of the information from the enterprises will be similar to "spam" emails. In order to solve the problem, it is important to capture and understand feasible patterns in these social media. Many research groups have been currently investigating *social data analytics*. Hence, various types of patterns can be discovered by analyzing the datasets collected from users in social media. Such datasets (e.g., "like" in FaceBook[1], "check in" in FourSquare[2], and "RT (retweet)" in Twitter[3]) can be generated by the explicit user reactions.

⋆ Corresponding author.

[1] http://www.facebook.com/
[2] http://foursquare.com/
[3] http://twitter.com/

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 266–274, 2013.

Particularly, the diffusion patterns within such social media can indicate temporal dynamics of information flow (i.e., how the information is propagated along social links among users). It is important to understand which factors can make any significant influence on a certain diffusion pattern. We can expect that such understanding is applicable for establishing optimal business strategies (e.g., marketing).

The goal of this study is to understand how the information from enterprises is propagated over time through social medias. More importantly, we want to reveal the relationships between the types of information and the diffusion patterns. In this work, we are focusing on such information diffusion patterns in Twitter. There are two main assumptions of this study, as follows;

- RT is assumed to be related to user contexts (e.g., relevance and preferences). When users find a certain tweet that they are interested in, they can take a RT action on this tweet.
- RT is assumed to increase the chance of next RT. As users are doing RT more, the corresponding tweets can get reached to more users.

For example, in Fig. 1, four users (of course, there can be more users) are following a business @Olleh_mobile. Two tweets are retweeted by these four users (i.e., twt_a by John and Anne, and twt_b by John, Paul, and Smith). We can assume that John and Anne are interested in the topic of twt_a. Also, we want to note that the tweets retweeted by them can be retweeted again by additional users who are following them.

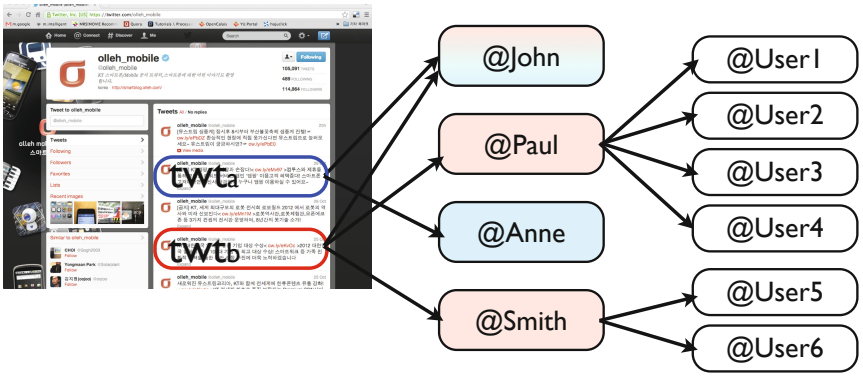


Fig. 1. An example of a friendship network by "following" in Twitter

The outline of this paper is as follows. In the following Sect. 2, the existing studies on information diffusion in social media are described. Sect. 3 introduces basic definition (with notations), and then formalizes diffusion patterns on social networks. In Sect. 4, we show our experiment environment and demonstrate the implemented system called TweetScope. Sect. 5 will match the discovered diffusion patterns with thematic features manually labelled by experts. Finally, in Sect. 6, we will draw the conclusion of this work and note several practical research limitation.

2 Related Work

Traditionally, *word of mouth* (WOM) has been regarded as an important communication phenomenon for designing marketing strategies. Many studies have claimed various factors which can make an important influence on *viral* effects. With emerging online social media, such WOM studies have been extended to electronic WOM (eWOM) [1]. In particular, efficient data processing schemes (e.g., customer segmentation, and network analysis) have been applied to eWOM. Text mining methods have been also applied for opinion mining from textual comments from customers [2].

Moreover, many studies have been presenting several interesting approaches to easily understand information diffusion phenomena in various types of social media (e.g., folksonomy [3]). Also, in [4], customer activities on social media can be aggregated to build a set of social pulses and the similar activities are visualized. Also, event diffusion [5] has been proposed by statistical analysis on textual information.

3 Diffusion Patterns

3.1 Network Formalization

To efficiently capture and understand diffusion patterns in social media, we want to formulate microtexts in social media and topological features of the social networks.

Definition 1 (Microtext). *A microtext is composed of i) text (a set of words), ii) a timestamp (when the microtext was generated), iii) a set of neighbors (who appear in the microtext), and iv) a set of URLs (a set of short hyperlinks).*

Definition 2 (Directed social network). *A directed social network $\mathcal{S}$ is represented as*

$$\mathcal{S} = \langle \mathcal{U}, \mathcal{N} \rangle \tag{1}$$

where $\mathcal{U}$ is a set of users and $\mathcal{N} \subseteq |\mathcal{U}| \times |\mathcal{U}|$ is a set of links between the users. Since this social network is a directed graph, these links can be easily represented as an asymmetric matrix.

Property 1 (Out-degree). Given a certain user $u_i \in \mathcal{U}$, his out-degree $Out(u_i)$ can be represented as a set of outgoing edges.

Property 2 (Distance). Given two users u_i and u_j, a distance $Dist(u_i, u_j)$ is represent as a length of the shortest path between them. Since the network is directed, $Dist(u_i, u_j) \neq Dist(u_j, u_i)$.

In case of Twitter, the microtexts are called 'tweets'. The social network among users can be constructed by 'following' actions which are directional links. Out-degree of each user can be regarded as a number of followers. In Fig. 1, $Out(@\text{Olleh\_mobile}) = 4$. Thus, once a certain target account u_X (e.g., businesses and users) is selected, it is mainly represented as two parts; i) a set of tweets, and ii) a set of followers.

3.2 Diffusion Patterns

Information diffusion in Twitter is expressed by user retweets (RT). Consequently, each tweet can has its own retweet network (if it is retweeted by followers). As this retweet actions by followers are occurred, the corresponding tweets are diffused into more users. This RT network $\mathcal{S}_{RT}$ is assumed to be a sub-network of $\mathcal{S}$.

Definition 3 (Retweet network). *Given a certain tweet twt at time t, a retweet (RT) network $\mathcal{S}^{twt}_{RT(t)}$ is represented as*

$$\mathcal{S}^{twt}_{RT(t)} = \langle \mathcal{U}^{twt}_{RT}, \mathcal{N}^{twt}_{RT}, \mathcal{T}^{twt}_{RT} \rangle \tag{2}$$

where $\mathcal{U}^{twt}_{RT} \subseteq \mathcal{U}$, $\mathcal{N}^{twt}_{RT} \subseteq |\mathcal{U}^{twt}_{RT}| \times |\mathcal{U}^{twt}_{RT}|$, and $\mathcal{T}^{twt}_{RT}$ is a set of timestamps when $\mathcal{U}^{twt}_{RT}$ have retweeted.

With this RT network, we can design several principal factors for indicating diffusion patterns.

Definition 4 (Coverage). *A coverage ρ is an extent where the target tweet had diffused. It is be computed as*

$$\rho^{twt}_{(t)} = \frac{|\mathcal{U}^{twt}_{RT}|}{|\mathcal{U}|} \tag{3}$$

where $|\mathcal{U}^{twt}_{RT}|$ is the number of users who has retweeted the tweet. By considering the distance, it can be extended to

$$\rho^{twt}_{(t)} = \frac{\sum_{u_i \in \mathcal{U}^{twt}_{RT}} (\eta \times Dist(u_0, u_i)^2)}{|\mathcal{U}|} \tag{4}$$

where $Dist(u_0, u_i)$ is the distance from the target account u_0. Heuristically, η is a weighting parameter for distance.

Definition 5 (Sensitivity). *A sensitivity τ is a response time since the target tweet has been generated. This time delay indicates how quickly users have retweeted. It can be measured by*

$$\tau^{twt}_{(t)} = \frac{\sum_{t^{twt}_i \in \mathcal{T}^{twt}_{RT}} |t^{twt}_i - t^{twt}_0|}{|\mathcal{U}^{twt}_{RT}|} \tag{5}$$

where t^{twt}_0 is the timestamp of the previous tweet (or retweet) directly influencing the retweet actions.

An original tweet from a target user has been retweeted by three users U_A, U_B, and U_C. The time delays of retweets by U_A and U_B are $|t_1 - t_0|$ and $|t_3 - t_0|$, since their distances are 1. On the other hand, the time delay of retweets by U_C is $|t_2 - t_1|$ (not $|t_2 - t_0|$).

Then, these two factors can be merged to compare the diffusion patterns between two arbitrary tweets. A coverage rate indicates how many users a certain tweet is diffused to within a unit time.

Definition 6 (Coverage rate). *A coverage rate ϕ can be measured as*

$$\phi_{(t)}^{twt} = \frac{\kappa \times \rho_{(t)}^{twt}}{(1-\kappa) \times \tau_{(t)}^{twt}} \tag{6}$$

where (t) is a certain timestamp for understanding temporal dynamics. Also, κ is a weighting parameter for emphasizing either coverage (i.e., $0.5 \leq \kappa \leq 1$) or sensitivity (i.e., $0 \leq \kappa < 0.5$).

4 Experiments

In order to justify the proposed indicators for discovering diffusion patterns, we have needed to collect real Twitter dataset. Thereby, a practical system (called TweetScope) has been implemented.

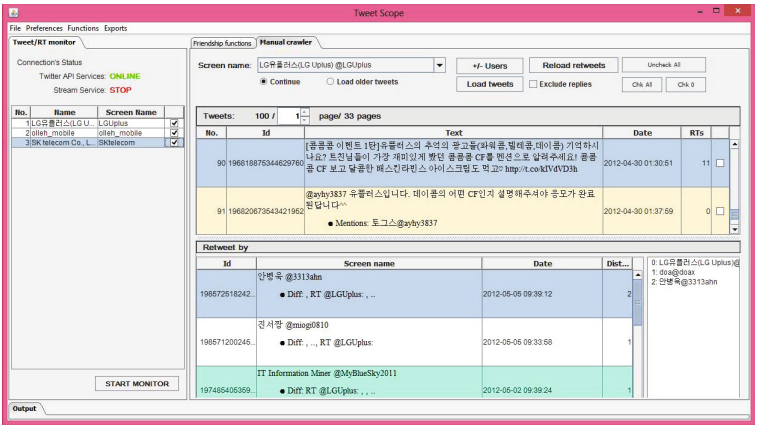


Fig. 2. A snapshot of user interface of TweetScope

4.1 Implementation

TweetScope is based on Twitter API version 1.1, but it does not use all the supported Twitter API functions. We only consider on fetching i) user-timelines, ii) retweets and also iii) friend relationships. Beside, by using Twitter Stream API we have tried to reduce the number of connections to Twitter Service. The collected dataset is stored in a Microsoft SQL Server database which can efficiently support all of the proposed data analytic processes. TweetScope has been implemented by combining JAVA programming language and T-SQL stored procedures as the main techniques for extracting and measuring tweet propagation patterns hidden in the raw data. Fig. 2 shows the user interface of TweetScope. As main features, TweetScope can manage tweet streams from multi-user accounts simultaneously in run-time. It can also allow users to fetch interesting information from Twitter Service such as: user information, relationship, measuring user distance in RT networks by showing the path of information diffusion and exporting the analyzed data to other formats (Excel, CSV, and GraphML).

4.2 Data Collection

We have selected three major telecommunication companies in Korea. All of the tweets from their Twitter accounts (i.e., @SKtelecom, @Olleh_mobile and @LGUplus) have been collected from 16 March 2012 to 30 October 2012. By using TweetScope, we have monitored customers' retweets for building their RT networks. The statistical specification of the collected dataset is shown in Table 1.

Table 1. Statistical specification of the collected dataset

Account	Number of followers	Number of tweets	Number of retweets
SKT (@SKtelecom)	42145	3720	2143
KT (@Olleh_mobile)	44325	3428	7470
LGU+ (@LGUplus)	30250	3213	3981

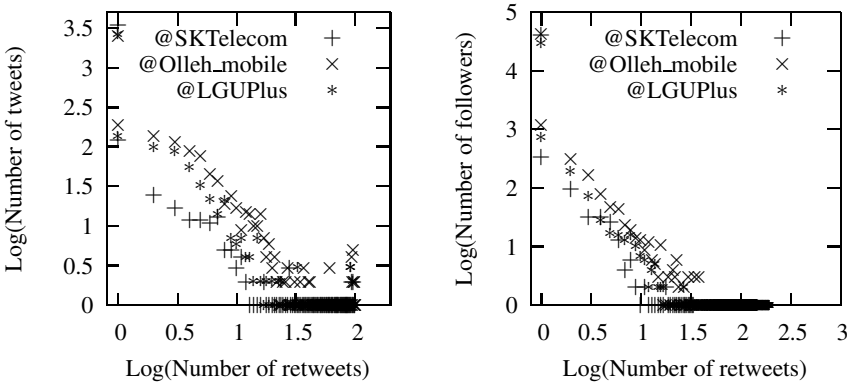


Fig. 3. A log-log plot of (a) the number of tweets and (b) the number of followers w.r.t. the number of retweets

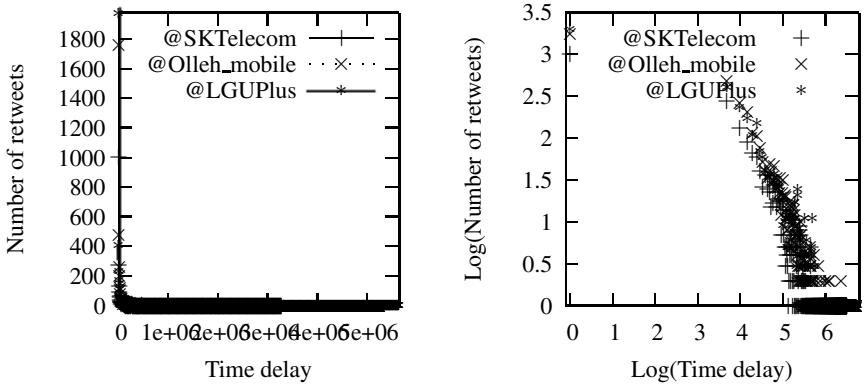


Fig. 4. (a) A plot of to the number of retweets w.r.t. time delay (b) a log-log scale

Additionally, we want to show statistical distribution on the collected datasets. Fig. 3 depicts *power low* distributions of the number of tweets and followers with respect to the number of retweets. It means that i) most of retweets are concentrated on a small set of tweets, and ii) most of retweets are conducted by a small set of followers. More importantly, similar to Fig. 3, Fig. 4 shows a power low distribution of the number of retweets with respect to the time delay (i.e., $|t_i^{twt} - t_0^{twt}|$). It means that most of retweets are intensively concentrated within a short time delay.

5 Evaluation and Discussion

In order to evidently present the diffusion patterns of the tweets, we have selected a set of tweets whose retweet counts are more than 20. (Finally, the numbers of tweets in @Olleh_mobile, @SKTelecom and @LGUplus are 65, 23, and 37, respectively.) Also, weighting parameters have been set as $\kappa = \frac{\mathcal{U}}{1+\mathcal{U}}$ and $\eta = 1$.

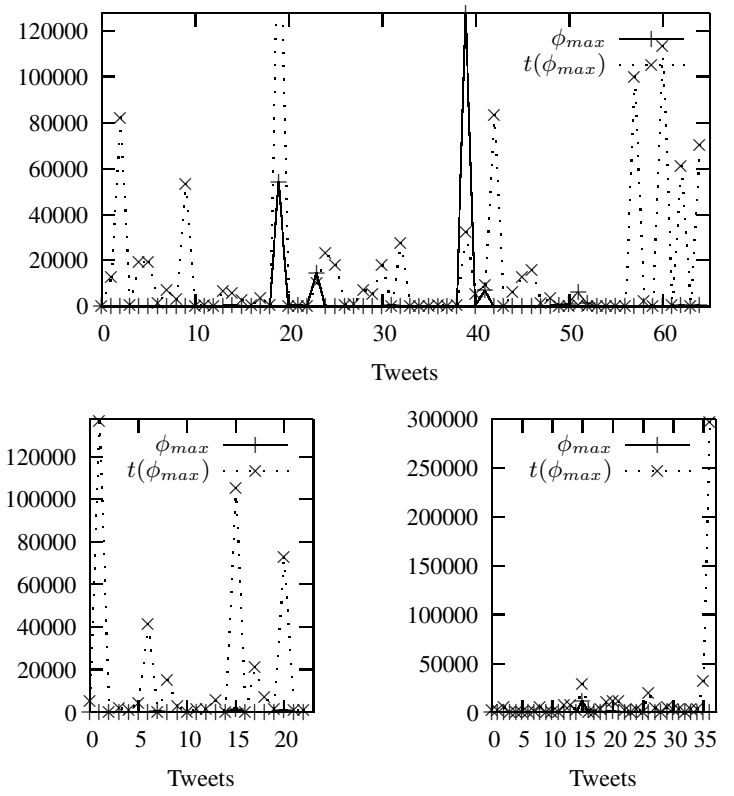


Fig. 5. Two indicators of the diffusion patterns by coverage rate; (a) @Olleh_mobile, (b) @SKTelecom, and (c) @LGUplus

5.1 Diffusion Patterns by Coverage Rate

With the coverage rate ϕ, we have tried to measure the following two kinds of indicators for designing diffusion patterns.

- The maximum value of ϕ_{max}^{twt}
- The time of the maximum value during the time period of retweets $t(\phi_{max}^{twt})$

Fig. 5 depicts two indicators of the diffusion patterns of the selected tweets. Each of these features can segment the tweets into three groups. Table 2 shows the range of . In total, we can classify the tweets into nine diffusion patterns.

5.2 Thematic Identification of Tweets

Once we have analyzed tweet counts (texts) in nine diffusion patterns manually, we have realized that the tweets can be thematically classified into three main topics, which are promotion, informative, and notification. They are denoted as P, I, and N, respectively. Table 3 shows the number of tweets with respect to the main topics.

Table 2. Range of the nine diffusion patterns

Range	ϕ_{max}^{twt}			$t(\phi_{max}^{twt})$		
	Strong (S)	Medium (M)	Weak (W)	Early (E)	Medium (M)	Late (L)
@Olleh_mobile	0.5 to 127613.5			7 to 193485		
	186.8 - 127613.5	53.6 - 175.7	0.5 - 46.7	7057 - 193485	444 - 6623	7 - 290
@SKtelecom	3.5 to 1897.9			81.0 to 136829.0		
	213.1 - 1897.9	82.4 - 205.8	3.5 - 71.6	5755 - 136829	1108 - 5168	81 - 1074
@LGUplus	2.9 to 11115.0			22.0 to 297124.0		
	188.5 - 11115	44.9 - 129.9	2.9 - 40	4370 - 297124	954 - 3978	22 - 602

Table 3. Thematic analysis with respect to the groups of ϕ^{twt}

Groups (ϕ_{max}^{twt}, $t(\phi_{max}^{twt})$)	@Olleh_mobile			@SKTelecom			@LGUPlus		
	P	I	N	P	I	N	P	I	N
G_1 (S, E)	2	7	1	3	1	0	4	0	0
G_2 (S, M)	4	1	0	0	0	1	3	0	0
G_3 (S, L)	5	0	0	0	0	2	4	1	0
G_4 (M, E)	7	1	1	0	0	2	4	0	1
G_5 (M, E)	7	3	2	3	1	1	3	0	1
G_6 (M, E)	2	0	0	1	0	0	2	0	1
G_7 (W, E)	2	1	1	2	0	0	3	0	1
G_8 (W, E)	6	0	0	2	0	0	5	0	0
G_9 (W, E)	9	3	1	2	1	1	3	0	0

6 Concluding Remark

As a conclusion, this work proposes an integrated platform (called TweetScope) to efficiently understand how the information can be spread in the social media (particularly, Twitter). Two main indicators have been designed to represent the diffusion patterns. Consequently, we have found out the following patterns;

- In all of the target account, the most frequently retweeted tweets are for promotions, regardless of $t(\phi_{max}^{twt})$.
- In @Olleh_mobile and @LGUPlus, the tweets for the informative topics have shown the strong ϕ_{max}^{twt}.

In this work, we have considered only the "retweet" network. However, this is a subnetwork of the real relationships among users, so that the diffusion pattern we have obtained in this work might be biased. More seriously, the results collected by Twitter API are practically limited. For example, the maximum number of retweets by the API is 99.

In future work, we have to consider how to exploit the discovered diffusion patterns. We can plan to apply the diffusion patterns to contextual synchronization [6] in social network environment.

Acknowledgement. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No. 2011-0017156).

References

1. Gruen, T.W., Osmonbekov, T., Czaplewski, A.J.: ewom: The impact of customer-to-customer online know-how exchange on customer value and loyalty. Journal of Business Research 59(4), 449–456 (2006)
2. Dey, L., Haque, S.M.: Opinion mining from noisy text data. International Journal on Document Analysis and Recognition 12(3), 205–226 (2009)
3. Jung, J.J.: Discovering community of lingual practice for matching multilingual tags from folksonomies. Computer Journal 55(3), 337–346 (2012)
4. Pham, X.H., Jung, J.J., Hwang, D.: Beating social pulse: Understanding information propagation via online social tagging systems. Journal of Universal Computer Science 18(8), 1022–1031 (2012)
5. Kim, M., Xie, L., Christen, P.: Event diffusion patterns in social media. In: Breslin, J.G., Ellison, N.B., Shanahan, J.G., Tufekci, Z. (eds.) Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, The AAAI Press (2012)
6. Jung, J.J.: Boosting social collaborations based on contextual synchronization: An empirical study. Expert Systems with Applications 38(5), 4809–4815 (2011)

A Practical Method for Compatibility Evaluation of Portable Document Formats

Dariusz Król[1,2] and Michał Łopatka[3]

[1] School of Design, Engineering and Computing, Bournemouth University, UK
[2] Institute of Informatics, Wrocław University of Technology, Poland
Dariusz.Krol@pwr.wroc.pl
[3] Faculty of Computer Science and Management,
Wrocław University of Technology, Poland

Abstract. This paper presents a method for verification of PDF documents for compatibility with publication models provided by scientific publishers. We first consider the problem of converting a document from PDF to XML format. Subsequently, we present an analysis of the document's graphical layout which operates in two phases. The first phase develops a model using a semi-automatic process with limited user interaction. This is followed by comparing and matching of submitted documents. The experimental results demonstrate the degree of document compatibility with the model along with a report of errors and warning messages.

Keywords: document image analysis, document semantic analysis, layout extraction, PDF, XML.

1 Introduction

Evaluating the technical correctness of electronic documents is a common problem of scientific conference organizers and publishers. The main difficulty is actual document verification because this is normally the responsibility of a specific person and thus a very time consuming process. For this reason it is necessary to develop methods allowing an automated evaluation of submitted PDF files [1].

There are several applications available which can assist with this process. However, only a limited number automatically verify submitted documents. Among them is IEEE which uses *PDF eXpress* to assist authors in converting their documents to PDF files that meet IEEE compliance criteria. Among other aspects of compatibility fonts, graphics, margins and document layout are checked. It is, however, a dedicated solution with a limited functionality that cannot be used in a standalone manner.

In this context, the aim of this paper is to present a effective method for analysing the compatibility of submitted documents with a given model. By document compatibility is meant the degree of correspondence between a model and the basic elements of a paper (author and title), their sequence, the formatting (the font name and size, margins) as well as any mandatory and repeatable

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 275–284, 2013.

occurrences. Furthermore, this is extended by taking into account parameters such as page size and total number of pages.

The document evaluation process should indicate the elements in the documents, which, when compared to the model, have inappropriate formatting. It should also indicate why these elements are wrong and how much they differ from the model. Taking this into consideration, a method was developed which merges the conversion of PDF files into XML [2,3] and JPG formats. An XML file generated by *pdf2xml* convertor (`pdf2xml.sourceforge.net`) contains a lot of details about the document such as layout and format characteristics of blocks. However, following this procedure, it is also necessary to correct shortcomings in the layout and content which arise during the first conversion phase [4]. Hence the application of a second conversion method to JPG format is required. To enable this conversion to take place, free software can be used such as ImageMagick (`www.imagemagick.org/script/index.php`). Just then relevant information in the form of blocks with defined positions, formatting and styles is sufficiently organised to enable more in-depth study and further processing.

The method applied to examine document compatibility requires two data sets as input. The first consists of a set of acceptable data assigned from the model such as logic types of blocks and their relationships (title, abstract, etc), block size (min/max), page positioning (min/max), block margins (min/max), fonts and their features, plus finally, whether it is a mandatory or a unique block. The second data set is made up of detected blocks from respective documents, which will be subjected to the compatibility evaluation process. Parameters for every block are compared with data from the model set. By defining tolerance ranges, the process generates a result giving the compatibility level expressed as a percentage. Moreover, identified errors and warnings are listed and include data regarding the page number as well as the specific block, in which they were found.

2 Data-Driven Approach to the Model

Before building the model, it is important to convert the publisher sample (normally as a PDF file), into XML format which is then converted into a JPG image. The XML conversion classifies input data into the following object hierarchy: *page* - a set of blocks with specified width and height, *block* - a set of texts with specified font name and size and *text* - a set of character strings with specified width, height and positioning. Once this action has been executed, the document is converted again, this time into JPG file. The document's pages are saved as black-and-white images. Next, a content analysis of each page is executed. This analysis gives data on blocks which are read from the document's image. In order to identify blocks, a segmentation-like algorithm is used which locates white lines [5,6]. This algorithm uses the following input parameters:

– *colour value* [0,255] defines the threshold applied for image pixels, the lower values indicate the content areas, so called 'black pixels', the higher values indicate irrelevant background, the white areas on the page,

- *white value* - the maximum number of 'black pixels' allowed to accept an acknowledgement of analysed line as white,
- *Y value* - maximum height of a white area, for lower or equal values the line is recognised as a component of an analysed area, for a higher value it is recognised as a new white area on the page,
- *X value* - maximum width of a white area, identical specifications to the *Y value* but regarding width.

The algorithm requires as input a set of pages of the document, which have been saved as JPG images. Initially, the pages are analysed vertically - from top to bottom, then horizontally - from left do right. The generated block set contains allotted page areas with content, i.e. those which are not white. Those are separated by white lines. In this way block definition is completed. The two resulting sets of generated data (the first from XML file analysis, the second from JPG image analysis) are combined into a final logarithm with parameters performed as follows on every block identified:

- *tolerance value* - this parameter determines the tolerance for disparity in positioning of the blocks generated from XML and JPG files,
- *common section value* - this parameter determines the common area for blocks derived from the XML and JPG files indicating the maximum difference in size between them.

Once the last algorithm is executed, a model can be created by applying data set which contains information about individual blocks derived from the pages of the analysed document. The first step in creating the model is supplementing with the semantic data on the logical significance of individual blocks in the document. An interactive web service was developed for this purpose. It was essential to link subsequent blocks with semantic elements and also to define whether they are mandatory or unique. After completing this indexing step, the model is supplemented with the following:

- *block type* - the semantic type of the block (title, abstract etc),
- *mandatory block* - a parameter determining whether a given block has to appear in the analysed document,
- *block uniqueness* - a parameter determining the number of times a given block appears.

Having completed the supplementary data indexing stage, it is necessary in a second stage to merge the block sets into one model. This is carried out by dividing blocks into four groups: unique and mandatory blocks, unique blocks, mandatory blocks and other blocks. The four sets of data constitute the model used subsequently to analyse submitted documents.

3 Analysing Document Compatibility

Applying the model, it is possible to execute a process checking document compatibility in a similar way, but this time it is done fully automatically. A block

classification algorithm is used in the first phase and for this purpose, each document is converted into XML format then into JPG file. The identification of the blocks from the subjected document occurs for each group developed by the model.

The first step of analysing is identifying the mandatory and unique blocks on the first page of the document. To identify these blocks, a similarity measure is applied between blocks. The values obtained in this way are then compared and the lowest value is considered as a candidate for a given block type. The algorithm has been developed to return the specific positioning of the block, which will have the highest similarity value. The algorithm sums up the absolute values of the differences occurring in the sequential parameters defined as X coordinate, Y coordinate, height and width. Identifying blocks which contain no fonts is the second move. In this phase blocks with no fonts are automatically categorized as images. Because of *pdf2xml* limitations such as inability to identify text areas within an image or a picture, the probability that font-less blocks will end up being categorized as images is very high.

The subsequent step consists of the recognition of first level headings. This recognition is based on the fact that such headings always occur on the first page of a paper. The procedure then calculates the similarity value in a manner identical to the one described above with one exception being that the value of the Y coordinate is omitted. It should be noted that the probability value up to which a given block is regarded as being a heading is one of the parameters used to configure the model. The block which contains a level 1 heading is recognized as the model block. It is then used to locate further level headings within the document. When searching for other such type headings within the document, the following data is taken into account: margins, similarity value and font size. Detecting captions is the following step in the method. By captions is meant the text which appears under pictures, images and charts. At this stage, a *tolerance value*) is determined up to which a block can be recognized as a caption.

The next step begins with detecting all the potential images. For every block there is a checking process during which it is inspected to see if the immediately preceding block has already been identified. If it has not been identified, the new block is recognized as an image. The following step consists of identifying so far unclassified text blocks. The difference in the parameters is that a greater emphasis is placed on the value of the X coordinate and the width of a block. This is due to the fact that text most often consists of blocks which are stretched across the entire width of the print area. The final round consists of detecting level 2 headings. Here, an identical procedure is followed to that used to recognize level 1 headings.

The complete recognition process allows for a block set to be created and then compared with the model to verify format correctness. The following values are determined for unique blocks and unique and mandatory blocks: X coordinate, Y coordinate, width, height, top and bottom margin, preceding blocks, subsequent blocks and fonts used in the block. The following values are determined for mandatory blocks and for other blocks: width, height, top and bottom margin,

their preceding blocks, their subsequent blocks and fonts used in the verified block. Apart from checking the formatting of the blocks, the overall parameters of the document such as page size, page margin width and number of pages are also inspected. *Tolerance values* are applied on the inspected values which define whether exceeding a given value will be flagged as an error or as a warning. Obviously, this will have an impact on the results when compatibility with model is determined.

The model against which submitted papers are validated consists of five elements: general data about the document, data on unique and mandatory blocks, data on unique blocks, data on mandatory blocks, data on other blocks. Because document data is divided into five parts, the compatibility evaluation begins with distributing the maximum compatibility values of these elements and a 20% maximum compatibility value is allocated for each part of the document. This value is then further distributed among respective components. For instance, in the case of general data, a 2.5% possible value is applied to every component. This is because the following components are compared: top margin on the first page, top margins on the remaining pages, bottom margin, left-hand margin, right-hand margin, page width and height, and number of pages.

The *tolerance value* determines whether a given result is registered as an error or as a warning. It determines the accepted deviation from the correct value. If the parameters fall within the extreme values of the above parameters, a given component is considered correct. Otherwise the following is deducted from the 20% final value: 2.5% in the case of an error and 1.25% in the case of a warning. The compatibility of remaining model components is calculated in an analogous way. For those blocks, which are permitted to occur frequently this value is evenly distributed after which it is calculated for each block and then summed up to generate the actual compatibility value.

4 Implementation Details

PDF documents are converted using the *pdf2xml* tool, which in the design is triggered by the *BASH* shell. Input is a PDF document whereas the converted document is saved as an XML file in a working folder. The *pdf2xml* tool does not convert the file completely because the programme is only capable of recognising text objects within the PDF file. Therefore, when converting a graphic image, the *ImageMagick convert* tool is used with the -type Grayscale switch. Because of this, converted images are saved as grayscale images. Converting to grayscale makes image analysis much simpler by enabling detection of white areas on a page. The result of the conversion is a list of JPG images equalling the number of pages in the document.

To have a better understanding of the difficulties involved in converting a PDF file to an XML file it is worth reviewing the results of using *pdf2xml* tool. The application divides the text in the document into blocks, which contain variable font parameters. Also, the method in which font families are saved should also be noted. Their names have prefixes and, what can be described as a draft name.

When looking closely at config_fonts.xml, one can notice that the fonts are saved as follows:

```
<font>Courier New,Italic</font>
<font>Courier New,Bold Italic</font>
<font>Times New Roman,Bold</font>
```

The first value is the name of the font, after the comma acceptable font-style values are shown as either bold or italics. The module responsible for categorizing font names from the XML file is a font classifier. It is a simple module, which matches the fonts from the XML file with the config_fonts.xml fonts. The matching algorithm exploits the Jaccard coefficient for trigrams obtained from the font names. In order to detect a given text block in the document, a method is used that identifies white lines. Initially, pages are analysed vertically (from top to bottom), and then for the blocks identified in a horizontal direction (from left to right). The algorithm returns satisfactory results. Usually, blocks detected in JPG image and in XML file overlap. However, this is not always the case.

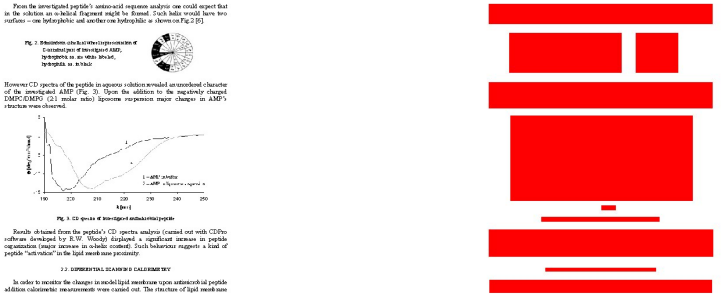


Fig. 1. An example of block identification

In Fig. 1 the algorithm mistakenly recognised the caption below one of the axis as a new block while deficiencies were observed in places with graphic elements when compared to the visualisation of blocks obtained from the XML conversion. This is due to the fact that the application is not able to convert images. The result of the document analysis was a set of XML files containing data about the blocks, appearing on the pages of the document. They were obtained through combining the blocks from both JPG files as well as XML files. Every page of the document was depicted by a separate XML file with the following excerpt of structure:

```
<block>
    <x>289</x>
    <y>197</y>
    <width>191</width>
    <height>30</height>
        <font>
            <family>Times New Roman</family>
            <size>9</size>
            <bold>false</bold>
        </font>
    </block>
```

The resulting XML file contains data on blocks on a given page. Every block is described by a set of attributes and every page of the converted document corresponds to one XML file.

5 Experiments and Results

Two sets of documents were used to evaluate the proposed method. The first consisted of a set of correct documents used to create the model (further referred to as *model set*). The second consisted of documents requiring verification as though they were papers submitted for a conference (further referred to as *test set*). The model set consisted of ten correctly formatted papers in PDF format. Those are taken from the annual Students' Science Conference held by the Wrocław University of Technology. The test set consisted of 50 randomly picked papers from among the papers which had been submitted for the conference. The first step was to determine blocks for the documents. The following were identified: level 1 heading, level 2 heading, captions, images, text, title, abstract, affiliation, key words, author, references, synopsis. The second step was to find out about document compatibility. 8% of the submitted documents were identified as *most compatible* and 14% as *least compatible*. Taking a closer look at the reasons for the very high disparity range identified by these percentages (extremes between 31.77% and 87.99%) confirmed that the method had correctly allocated a low rating to some of the documents. The lowest rated paper was a document, with the following compatibility values: conformity 31.77%, errors 385, warnings 7. The most important of these were: unclassified blocks as keywords, abstract, affiliation, author; incorrect classification of the actual block keywords as title; incorrectly identifying references and summary; incorrectly classifying blocks as image, heading 1 and heading 2. When analysing this document it was clear that it's formatting was incompatible with the model set.

The differences were not only in the location of the individual blocks, they are also present in the attributes used to format them. Due to too wide sentence spacing when compared with the model set, the validation method had divided a cohesive text into a number of small blocks. Analysing the results one could notice that the validation method had correctly identified the following errors:

```
- Page number
ERROR [1] Number of pages:7, expected:[6,6]
- Margins on given pages
ERROR [2] Page [1] right margin:92, expected:[111,114]
ERROR [3] Page [3] right margin:98, expected:[111,114]
ERROR [4] Page [3] left margin:85, expected:[98,99]
- Block size and location
ERROR [5] Block [0,1] author: Width:225, expected:[94,198]
ERROR [6] Block [0,2] title: Height:45, expected:[24,43]
WARNING [1] Block [0,3] abstract: X:112, expected:[113,113]
```

Due to the fact that the summary block was absent from the paper, the validation method incorrectly classified the blocks on the last page of the paper. An actual text was recognised as references, while the references were recognised as summary. The highest rated paper had the following results: conformity 87.99% errors 25, warnings 5. The results would have been higher if it were not for the fact that some blocks have been incorrectly recognised. The process incorrectly identified the following blocks: references and summary. At times it has also happened that the process falsely recognised some blocks as image. Also, mistaking caption for heading 2 was a common occurrence.

Further tests were conducted in order to verify the correctness of the validation method. The first test was to define the error value caused by incorrect block identification. The compatibility test was carried out on a set of model documents. The average compatibility result was 77.88% where the highest was 93.99% and the lowest was 62.60%. The most irregularities were found in two block types: images and captions. Taking this result into account, the margin of error for the validation method is 22.12%. This value represents the average compatibility value (CV). Based on this, it is possible to create the following compatibility scale: $CV \geq 77\%$ - the document is fully compatible and $77\% > CV > 62\%$ - the document is probably compatible. A subsequent test was designed in order to further evaluate the correctness of the validation method. One document was taken from the model set and a few of its elements were modified. Initially, this document had a 93.34% compatibility rate. After modification, the following results for every round were generated:

- amending the typeface and font style for the keywords block: 92.30%,
- missing author and keywords blocks: 81.63%,
- missing only keywords block: 86.71%,
- using Arial font for the whole document: 62.85%,
- more than 12 changes in font style and font size interspersed throughout the document: 77.91%.

In the case of applying the model set, all the modified documents would have been recognised as either compatible or probably compatible with the model set.

The last test was to assess the degree to which unique semantic elements appearing on the title page were incorrectly classified. In order to generate a model set, a test was conducted on those documents, which had a compatibility

result > 62%. To evaluate classification correctness the following parameters were taken into account:

- true positive (TP) - the correctly classified semantic element is located in the examined document, has been recognised by the validation method and it has also been identified;
- true negative (TN) - the correctly unclassified (not found) semantic element in the examined document has not been recognised by the validation method and it has also not been identified;
- false positive (FP) - the incorrectly classified semantic element is located in the examined document, has been recognised by the validation method but it has not been identified;
- false negative (FN) - the incorrectly unclassified (not found) semantic element in the examined document has not been recognised by the validation method but has been identified.

Individual values for the elements are illustrated in Tab. 1 by percentages and additionally by F-score. F-score is defined as the harmonic mean of precision $P = \frac{TP}{TP+FP}$ and recall $R = \frac{TP}{TP+FN}$. After analysing the results the most recognised unique elements were *keywords* and *author*, while *affiliation* had the worst values.

Table 1. The experimental results for the elements into percentages and F-score

Elements	TP	TN	FP	FN	F-score
keywords	95.65	4.35	0	0	1
author	95.65	0	4.35	0	0.98
title	86.96	0	0	13.04	0.93
abstract	91.30	0	8.7	0	0.96
affiliation	73.92	13.04	0	13.04	0.92

6 Conclusion

It is generally well known that model-based identification does not give a 100% accuracy rate. This is because of similarities between specific blocks and the impossibility to conduct a complete analysis of block content. However, when evaluating the validation method, it can be stated that final results were satisfactory. The list of errors and warnings produced allowed identification and correction of the elements, which had been wrongly formatted and used. In most cases, detailed data was obtained regarding the paper's first page. Specifically with regard to title, keywords, authors, abstract or affiliations. First level headings are also correctly identified. The heading on the first page are becoming a model for recognising other headings. The developed method does, however, have its drawbacks. Its weakest point is the automatic type identification of blocks. The

problem is not so noticeable when classifying mandatory and unique blocks on the first page, because both their size and their location differ from the other elements. Thus, their *recognition rate* is high. The score for the classifying blocks on the last page of the document is however very clearly much lower. This is due to the similarity between the following blocks: *text*, *image*, *references* and *summary*.

To sum up, in order to receive better results, there are possibilities to create a recursive version of the algorithm which better detects white lines and blocks in the document. Following this path, it could be tempting to analyse blocks for table occurrence based on the fact that tables are framed and have regular intervals between the white areas [7]. When analysing the converted XML file, it is noticeable that some words and phrases are incorrectly read from the PDF file. It may be possible to solve this by analysing letter locations and sizes. Debugging XML file content may also enable better extraction and in this way would bring a new level of quality with regard to the applicability of using machine learning [8]. Potential applications of this method are not limited solely to compatibility evaluation of scientific papers, it could also provide a very useful tool for analysis of the enormous electronic archives of legacy PDF documents.

References

1. Beel, J., Gipp, B., Shaker, A., Friedrich, N.: SciPlore Xtract: Extracting Titles from Scientific PDF Documents by Analyzing Style Information (Font Size). In: Lalmas, M., Jose, J., Rauber, A., Sebastiani, F., Frommholz, I. (eds.) ECDL 2010. LNCS, vol. 6273, pp. 413–416. Springer, Heidelberg (2010)
2. Hardy, M., Brailsford, D.: Mapping and displaying structural transformations between XML and PDF. In: ACM Symposium on Document Engineering, pp. 95–102. ACM, New York (2002)
3. Déjean, H., Meunier, J.: A System for Converting PDF Documents into Structured XML Format. In: Bunke, H., Spitz, A.L. (eds.) DAS 2006. LNCS, vol. 3872, pp. 129–140. Springer, Heidelberg (2006)
4. Cesarini, F., Gori, M., Marinai, S., Soda, G.: Structured document segmentation and representation by the modified X-Y tree. In: Int. Conf. on Document Analysis and Recognition, pp. 563–566. IEEE Press, New York (1999)
5. Liu, Y., Bai, K., Mitra, F., Giles, C.L.: Improving the Table Boundary Detection in PDFs by Fixing the Sequence Error of the Sparse Lines. In: Int. Conf. on Document Analysis and Recognition, pp. 1006–1010. IEEE Computer Society, New York (2009)
6. Oro, E., Ruffolo, M.: PDF-TREX: An Approach for Recognizing and Extracting Tables from PDF Documents. In: Int. Conf. on Document Analysis and Recognition, pp. 906–910. IEEE Computer Society, New York (2009)
7. Yildiz, B., Kaiser, K., Miksch, S.: pdf2table: A method to extract table information from PDF files. In: Indian Int. Conf. on AI, IICAI, pp. 1773–1785 (2005)
8. Hu, Y., Li, H., Cao, Y., Teng, L., Meyerzon, D., Zheng, Q.: Automatic extraction of titles from general documents using machine learning. Information Processing and Management 42, 1276–1293 (2006)

Sentiment Analysis for Tracking Breaking Events: A Case Study on Twitter

Dongjin Choi and Pankoo Kim*

Department of Computer Engineering Chosun University
375 Seoseok-dong, Dong-gu, Gwangju, Republic of Korea
Dongjin.Choi84@gmail.com, pkkim@chosun.ac.kr

Abstract. Social media such as Twitter and Facebook can be considered as a new media different from the typical media group. The information on social media spread much faster than any other traditional news media due to the fact that people can upload information with no constrain to time or location. People also express their emotional status to let others know what they feel about information. For this reason many studies have been testing social media data to uncover hidden information under textual sentences. Analyzing social media is not simple due to the huge volume and variety of data. Many researches dealt with limited domain area to overcome the size issue. This study focuses on how the flow of sentiments and frequency of tweets are changed from November to December in 2009. We analyzed 110 million tweets collected by Stanford University and LIWC (Linguistic Inquiry Word Count) for sentiment analysis. We did find that people were not happy in afternoon but they were happy in night time as many psychologists suggested before. After analyzing large volume of tweets, we were also able to find the precise day when breaking events occurred. This study offer diverse evidence to prove that Twitter has valuable information for tracking breaking news over the world.

Keywords: Social network services, Sentiment analysis, Twitter.

1 Introduction

Twitter and Facebook are one of the most popular social network services (SNS) around world. SNS is an online platform to provide social networks or social relations among people in order to share interests, activities, information, and etc. Due to the dramatic development of wireless internet infrastructure and Smartphone devices, people can find and share information with no constrain to time or location. This fact changes entire online system we had experienced over the year. We no longer have to go back to home or internet cafe to search information or upload photos. We can simply find and share information using Smartphone via SNS within few seconds. There was an event that made Twitter very popular after U.S. Air ways jet crashes into the Hudson River on 15th of January 2009. The first photograph of the accident

* Corresponding author.

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 285–294, 2013.

appeared on Twitter before any even local news media arrived the place. Moreover, there was another well-known event in Iran that many people on the street tweeted about what was happening on the street when all the major news media in Iran were not reporting what had been happened. These events bring an aspect that Twitter is not just social communities but it is social media. Because of the high popularity for Twitter, people spread diverse information with their personal emotional states concerning the earthquake, terrorism, weather, celebrities, and more when breaking events had happened. For this reason, the volume of Twitter data has been increasing dramatically with diverse information. The total number of tweets was approximately 29 billion counted by gigatweet[1] on 6th of November 2010. Many researchers have the same idea that SNS is the crude coal which has to be turned into diamonds. There were lots of researches have been studying for analyzing Twitter in order to find who affect others, what issues people were excited using small amount of data in limited domains. However, our study dealt with entire tweets collected from November to December in 2009 by Stanford University [1] in order to detect what breaking events are by focusing on how the flow of sentiments and frequency of tweets are changed. We analyzed 110 million tweets by each date and each hour using LIWC (Linguistic Inquiry Word Count)[2] which is a program designed by James W. Pennebaker *et al*. to calculate the degree of 70 categories including positive or negative emotions. We did not find any meaningful statistic and emotional flow when we analyzed by each date. However, we did find a significant fact that people were not happy in working time but they were happy in night-time as many psychologists suggested before when we analyzed tweets by each hour. We were also able to find the precise date when breaking events occurred.

The reminder of the paper is organized as follows: Section 2 describes related works and what LIWC is; Section 3 explains the flow of sentiments and frequency of tweets using LIWC and how we conducted experiments; Section 4 gives example for tracking breaking-events using Twitter; and finally Section 5 presents a conclusion to this work and makes suggestions for the future work.

2 Related Works

Sentiment analysis or opinion mining are the sub-area of Natural language processing and text analytics to determine the attitude of speakers or writers to reveal what other people think [2]. It has been applied to online review systems, personal blogs, or surveys for analyzing huge amount of data, automatically. People are starting to examine with Twitter data because of its great possibilities. Users on Twitter are likely to express their personal feelings with no constrains. So, there are full of emotional texts related to various topics over the Twitter. There was a famous research to study whether Twitter can be considered as a news media or not by analyzing entire data collected from June to September in 2009 which was approximately 106 million tweets [3]. This study found reliable evidences to prove

[1] http://gigatweeter.com/
[2] http://www.liwc.net/

that twitter is a news media due to the fact that any retweeted tweets spread broadly no matter what the number of followers of the original tweet is. One of Yahoo research teams studied long-running structure-rich events such as football rather than one-shot events or headline news such as earthquakes [4]. However, detecting and tracking the headline events on Twitter are still receiving attentions from many scientists. In order to detect breaking news on Twitter, many researchers have been based on the word frequency and sentiment degrees in texts [5, 6]. The first research focused on proper nouns frequency rather than sentiments analysis using LIWC which was conducted in the second research. LIWC is a powerful program to calculate the degrees of more than 70 categories including positive or negative emotions in given sentences [7]. The following table 1 shows an example of psychological categories for affective processes.

Table 1. Psychological texts processing categories in LIWC

Psychological Processes	Grand Means
Affective processes	4.41
Positive emotion	2.74
Negative emotion	1.63
Anxiety	0.33
Anger	0.47
Sadness	0.37
…	…

There are four main categories available to determine how the sentiments degrees are in designated text files using LIWC. The first category is a linguistic process which includes word count, total pronouns, auxiliary verbs, past tense, and so on. This process is a step to calculate linguistic structure of given texts based on predefined word dictionary. The second category is a psychological process which is the most important category in LIWC. This process is able to estimate social ratio, positive, negative emotion, insight, causation, discrepancy, tentative, and more by comparing with psychological word dictionary. For example, if the given text includes positive words such as “love,” “nice,” “sweet,” and etc. rather than negative words such as “hurt,” “ugly,” “nasty,” and more, the degree of positive emotion for the given text goes higher than 2.74 which is the grand means of positive emotion. The other processes are personal concerns and spoken categories refer to the “the development and psychometric properties of LIWC 2007.”

Because of the fact that LIWC can easily calculate positive or negative emotions in given texts, it was applied to capture real-time moods of users in the Twitter in order to find who has depressive mood [8]. It was also applied to analyze public opinion and mood for election [9]. LIWC was proved its performance by many researches but still has a limitation that it fails to calculate degrees for given texts due to technological issues. It returns zero to all of degrees when LIWC failed. Besides, the degree of positive of negative emotion goes too high when test sentences are short. In this paper, we conduct experiment for analyzing sentiments flow from November to December in 2009 using LIWC but we will ignore if tweets are less than 100 when we analyze tweets by each hour.

3 Sentiment Analysis

This section describes Twitter sentiment analysis from November to December in 2009 using LIWC. The data which was collected by Stanford University consists of 110 million tweets concerning diverse area information. At first, we simply calculate the number of tweets by each date and sentiments scores by each date as shown in figure 2. We only focused on the affective processes which were shown in table 1. It is clear to see that there is no meaningful events can be found when we looked into the flow. However, we did find the issue which was not expected that the positive and negative graphs just follow the graph of the frequency of tweets. In other words, the emotional degree of LIWC is somehow affected by the number of test sentences. It can be concluded that there is higher possibility to have many positive words if test sentences are longer than others. We hereby, give great attention to analyzing tweets by each hour. We separated entire test tweets by each hour to see what will be happened. The process of our sentiment analysis will be followed by the sentiment analysis process described in figure 1.

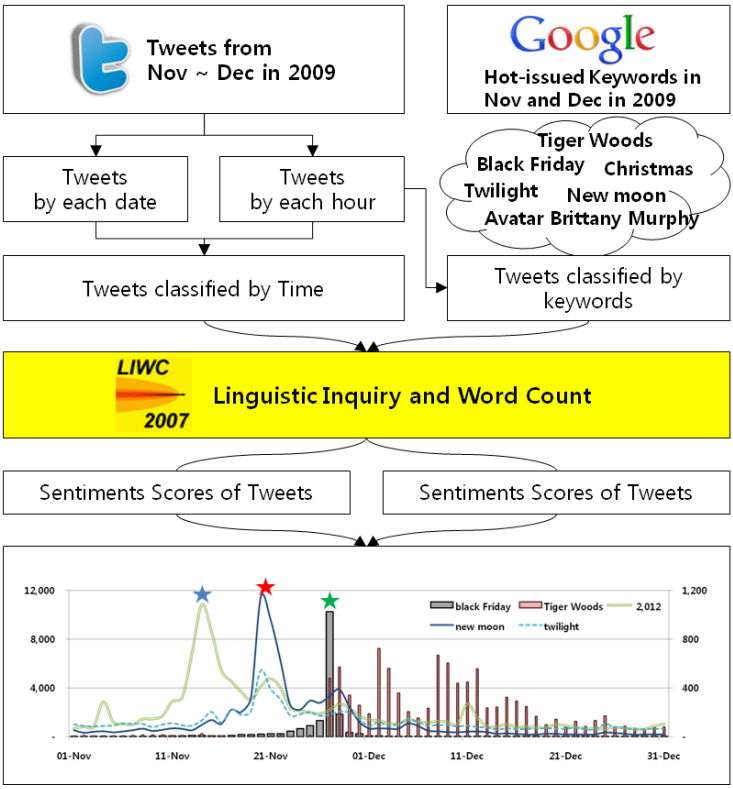


Fig. 1. The process of sentiment analysis of Twitter

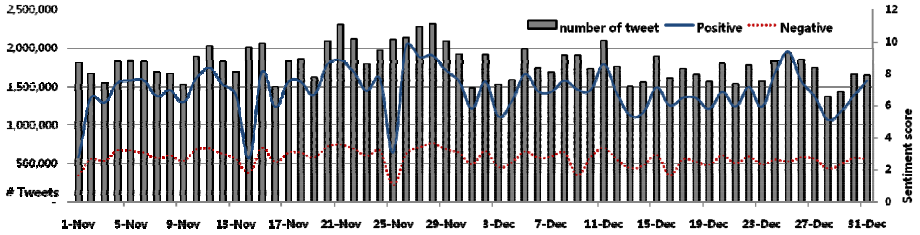


Fig. 2. Tweets and their sentiment flow per day from November to December in 2009

We conducted the same sentiments analysis as tested in figure 2. Afterward, we calculated the average value of tweets, positive sores, and negative scores by each hour. The next figure 3 and 4 show the results for analyzing tweets by each hour during November and December in 2009. When we analyzed sentiments in tweets by each date, the results somehow correlated with the number of tweets. However, we finally obtained hidden patterns under each hour that people are happy in the morning and night-time but not in working time.

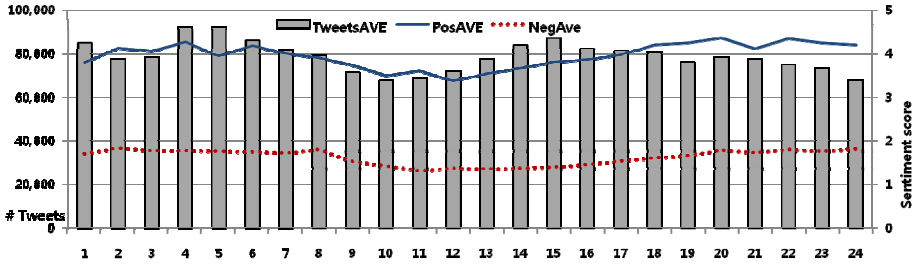


Fig. 3. The average tweets and their sentiment flow per hour during November in 2009

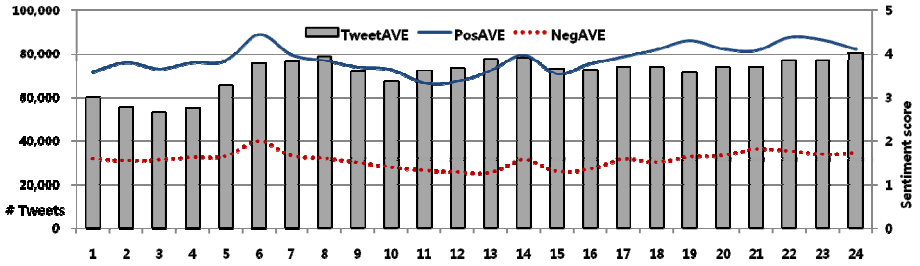


Fig. 4. The average tweets and their sentiment flow per hour during December in 2009

According to what we understanding from the figure 3 and 4, people are normally express positive emotion than negative emotion considering that the mean of positive and negative categories are 2.74 and 1.63 respectively. These results may be based on the reason that people text to others “good morning” and “good night” in that time.

4 Sentiment Analysis for Tracking Breaking Events

This section describes basic experiments whether Twitter data can be applied to detect breaking events or not based on sentiment analysis using LIWC. The hot issued keywords provided by Google trends[3] on November in 2009 were "Black Friday," "Tiger Woods," "modern warface 2," "watch-movies.net," "2012," "e okul," "new moon," "toys r us," "twilight," and "best buy" ordered by a rank. And the hot issued keywords on December in 2009 were "Brittany Murphy," "Tiger Woods,' "avatar," "e okul," "Christmas," "meteo," "farmville," "weather," "weather forecast," and "wii" also ordered by a rank. We want to explain meanings that theses keywords have at first in order to make understand this research more easily. "Black Friday" is the Thanksgiving Day in United States often regarded as the beginning of the Christmas shopping season. Due to the big discount on this day, the number one hot issued keyword was "Black Friday" and the precise date in 2009 was 27th of November. There were another shopping related keywords on November correlated with Black Friday. The keywords were "toys r us" and "best buy." People gave queries into Google to find and share information with others related to which products were worth to buy on Black Friday. The next keyword was "Tiger Woods" who is the most successful professional golfer in United States. The reason why his name came to hot issued keywords was the fact that he had a car accident on 27th of November. The movie related keywords "2012" and "New moon" also appeared on November because those movies were released to public on November. Another keyword we have to give big attention is "Brittany Murphy" who was an American actress and singer. She was passed away on 20th of December. Moreover, keywords "Christmas" and "wii" are similar keywords to "Black Friday," "toys r us," and "best buy." "Wii" is a home video game console released by Nintendo and it was ranked on the top in survey to children that which things you most want to get on Christmas.

We hereby start sentiment analysis to tracking these interesting keywords by classifying 110 million tweets into corresponding hot issued keywords on November and December based on simple text matching method.

As we can see in the figure 5 that the tweets frequency flow has significant meanings that we have to study. The tweet frequency of Black Friday increased suddenly on 27th of November which was the precise day of Black Friday in 2009. However, the number of tweets for Black Friday dropped just after one day. This is one of the one-shot events which can be decided that this type of event only got famous in few days. The other frequency of tweets related to keywords "2012," "new moon," and "twilight" increased quickly on some days. Those days were the day these movies were released to public in the United States which was 12th and 20th of November. However, in the case for "Tiger Woods" has three peaks which was not easy to determine which date was something happened to him. Therefore, the sentiments analysis for tweets corresponding to "Tiger Woods" is necessary in order to trace what types of events occurred to him and when was the precise date of the event.

[3] http://www.google.com/trends

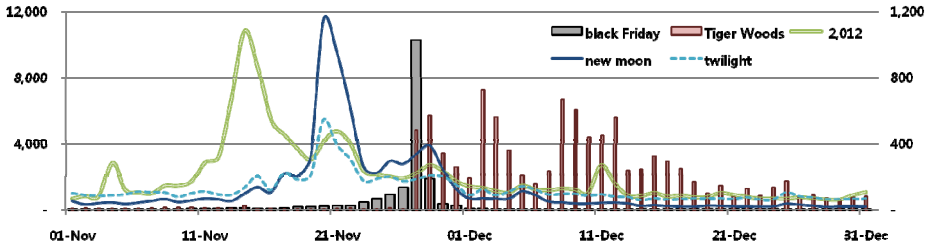


Fig. 5. The hot-issue of November tweets frequency flow per day during November and December in 2009

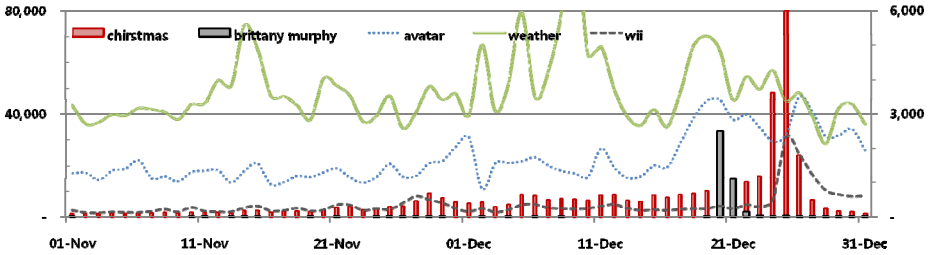
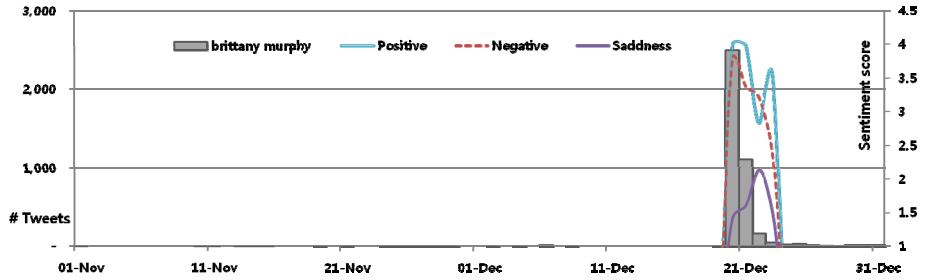
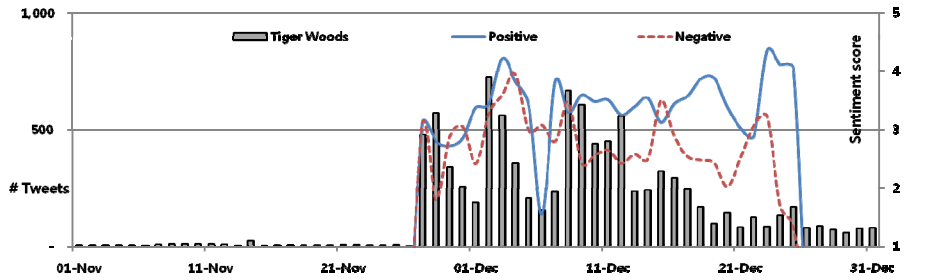


Fig. 6. The hot-issue of December tweets frequency flow per day during November and December in 2009



(a) Sentiment analysis graph for "Brittany Murphy"

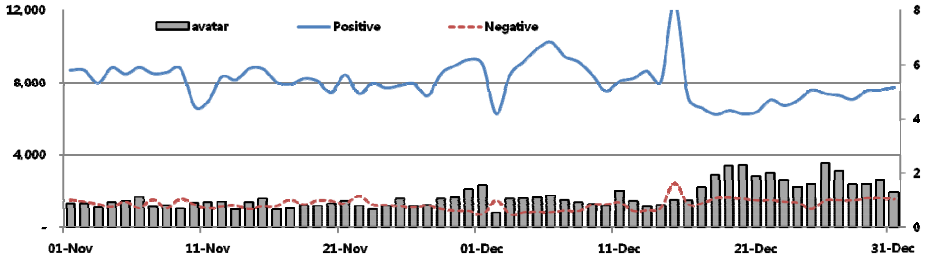


(b) Sentiment analysis graph for "Tiger Woods"

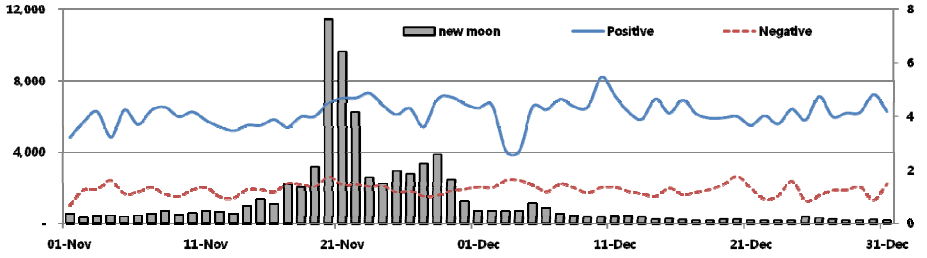
Fig. 7. The hot-issued tweets frequency and sentiments flow per a day during November and December in 2009

Figure 6 indicates that there was an event related to Brittany Murphy on 20th of December. Also, the frequencies of tweets have 'wii' increased suddenly on Christmas. However, it was not easy to estimate precise date for 'avatar' and 'weather' which is commonly interested by many others. We hereby analyze sentiment scores of hot-issued tweets by each date using LIWC. The following figure 7 shows the results that the positive and negative emotion for "Brittany Murphy" and "Tiger Woods" increased higher than grand means on specific day. Because of the reason that Tiger Woods got a car accident on 27th of November and Brittany Murphy was passed away on 20th of December. Interesting thing is that Brittany Murphy just only had been tweeted for less than three days. This might be depending on how celebrities are popular to publics.

There is another example that related to movies. People tweeted about how movies are to share information. The following figure 8 is the results for "New Moon" and "Avatar" which was released on 20th of November and 17th of December in 200.



(a) Sentiment analysis graph for the movie "Avatar"



(b) Sentiment analysis graph for the movie "New Moon"

Fig. 8. The tweets frequency and their sentiments flow per a day during November and December in 2009

As described in figure 8, it is possible to determine that the movie "avatar" got more positive degree than "New Moon." The movie "New moon" was interested to the people in few days based on the figure (b) in 8. Before it was released to the public, the number of tweets was less than 1,000. After 20th of November, the number of tweets was increased to over 10,000 but it was decreased after ten days to less than 500. The positive degree during these days was not low but there was no

significantly interesting peak on the flow. On the contrary, the movie "Avatar" had been tweeted more than 1,000 before the movie was released to public. The degree of positive emotion was higher than "New moon." Moreover, the number of tweets after 17th of December was still higher than 2,000. It can be considered that the "Avatar" got more satisfactions from people. We have concluded that if we analyze sentiments and frequency of tweets for given movies, we can possibly obtain information whether the movies were pleased by people or not. Considering that the Twitter was not famous in 2009 comparing to these days, if we analyze sentiment score for current tweets for the movies, the results will be more attractive than this. The problem is that the size of tweets is so huge that it will take many costs to conduct the experiment.

5 Conclusion and Future works

This study conducted sentiments analysis in case of Twitter using LIWC to prove that it is possible to use online social network data for tracking when breaking-events were occurred and how the events were satisfied by people. We tested 110 million tweets collected by Stanford University from November to December in 2009. We did not find any meaningful information when we analyzed sentiments in tweets by each date. However, we did find when we looked into tweets by each hour that there is a tendency that people express positive emotions in night-time rather than working-time. This might because of greeting to each other in the morning, evening and before go to bed. When we only focused on tweets related to hot-keywords on November to December in 2009 provided by Google rather than entire 110 million tweets, it is possible to trace precise date when the events were occurred by analyzing the number of tweets and their sentiments scores. People express not only positive but also negative feelings when a tragedy related to celebrities was happened. Moreover, it is also possible to trace how people were considering about specific movies. This study offer diverse evidence to prove that Twitter has valuable information for tracking breaking news over the world. However, it is still challenging task that the size of current Twitter is so huge to conduct this type of experiments. This is why many researchers studied Twitter in limited domain.

In the nearest future, we plane to analyze negative tweets in detail in order to find who is isolated, why, and how to overcome. Many researchers focused on the study to find who effect to whom and how these people famous than others. However, it has come to our attention that there are many unknown users who were isolated and were under suffering psychological problems. These people are likely to express negative emotion than positive, theoretically. Our future goal will be aimed to find isolated online group to help them up and make them to combine with normal user group.

Acknowledgments. This research was financially supported by the Ministry of Education, Science Technology (MEST) and National Research Foundation of Korea (NRF) through the Human Resource Training Project for Regional Innovation.

References

1. Yang, J., Leskovec, J.: Patterns of Temporal Variation in Online Media. In: ACM International Conference on Web Search and Data Mining, pp. 177–186 (2011)
2. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2008)
3. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a Social Network or a News Media. In: 19th International Conference on World Wide Web, pp. 591–600 (2010)
4. Chakrabarti, D., Punera, K.: Event Summarization using Tweets. In: 5th AAAI Conference on Weblogs and Social Media, pp. 66–73 (2011)
5. Phuvipadawat, S., Murata, T.: Breaking News Detection and Tracking in Twitter. Web Intelligence and Intelligent Agent Technology 3, 120–123 (2010)
6. Park, J., Cha, M., Kim, H., Jeong, J.: Managing Bad News in Social Media: A Case Study on Domino's Pizza Crisis. In: 6th International AAAI Conference on Weblogs and Social Media, pp. 282–289 (2012)
7. Pennebaker, J.W., Booth, R.J., Francis, M.E.: Linguistic Inquiry and Word Count, LIWC (2007)
8. Park, M., Cha, C., Cha, M.: Depressive Moods of Users Portrayed in Twitter. In: Proc. of the ACM SIGKDD Workshop on Healthcare Informatics, HI-KDD (2012)
9. Elson, S.B., Yeung, D., Roshan, P., Bohandy, S.R., Nader, A.: Using Social Media to Gauge Iranian Public Opinion and Mood after the 2009 Election (2012)

Classification of Plantar Dermatoglyphic Patterns for the Diagnosis of Down's Syndrome

Hubert Wojtowicz[1] and Wieslaw Wajs[2]

[1] University of Rzeszow, Faculty of Mathematics and Nature, Institute of Computer Science, Rzeszow, Poland
[2] AGH University of Science and Technology, Faculty of Electrical Engineering, Institute of Automatics, Cracow, Poland

Abstract. Classification of patterns of the hallucal area of sole is one of the tasks of dermatoglyphic analysis. The paper describes pattern recognition and image processing methods applied to the problem of the hallucal area of sole patterns classification. Contrast enhancement, segmentation and contextual filtration techniques are used to enhance quality of the images. Application of an algorithm based on multi-scale pyramid decomposition of an image is proposed for ridge orientation calculation. Hallucal area pattern classifiers, which are part of an automatic system for rapid screen diagnosing of trisomy 21 (Down's Syndrome) in infants, are created and discussed. The system is a tool supporting medical decision by automatic processing of dermatoglyphic prints and detecting features indicating presence of genetic disorder. Images of dermatoglyphic prints are pre-processed before the classification stage to extract features analyzed by Support Vector Machines algorithm. RBF kernel type is used in the training of SVM multi-class systems generated with one-vs-one scheme. Experiments conducted on the database of Collegium Medicum of the Jagiellonian University in Cracow show effectiveness of the proposed approach in classification of infants' dermatoglyphs.

1 Introduction

It is often difficult to reach a conclusive diagnosis, on the basis of clinical evidence alone, as to whether a child has Down's syndrome. This diagnosis is even more difficult in the case of a newborn infant. The study of the complex patterns of parallel ridges and furrows found on the digits, palms and soles of infants, termed dermatoglyphics, helps in establishing the diagnosis of Down's syndrome. Association of unusual dermatoglyphics with Down's syndrome was first reported by Cummins [2]. The first test for the diagnosis of Down's syndrome was developed by Walker and later extended and improved by Reed [5] and others. Since then dermatoglyphics have been a valuable aid when clinical diagnosis was in doubt. Association of unusual dermatoglyphics with Down's syndrome also led investigators to examine the dermatoglyphic traits occurring in other cases of trisomy. Other syndromes have been found with associated unusual dermatoglyphics among them Turner's [4] and Klinefelter's [7] syndromes. Several

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 295–304, 2013.

dermatoglyphic traits discriminate the cases of suspected Down's syndrome from normal infants. The most characteristic traits for Down's syndrome have been determined using statistical analysis. These traits are used in the dermatogram diagnostic index developed by Reed [6]. Detection of Down's syndrome using this index relies on examination of the epidermal ridges on the digits, palms and soles of infants by the naked-eye of an anthropologist.

2 The Aim of the Work

Medical imaging found a permanent place in medical practice as a valuable tool in supporting diagnostic decision-making processes for many diseases. The problem of processing large data sets, containing collections of images related to the particular diseases, is an important and interesting issue from the scientific point of view. With the development of methods and algorithms for image pattern recognition and image understanding new issues are taken up, which may help in diagnosis and classification of diseases. The area of authors' research interests encompass issues related to the classification of genetic disorders like Down's or Turner's syndrome. In most of medical cases, a diagnosis is made based not on a single image but often on the basis of several images, which are acquired using various techniques of medical imaging. This observation is strongly related to the fact, that the diagnosis is a result of accurate differentiation of features specific to the particular disorder, which are inferred from the images. An outstanding specialist generally is able to discern subtle differences, but a person with less experience and knowledge may require help, which can be provided by means of advanced telemedicine technologies. The occurrence of particular disorders by its nature pertains to the more or less uniform distribution for the particular area or continent. Nowadays remote transmission of data, even large sets of data, is not a challenge from a technical point of view. By means of telecommunications, image collections of many patients can be sent to the diagnostic support center for detailed analysis. In special cases data may be analyzed by the human specialist. In the case of screening tests the role of the computer system is different in regard to the support of medical diagnosis. In this case the computer system may perform an initial classification of the collection of images. Such a system can be called a medical decision support system for the initial classification of a particular disorder. In many cases this system supporting diagnosis can be an important aid for the medical practitioners, and for the training of medical students. Selected cases difficult for the computer system can be evaluated by the human specialist, whose work can be ineffective when dealing with large number of medical images. The issue investigated in the work concerns classification of Down's syndrome on the basis of a set of dermatoglyphic images of palms, soles and fingers without results of genetic tests. It seems that the proposed approach can be extended for the problem of supporting diagnosis decision in case of Turner's syndrome.

3 Characteristics of the Dermatoglyphic Nomogram

A diagnostic index for the diagnosis of Down's syndrome was designed, taking advantage of discriminant function analysis to obtain the combinations of variables representing patterns of dermatoglyphic areas, which reliably separate groups of infants with Down's syndrome and healthy infants [6]. In the dermatoglyphic nomogram log-odds were calculated and used to assign weights to all the possible patterns, which allowed placing patterns on the scale differentiating between the healthy and Down's syndrome cases. A set of 32 pattern areas was used in the calculation of stepwise discriminant function on the log-odd weights. The results of the calculations were basis for the selection of four most significant variables and the design of the dermatoglyphic nomogram in its graphical form. These variables correspond to the following four dermatoglyphic traits: pattern types of the right and left index fingers, pattern type of the hallucal area of the right sole and the ATD angle of the right hand. Three outcomes of the diagnosis correspond to the appropriate segments of the main diagnostic line. On the basis of the determined index score infants are qualified to the healthy class, class with Down's syndrome or to the class located in the overlap area, which doesn't give clear indication of whether an infant is healthy or not.

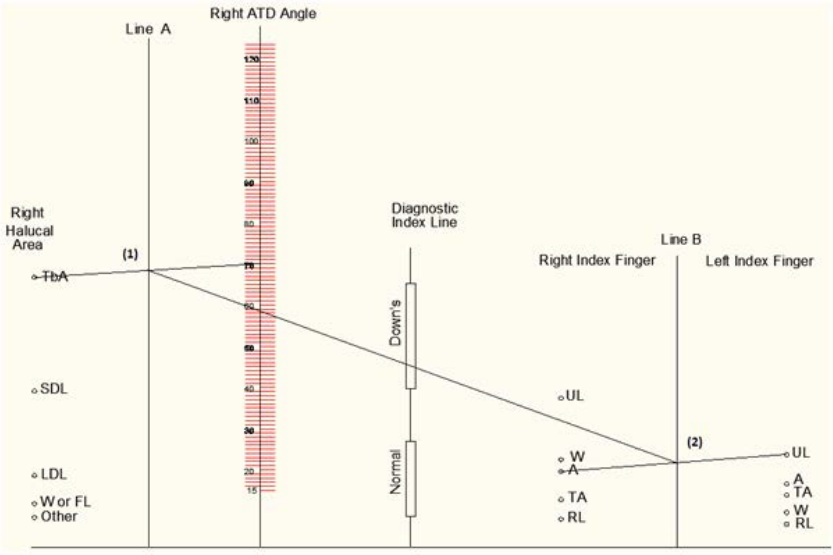


Fig. 1. Dermatoglyphic diagnostic index score for the particular case of infant determined on the basis of the nomogram. Abbreviations of the descriptions of the points located along the diagnostic lines denote the names of dermatoglyphic patterns: ulnar loop (UL), radial loop (RL), plain arch (A), tented arch (TA), whorl (W), fibular loop (FL), small distal loop (SDL), large distal loop (LDL).

Fig. 2. The green rectangle marks the location of the hallucal area. The print of the sole is taken from male child with Down's syndrome. The pattern within the rectangle belongs to the tibial arch class.

The schematic model of diagnosis used as the basis of the decision support system, is subject to a hierarchy and discipline of points. The diagnosis based on the dermatoglyphic nomogram reflects the course of the inference on the basis of explicit premises and leads to the conclusion involving scientifically justified degree of probability of the genetic syndrome under study.

The analysis of the nomogram index score for the particular case (Fig.1) requires in the first phase ascertaining patterns types of the left (α) and right (β) index fingers, pattern type of the hallucal area of the right sole (γ) and the value of the ATD angle (δ) of the right hand. The conjunction of the incidences α and β gives the diagnostic value of point "1" found on the line "A". The conjunction of the incidences γ and δ allows determining the diagnostic value

of point "2" located on the line "B". Jointly the conjunctions of the values in point denoted as "1" (α and β) and in point denoted as "2" (γ and δ) create the diagnostic line, which intersects the Diagnostic Index Line and sets the value of the diagnostic score. In the example case the determined score corresponds to the boundary dermatoglyphic configuration, which is located between the configuration typical for healthy infants and the configuration typical for infants with Down's syndrome.

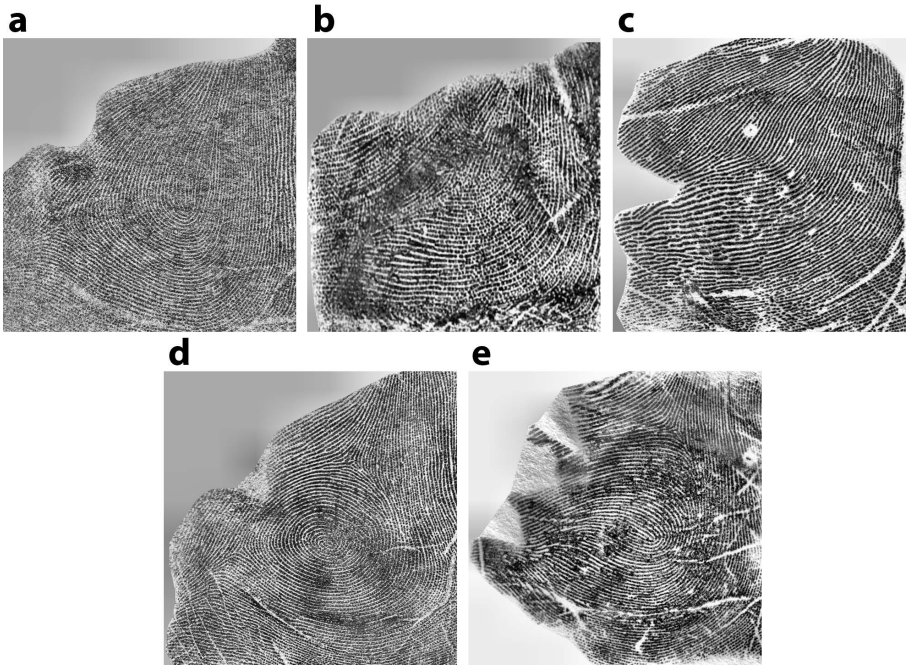


Fig. 3. Types of the local patterns in the hallucal area: (a) large distal loop (LDL), (b) small distal loop (SDL), (c) tibial arch (TA), (d) whorl (W), (e) tibial loop (TL)

On the basis of dermatoglyphic nomogram an expert system is created comprising of a set of rules. Each rule in the set corresponds to the particular combination of four dermatoglyphic pattern types. The set consists of 125 rules.

4 Patterns of the Hallucal Area of the Sole

The hallucal area is located in the tibial part of the ball region of the sole and represents two combined plantar configurational areas: distal thenar and first interdigital. In colloquial language it is the area below the big toe of the foot (Fig. 2).

The local patterns in the hallucal area are classified into five types: small distal loop (SDL), large distal loop (LDL), tibial loop (TL), tibial arch (TA) and whorl (W) (Fig. 3). The sole patterns are large, very intricate and show extensive variety. The fact of high interclass similarity between the different classes of patterns and simultaneously low intraclass similarity of patterns in the same class makes the automatic recognition of plantar patterns a very challenging problem. In Fig. 4 the upper series of images three varieties of the whorl class patterns are presented, the lower series of images in Fig. 4 shows from the left patterns of tibial arch and small distal loop classes, respectively.

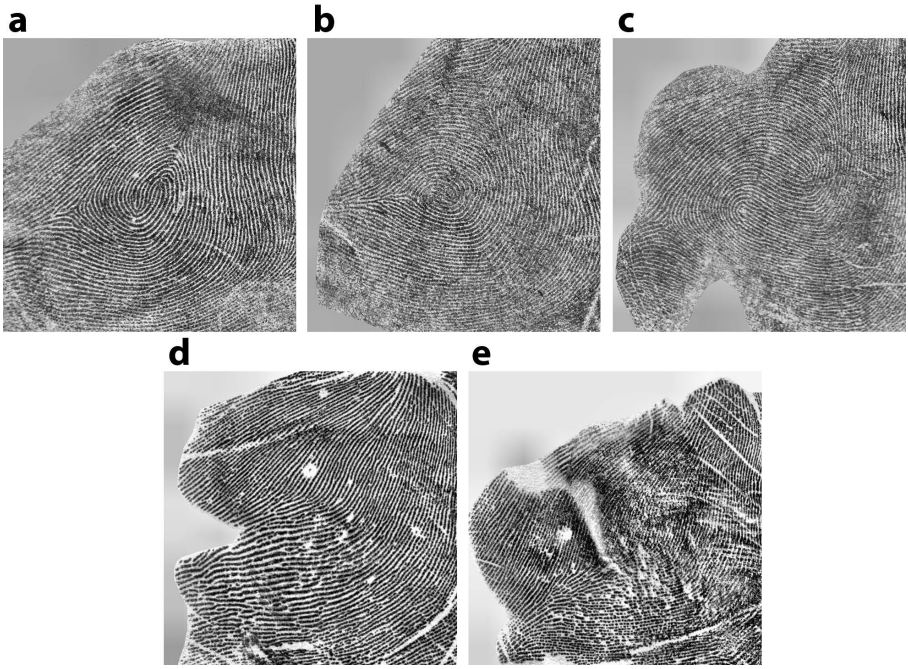


Fig. 4. The upper series of images shows three varieties of the whorl class patterns (low intraclass similarity), the lower series of images shows patterns of tibial arch and small distal loop classes (high interclass similarity)

5 Feature Extraction

Fingerprint impressions of low quality complicate the learning process of computational intelligence algorithms, and negatively influence their ability to accurately recognize the patterns. The classification accuracy can be improved by pre-processing of the images analyzed, which enhances their quality. In our approach dermatoglyphs of infants are subjected to the image enhancement procedure consisting of the following steps:

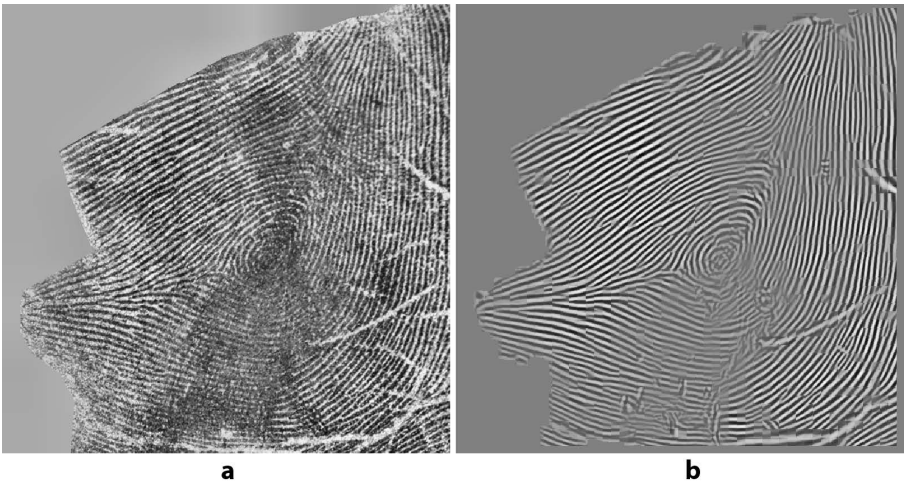


Fig. 5. An image of impression of the hallucal area belonging to the whorl class preprocessed using a contrast enhancement algorithm CLAHE (a), and then by filtration algorithm STFT (b)

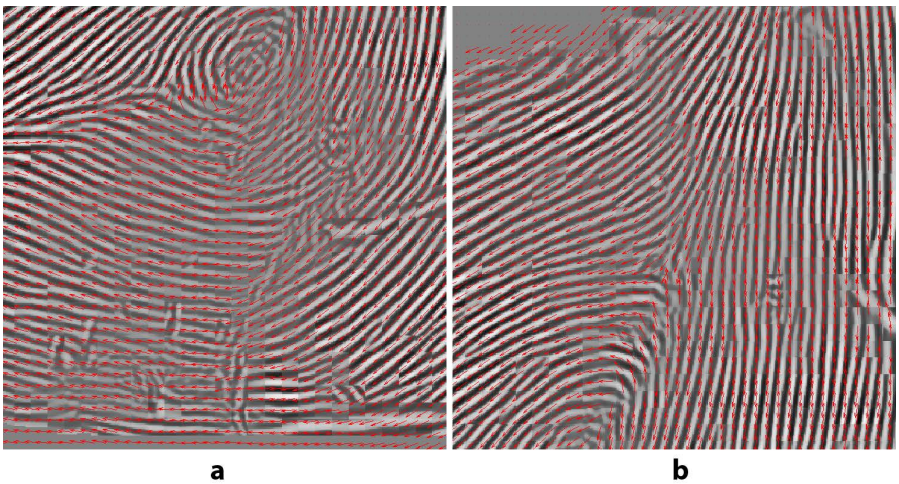


Fig. 6. Illustration of the orientation fields in low-quality regions of the impression of the hallucal area of the sole: bottom area of the pattern containing lower core of the whorl, top area of the pattern containing upper triradius and upper core of the whorl

1. Determination of the image mask - the region mask separates the image into the foreground enclosing the distinctive pattern area and the background containing noise.
2. Pattern area size normalization - the entire image is truncated using the coordinates of the foreground boundaries and resized to the square frame of fixed dimension.

3. Contrast enhancement using a CLAHE algorithm (Fig. 5a).
4. Determination of local frequency and orientation of ridges (Fig. 5b).
5. Adaptive filtration using a STFT algorithm [1].

Estimation of ridges' orientations is realized by the algorithm taking advantage of Principal Component Analysis and pyramid-based image decomposition into multiple scales [3]. Application of this algorithm helps in overcoming problems with reliable calculation of ridges' directions in the noisy areas of the impressions. Maps of the orientations calculated in the noised regions of the hallucal patterns from Fig. 5 are shown in Fig. 6a and Fig. 6b.

6 Description of the Data Set and Classification Using Support Vector Machines

In our research 240 8-bit grey scale images of the hallucal area prints from the database of CMUJ Cracow are used for the classification. The patterns of the hallucal area are slightly larger than the fingerprint patterns therefore the captured images size is 640 x 640 pixels. The data set contains 60 images for

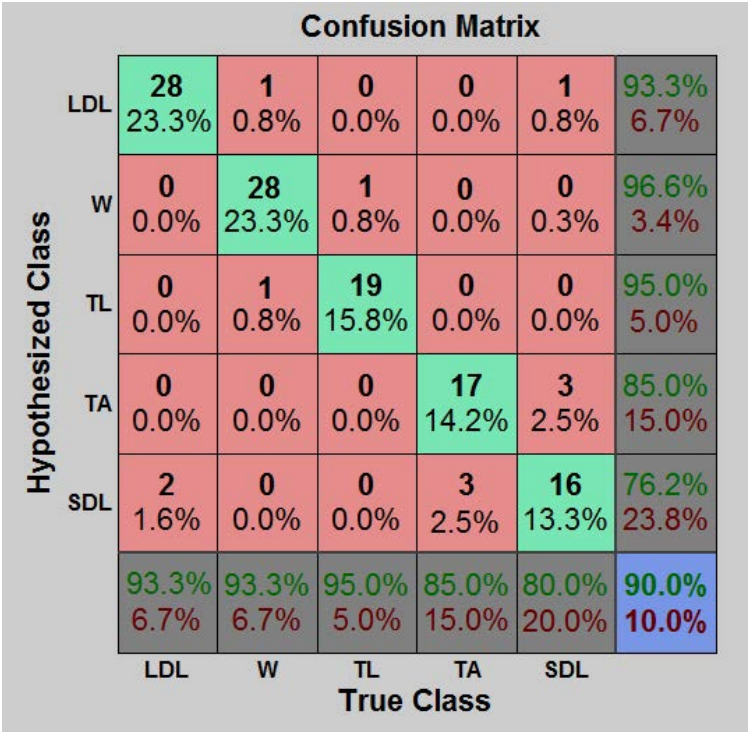


Fig. 7. Confusion matrix of the SVM classification accuracy

whorl class and the same number for large distal loop class. The database also contains 40 images for each of the remaining three classes: small distal loop, tibial arch and tibial loop. In the process of creating the vector data all of the images were enhanced using SFTF algorithm. Vectors containing the orientation maps were computed using two stage approach taking advantage of PCA and pyramid-based image decomposition algorithm [3]. Resulting ridge orientation maps are features which are classified by an ensemble of SVMs with one vs one voting strategy. For the training of classifiers Radial Basis Function kernels were chosen. The data set is split evenly in two halves according to the number of images representing particular classes. Training and testing vectors are both comprised of 120 examples. The optimal values of the SVMs kernel hyper-parameters were computed using grid search algorithm and cross-validation. The accuracy rate obtained in the testing phase is 90,0 %. The confusion matrix with test results is shown in Fig.7.

7 Summary

In the paper methods for enhancement, feature extraction and classification of plantar patterns have been presented. The classifier is part of an automatic system supporting medical diagnosis of Down's syndrome. The system is comprised of an expert system module and three modules realizing pattern recognition tasks. Rules of the expert system are created on the basis of the knowledge found in the scientific literature describing dermatoglyphic peculiarities of Down's syndrome. Pattern recognition modules implement three tasks required for the detection of traits characteristic of Down's syndrome and solution of the dermatoglyphic nomogram. These tasks are classification of fingerprint patterns [8], classification of plantar patterns of the hallucal area and calculation of palm print's ATD angle, respectively. Results of the recognitions are used as the values of the premises of the expert system, which allows calculation of the dermatogram index score and automatic diagnosis. In our future work we intend to apply the proposed methods to create decision support systems for the dermatoglyphic diagnosis of other genetic disorders.

References

1. Chikkerur, S., Cartwright, A.N., Govindaraju, V.: Fingerprint image enhancement using STFT analysis. Pattern Recognition 40, 198–211 (2007)
2. Cummins, H.: Dermatoglyphic stigmata in mongolian idiocy. Anat. Rec. 64(suppl. 2), 11 (1936) (abstract)
3. Feng, X.G., Milanfar, P.: Multiscale principal components analysis for image local orientation estimation. In: Proc. of the 36th Asilomar Conf. on Signals, Systems and Computers, vol. 1, pp. 478–482 (2002)
4. Holt, S.B., Lindsten, J.: Dermatoglyphic anomalies in Turner's syndrome. Ann. Hum. Genet. Lond. 28, 87–100 (1964)
5. Reed, T.: Dermatoglyphics in Down's syndrome. Clinical Genetics (6), 236 (1974)

6. Reed, T.E., et al.: Dermatoglyphic nomogram for the diagnosis of Down's syndrome. J. Pediat. (77), 1024–1032 (1970)
7. Uchida, I.A., Miller, J.R., Soltan, H.C.: Dermatoglyphics associated with XXYY chromosome complement. Amer. J. Hum. Genet. 16, 284–291 (1964)
8. Wojtowicz, H., Wajs, W.: Intelligent Information System for Interpretation of Dermatoglyphic Patterns of Down's Syndrome in Infants. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ACIIDS 2012, Part II. LNCS (LNAI), vol. 7197, pp. 284–293. Springer, Heidelberg (2012)

Adaptive Cumulative Voting-Based Aggregation Algorithm for Combining Multiple Clusterings of Chemical Structures

Faisal Saeed[1,2,*], Naomie Salim[1], Ammar Abdo[3,4], and Hamza Hentabli[1]

[1] Faculty of Computing, Universiti Teknologi Malaysia, Malaysia
[2] Information Technology Department, Sanhan Community College, Sana'a, Yemen
[3] Computer Science Department, Hodeidah University, Hodeidah, Yemen
[4] LIFL UMR CNRS 8022 Universite´ Lille 1 and INRIA Lille Nord Europe, 59655 Villeneuve d'Ascq cedex, France
alsamet.faisal@gmail.com

Abstract. Many consensus clustering methods have been studied and applied in many areas such as pattern recognition, machine learning, information theory and bioinformatics. However, few methods have been used for chemical compounds clustering. In this paper, Adaptive Cumulative Voting-based Aggregation Algorithm (A-CVAA) was examined for combining multiple clusterings of chemical structures. The effectiveness of clusterings was evaluated based on the ability of clustering to separate active from inactive molecules in each cluster and the results were compared to the Ward's method. The chemical dataset MDL Drug Data Report (MDDR) database was used. Experiments suggest that the adaptive cumulative voting-based consensus method can efficiently improve the effectiveness of combining multiple clustering of chemical structures.

Keywords: Co-association matrix, Compound selection, Cumulative voting, Ensemble clustering, Molecular datasets.

1 Introduction

The main objective of clustering is to subdivide a data objects into smaller groups known as clusters so that each group exhibits a high degree of intra-cluster similarity and inter-cluster dissimilarity [1]. Many types of clustering techniques for chemical structures datasets have been used in the literature [2-9].

In chemoinformatics, the clustering is used to reduce the high costs and lengthy time needed to discover new drugs, especially in the process of High-Throughput Screening (HTS), in which hundreds of thousands of chemical compounds are screened for testing the biological activity. The clustering helps the pharmaceutical industries to find faster and more effective ways of discovering and producing chemical compounds that can effectively react to the examined disease [10].

* Corresponding author.

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 305–314, 2013.

There are different types of clustering that can be grouped based on the problem they intend to solve, the general strategy they use, or others. For example Jain *et al.* [11] have organized the clustering methods into five opposing approaches which are agglomerative verses divisive, hard verses soft, monothetic verses polythetic, deterministic verses stochastic and incremental verses non incremental. Recently, individual clustering has been used verses consensus clustering.

Consensus clustering involves two main steps: (i) partitions generation and (ii) combination using the consensus function. In the first step, as many as possible individual partitions will be generated. In the combination step, there are two main approaches: objects co-occurrence based and median partition based approaches. In the first approach, the idea is to determine which must be the cluster label associated to each object in the consensus partition. To do that, it is analyzed how many times an object belongs to one cluster or how many times two objects belong together to the same cluster. The consensus is obtained through a voting process among the objects. So, each object should vote for the cluster to which it will belong in the consensus partition. Co-association matrix and Voting based methods are examples of this approach. In the second consensus function approach, the consensus partition is obtained by the solution of an optimization problem, the problem of finding the median partition with respect to the cluster ensemble [12].

There are many voting-based consensus clustering methods [13-19], in which the consensus partition is derived by seeking an optimal relabeling of the ensemble partitions. In general, the optimal relabeling of the ensemble partitions is addressed through a pairwise relabeling of each ensemble partition with respect to a representative partition [19]. Then, the voting process is used to assign object to the higher voted cluster in order to obtain the consensus partition [18-19].

For clustering of chemical structures datasets, it is most unlikely that any single method will yield the best classification under all circumstances, even if attention is restricted to a single type of application [20]. Chu, *et al.* [20] used consensus similarity matrix methods on sets of chemical structures and concluded that the consensus clustering methods can outperform the Ward's method. However, based on the implemented methods, it was not the case if the clustering is restricted to a single consensus method. In addition, Saeed *et al.* [21] examined the use of the graph-based consensus clustering method, Cluster-based Similarity Partitioning Algorithm (CSPA) [22], for clustering of MDDR dataset and concluded that it can improve the effectiveness of individual clusterings and provide robust and stable clustering. However, an adaptive cumulative voting-based aggregation algorithm is used in this paper to improve the effectiveness of combining multiple clusterings of chemical structures. The algorithm is efficient and has linear computational complexity in the number of data objects n, whereas co-association-based consensus methods are quadratic in n [12].

2 Materials and Methods

2.1 Dataset

The MDL Drug Data Report (MDDR) database [23] was used for experiments. This database consists of 102516 molecules. The MDDR subset dataset was chosen from

the MDDR database which has been used for many virtual screening experiments [24-26]. The MDDR dataset contains eleven activity classes (8294 molecules), which involves homogeneous and heterogeneous (i.e., structurally diverse) active molecules. Details of this dataset are listed in Table 1. Each row in the table contains an activity class, the number of molecules belonging to the class, and the diversity of the class, which was computed as the mean pairwise Tanimoto similarity calculated across all pairs of molecules in the class. For the clustering experiments, two 2D fingerprint descriptors were used which were developed by Scitegic's Pipeline Pilot [27]. These were 120-bit ALOGP and 1024-bit ECFP_4 fingerprints.

Table 1. MDDR Activity Classes for DS1 Data Set

Activity Index	Activity class	Active molecules	Pairwise similarity
			Mean
31420	Renin Inhibitors	1130	0.290
71523	HIV Protease Inhibitors	750	0.198
37110	Thrombin Inhibitors	803	0.180
31432	Angiotensin II AT1 Antagonists	943	0.229
42731	Substance P Antagonists	1246	0.149
06233	Substance P Antagonists	752	0.140
06245	5HT Reuptake Inhibitors	359	0.122
07701	D2 Antagonists	395	0.138
06235	5HT1A Agonists	827	0.133
78374	Protein Kinase C Inhibitors	453	0.120
78331	Cyclooxygenase Inhibitors	636	0.108

2.2 Partitions Generation

The partitions were generated by using six individual clustering algorithms on each 2D fingerprint. These algorithms were single linkage, complete linkage, average linkage, weighted average distance, Ward and K-means clustering methods. The thresholds of 500, 600, 700, 800, 900 and 1000 were used to generate partitions with different sizes (number of clusters). Every individual clustering method was applied by using Jaccard distance measures in order to generate one ensemble (includes six partitions) for each 2D fingerprint.

2.3 Adaptive Cumulative Voting-Based Aggregation Algorithm

The cumulative voting based aggregation algorithm consists of two steps; the first one is to obtain the optimal relabeling for all partitions, which is known as the voting problem. Then, the voting-based aggregation algorithm is used to obtain the aggregated (consensus) partition. The adaptive voting-based aggregation algorithm, which was described by Ayed and Kamel [18-19], was used for combining the ensembles that were generated in the previous step.

Let χ denote a set of n data objects, and let a partition of χ into k clusters be represented by an $n \times k$ matrix $\mathbf{U}$ such that $\sum_{q=1}^{k} ujq = 1$, for $\forall$ j. Let $u = \{U_i\}_{i=1}^{b}$ denote an ensemble of partitions. The voting-based aggregation problem is concerned with searching for an optimal relabeling for each partition $\mathbf{V}^i$ with respect to

representative partition $\mathbf{U^0}$ (with k^0 clusters) and for a central aggregated partition denoted as $\bar{\mathbf{U}}$ that summarises the ensemble partitions. The matrix of coefficients $\mathbf{W}^i$, which is a $k^i \times k^0$ matrix of $w_{lq}{}^i$ coefficients, is used to obtain the optimal relabeling for ensemble partitions.

Let *H(C)* denote the Shannon entropy associated with cluster *C*, which is sometimes written as *H(p(c))*. Defined over the cluster labels of the partition $\bar{\mathbf{U}}$, *H(C)* measures the average amount of information associated with *C* and is defined as a function of its distribution *p(c)* as follows [28]:

$$H(C) = - \sum_{c \in C} p(c) \log p(c) \quad (1)$$

Let *I (C;X)* denote the mutual information between *C* and *X*. *I (C;X)* measures the amount of information that the random variable *C* contains about *X*, and vice versa. It is defined as:

$$I(C;X) = H(C) - H(C|X) \quad (2)$$

It is noted that for a hard partition $\mathbf{U}^i$, we have $I(C^i;X) = H(C^i)$, since the value of C^i is completely determined by the value of *X* (i.e., $H(C^i|X) = 0$). It follows that *I(C;X)* is bounded from above *by H(C)*; $I(C;X) \leq H(C)$, where $H(C) = H(C^0)$. Thus, the initially selected reference partition determines the following measures: the entropy associated with the aggregated clusters, the initial value of the mutual information $I(C^0;X)$, and the upper bound on the amount of information that random variable *C* contains about *X*. This result motivates the use of a selection criterion for the initial reference partition based on the mutual information $I(C^i;X)$, which is equal to $H(C^i)$ for hard partitions (all individual clusterings used in this paper are hard clusterings).

Therefore, the fixed-reference approach is used, whereby an initial reference partition is used as a common representative partition for all the ensemble partitions and remains unchanged throughout the aggregation process. The algorithm incorporates the selection criterion for the initial reference partition, which is the one with highest entropy $H(C^i)$, then sorting the ensemble partitions in descending order of their entropies. The adaptive cumulative voting based aggregation algorithm is described as follows:

Adaptive Cumulative Voting-based Aggregation Algorithm

1: Re-order *u*, s.t. $\mathbf{U^i}$ are sorted in decreasing order of $H(C^i)$ based on Eq. 1
2: Assign U^1 to U^0
3: for $i = 2$ to b do
4: $\mathbf{W}^i = (\mathbf{U}i^{\mathrm{T}}\, \mathbf{U}i)^{-1}\, \mathbf{U}i^{\mathrm{T}}\, \mathbf{U^0}$
5: $\mathbf{V}^i = \mathbf{U}i\, \mathbf{W}^i$
6: $\mathbf{U}^0 = \frac{i-1}{i}\, \mathbf{U}^0 + \frac{1}{i}\, \mathbf{V}^i$
7: end for
8: $\bar{\mathbf{U}} = \mathbf{U}^0$.

2.4 Performance Evaluation

The results were evaluated based on the effectiveness of the methods to separate active from inactive molecules using two measures: the F-measure [29] and Quality Partition Index (QPI) measure [30]. If the cluster contains n compounds, that a of these are active and that there is a total of A compounds with the chosen Activity. The precision, P, and the recall, R, for that cluster are [20]:

$$P = \frac{a}{n} \tag{3}$$

$$R = \frac{a}{A} \tag{4}$$

$$F = \frac{2PR}{P + R} \tag{5}$$

This calculation is carried on each cluster and the F-measure is the maximum value across all clusters.

Also, an active cluster is defined as a non-singleton cluster for which the percentage of active molecules in the cluster is greater than the percentage of active molecules in the dataset as a whole. Let p be the number of actives in active clusters, q be the number of inactives in active clusters, r be the number of actives in inactive clusters (i.e., clusters that are not active clusters) and s be the number of singleton actives. The high value occurs when the actives are clustered tightly together and separated from the inactive molecules. The QPI is defined to be [8]:

$$QPI = \frac{p}{p + q + r + s} \tag{6}$$

3 Results and Discussion

The two ensembles were combined using the adaptive cumulative voting-based consensus clustering (A-CVAA) and graph-based consensus clustering (CSPA).

The mean of F-measure and QPI values were averaged over the eleven activity classes of the dataset. Figures 1-4 show the effectiveness of MDDR dataset clustering for ALOGP and ECFP_4 fingerprints.

Visual inspection of F-measure and QPI values in Figures 1-4 enables comparisons to be made between the effectiveness of two consensus clustering methods and the Ward's method for clustering of chemical structures. In addition, two fingerprints were used for the experiments in order to study the effectiveness of consensus clustering on different representations of molecular dataset.

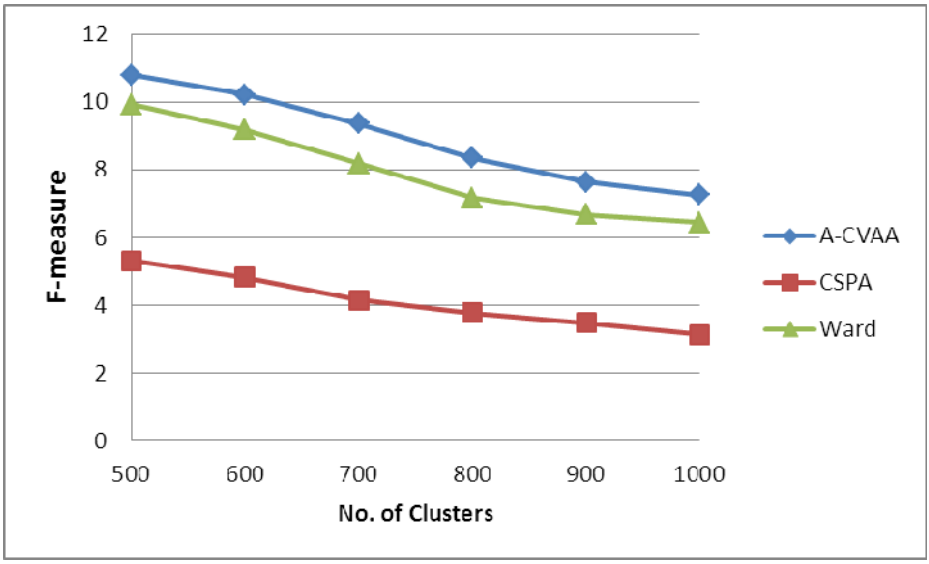


Fig. 1. Effectiveness of clustering MDDR dataset using F-Measure: ALOGP Fingerprint

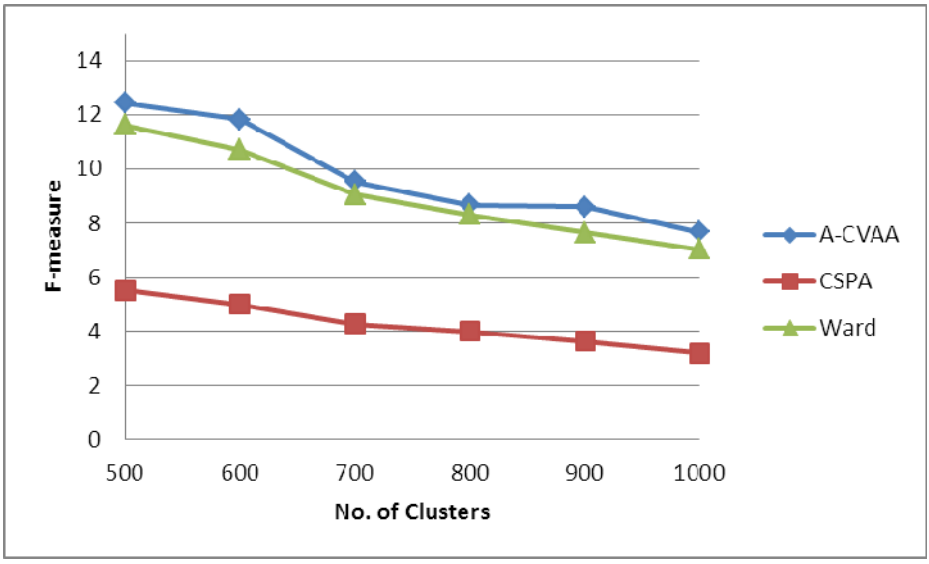


Fig. 2. Effectiveness of clustering MDDR dataset using F-Measure: ECFP_4 Fingerprint

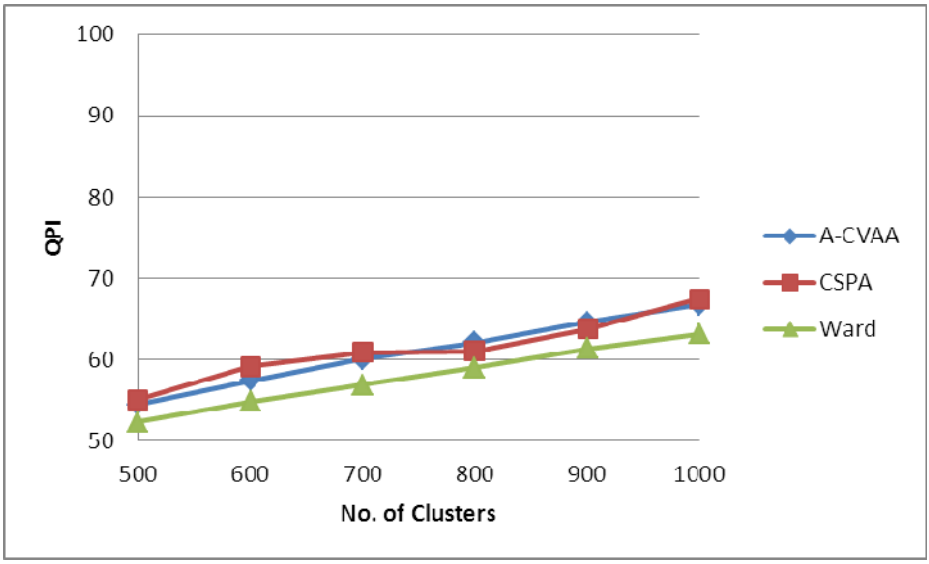


Fig. 3. Effectiveness of clustering MDDR dataset using QPI: ALOGP Fingerprint

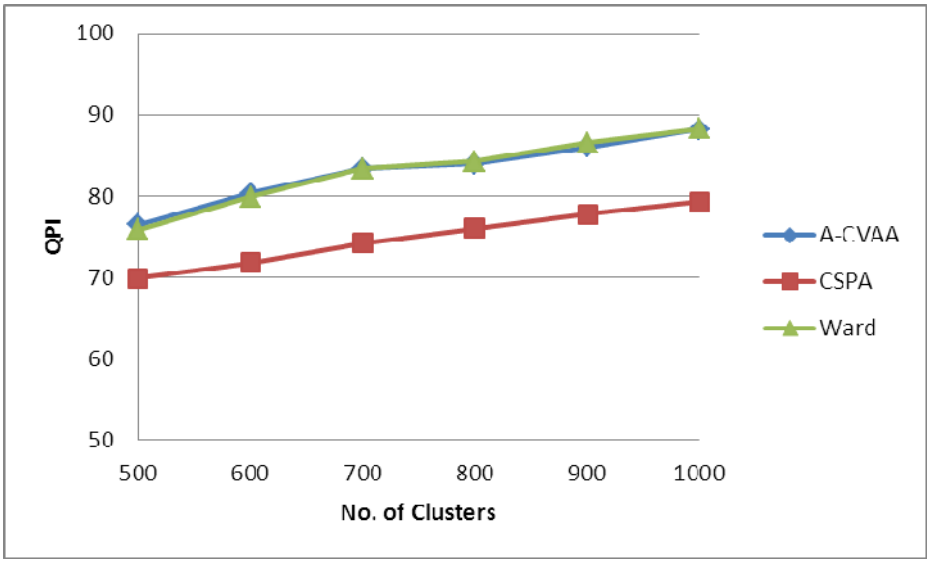


Fig. 4. Effectiveness of clustering MDDR dataset using QPI: ECFP_4 Fingerprint

In Figures 1-2, the performance of adaptive cumulative voting-based aggregation algorithm (A-CVAA) outperformed the Ward's method and the graph-based consensus clustering method (CSPA) using F-measure for both ALOGP and ECFP_4 fingerprints. The performance of CSPA was inferior to others methods.

In addition, both A-CVAA and CSPA showed similar performance and they outperformed the Ward's method using QPI measure for ALOGP (Figure 3). However, when using ECFP_4 fingerprint, A-CVAA consensus clustering outperformed the CSPA method using QPI measures and showed a similar performance to the Ward's method (Figure 4).

4 Conclusion and Future Work

The results of the experiments show that the adaptive cumulative voting-based aggregation algorithm (A-CVAA) can efficiently improve the effectiveness of combining multiple clusterings of chemical structures with linear computational complexity $O(n)$, rather than $O(n^2)$ using CSPA or other co-association based consensus methods. The performance of A-CVAA consensus clustering outperformed the Ward's method using both F and QPI measures for ALOGP fingerprint. While, in case of using ECFP_4, A-CVAA outperformed the Ward's method using F-measure and both methods showed similar performance for QPI measure. In the future work, other voting-based consensus clustering methods will be examined for combining multiple clusterings of chemical structures.

Acknowledgment. This work is supported by the Ministry of Higher Education (MOHE) and Research Management Centre (RMC) at the Universiti Teknologi Malaysia(UTM) under Research University Grant Category (VOT Q.J130000.7826.4F011). We also would like to thank MIS-MOHE for sponsoring the first author.

References

1. Everitt, B.S., Landau, S., Leese, M.: Cluster Analysis, 4th edn. Edward Arnold, London (2001)
2. Downs, G.M., Barnard, J.M.: Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures. Journal of chemical information and computer science 32, 644–649 (1992)
3. Willett, P.: Similarity and Clustering in Chemical Information Systems. Research Studies Press, Letchworth (1987)
4. Downs, G.M., Willett, P., Fisanick, W.: Similarity searching and clustering of chemical-structure databases using molecular property data. J. Chem. Inf. Comput. Sci. 34, 1094–1102 (1994)
5. Brown, R.D., Martin, Y.C.: The information content of 2D and 3D structural descriptors relevant to ligand–receptor binding. J. Chem. Inf. Comput. Sci. 37, 1–9 (1997)

6. Downs, G.M., Barnard, J.M.: Clustering methods and their uses in computational Chemistry. In: Lipkowitz, K.B., Boyd, D.B. (eds.) Reviews in Computational Chemistry, vol. 18. John Wiley (2002)
7. Holliday, J.D., Rodgers, S.L., Willet, P.: Clustering Files of chemical Structures Using the Fuzzy k-means Clustering Method. Journal of Chemical Information and Computer Science 44, 894–902 (2004)
8. Varin, T., Bureau, R., Mueller, C., Willett, P.: Clustering files of chemical structures using the Székely–Rizzo generalization of Ward's method. Journal of Molecular Graphics and Modeling 28(12), 187–195 (2009)
9. Brown, R.D., Martin, Y.C.: Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. J. Chem. Inf. Compute. Sci. 36, 572–584 (1996)
10. Salim, N.: Analysis and Comparison of Molecular Similarity Measures. University of Sheffield. PhD Thesis (2003)
11. Jain, A.K., Murty, M.N., Flynn, P.J.: Data Clustering: a review. ACM Computing Surveys 31 (1999)
12. Vega-Pons, S., Ruiz-Schulcloper, J.: A survey of clustering ensemble algorithms. International Journal of Pattern Recognition and Artificial Intelligence 25(3), 337–372 (2011)
13. Fischer, B., Buhmann, J.M.: Bagging for path-based clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(11), 1411–1415 (2003)
14. Dudoit, S., Fridlyand, J.: Bagging to improve the accuracy of a clustering procedure. Bioinformatics 19(9), 1090–1099 (2003)
15. Dimitriadou, E., Weingessel, A., Hornik, K.: A combination scheme for fuzzy clustering. International Journal of Pattern Recognition and Artificial Intelligence 16(7), 901–912 (2002)
16. Gordon, A.D., Vichi, M.: Fuzzy partition models for fitting a set of partitions. Psychometrika 66(2), 229–248 (2001)
17. Topchy, A., Law, M., Jain, A.K., Fred, A.: Analysis of consensus partition in clustering ensemble. In: Proceedings of the IEEE Intl. Conf. on Data Mining 2004, Brighton, UK, pp. 225–232 (2004)
18. Ayad, H.G., Kamel, M.S.: Cumulative voting consensus method for partitions with a variable number of clusters. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(1), 160–173 (2008)
19. Ayad, H.G., Kamel, M.S.: On voting-based consensus of cluster ensembles. Patt. Recogn. 43, 1943–1953 (2010)
20. Chu, C.-W., Holliday, J., Willett, P.: Combining multiple classifications of chemical structures using consensus clustering. Bioorganic & Medicinal Chemistry (available online March 10, 2012)
21. Saeed, F., Salim, N., Abdo, A., Hentabli, H.: Combining Multiple Individual Clusterings of Chemical Structures Using Cluster-Based Similarity Partitioning Algorithm. In: Hassanien, A.E., Salem, A.-B.M., Ramadan, R., Kim, T.-h. (eds.) AMLTA 2012. CCIS, vol. 322, pp. 276–284. Springer, Heidelberg (2012)
22. Strehl, A., Ghosh, J.: Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions. J. Machine Learning Research 3, 583–617 (2002)
23. Sci Tegic Accelrys Inc., the MDL Drug Data Report (MDDR) database is available from at `http://www.accelrys.com/` (accessed November 1, 2012)
24. Abdo, A., Chen, B., Mueller, C., Salim, N., Willett, P.: Ligand-Based Virtual Screening Using Bayesian Networks. J. Chem. Inf. Model. 50, 1012–1020 (2010)

25. Abdo, A., Salim, N.: New Fragment Weighting Scheme for the Bayesian Inference Network in Ligand-Based Virtual Screening. J. Chem. Inf. Model. 51, 25–32 (2011)
26. Abdo, A., Saeed, F., Hentabli, H., Ali, A., Salim, N.: Ligand expansion in ligand-based virtual screening using relevance feedback. Journal of Computer-Aided Molecular Design 26, 279–287 (2012)
27. Pipeline Pilot, Accelrys Software Inc., San Diego (2008)
28. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New York (1991)
29. Van Rijsbergen, C.J.: Information Retrieval. Butterworth, London (1979)
30. Varin, T., Saettel, N., Villain, J., Lesnard, A., Dauphin, F., Bureau, R., Rault, S.J.: 3D Pharmacophore, hierarchical methods, and 5-HT4 receptor binding data. Enzyme Inhib. Med. Chem. 23, 593–603 (2008)

LINGO-DOSM: LINGO for Descriptors of Outline Shape of Molecules

Hamza Hentabli, Naomie Salim, Ammar Abdo, and Faisal Saeed

Faculty of Computer Science and Information System,
Universiti Teknologi Malaysia
Hentabli_hamza@yahoo.fr

Abstract. The linear notations are more compact than connection tables so they can be useful for storing and transmitting large number of chemical structures. Implicitly they contain the information needed to compute all kinds of molecular structures and, thus, molecular properties derived from these structures. In this DOSM is a new method of obtaining a rough description of 2D molecular structure from its 2D connection graph in the form of character string. Our method is based on the fragmentation of DOSM strings into overlapping substrings of a defined size that we call LINGO-DOSM. The integral set of LINGO-DOSM derived from a given DOSM string, LINGO-DOSM allows rigorous structure specification using very small and simple rule. In this paper, we study the possibility of using the textual descriptor for describing the 2D structure of the molecule. Simulated virtual screening experiments with the MDDR database show clearly the superiority of the LINGO-DOSM descriptor compared to many standard descriptors tested in this paper.

Keywords: Molecular database retrieval, Molecular descriptor, Molecular similarity Structures and substructure searching.

1 Introduction

The search for compounds similar to a given target ligand structure and compounds with defined biophysical profiles are two important principles of the modern drug discovery process. Both tasks make use of molecular descriptors with differing complexity (atomic, topographic, substructure fingerprints, 3D, biophysical properties, etc.) leading to different representations of the same molecule [1]. such descriptors can then be used as input for QSPR models and intermolecular distance calculations [2]. The development, implementation, and application of molecular descriptors and the subsequent mathematical treatment of the information contained in these descriptors have become an area of intense theoretical and practical interest in recent [2-4].

Molecular databases are searched for molecules similar to those with known bioactivities in the hope that they will exhibit similar biological profiles. This concept is commonly termed bioisosterism. The intermolecular similarity value depends hereby on the molecular description and the distance calculation employed, and relatively small structural changes, especially in ring systems, can cause large deviations in the

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 315–324, 2013.

similarity values. The biological similarity metrics depends finally on the problem given which means that the same small structural changes introduced in a given molecule might have different effects in the biological activities depending on the target. In our experience different methods work better for different targets, and various methods should be employed independently instead of using consensus scoring.

A wide range of structural descriptors are based on 2D and 3D structures [5].To process large databases 2D descriptors are preferred for computation speed reasons. Such descriptors, for example MDL fingerprints [6] are built on a predefined set of substructures, whose presence or absence is recorded as one bit in a bit-string. The building of such keys is a computationally expensive graph-theoretical NP-complete problem. However, the resulting keys allow very rapid molecular comparisons. It must be underlined that structural information is lost due to the limited number of predefined substructures and the binary representation. Hashed fingerprints overcome the latter problem by generating an exhaustive list of structural fragments according to a certain rule set [7]. The very large number of fragments potentially generated does not allow the assignment of an individual bit to each fragment, and, instead, several different substructures are represented by one bit using a pseudo randomizing algorithm. This leads to a new problem described as fragment collision [8]. This process reduces the accuracy of the fingerprint but allows the use of a much larger number of fragments.

However, the most common similarity approaches use molecules characterized by 2D fingerprints that encode the presence of 2D fragment substructures in a molecule. The similarity between two molecules is then computed using the number of substructure fragments common to a pair of structures and a simple association coefficient[9].

The shape similarity between two molecules can be determined by comparing the shapes of those molecules, finding the overlap volume between them and then using a similarity measure (e.g. Tanimoto) to calculate the similarity between the molecules. However, most of the works in shape-based similarity approaches have depended on the 3D molecular shape[6]. Recently, the use of field-based or shape-based approaches has been increased [10]. The shape comparison program Rapid Overlay of Chemical Structures (ROCS) [11] is used to perceive similarity between molecules based on their 3D shape. The objective of this approach is to find molecules with similar bioactivity to a target molecule but with different chemotypes, i.e. scaffold hopping. However, a disadvantage of 3D similarity methods is that the conformational properties of the molecules should be considered and therefore these methods are more computationally intensive than methods based on 2D structure representation. The complexity increases considerably if conformational flexibility is taken into account. In 2D structure representation, the molecular structure is represented by a large number of structural descriptors in a numerical form (integer or real). Among these, descriptors computed based on a molecule graph are widely used in modeling physical, chemical, or biological properties. The simplest 2D descriptors are based on simple counts of features such as hydrogen donors, hydrogen bond acceptors, ring systems (such as aromatic rings) and rotatable bonds, whereas the complex 2D descriptors are computed from complex mathematical equations such as 2D fingerprints and topological indices 12-13. They characterize molecular structures according to their size, degree

of branching and overall shape where the structural diagram of molecules is considered as a mathematical graph, but not the contour of molecule shape.

Due to the multi-faceted nature of biological activities, there is a high possibility that there are no single and best molecular descriptors that can uniquely represent the molecules [14]. This possibility has encouraged many researchers to continue to develop new molecular descriptors. Therefore, developing new molecular descriptors that can give a comparable or better result than the existing descriptors is highly desirable.

In this paper, we introduced a new Descriptors of Outline Shape of Molecules (DOSM) that was inspired by research in information retrieval on the use of contour-based shape descriptor for image retrieval systems. DOSM is a new method to obtain a rough description of the 2-D molecular structure from its outline shape. DOSM is a textual descriptor which allows rigorous structure specification by use of a very small and natural grammar. Our method is based on the fragmentation of DOSM strings into overlapping substrings of a defined size that we call LINGO-DOSM. The integral set of LINGO-DOSM derived from a given DOSM string.

2 Materials and Methods

The new descriptor DOSM is a textual descriptor using printable characters for representing molecules based on their shapes. In this paper, the outline shape (for the whole molecule) and the internal region (inside molecule rings) are exploited to calculate a rough description of the 2-D structure molecule. The proposed method uses a connection table to extract the information needed to represent the molecule shape. A specific language has been developed to describe the shape features; descriptors written in this language are invariants to scale change and rotation. DOSM is a true language, albeit with a simple vocabulary (atom and bond symbols) and only a few grammar rules. However, part of the power of the DOSM is that it is highly sensitive to molecular structure changes. In this work, the graph denotes the 2D molecular structure. This is essentially the 2D image chemists draw to describe the molecule. Here, only the labeled molecular graph (i.e. atoms and bonds) and all possible paths between every atom pair are taken into account.

A corresponding shape to a 2D molecule structure is generally composed by a main region (representing the outline shape) and one or many internal regions (representing areas inside rings) obtained after visiting all the atoms in the connection table of a molecule. In addition to the geometry of its outline, we take into account the geometry and the position of its internal regions. This additional information for shape description is important to identify and represent the molecule rings in the DOSM descriptor context. It is also very useful for shape comparison in the similarity calculation between two molecules from their 2D graph.

The process of generating the shape descriptor of any molecule starts with determining the top left atom in the molecule graph. The atom name is represented in the descriptor as the grammar described below. Then, we move in a clockwise direction to the next atom. The bond type and direction of the movement are represented before

we visit and represent the next atom. The same procedure is repeated until we visit again the starting atom.

Once the starting atom is visited again, the description of the outline shape of the molecule graph is completed. The atoms and bonds within the internal regions (rings) need to be visited again to represent the internal region. However, the process of generating the DOSM is composed of a number of specification rules as explained below.

The language used for writing the DOSM descriptor consists of a series of characters and symbols. There are five generic encoding rules corresponding to the specification of atoms, bonds, ring closure, direction angle and disconnected parts. Some of these rules are similar to the rules used in SMILES strings [15]. Atoms are represented by their atomic symbols, usually two characters. The second character of the atomic symbol must be entered in lower case. If the atomic symbol is just one letter we add a blank space to the end of the atomic symbol, e.g., "Br", "C1","N ", "O ".

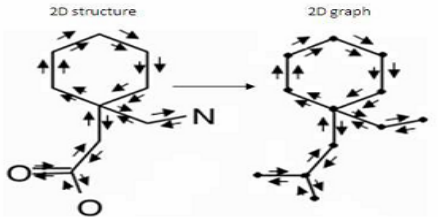


Fig. 1. The external and internal visiting movement in a 2D graph

The single, double, and triple bonds are represented by the symbols "-", "=", and "#", respectively. The internal regions (rings or cycles structures) are specified by enclosing their description in parentheses (Figure 2 for example).

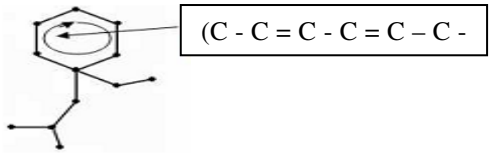


Fig. 2. Example of the representation of internal regions (cyclic structures)

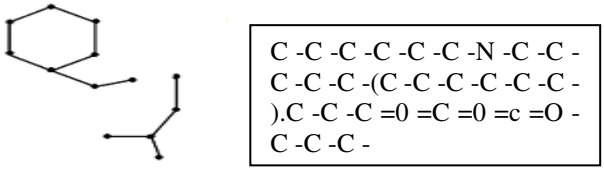
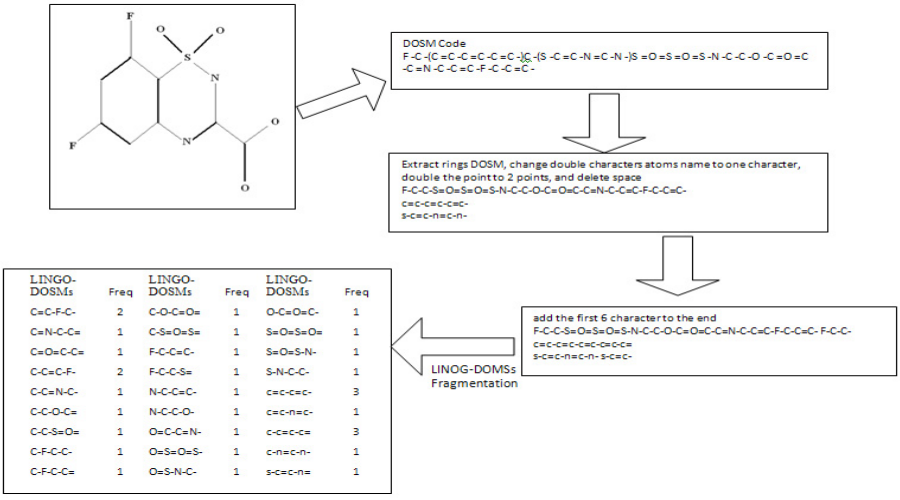


Fig. 3. The disconnect parts representation

If the molecule graph is composed of more than one part (disconnected structures), the description of the disconnected compound is written as individual structures separated by "." (Period) as shown in Figure 3.

LINGO-DOSM Generation: A q-LINGO is a q-word string with word is 2 character lengths, including letters, and symbols, obtained by stepwise fragmentation of a DOSM molecular representation. For a given molecule, A total number of (n/q) substrings of length q are extracted from a DOSM string of length n. In this work we use q= 4 unless indicated otherwise and the q- prefix is omitted. The molecule-specific LINGO-DOSM profile is defined as the ensemble of LINGO-DOSMs and their corresponding number of occurrences and does not depend on the order of appearance of LINGO-DOSMs in the DOSM string. The LINGO-DOSM generation process is summarized in Figure 4.



LINGO-DOSMs	Freq	LINGO-DOSMs	Freq	LINGO-DOSMs	Freq
C=C-F-C-	2	C-O-C=O=	1	O-C=O=C-	1
C=N-C-C=	1	C-S=O=S=	1	S=O=S=O=	1
C=O=C-C=	1	F-C-C=C-	1	S=O=S-N-	1
C-C=C-F-	2	F-C-C-S=	1	S-N-C-C-	1
C-C=N-C-	1	N-C-C=C-	1	c=c-c=c-	3
C-C-O-C=	1	N-C-C-O-	1	c=c-n=c-	1
C-C-S=O=	1	O=C-C=N-	1	c-c=c-c=	3
C-F-C-C-	1	O=S=O=S-	1	c-n=c-n-	1
C-F-C-C=	1	O=S-N-C-	1	s-c=c-n=	1

Fig. 4. LINGO-DOSMs generation process

Some changes in the original DOSM code are introduced. Thus, “Cl” and “Br” are substituted by “L” and “R”, respectively, and delete all space between character, and double the point represented the disconnected part, after that extract all the rings between “(” and “)”, and convert all the character represented the rings from capital letter to small letter for example “(C -C =C -C -C =C -)” become “(c-c=c-c-c=c-)” and delete the parentheses, the last step is to add the first (q-1)words at the end of each string we get, for example “c-c=c-c-c=c-“ become “c-c=c-c-c=c-c-c=c-” This steps reduces the number of possible LINGO-DOSMs and improves statistical sampling in the QSPR models.

3 Experimental Design

In this section, we present experiments that show the usefulness of the new descriptor LINGO-DOSMs, when used for similarity-based virtual screening. To evaluate the LINGO-DOSMs descriptor, LINGO-DOSMs was compared with six different descriptors (fingerprints) from Scitegic's Pipeline Pilot16 and PaDEL-descriptor [19]software. These were 120-bit ALOGP, 166-bit MACCS and 1024-bit Path fingerprints (EPFP4) from Scitegic's Pipeline Pilot and 1024-bit CDK (CDKFP), 1024-bit CDK graph only (GOFP), and 881-bit Pubchem fingerprints (PCFP) from the PaDEL software.

Experiments were conducted over the most popular cheminformatics database: the MDL Drug Data Report (MDDR) [10] which has been used in our previous studies [19]. This database consisted of 102516 molecules and contains 11 activity classes, which involve structurally homogeneous and heterogeneous actives, as shown in Table 1. Each row in the tables contains an activity class, the number of molecules belonging to the class, and the diversity of the class, which was computed as the mean pairwise Tanimoto similarity calculated across all pairs of molecules in the class.

Table 1. MDDR Activity Classes for MDDR Data Set

Activity	Activityclass	Active	Pairwise similarity
31420	renin inhibitors	1130	0.290
71523	HIV protease inhibitors	750	0.198
37110	thrombin inhibitors	803	0.180
31432	angiotensin II AT1 antagonists	943	0.229
42731	substance P antagonists	1246	0.149
06233	substance P antagonists	752	0.140
06245	5HT reuptake inhibitors	359	0.122
07701	D2 antagonists	395	0.138
06235	5HT1A agonists	827	0.133
78374	protein kinase C inhibitors	453	0.120
78331	cyclooxygenase inhibitors	636	0.108

Intermolecular Similarity Calculation: Based on the LINGO-DOSMs profiles of the two molecules A and B to be compared, intermolecular similarities were computed using the integral Tanimoto coefficient of the form

$$T_c = \frac{\sum_{i=1}^{l} 1 - \frac{|N_{A,i} - N_{B,i}|}{N_{A,i} + N_{B,i}}}{l}$$

Where $N_{A,i}$ is the number of LINGO-DOSM of type i in molecule A, $N_{B,i}$ is the number of LINGO-DOSM of type i in B, and l is the number of LINGO-DOSM contained in either molecule A or B. Molecules having the same types and numbers of LINGO-DOSM will show $T_c = 1$. We call the L LINGO-DOSM profile based T_c "LINGO-$DOSM_{sim}$".

The second screening system was based on the Tanimoto (TAN) coefficient, which has been used for ligand based virtual screening for many years and which can hence be considered as a reference standard. TAN was used with six types of descriptors (fingerprints) in this study.

The screening experiments were performed with 10 reference structures selected randomly from each activity class. The recall results were averaged over each such set of active molecules, where the recall is the percentage of the actives retrieved in the top-1% or the top-5% of the ranked list resulting from a similarity search.

4 Results and Discussion

The main objective of this work is to identify the possibility of using the LINGO-DOSM descriptor in similarity-based virtual screening and then identifying the retrieval effectiveness of using such a descriptor. In this study, we compared the retrieval effectiveness of LINGO-DOSM against six different types of descriptors on the MDDR database. Selecting the best descriptors is based on their use in predicting the property/activity of a molecule from another molecule that is considered similar to it, either by using a certain similarity method, clustering or using its k-nearest neighbours 21.For those descriptors, and for predicting the activity class of molecules, the best descriptors are those yielding the highest number of correct predictions (molecules with similar activity class), taking into account the total number of molecules having that activity class in the database used.The results for the searches of MDDR are shown in Tables 2-3, using cutoffs of both 1% and 5% respectively.

Each row in a table corresponds to one activity class; shows the recall for the top 1% and 5% of a sorted ranking when averaged over the ten searches for this activity class. The penultimate row in a table corresponds to the mean value for that descriptor when averaged over all of the activity classes for a dataset. The descriptor with the best recall rate in each row is strongly shaded, and the recall value is bold-faced, any descriptor with an average recall within 5% of the value for the best descriptor is shown lightly shaded. The bottom row in a table corresponds to the total number of shaded cells for each descriptor type across the full set of activity classes.

Table 2. Retrieval results of top 1% for data set MDDR

Activity Index	LINGO-DOSM	GRFP	PCFP	ALOGP	MACCS	EPFP4	CDKFP
31420	61.10	12.17	26.13	22.06	28.65	34.75	41.8
71523	26.10	8.68	9.61	13.72	14.71	14.29	19.6
37110	17.37	14.89	12.38	9.26	17.99	18.8	18.74
31432	38.63	15.12	15.55	16.52	24.52	22.81	25.75
42731	11.86	7.71	9.63	6.05	8.18	10.08	12.27
6233	11.46	5.58	6.8	7.98	8.8	8.35	9.47
6245	4.66	3.94	4.11	3.66	4.94	5.61	7.21
7701	10.38	4.19	4.62	5.86	7.39	6.75	7.77
6235	10.34	4.37	4.27	6.22	6.91	6.55	8.29
78374	12.01	6.88	13.16	7.81	6.02	8.01	10.64
78331	5.8	3.94	5.13	4.11	6.33	4.94	5.72
Mean	19.06	7.95	10.13	9.39	12.22	12.81	15.21
Shaded	8	0	1	0	2	1	2

Table 3. Retrieval results of top 5% for data set MDDR

Activity Index	LINGO-DOSM	GRFP	PCFP	ALOGP	MACCS	EPFP4	CDKFP
31420	84.82	30.59	45.95	45.08	55.41	76.76	80.27
71523	50.11	20.17	19.73	33.38	29.97	33.31	37.92
37110	28.19	27.83	27.99	26.71	34.7	39.96	37.26
31432	75.27	33.91	33.73	39.37	48.29	41.01	51.46
42731	21.62	14.92	19.32	12.91	19.36	20.71	23.2
6233	25.73	14.34	17	20.47	24.07	20	19.92
6245	10.92	9.89	10.08	10.59	11.06	12.65	17.88
7701	22.99	9.92	11.62	13.6	22.34	17.69	18.86
6235	26.43	13.84	13.51	14.71	20.33	17.82	19.21
78374	18.30	13.74	18.1	14.71	11.73	12.59	15.11
78331	10.16	8.87	11.23	9.97	14.35	9.37	10.55
Mean	34.04	18	20.75	21.95	26.51	27.44	30.15
Shaded	9	0	0	0	1	1	1

Visual inspection of the recall values and the number of shaded cells in Tables 1 and 2 enables comparisons to be made between the effectiveness of the LINGO-DOSM descriptor and the various other descriptors. In addition, a more quantitative approach using the Kendall W test of concordance was used to determine which of the descriptors performed best 22. This test was developed to quantify the level of agreement between multiple sets of rankings of the same set of objects, here and in previous works 19]. We used this approach to rank the effectiveness of different descriptor types. In the present context, the activity classes were considered as judges and the recall rates of the various descriptor types as objects. The outputs of the test are the value of the Kendall coefficient and the associated significance level, which indicates whether this value of the coefficient could have occurred by chance. If the value is significant (for which we used cut-off values of (0.01 or 0.05), then it is possible to give an overall ranking of the objects that have been ranked. The results of the Kendall analyses are reported in Table 4 and describe the top 1% and 5% ranking for the various descriptor types. In Table 5, the columns show the data set type, the value of the coefficient, the associated probability, and the ranking of the descriptor. The descriptors are ranked in decreasing order of screening effectiveness (if two descriptors have the same rank then they are ordered on the basis of the mean recall, i.e. the mean values from the main tables of results). We shall use the 5% MDDR results (in Table 4) to illustrate the processing that took place. Here, the mean figures suggest that the LINGO-DOSM descriptor has the best overall performance at the 5% cut-off. In addition, according to the total number of shaded cells in Table 4, LINGO-DOSM is the best performing descriptor across the 11 activity classes. We can hence conclude that the overall ranking of the seven descriptors are:

LINGO-DOSM >CDKFP>MACCS>EPFP4>ALOGP>CFP>GRFP.

Table 4. Rankings of various types of descriptors Based on Kendall W Test Results: Top 1&5%

Recall type	W	P	Ranking
Top 1 %	0.642	<0.01	LINGO-DOSM> CDKFP> MACCS> EPFP4 >PCFP>ALOGP>GRFP
Top 5%	0.473	<0.01	LINGO-DOSM >CDKFP>MACCS>EPFP4>ALOGP>PCFP>GRFP

Table 5. Numbers of Shaded Cells for Mean Recall of Actives Using Different Descriptors: Top 1% and 5%

	LINGO-DOSM>	GRF	PCFP	ALOGP	MACC	EPFP	CDKF
Top 1%	8	0	1	0	2	1	2
Top 5%	9	0	0	0	1	1	1

The good performance of LINGO-DOSM is not restricted to the top 5% for MDDR, since it also gives one the best results for the top-1% for MDDR. Using the mean recall value as an evaluation criterion could be impartial to some descriptor type but not others, and that is because some of the activity classes may contribute disproportionally to the overall value of mean recall. To avoid this bias, the effectiveness performance of different descriptors has been further investigated based on the total number of shaded cells for each descriptor across the full set of activity classes, as shown in the bottom rows of Tables 4. These shaded cell results are listed in Table 5.

Visual inspection of the results in Table 5 (left-hand column) shows very clearly that the LINGO-DOSM descriptor can provide a level of performance that is generally superior to the other descriptors. Finally, it should be noted here in this paper that the main purpose of using several types of descriptor in the experiments was not a performance comparison, but to show that our new descriptor LINGO-DOSM is capable of representing and characterizing the molecule structure, and to show the possibility and feasibility of its use for similarity-based virtual screening. However, the retrieval performance for any descriptor depends on the type of similarity approach used. Hence, we believe that using different text similarity searching approaches with the LINGO-DOSM descriptor will yield different results which may be much better than the current results.

5 Conclusions

In this paper, we present a new shape-based 2D molecular descriptor, LINGO-DOSM that represents a rough description of 2D molecular structure from its outline shape in a textual form. Experiments with the MDDR database show clearly the superiority of the LINGO-DOSM compared to many standard descriptors tested in this study. Experiments also show that the LINGO-DOSM allows for an effective screening search to be carried out.

Acknowledgment. This work is supported by Ministry of Higher Education (MOHE) and Research Management Centre (RMC) at the Universiti Teknologi Malaysia (UTM) under Research University Grant Category (VOT Q.J130000. 7826.4F011).

References

1. Agrafiotis, D.K., Myslik, J.C., Salemme, F.R.: Advances in diversity profiling and combinatorial series design. Mol. Diversity 4, 1–22 (1999)
2. Jorgensen, W.L.: The many roles of computation in drug discovery. Science 303 (2004)

3. Flower, D.R.: On the properties of bit-string-based measures of chemical similarity. J. Chem. Inf. Comput. Sci. 38, 379–386 (1998)
4. Willett, P., Barnard, J.M., Downs, G.M.: Chemical similarity searching. J. Chem. Inf. Comput. Sci. 38, 983–996 (1998)
5. Brown, R.D.: Descriptors for Diversity Analysis. Perspect. Drug. Discovery Des. 7/8, 31–49 (1997)
6. David, V., Michael, T., Miquel, P.: LINGO, an Efficient Holographic Text Based Method To Calculate Biophysical Properties and Intermolecular Similarities. J. Chem. Inf. Model. 45, 386–393 (2005)
7. UNITY Reference Manual, Tripos Inc., St. Louis, MO (1995)
8. Winkler, D.A., Burden, F.R.: Holographic QSAR of benzodiazepine. Quant. Struct.-Act. Relat. 17, 224–231 (1998)
9. Leach, A.R., Gillet, V.J.: An Introduction to Chemoinformatics. Kluwer, Dordrecht (2003)
10. Wild, D.J., Willett, P.: Similarity Searching in Files of Three-Dimensional Chemical Structures. Alignment of Molecular Electrostatic Potential Fields with a Genetic Algorithm. J. Chem. Inf. Comput. Sci. 36, 159–167 (1996)
11. Kirchmair, J., Distinto, S., Markt, P., Schuster, D., Spitzer, G.M., Liedl, K.R., Wolber, G.: How To Optimize Shape-Based Virtual Screening: Choosing the Right Query and Including Chemical Information. J. Chem. Inf. Model. 49, 678–692 (2009)
12. Rush, T.S., Grant, J.A., Mosyak, L., Nicholls, A.: A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein–Protein Interaction. J. Med. Chem. 48, 1489–1495 (2005)
13. Warr, W.A.: Representation of chemical structures. Wiley Interdisciplinary Reviews: Computational Molecular Science 1, 557–579 (2011)
14. Hall, L.H., Kier, L.B.: Issues in representation of molecular structure: The development of molecular connectivity. J. Mol. Graph. 20, 4–18 (2001)
15. Kogej, T., Engkvist, O., Blomberg, N., Muresan, S.: Multifingerprint Based Similarity Searches for Targeted Class Compound Selection. J. Chem. Inf. Model. 46, 1201–1213 (2006)
16. Weininger, D.: SMILES, A chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Comp. Sci. 28, 31–36 (1988)
17. SciTegicAccelrys Inc.
18. Yap, C.W.: PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. 32, 1466–1474 (2011)
19. Abdo, A., Chen, B., Mueller, C., Salim, N., Willett, P.: Ligand-Based Virtual Screening Using Bayesian Networks. J. Chem. Inf. Model. 50, 1012–1020 (2010)
20. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402 (1997)
21. Brown, R.D., Martin, Y.C.: Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. J. Chem. Inf. Comput. Sci. 36, 572–584 (1996)
22. Siegel, S., Castellan, N.J.: Nonparametric Statistics for The Behavioral Sciences. McGraw-Hill, New York (1988)

Prediction of Mouse Senescence from HE-Stain Liver Images Using an Ensemble SVM Classifier

Hui-Ling Huang[1,2], Ming-Hsin Hsu[2], Hua-Chin Lee[1], Phasit Charoenkwan[2], Shinn-Jang Ho[3], and Shinn-Ying Ho[1,2,*]

[1] Department of Biological Science and Technology,
National Chiao Tung University, Hsinchu, Taiwan
[2] Institute of Bioinformatics and Systems Biology,
National Chiao Tung University, Hsinchu, Taiwan
[3] Department of Automation Engineering,
National Huwei Institute of Technology, Yunlin, Taiwan
hlhuang@mail.nctu.edu.tw, morris679@yahoo.com.tw,
huachinlee@g2.nctu.edu.tw, phaznexus@hotmail.com,
sjho@nfu.edu.tw, syho@mail.nctu.edu.tw

Abstract. Study of cellular senescence from images in molecular level plays an important role in understanding the molecular basis of ageing. It is desirable to know the morphological variation between young and senescent cells. This study proposes an ensemble support vector machine (SVM) based classifier with a novel set of image features to predict mouse senescence from HE-stain liver images categorized into four classes. For the across-subject prediction that all images of the same mouse are divided into training and test images, the test accuracy is as high as 97.01% by selecting an optimal set of informative image features using an intelligent genetic algorithm. For the leave-one-subject-out prediction that the test mouse is not involved in the training images of 20 mice, we identified eight informative feature sets and established eight SVM classifiers with a single feature set. The best accuracy of using an SVM classifier is 71.73% and the ensemble classifier consisting of these eight SVM classifiers can advance performance with accuracy of 80.95%. The best two feature sets are the gray level correlation matrix for describing texture and Haralick texture set, which are good morphological features in studying cellular senescence.

Keywords: Aging, cellular senescence, feature selection, genetic algorithm, HE-stain, image analysis, prediction, SVM.

1 Introduction

Research of cellular senescence has been studied for more than 40 years. There are many literatures which point the differences between young and senescent cells in molecular level. In morphological way, the changes on surface of senescent cells are observed. With quantification of variation, seldom studies were proposed to solve the problem of identifying senescence from images in molecular level. Predicting the hierarchy of senescence by HE-stain images plays an important role in studying cellular senescence.

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 325–334, 2013.

An increasing proportion of the healthcare budget is devoted to the growing geriatric population, so it is essential to understand the molecular basis of ageing and identify possible avenues for therapeutic intervention [1]. Cellular senescence has been studied since the research [2] which showed that the senescent cells enlarged more than twofold relative to non-senescent counterpart. In addition to losing the ability to divide, cells in the senescent state exhibit dramatic alterations in morphology, mass, and dynamics of their subcellular organelles, and thereby display structural and functional differences compared to proliferating cells. These differences include an enlarged and flat cellular morphology, increased reactive oxygen species (ROS) production [3], and the accumulation of resultant ROS-mediated damage products such as: (1) lipofuscins and granular particles [4], (2) altered mass and functionality of mitochondria and lysosomes [5], and (3) certain cytosolic and nuclear markers such as senescence associated-b-galactosidase activity (SA-b-Gal) and senescence-associated heterochromatin foci (SAHF) [6].

For predicting mouse senescence and further analyzing the molecular basis of ageing, we used HE-stain liver images categorized into four classes (1 month, 6 months, 16 months, and 24 months) obtained from a public dataset. To design an accurate classifier for senescence prediction, there are three essential tasks: 1) identification of an informative image feature set, 2) determination of selecting an efficient classifier, and 3) development of an effective strategy for coping with the leave-one-subject-out prediction considering the variation of mouse aging.

Considering the aforementioned three tasks, we propose an optimized image feature selection method using an intelligent genetic algorithm for automatically identifying an informative image feature set. At first, we established a number of feature sets which maybe relate to morphology of senescent cells. The feature selection and classifier design are done at the same time. By comparing the widely-used support vector machine (SVM) with the selected feature set, SVM is adopted in the consequent study. Finally, this study proposes an SVM-based classifier with a novel set of grey-level image features to predict mouse senescence from HE-stain liver images.

2 Related Work

In these alterations showed in senescent cells, some could be displayed by hematoxylin and eosin staining, i.e. lipofuscin [7]. The accumulations of highly cross-linked protein are thought to relate to chronic oxidative stress and a failure to degrade damaged and denatured proteins [8]. Besides the highly oxidized insoluble proteins, senescence-induced nuclear defects resulted from accumulation of lamin A/C show the differences between young and old individuals [9]. In morphological perception, the large cell change frequently found in liver biopsy specimens from old subjects would represent senescent hepatocytes with HE stain [10].

Pasquinelli et al verified the possible influence of age in healthy liver parenchyma on DwI-related parameters: apparent diffusion coefficient, perfusion fraction, diffusion and pseudodiffusion coefficient [11]. Fonseca et al identified structural remodeling in human saphenous veins by applying histochemistry, fluorescence staining and

quantitative image analysis to specifically assess intimal area, intimal cellularity and intimal collagen content and organization [12]. In the study of Udono et al, they mentioned that high content screening (HCS)-based image analysis is becoming an important and powerful research tool. An automated and quantitative cellular image-analysis system was employed in their study to quantify the cellular senescence phenotypes induced in normal human diploid fibroblasts, TIG-1 cells, and found to be a powerful tool in the cellular senescence study [13].

On the other hand, there are some literatures that proposed a computational method for both quantitatively predicting and analyzing. Driscoll et al show a novel, automated and high throughput nuclear shape analysis that quantitatively measures curvature, area, perimeter, eccentricity and additional metrics of nuclear morphology for large populations of cells [14]. Choi et al used automated segmentation and shape analyses, with pre-defined features and with computer generated components, to compare nuclei from various premature aging disorders caused by alterations in nuclear proteins [15]. Shamir et al proposed to use the image texture entropy as an objective measurement that reflects the structural deterioration of the C. elegans muscle tissues during aging [16]. Johnston et al considered changes in overall morphology to quantitatively track tissue architecture during adulthood and aging in the C. elegans pharynx, the neuromuscular feeding organ [17]. All of the related works are listed in Table 1.

Table 1. Related work of studying cellular senescence

References	Aging related	Main type of method	Prediction	Objects of study
[11]	Yes	Biological	No	Human saphenous vein
[12]	Yes	Biological	No	Human diploid fibroblast
[13]	Yes	Biological and mathematical	No	Human liver parenchymas
[14]	Yes	Biological	No	Human diploid fibroblast and mouse tail fibroblast
[15]	Yes	Biological and computational	No	Human dermal fibroblast
[16]	Yes	Computational	No	Nuclei of human cells and mouse model
[17]	Yes	Computational	No	Pharynx of C. elegans
Ours	Yes	Computational	Yes	Mouse liver biopsy specimen images of HE stain

3 Design Procedure

The original HE-stain images were preprocessed such as noise remove and grey level transformation. From Fig. 1, the core technique IBCGA and SVM with feature selection were used for determining the optimized feature set with best accuracy. The feature extraction tool is developed for extracting the features from the bio-medical or tissue images by using the inheritable bi-objective combinatorial genetic algorithm (IBCGA) [18][19] as the core technique. The feature extraction tool designed by authors is a general purpose tool in the image informatics field and can automatically extract the image features categorized into three parts: shape features, texture features and wavelet features. The shape features include boundary moment, Zernike moment,

Morphological, Regional property and Tchebichef moment. The texture features include Haralick, Gray level co-occurrence matrix (GLCM), Gabor filter, Tamura, Granularity, Intensity and Fourier; wavelet feature includes Haar and Daubichies. In addition, each feature set has its own parameter setting to be defined depending on usages and applications, such as degree, radius, histogram bin, max distance or distance gap. For different predicting aims, across subject and leave one subject out are designed to prediction and analysis HE-stain images.

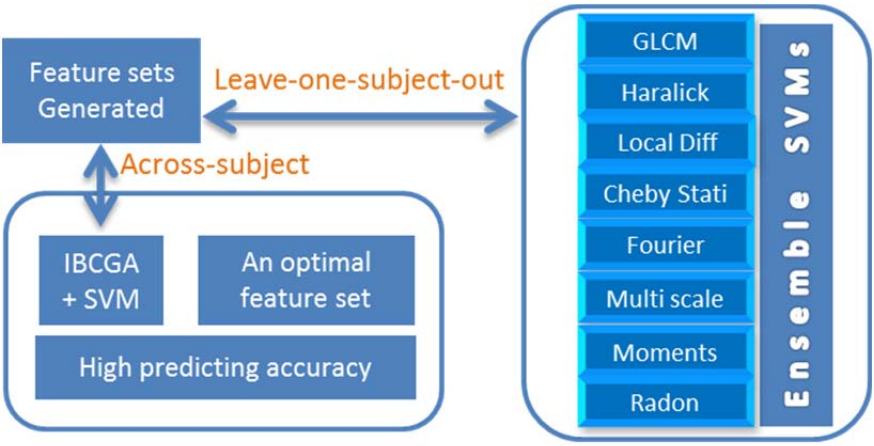


Fig. 1. Design procedures of the system. The proposed feature sets are used to design and evaluate the proposed method.

3.1 Selecting an Informative Feature Set

The core of IBCGA optimizes classification accuracy and minimizes number of features [18]. The intelligent genetic algorithm uses a divide-and-conquer strategy and an orthogonal array crossover to efficiently solve large-scale parameter optimization problems. IBCGA can efficiently explore and exploit the search space of $C(n, r) = n!/[(n-r)!r!]$. IBCGA can efficiently search the space of $C(n, r \pm 1)$ by inheriting a good solution in the space of $C(n, r)$ [18]. The proposed chromosome encoding scheme of IBCGA consists of both binary genes for feature selection and parametric genes for tuning SVM parameters.

The fitness function is the guide for IBCGA to acquire solutions with best performance. The fitness function of IBCGA is the 5-CV overall accuracy. IBCGA with the fitness function $f(X)$ can simultaneously obtain a set of solutions, X_r, where $r=r_{start}$, $r_{start}+1$, …, r_{end} in a single run. The algorithm of IBCGA with the given values r_{start} and r_{end} is described as follows:

Step 1) (Initiation) Randomly generate the initial population of N_{pop} individuals. All the n binary genes have r 1's and n-r 0's where $r = r_{start}$.

Step 2) (Evaluation) Evaluate the values of fitness for all individuals using $f(X)$.

Step 3) (Selection) Use the traditional tournament selection which chooses the winner from two randomly selected individuals to form a mating pool.

Step 4) (Crossover) Select $p_c \cdot N_{pop}$ parents from the mating pool and perform orthogonal array crossover on the selected pairs of parents where p_c is the crossover probability.

Step 5) (Mutation) Apply the swap mutation operator to the randomly selected $p_m \cdot N_{pop}$ individuals in the new population where p_m is the mutation probability. To prevent the best fitness value from deteriorating, mutation is not applied to the best individual.

Step 6) (Termination test) If the stopping condition for obtaining the solution X_r is satisfied, output the best individual as X_r. Otherwise, go to Step 2). In this study, the stopping condition is to perform 60 generations.

Step 7) (Inheritance) If $r < r_{end}$, randomly change one bit in the binary genes for each individual from 0 to 1; increase the number r by one, and go to Step 2). Otherwise, stop the algorithm.

3.2 Ensemble SVM Classifiers

In this study, the number of mice is relatively small, compared with the complex of recognition problems in the HE-stain liver images. For coping with the unknown subject problem, the ensemble SVM classifier consisting of k SVM classifiers and a voting method is developed to advance the prediction accuracy.

(1) SVM classifiers: The prediction accuracy is highly related to the selected feature set for every SVM classifier. The used types of features include: 1) GLCM, 2) Haralick, 3) Local Diff, 4) Cheby Statis, 5), Fourier 6) Multi scale, 7) Moments, and 8) Radon. Each SVM classifier uses one type of features.

(2) Voting method: Different classification results of the query sequences will be obtained from the outputs of the k independent SVM classifiers, and then these results are integrated using the simple voting method.

$$\mathrm{FS}_j = \sum_{i=1}^{k} \tau, \quad \tau = \begin{cases} 1, & C_i = j \\ 0, & otherwise \end{cases}, \tag{1}$$

where k (=8 in this study) is the number of SVM classifiers, j=1, 2, …, C (C=4 in this study) is the class label, C_i is the predicted class label by the ith SVM classifier. The final class is determined by argmax $\{\mathrm{FS}_1, \mathrm{FS}_2, \mathrm{FS}_3, \mathrm{FS}_4\}$.

4 Dataset

The mouse liver biopsy specimen images are provided by IICBU 2008 [19]. Fifty color images per liver were manually acquired using a Carl Zeiss Axiovert 200

microscope and 40x objective. Therefore, 1500 images from 30 livers were collected, and each image was converted into a gray-scale TIFF image. We use both male- and female-mouse of 1 month, 6 months, 16 months and 24 months in ad-libitum diet (Table 2), and all the images were HE-stain images. A single mouse has prepared all imaging and staining for leading to a small variability. Finally, there are 1027 images from 21 mice to be used for analysis.

In this study, two prediction methods are investigated for understanding the informative image features and the relationship between cell morphology and cellular senescence. For the across-subject prediction that all images of the same mouse are randomly divided into training (5/6) and test images (1/6). For the leave-one-subject-out prediction that the test mouse is not involved in the training images of 20 mice. The prediction accuracy is the mean of 21 predictions where each mouse is served as a test mouse.

Table 2. The number of individuals in each classes and number of images for each individauls

	Class no.										
	1		2			3			4		
Individual	4		6			5			6		
Images	50	50	50	50	15	50	61	51	50	51	50
	50	50	50	50	50	50	50		49	50	50

5 Results

5.1 Across-Subject Prediction

Table 3 shows the results of 30 independent runs. The averaged training and test accuracies are 98.80% and 97.01%, respectively. With the same dataset, the results of using the existing Wnd-charm feature set are shown in Table 4. The results show that the averaged training and test accuracies are 93% and 88%, respectively. The result reveals that the proposed method is better than the Wnd-charm method due to the feature selection of IBCGA. In the Wnd-charm method, the 601 features were chosen from default 4008 characteristics by using the Fisher score. However, IBCGA chooses 89.5 features on average as a feature set. From the features selected in the 30 runs (Table 4), there are four features which appear more than 15 times. The top 4 frequently-selected features are listed in Table 5.

Table 3. Results of 30 independent runs with the number of features selected by IBCGA

Run number	Training accuracy	Test accuracy	No. of elected features
1	99.07%	97.56%	42
2	98.84%	96.34%	94
3	98.38%	96.95%	133
4	99.19%	95.12%	75
5	98.03%	96.95%	111
6	98.73%	96.95%	57
7	98.61%	96.95%	74
8	98.61%	96.34%	38
9	98.96%	96.95%	139
10	98.84%	98.78%	100
11	98.73%	97.56%	78
12	98.73%	97.56%	88
13	98.49%	98.78%	123
14	98.84%	97.56%	104
15	98.03%	96.34%	69
16	98.73%	97.56%	148
17	98.73%	97.56%	111
18	99.07%	98.17%	39
19	99.07%	98.17%	49
20	98.38%	96.34%	98
21	99.30%	98.17%	33
22	98.73%	95.73%	149
23	99.19%	95.12%	40
24	99.07%	95.73%	52
25	99.54%	97.56%	141
26	98.84%	96.34%	133
27	98.61%	95.12%	114
28	99.19%	97.56%	47
29	98.84%	95.73%	79
30	98.73%	98.78%	126
Average	98.80%	97.01%	89.5

Table 4. The training and test results by using the Wnd-charm method

	Training	Independent test
Accuracy	93%	88%

Table 5. The top 4 feature sets with the highest selection frequency

Feature name (feature ID)	No. of appearances in the feature selection
Gabor_Standard deviation 3 (982)	16
Debeucies4_'W_h6' (1025)	16
Granularity_Standard deviation pixel distance 1 (1071)	19
Granularity_Standard deviation pixel distance 3 (1073)	17

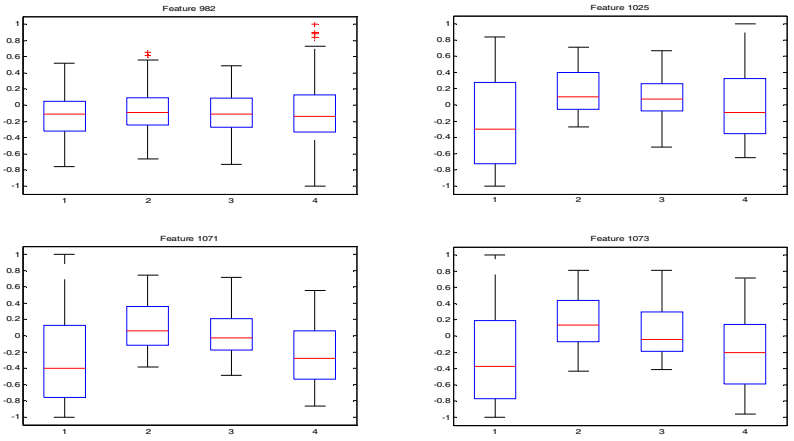


Fig. 2. Box plots of the four top-ranked features

Fig. 2 shows the box plot statistics of the top 4 features, and two of the top 4 features (features 1071 and 1073) are the texture characteristics from the dataset.

5.2 Feature Selection

IBCGA can automatically choose feature sets with preferred sizes from all candidate features. Fig. 3 shows the feature selection results with accuracy. Finally, we choose 10 features as one feature set with accuracy of 93.29%.

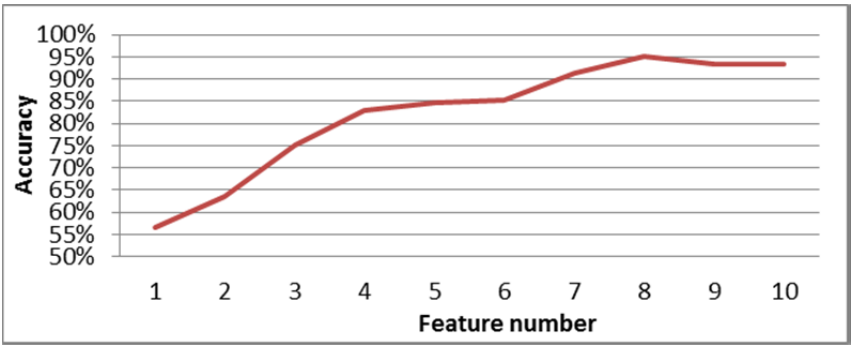


Fig. 3. Feature selection from 1 to 10 features using IBCGA

5.3 Ensemble Classifier

For the leave-one-subject-out prediction that the test mouse is not involved in the training images of 20 mice, we identified eight informative feature sets and established eight SVM classifiers with a single feature set, shown in Table 6. The best accuracy of using an SVM classifier is 61.90% and the ensemble classifier consisting of these eight SVM classifiers can advance performance with accuracy of 80.95%.

The best two feature sets are the gray level correlation matrix for describing texture and Haralick texture set, which are good morphological features in studying cellular senescence.

Table 6. Leave-one-subject-out prediction results

Feature	GLCM	Haralick	Local Diff.	Cheby Statis
ACC	61.90%	57.14%	19.05%	47.62%

Feature	Cheby Fourier	Multi scale	First 4 moments	Radon	Ensemble
ACC	19.05%	38.10%	28.57%	33.33%	80.95%

The two best features are discussed as follows:

1. Gray level correlation matrices (GLCM)
GLCM set was used for describing texture of tissue [21]. It could express the relation between pixels. The probability value in GLCM will directly use as texture features. The level for histogram with 32 bins is usually the multiple times of 8. So here we set level to 8 and maximum distance to 5, which allows the total feature number to 1280.

2. Haralick texture set
Haralick texture set was published by Haralick, R.M. at 1973 [21]. Haralick defines fourteen statistical measurements calculated from GLCM. In this study, additional four statistical formulae are also used. Because it is based on GLCM, so the parameter settings are the same with those of GLCM.

6 Conclusions

We have proposed an ensemble support vector machine (SVM) based classifier with a novel set of image features to predict mouse senescence from HE-stain liver images categorized into four classes. The informative image features are obtained using the optimized feature selection algorithm IBCGA. The GLCM and GLCM-based Haralick texture set are good morphological features in studying cellular senescence. The ensemble SVM classifier performs well, compared with existing methods. For the leave-one-subject-out prediction, the ensemble approach can efficiently advance accuracy, compared with the non-ensemble approach. This study investigates classifiers and various sets of features on grey-level images. The future work is to investigate color-based feature sets for further analyzing the senescent cells.

References

1. Martin, J.E., Sheaff, M.T.: The pathology of ageing: concepts and mechanisms. J. Pathol 211(2), 111–113 (2007)
2. Hayflick, L.: The Limited in Vitro Lifetime of Human Diploid Cell Strains. Exp. Cell Res. 37, 614–636 (1965)
3. Kurz, T., et al.: Lysosomes and oxidative stress in aging and apoptosis. Biochim. Biophys. Acta 1780(11), 1291–1303 (2008)

4. Schmucker, D.L., Sachs, H.: Quantifying dense bodies and lipofuscin during aging: a morphologist's perspective. Arch Gerontol Geriatr 34(3), 249–261 (2002)
5. Terman, A., et al.: Mitochondrial recycling and aging of cardiac myocytes: the role of autophagocytosis. Exp. Gerontol. 38(8), 863–876 (2003)
6. Braig, M., et al.: Oncogene-induced senescence as an initial barrier in lymphoma development. Nature 436(7051), 660–665 (2005)
7. Hoare, M., Das, T., Alexander, G.: Ageing, telomeres, senescence, and liver injury. J. Hepatol. 53(5), 950–961 (2010)
8. Jung, T., Bader, N., Grune, T.: Lipofuscin: formation, distribution, and metabolic consequences. Ann. N. Y. Acad. Sci. 1119, 97–111 (2007)
9. Scaffidi, P., Misteli, T.: Lamin A-dependent nuclear defects in human aging. Science 312(5776), 1059–1063 (2006)
10. Ikeda, H., et al.: Large cell change of hepatocytes in chronic viral hepatitis represents a senescent-related lesion. Human Pathology 40(12), 1774–1782 (2009)
11. Pasquinelli, F., et al.: Magnetic resonance diffusion-weighted imaging: quantitative evaluation of age-related changes in healthy liver parenchyma. Magnetic Resonance Imaging 29(6), 805–812 (2011)
12. Fonseca, C., et al.: The effects of aging on the intimal region of the human saphenous vein: insights from multimodal microscopy and quantitative image analysis. Histochem. Cell Biol. (2012)
13. Udono, M., et al.: Quantitative analysis of cellular senescence phenotypes using an imaging cytometer. Methods 56(3), 383–388 (2012)
14. Driscoll, M.K., et al.: Automated image analysis of nuclear shape: what can we learn from a prematurely aged cell? Aging (Albany NY) 4(2), 119–132 (2012)
15. Choi, S., et al.: Computational image analysis of nuclear morphology associated with various nuclear-specific aging disorders. Nucleus 2(6), 570–579 (2011)
16. Shamir, L., Wolkow, C.A., Goldberg, I.G.: Quantitative measurement of aging using image texture entropy. Bioinformatics 25(23), 3060–3063 (2009)
17. Johnston, J., et al.: Quantitative Image Analysis Reveals Distinct Structural Transitions during Aging in Caenorhabditis elegans Tissues. PLoS One 3(7) (2008)
18. Ho, S.-Y., Chen, J.-H., Huang, M.-H.: Inheritable genetic algorithm for biobjective 0/1 combinatorial optimization problems and its applications. IEEE Trans. Syst. Man Cybern. B Cybern. 34(1), 609–620 (2004)
19. Ho, S.-Y., Shu, L.-S., Chen, J.-H.: Intelligent evolutionary algorithms for large parameter optimization problems. IEEE Trans. on Evol. Comp. 8(6), 522–541 (2004)
20. Shamir, L., et al.: IICBU 2008: a proposed benchmark suite for biological image analysis. Med. Biol. Eng. Comput. 46(9), 943–947 (2008)
21. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Trans. on Systems, Man, and Cybernetics 6, 269–285 (1973)

An Introduction to Yoyo Blind Man Algorithm (YOYO-BMA)

Mohammad Amin Soltani-Sarvestani[1] and Shahriar Lotfi[2]

[1] Computer Engineering Department, University College of Nabi Akram, Tabriz, Iran
soltani_mohammadamin@yahoo.com
[2] Computer Science Department, University of Tabriz, Tabriz, Iran
shahriar_lotfi@tabrizu.ac.ir

Abstract. In this paper, a new algorithm is proposed which is inspired by human intelligence named YOYO Blind Man Algorithms (YOYO-BMA). The main idea of YOYO-BMA is the combination of human intelligence with features of yoyo. In the proposed algorithm, it is supposed that there are some men in a dark room, which are named *blind men*. They look for the optimum. Each man has at least a yoyo to use as assistant. Men search problem space using their yoyos. This new algorithm is compared with 5 other different algorithms and the results show the better performance of YOYO-BMA compared with the other ones.

Keywords: Optimization, Evolutionary Algorithm, YOYO Blind Man Algorithm, YOYO-BMA, Blindness Operator, Human intelligence, yoyo, Meta-heuristic algorithm.

1 Introduction

Evolutionary algorithms (EA) [1] are inspired by Darwin's evolution theory. They have many applications to solve NP-hard[1] problems in various fields of science. Some of the well-known proposed evolutionary algorithms are as follow; First, genetic algorithms that inspired by natural genetic variation and natural selection [2, 8], second, simulated annealing is a technique that simulates the annealing process [7]. Third, the other algorithm, particle swarm optimization (PSO) is proposed by Edward and Kennedy in 1995 which is inspired by social intelligence of animals, such as bird flocking or fish schooling [6]. Artificial bee colony algorithm is based on the intelligence of honey bee swarm [5]. The evolutionary algorithms are inspired by nature; while the YOYO-BMA is inspired by human intelligence .It is believed that the power of human intelligence is stronger than evolution. So, the algorithms inspired by human intelligence may have better performance compared to evolutionary algorithms. This paper is the first one that introduces an algorithm used human intelligences as its operators directly. This paper proposes a new class of algorithms for optimization which are inspired by human intelligence, unlike the

[1] Non Polynomial hard.

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 335–344, 2013.

evolutionary algorithms which are inspired by the nature. The YOYO-BMA is combinations of human intelligence and features of yoyo. In fact, some men try to solve a problem using their talent and some assistant objects. The new algorithm is called YOYO Blind Man Algorithm (YOYO-BMA).

The rest of this paper is arranged in this order; section 2 is devoted to explain the YOYO-BMA. The YOYO-BMA is tested with some benchmark functions and analyzed in section 3, and finally the conclusion is presented in section 4.

2 The Yoyo Blind Man Algorithm

In this part, the YOYO-BMA is introduced. YOYO-BMA is a new optimization algorithm which is inspired by human intelligence used features of yoyo to solve problems. In the proposed algorithm, it is supposed that there are some men in a dark room. There is no light in the room and men can't see their environment. So, we call them *blind men*, but they are not blind necessarily. Each man has some instruments in his hands which are used to search in the room. They have to find optimum using their instruments. Instruments are some utilities which guide men to find optima. Instrument can be any arbitrary object such as Stick, Billy, Cane, Ball, Rope, Yoyo and etc. In this paper, we plan to use yoyo as men's instruments because of its flexible movement. In fact, yoyo is selected only as an example.

Figure 1 illustrates the flowchart of Yoyo-BMA. The algorithm is explained in this part in order to the flowchart. The BMA starts with an initial random population. Each individual is a $1 \times D$ array, each element of which is corresponded to one dimension of problem. The individuals are defined in this order

$$Individual = (I_1, I_2, ..., I_D) \tag{1}$$

D is the number of dimensions of the problem, and I_i indicates the i^{th} dimension of the problem. Next, the population should be classified into two categories; men and yoyos. Therefore, in order to divide the population, the cost of each individual is calculated based on a cost function.

$$Cost_i = Cost\_function(Individual_i(I_1, I_2, ..., I_D)) \tag{2}$$

$Cost_i$ reflects the cost of i^{th} individual. When the costs of individuals are calculated, N_{Men} of them with the least cost are selected as men and the rest of individuals are considered as yoyos. Among men, one with the least cost is selected as *boss.* In fact, *boss* is the most powerful individual. It is supposed that *boss* is probably near to the global optima. So, *boss* is the main leader of the algorithm. The ordinary men are useful to escape from local traps.

After that, yoyos distribute among men. In the proposed algorithm, the arbitrary numbers of hands are considered for each man. Therefore, yoyos are distributed randomly among men based on the number of their hands. These yoyos are called *hand-yoyos*. After distributing *hand-yoyos*, some extra yoyos are remained. The additional yoyos are possessed by *boss* called *hanging-yoyos.* This kind of yoyos are located around the *boss* in order to explore his neighbors more than other spaces.

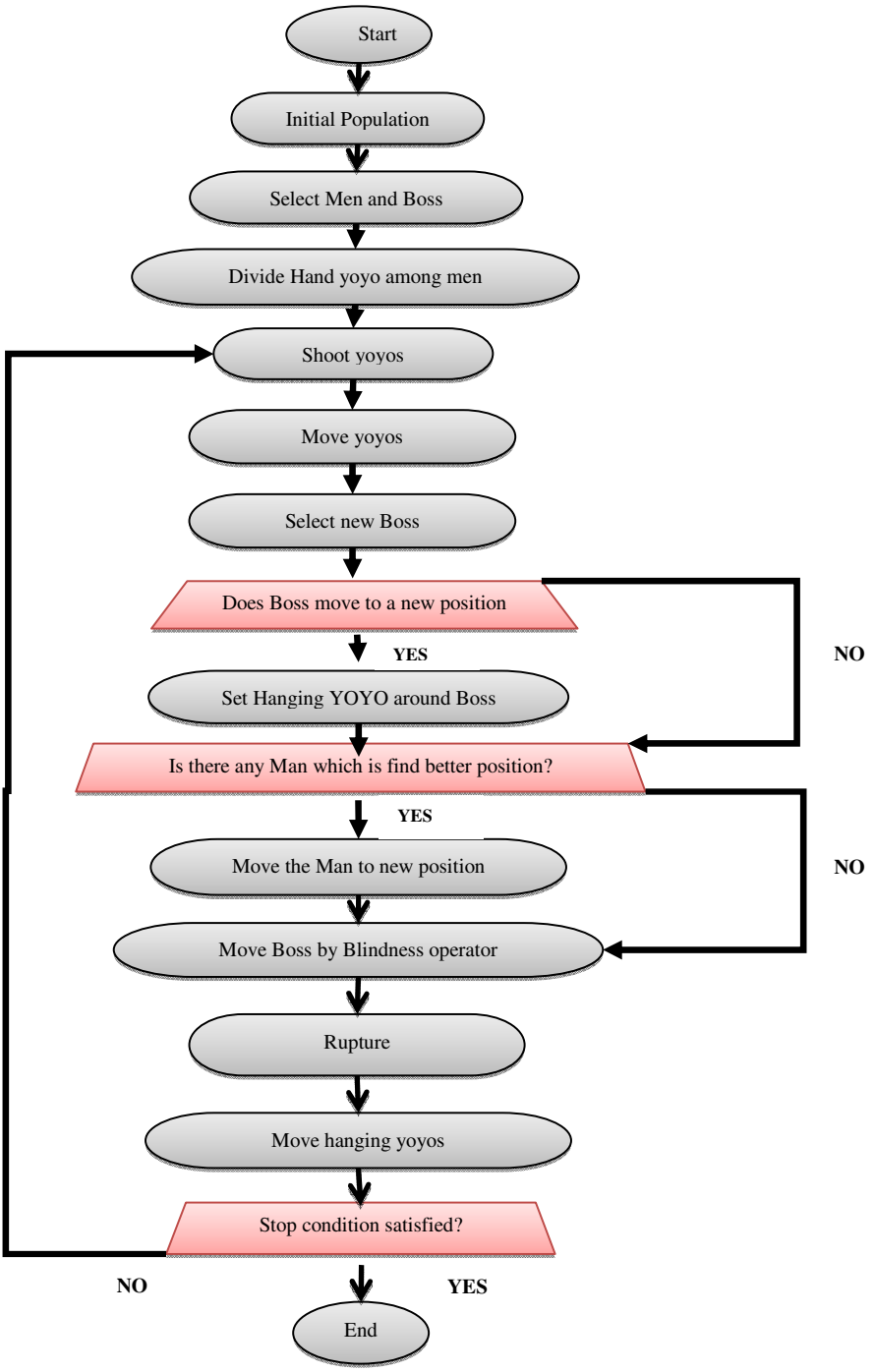


Fig. 1. The flowchart of YOYO-BMA

Then yoyos start to move toward men. But, it is essential to explain some key points, before defining the yoyos movement. A yoyo has two different directions of movement. Sometimes it gets far away from its player and sometimes it gets closer. Therefore, a player first shoots yoyo, and then yoyo move back toward the player. In fact, yoyos move to explore different directions in the search space. Figure 2 illustrates yoyo's different directions of movement.

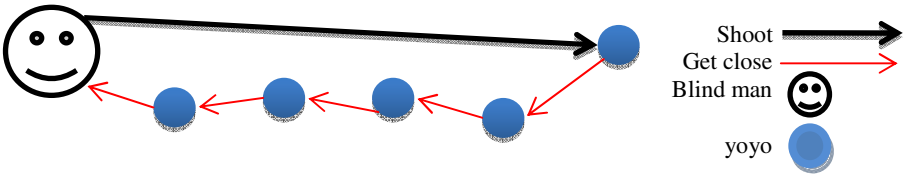


Fig. 2. Different direction of yoyo's movement

Yoyos are shot when they are distributed among men. In fact, when yoyos get close enough to men, each one shoots all his *hand*-yoyos together. The *hanging*-yoyos are not considered to be shot. The *shoot* operator is defined to disperse the population when they converge to men's position. The *shoot* operator acts in this order.

$$\begin{aligned} &if \quad Rand() < Shoot\_threshold \quad then \\ &yoyo_{\,i} = Random\_Position() \qquad i \in [1, N_{YOYO}] \end{aligned} \tag{3}$$

Rand() produces a random positive value less than 1, *Shoot_threshold* is an arbitrary positive number less than 1, *Random_Position()* is a function that sets yoyos coordinate randomly, and N_{YOYO} is the number of yoyos. In fact, *shoot* operator specifics a new random position for a yoyo.

As it mentioned, yoyos start to move toward their owner after shooting. While men are playing with their yoyos, their positions are motionless and they move yoyos through different dimensions. The movement of yoyos toward men is simulated in this order:

$$\begin{aligned} &dist = (yoyo_i(k) - Man_j(k)) \\ &move\_step = dist \times rand() \times Step\_size \\ &new\_yoyo_i(k) = yoyo_i(k) - move\_step \end{aligned} \tag{4}$$

$yoyo_i(k)$ is the k^{th} dimension of i^{th} yoyo, $Man_i(k)$ is the k^{th} dimension of j^{th} Man. It is considerable that $yoyo_i$ belonges to Man_j. *dist* indivates the distance between man and yoyo, *Step_size* is a positive number less than one, *move_step* indicates the maximum step size of yoyo movement. New_yoyo_i is the new position of i^{th} yoyo.

Since the *boss* play an important leadering role in the algorithm, he should be chosen precisely. In order to make a good choice, men are compared to each other and the best one is chosen as new *boss* in each iteration. New *boss* get the control of *hanging-yoyos*. The *boss* hangs up the *hanging*-yoyo around him in this order:

$$i = randi(D)$$
$$Hanging\_yoyo = Boss(I_1, I_2, ..., I_i + \alpha \times rand(), ..., I_D) \tag{5}$$

randi(D) is a function which returns an integer number over [1,*D*], α is an arbitrary value that determines the distance between the boss and *hanging_yoyo*, *rand()* is a function which returns a random float value over [0,1], and *hanging_yoyo* is an additional yoyo which is hung to the boss. In fact, coordinate of each *hanging_yoyo* is the same as the boss, except in one dimension. Therefore, each *hanging_yoyo* moves toward the boss just from this dimension. Suppose that this dimension is kth dimension.

$$\begin{aligned} dist &= (yoyo_i(k) - Boss(k)) \\ move\_step &= dist \times (rand() + Size\_step) \\ new\_yoyo_i(k) &= yoyo_i(k) + move\_step \end{aligned} \tag{6}$$

$yoyo_i(k)$ is the kth dimension of ith yoyo, *Boss(k)* is the kth dimension of boss. It is considerable that $yoyo_i$ belonges to *boss*. *dist* is the distance between boss and yoyo, *Step_size* is a positive number less than one, *move_step* indicates the extent of yoyo movement. New_yoyo_i is the new position of ith yoyo. The movement of *hanging-yoyo* and *hand-yoyo* are nearly the same. The difference is that *hand-yoyos* move on a curve, but *hanging-yoyos* move on a line. The other difference is that *hand-yoyos* move by longer steps compared to the *hanging-yoyos*.

It is possible for the algorithm to face situations in those a yoyo reaches a position with less cost compared to its owner. In this case, the owner changes his position to the position of the yoyo.

The process of moving is continued until the end of algorithm. In the middle of this process, the *blindness* and *rupture* operators are processing too. In the proposed algorithm, it is supposed that men are in a dark room, so they can't see the environment. Therefore they should act like a blind man when they start to move. The *blindness* operator is usually used only by boss, and there is no necessity to be used by other men. It is because that the goal of this operator is more exploration around the best obtained result. In order to understand the *blindness* operator action, close your eyes and try to walk. A blind man usually moves one of his legs around himself and looks for a safe step. When he finds a safe step, he takes one step straight ahead. Otherwise, he looks for a safe step in the other directions. In an optimization problem a position with less cost compared to the current position is considered as a safe place. So, a man should look for a less cost position compared to its current position as a safe step. In order to simulate the *blindness* operator, N_{Blind} of dimensions are selected randomly based on equation (7).

$$\begin{aligned} N_{Blind} &= Randi(N_{Dim} / 10 \times \omega) \\ selected\_Dim &= RandDim(N_{Blind}, N_{Dim}) \end{aligned} \tag{7}$$

N_{Blind} is the number of dimensions which are selected to take a step. *Randi(A)* is a function which returns a random integer value on [*1*, *A*]. ω is a positive number less than 100 which specifies the percentage of the number of dimensions through which the boss has to walk that determines the maximum number of N_{Blind}. *RandDim*(N_{Blind}, N_{Dim}) is a function which selects N_{Blind} integer random number over [*1, D*], and *selected_Dim* is a $1 \times N_{Blind}$ array which holds this random numbers. Then, N_{Blind} number of *blindness* steps are created based on equation (8).

$$\begin{aligned} Blind\_step_i &= Blindness(Boss, Selected\_Dim, \gamma, \lambda) \\ &= Boss[j] + (Boss[j] \times Rand(-1,1) \times \gamma) \wedge Randi(1, \lambda) \\ i &\in [1, N_{Blind}] \qquad j = Selected\_Dim(i, N_{Dim}) \end{aligned} \tag{8}$$

Blindness is an operator which calculates *Blind_Step*s. γ is step size, *Rand(-1, 1)* returns a value over [-1, 1], *Randi*(1, λ) returns a positive value less than λ, and *Boss[j]* indicates the j^{th} dimension of the boss. There are N_{Blind} different *Blind_Step*s, each moves through one dimension. *Safe_step*s are steps that satisfy equation (9). In fact, *Safe_step*s are steps which have less cost compared to the boss.

$$Safe\_step = \{Blind\_step(i) \mid \cos t\_function(Blind\_step(i)) < \cos t\_function(Boss)\} \tag{9}$$

When *Safe_steps* are calculated, only the best of them will be performed. The best *safe_step* is the least cost step in comparison with the other steps.

$$Best\_Safe\_step = Min \quad (\cos t\_function(Safe\_step)) \tag{10}$$

Finally the *Best_safe_step* will be applied by the boss. The *blindness* operator is the heart of the YOYO-BMA. In the proposed algorithm in order to search better and scape from local optima, there is a probability for each yoyo that may be ruptured. A yoyo move to a random position when it is ruptured. In other word a yoyo position is changed randomly if it is ruptured and the *rupture* operator act in the same way. This operator act the same as *mutation* in genetic algorithm.

The yoyo algorithm stops when its stop condition is satisfied. Stop condition can be anything, but in the proposed algorithm, it is considered as the number of generations. Figure 3 illustrates the pseudo code of YOYO-BMA.

YOYO Blind Man Algorithm

1. Create a random population
2. Divide population to *men* and *yoyos, and select Boss*
3. Distribute *yoyo* among *men*
4. *Shoot* yoyos
5. yoyos move toward their owner
6. select new *boss*
7. if *boss* move to a new position then set *hanging*-yoyos again around him
8. If a yoyo reach a better position than its owner's position, the owner move to *yoyo*'s position
9. *Move boss* by *blindness* operator
10. *rupture*
11. if the stop condition is not satisfied go to 4
12. end

Fig. 3. pseudo code of YOYO-BMA

3 Evaluation and Experimental Result

In this paper, a new algorithm is introduced, named YOYO Blind Man Algorithm. The proposed algorithm was tested on 20 benchmark functions provided by CEC2010 special session and competition on large-scale global optimization [11]. The benchmark suite includes separable, partially separable, and fully non-separable functions. Some of the used parameters are selected differently in a function from the other functions. The dimension of function was D=1000 and 25 runs of algorithm were needed for each function.

The obtained results (error values $f(x) - f(x^*)$) are presented in table 1. The algorithm tested with three Function Evaluations (FEs); 1.2E+5, 6.0E+5, and 3.0E+6. FEs acts nearly the same as generation. From the results, it can be seen that YOYO-BMA achieves desirable results on 6 functions, F1, F2, F3, F7, F12, and F17. The error of obtained results for these 6 functions is nearly zero. The proposed algorithm achieves acceptable results on 8 functions, F8, F10, F11, F13, F15, F16, F18, and F20.

We compare the mean results obtained by YOYO-BMA with the ones obtained by EOEA [12], MA-SW-Chains [9], DMS-PSO-SHS [14], DECC-G [13], and iDElsgo [4]. Table 2 presents a comparison on YOYO-BMA with 5 other algorithms. As can be seen, YOYO-BMA can find the best Result on eleven functions F1, F2, F7, F8, F9, F12, F13, F14, F17, F19, and F20, and it finds the absolute result in F1 and F2. As can be seen in table 2, YOYO-BMA can averagely achieve the best result.

Table 1. Results of YOYO-BMA

Metric		FEs= 1.2E+05	FEs= 6.0E+05	FEs= 3.0E+06	Metric		FEs= 1.2E+05	FEs= 6.0E+05	FEs= 3.0E+06
Fun1	Mean	4.82E+05	1.81E-17	0	Fun11	Mean	1.94E+02	1.94E+02	1.94E+02
	SD	2.69E+05	7.32E-17	0		SD	7.05E-01	7.05E-01	7.05E-01
Fun2	Mean	2.66E-11	0	0	Fun12	Mean	1.61E+03	2.77E-01	1.21E-09
	SD	3.54E-11	0	0		SD	3.77E+02	5.66E-02	1.15E-10
Fun3	Mean	4.38E-09	1.87E-12	1.62E-12	Fun13	Mean	2.32E+03	4.10E+02	3.36E+02
	SD	3.97E-09	2.56E-13	2.76E-13		SD	5.92E+03	5.97E+02	5.99E+02
Fun4	Mean	2.56E+12	4.68E+11	9.55E+10	Fun14	Mean	1.18E+08	3.73E+07	9.80E+06
	SD	4.40E+11	3.39E+11	9.52E+10		SD	1.59E+07	4.14E+06	2.17E+06
Fun5	Mean	4.23E+08	4.23E+08	4.23E+08	Fun15	Mean	8.08E+03	8.08E+03	8.08E+03
	SD	6.08E+07	6.08E+07	6.08E+07		SD	2.84E+02	2.84E+02	2.84E+02
Fun6	Mean	1.83E+07	1.79E+07	1.73E+07	Fun16	Mean	3.88E+02	3.88E+02	3.88E+02
	SD	7.23E+05	7.45E+05	9.49E+05		SD	9.91E-01	9.91E-01	9.91E-01
Fun7	Mean	1.06E+08	1.25E+03	3.22E-04	Fun17	Mean	1.19E+04	1.22E+01	8.23E-06
	SD	1.52E+08	6.73E+02	4.65E-05		SD	3.49E+03	6.97E+00	9.87E-06
Fun8	Mean	2.15E+07	3.02E+05	3.37E+02	Fun18	Mean	7.52E+04	5.01E+04	1.66E+03
	SD	3.23E+07	5.55E+05	3.93E+02		SD	1.00E+05	6.94E+04	1.71E+03
Fun9	Mean	4.99E+07	1.47E+07	4.33E+06	Fun19	Mean	1.58E+06	3.57E+05	2.13E+04
	SD	1.07E+07	4.43E+06	1.57E+06		SD	3.75E+05	4.76E+04	5.16E+03
Fun10	Mean	4.29E+03	4.29E+03	4.29E+03	Fun20	Mean	1.39E+03	1.11E+02	6.76E+00
	SD	2.92E+02	2.92E+02	2.92E+02		SD	1.72E+02	1.37E+02	1.17E+01

Table 2. Algorithm Comparison

	Error (average) at FEs=3.0E+06						Rank					
	Alg.1 EOEA	Alg. 2 MA-SW CHAINS	Alg. 3 iDElsgo	Alg. 4 Dms-PSO-SHS	Alg. 5 DECC-G	Alg. 6 YOYO-BMA	Fun 1	Fun 2	Fun 3	Fun 4	Fun 5	Fun 6
F1	2.20E-23	2.10E-14	8.86E-20	5.51E-15	2.86E-07	0	2	5	3	4	6	1
F2	3.62E-01	8.10E+02	1.25E-01	8.51E+01	1.31E+03	0	3	5	2	4	6	1
F3	1.67E-13	7.28E-13	3.81E-12	5.52E-11	1.39E+00	1.62E-12	1	2	4	5	6	3
F4	3.09E+12	3.53E+11	8.06E+10	2.45E+11	1.51E+13	9.55E+10	5	4	1	3	6	2
F5	2.24E+07	1.68E+08	9.72E+07	8.35E+07	2.38E+08	4.23E+08	1	4	3	2	5	6
F6	3.85E+06	8.14E+04	1.70E-08	8.27E-02	4.80E+06	1.73E+07	4	3	1	2	5	6
F7	1.24E+02	1.03E+02	1.31E-02	1.95E+03	1.07E+08	3.22E-04	4	3	2	5	6	1
F8	1.01E+07	1.41E+07	3.15E+06	1.29E+07	6.70E+07	3.37E+02	3	5	2	4	6	1
F9	4.63E+07	1.41E+07	3.11E+07	8.72E+06	3.18E+08	4.33E+06	5	3	4	2	6	1
F10	1.08E+03	2.07E+03	2.64E+03	5.53E+03	1.07E+04	4.29E+03	1	2	3	5	6	4
F11	3.86E+01	3.80E+01	2.20E+01	3.24E+01	2.33E+01	1.94E+02	5	4	1	3	2	6
F12	1.37E+04	3.62E-06	1.21E+04	6.12E+02	8.87E+04	1.21E-09	5	2	4	3	6	1
F13	1.24E+03	1.25E+03	7.11E+02	1.12E+03	3.00E+03	3.36E+02	4	5	2	3	6	1
F14	1.65E+08	3.11E+07	1.69E+08	1.75E+07	8.07E+08	9.80E+06	4	3	5	2	6	1
F15	2.14E+03	2.74E+03	5.84E+03	4.07E+03	1.18E+04	8.08E+03	1	2	4	3	6	5
F16	8.26E+01	9.98E+01	1.44E+02	9.67E+01	7.51E+01	3.88E+02	2	4	5	3	1	6
F17	7.93E+04	1.24E+00	1.02E+05	3.83E+03	2.89E+05	8.23E-06	4	2	5	3	6	1
F18	2.94E+03	1.30E+03	1.85E+03	2.25E+03	2.30E+04	1.66E+03	5	1	3	4	6	2
F19	1.84E+06	2.85E+05	2.74E+05	1.16E+06	1.11E+06	2.13E+04	6	3	2	5	4	1
F20	1.97E+03	1.07E+03	1.53E+03	3.51E+02	3.98E+03	6.76E+00	5	3	4	2	6	1
						Average	**3.5**	**3.25**	**3**	**3.35**	**5.35**	**2.55**

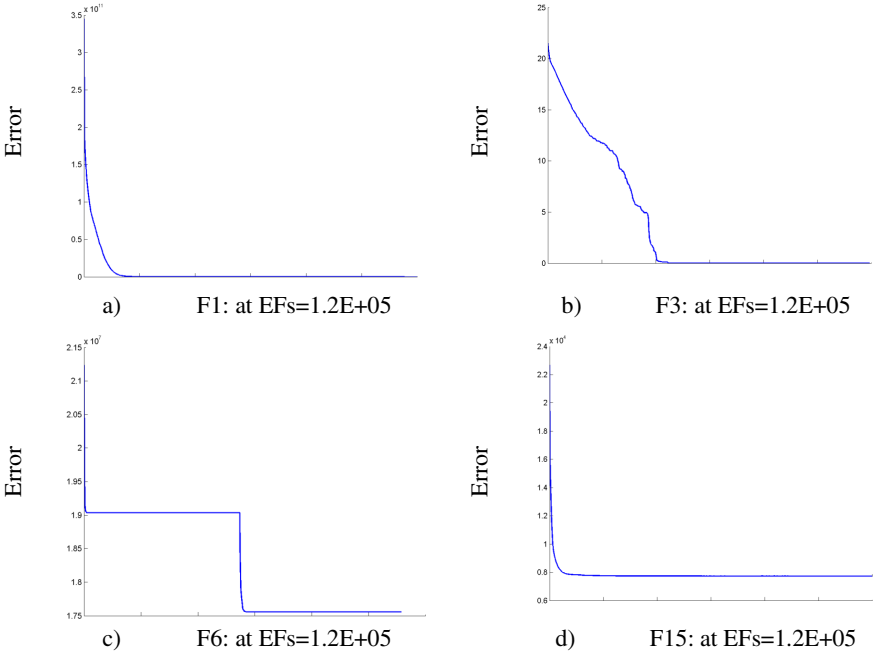


Fig. 4. The Convergence graph at FES=1.2E+05

The convergence curves of YOYO-BMA for the following four selected problems: F1, F3, F6, and F15, are presented in figure 4. The convergence graphs for the other benchmarks were nearly the same as these four graphs. As can be seen in figure 4, the convergence speed of YOYO-BMA algorithm is very fast and it can achieve its result in a desirable time. In this paper the convergence graphs is presented at FEs= 1.2E+05 because the speed of convergence is high and the graphs at EFs=3.0E+06 are not obvious.

From the obtained results, it can be seen that YOYO-BMA achieves the desirable results for the scalable group, but the results of the fully-nonseparable group are not that good. But, by comparing the results of fully-nonseparable group obtained by YOYO-BMA with the other algorithm, it can be seen that YOYO-BMA achieves better results.

4 Conclusion and Future Works

The presented paper introduced a new algorithm inspired directly by human intelligence which is called YOYO-BMA. The proposed algorithm supposes that there are some men in a dark room, and they look for optima. The men are not blind necessarily, but they cannot see their environment, because of which they are called *blind men*. In other word, YOYO-BMA is an algorithm which simulates treatment of human in lightless situations. There are some assistant yoyos possessed by men used to find optima. In fact, the proposed algorithm gets the benefit of the entire intelligence of human. Using their mind power and their yoyos, men try to find optima. The results show that the proposed algorithm solved all benchmark suites with high consistency. The results of YOYO-BMA were compared with 5 different algorithms. From the comparison, it can be seen that YOYO-BMA can averagely achieve better results compared to the other algorithms in all functions. Therefore, it can be seen that maping the human intelligence in search algorithm may lead to better results compared to the other intelligence methods such as evolutionary algorithm inspired from nature.

In the experiments, we consider yoyos as men assistant instruments, but there are many different instruments in the real world. By defining new assistants for men or even using a combination of difference assistants, some new algorithms can be proposed. More investigations will be conducted to define new instruments for the men in the future works.

References

1. Back, T.: Evolutionary Algorithm in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms. Oxford University Press, USA (1996)
2. Bajpai, P., Kumar, M.: Genetic Algorithm- an Approach to Solve Global Optimization Problems. Indian Journal of Computer Science and Engineering 1(3), 199–206
3. Bergh, V.D., Engelbrecht, A.P.: A cooperative approach to particle swarm optimization. IEEE Trans. Evol. Comput. 8(3), 225–239 (2004)

4. Brest, J., Zamuda, A., Fister, I., Maucec, S.: Large Scale Global Optimization using Self-adaptive Differential Evolution Algorithm. In: Proc. IEEE World Congress on Computational Intelligence, Spain, pp. 3097–3104 (2010)
5. Karaboga, D., Basturk, B.: A powerfull and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. Springer Sience+ Business Media (2007)
6. Kennedy, J., Eberhart, R.C.: Particle Swarm Optimization. In: Proceeding of the IEEE International Joint Conference on Neural Networks, pp. 1942–1948. IEEE Press (1995)
7. Laarhovan, P.J.M., Aarts, E.H.L.: Simulated Annealing: Theory and Applications. Kluwer Academic Publisher (1987)
8. Melanie, M.: An introduction to Genetic Algorithms. MIT Press, Massachusetts (1999)
9. Molina, D., Lozano, M., Herrera, F.: MA-SW-Chains: Memetic Algorithm Based on Local Search Chains for Large Scale Continuous Global Optimization. In: Proc. IEEE World Congress on Computational Intelligence, Spain, pp. 3153–3160 (2010)
10. Omidvar, M.N., Li, X., Yao, X.: Cooperative Co-evolution with Delta Grouping for Large Scale Non-separable Function Optimization. In: Proc. IEEE World Congress on Computational Intelligence, Spain, pp. 1762–1769 (2010)
11. Tang, K., Li, X., Suganthan, P.N., Yang, Z., Waise, T.: Benchmark Function for the CEC 2010 Special Session and Competition on Large Scale Global Optimization. Technical Report, Nature Inspired Computation and Applicayions Laboratory, USTC, China (2010)
12. Wang, Y., Li, B.: Two-stage based Ensemble Optimization for Large-Scale Global Optimization. In: Proc. IEEE World Congress on Computational Intelligence, Spain, pp. 4488–4495 (2010)
13. Yang, Z., Tang, K., Yao, X.: Large scale evolutionary optimization using cooperative coevolution. Information Sciences 178(15), 2985–2999 (2008)
14. Zhao, S.Z., Suganthan, P.N., Das, S.: Dynamic Multi-Swarm Particle Swarm Optimizer with Sub-regional Harmony Search. In: Proc. IEEE World Congress on Computational Intelligence, Spain, pp. 1983–1990 (2010)

A New Method for Job Scheduling in Two-Levels Hierarchical Systems

Amin Shokripour, Mohamed Othman*, Hamidah Ibrahim, and Shamala Subramaniam

Department of Communication Technology and Network, Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor D.E., Malaysia
shokripour@gmail.com, mothman@fsktm.upm.edu.my

Abstract. The use of parallel and distributed systems has become very common in the last decade. Dividing data is one of the challenges in these types of systems. Divisible load theory (DLT) is one of the proposed methods for scheduling data distribution in parallel or distributed systems. Many researches have been done in this field, but scheduling a multi-installment heterogeneous system with two-level hierarchical topology in which communication mode is blocking has not been addressed. In this paper, we find the proper size of task for each sub tree. Finally, in the experiments section, we show that the proposed methods work correctly and give us the best scheduling.

1 Introduction

Methodologies, such as Markovian queuing theory, circuit theory and divisible load theory (DLT), have been presented for data distribution in parallel and distributed systems. DLT is based on a specific type of data known as divisible load data [1]. Arbitrarily divisible data are a large amount of data that can be arbitrarily divided into desirable independent parts as each part can be processed separately from the others. Many scientific and engineering applications have this characteristic. Some of these applications include massive experimental data processing, image processing, video processing, mathematical applications, network scheduling, and biological applications [2].

Different models were investigated in DLT researches, each of which made some assumptions. An installment system with blocking and non-blocking mode communication [3], a system with different processor available time (SDPAT) [4], non-dedicated systems [5], and others are some examples of the investigated models. One of the new subjects in DLT research is job scheduling in a system with two-level hierarchical topology.

In this paper, we present a scheduling method to schedule jobs in a heterogeneous multi-installment system with two-level hierarchical topology. A formula

* The author is also an associate researcher at the Lab of Computational Science and Informatics, Institute of Mathematical Research (INSPEM), Universiti Putra Malaysia.

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 345–355, 2013.

for calculating the proper size of the task assigned to each sub tree is proposed. Knowing the size of task for each sub tree, we can independently schedule each sub tree by using the presented closed-form formulas. Finally, we will show that the proposed formulas are true, and the system's performance is the best.

The remainder of this paper is organized as detailed below. Section 2 presents related works. Our model and notations are introduced in the third section. In section 4, the proposed method, a formula for calculating the proper size of task for each sub tree is presented. The results of experiments and their analysis appear in section 5. The last section provides the conclusion.

2 Related Works

During the last decade, many studies regarding DLT have been performed. Each of these researches focused on some properties of DLT and defined some assumptions about the environment. Closed-form formulas are more preferred because we can use them in more complex systems (such as multi-source, hierarchical and non-dedicated systems) to achieve new formulas.

Multi-source systems are similar to tree system. Because each of branches is an independent source. In 2009, two new researches were carried out regarding job scheduling in multi-source systems [6,7]. Moges and his colleagues presented a strategy for job scheduling in a system with two sources in which the network topology is one-level tree. They proposed a closed-form formula for job scheduling in a system with concurrent data transfer. In their model, all the workers are connected to all the sources, and during data transferring, workers get each part of their data from one of the sources. Their method can only be applied on a system with two sources. It is not usable for more sources because they do data transferring concurrently where the communication method is in non-blocking mode. Their system is one installment, and because the transferring is done concurrently, the order of communication is not important. Also, they do not have any strategy for limiting the number of used processors because their workers do not have overhead; hence, they can use all the processors without any restriction.

Jia et al., proposed a method for job scheduling in a multi-source system with arbitrary network topologies [6]. They addressed a generalized divisible load scheduling problem for handling loads from multiple sources on arbitrary networks. They proposed two strategies, static scheduling strategy (SSS), for scheduling the systems in which as the time progresses, no new loads arrive, and dynamic scheduling strategy (DSS), in which some new load perhaps arrives during the task processing. Their research was based on graph partitioning (GP), and all the tasks are a graph for their system. They used non-blocking mode communication, but the communication was done by one port for both sending and receiving. Their system was one installment and network topology was arbitrary.

Table 1. Notations

Notation	Description
W	Total size of data
W_i	Total size of data for the P_i^s
V_i	Size of each installment for the P_i^s
n_i	Number of installments for the P_i^s
m_i	Number of processors for the P_i^s
l	Number of node in level 1
α_i^j	The size of allocated fraction to processor P_i in each internal installment in the branch j
β_i^j	The size of allocated fraction to processor P_i in the last installment in the branch j
w_i^j	Ratio of the time taken by processor P_i in the group j, to compute a given load, to the time taken by a standard processor, to compute the same load
z_i^j	Ratio of the time taken by P_i in the group j, to communicate a given load, to the time taken by a standard link, to communicate the same load
s_i^j	Computation overhead for processor P_i in the group j
o_i^j	Communication overhead for processor P_i in the group j
$T(W_i)$	The response time for a task with size W_i

3 Preliminaries

3.1 Notations

In this paper, several notations are used. The notations and their definitions are described in Table 1.

3.2 Model

In this research, we assume that we have some groups of clients. The node of branch number i is called P_i^s.

We investigate scheduling a two-level tree network topology. For the two-level tree network topology, we use a model as shown in Fig. 1.

3.3 A Review on the Presented Formula for Multi-installment Systems

In our previous paper [8], some closed-form formulas for different steps of job scheduling in a heterogeneous multi-installment system, which includes overheads, were presented. In this section, we review them as we need them in the next few sections.

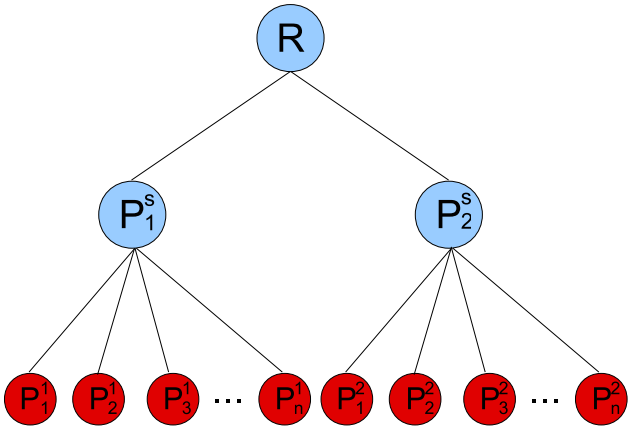


Fig. 1. Two-Level Hierarchical Network Topology

For scheduling internal installments Eq.(1) can be used:

$$\begin{cases} \Delta_i = \frac{z_1+w_1}{z_i+w_i} \\ \Phi_i = \frac{o_1+s_1-o_i-s_i}{z_i+w_i} \\ \alpha_1 = \frac{1-\frac{1}{V}\sum_{i=2}^{m}\Phi_i}{1+\sum_{i=2}^{m}\Delta_i} \\ \alpha_i = \alpha_1\Delta_i + \frac{1}{V}\Phi_i \end{cases} \tag{1}$$

The last installment can be scheduled using Eq.(2).

$$\begin{cases} \delta_{i+1} = \frac{s_i-(s_{i+1}+o_{i+1})}{w_{i+1}+z_{i+1}} \\ \varepsilon_{i+1} = \frac{w_i}{w_{i+1}+z_{i+1}} \\ \mathrm{E}_i = \prod_{j=2}^{i}\varepsilon_j \\ \Gamma_i = \sum_{j=2}^{i}(\delta_j \prod_{k=j+1}^{i}\varepsilon_k) \\ \beta_i = \mathrm{E}_i\beta_1 + \frac{1}{V}\Gamma_i, i = 2,\ldots,m \\ \beta_1 = \frac{1-\frac{1}{V}\sum_{i=2}^{m}\Gamma_i}{1+\sum_{i=2}^{m}\mathrm{E}_i} \end{cases} \tag{2}$$

In the above formulas, m is the proper number of processors. We use Eq.(3) for calculating its value.

$$W \geq \left(\frac{(1+\sum_{i=2}^{m} \Delta_i)(\sum_{j=2}^{m}(\Phi_j z_j + o_j) - s_1)}{w_1 - \sum_{j=2}^{m}(\Delta_j z_j)} + \sum_{i=2}^{m} \Phi_i \right)^2 . \\ \frac{(z_1+w_1)(\sum_{i=2}^{m-1} \mathrm{E}_i - \sum_{i=2}^{m} \Delta_i)}{(1+\sum_{i=2}^{m-1} \mathrm{E}_i)[(z_1+w_1)\sum_{i=2}^{m} \Phi_i - (1+\sum_{i=2}^{m} \Delta_i)(o_1+s_1)]} \tag{3}$$

The proper number of installment is another important parameter for job scheduling in a multi-installment system. The formula used for calculating this parameter is given by Eq.(4).

$$\frac{W(z_1+w_1)(\sum_{i=2}^{m-1} \mathrm{E}_i - \sum_{i=2}^{m} \Delta_i)}{(1+\sum_{i=2}^{m-1} \mathrm{E}_i)[(z_1+w_1)\sum_{i=2}^{m} \Phi_i - (1+\sum_{i=2}^{m} \Delta_i)(o_1+s_1)]} = (n+1)^2 \tag{4}$$

4 The Proposed Method

One question should be answered in this research; 2. How much is the data assigned to each root for processing?. Before answering these questions, we must have a closed-form formula to calculate the response time of a scheduled general system.

4.1 Response Time Function for a General System

DLT is based on the concept that for the best response time, all the processors should finish their tasks at the same time instant. Hence, we need to know the response time in a general multi-installment system. The response time for a multi-installment system includes required time for internal installments and last installment.

$$T(W_1) = n_1 \left(\sum_{j=2}^{m_1} (\alpha_j^1 V_1 z_j^1 + o_j^1) + \alpha_1^1 V_1 z_1^1 + o_1^1 \right) \\ + \left(\frac{V_1 - \sum_{i=2}^{m_1} \Gamma_i^1}{1+\sum_{i=2}^{m_1} \mathrm{E}_i^1} \right)(z_1^1 + w_1^1) + (o_1^1 + s_1^1) \tag{5}$$

As one of the unknown variables in this research is the size of assigned data to each group, we change this formula to one based on W_1. We define a new variable as

$$\tau_1 = \sqrt{\frac{1}{(1+\sum_{i=2}^{m_1} \mathrm{E}_i)}} \cdot \\ \sqrt{\frac{(1+\sum_{i=2}^{m_1} \Delta_i)(z_1+w_1) - (1+\sum_{i=2}^{m_1} \mathrm{E}_i)\sum_{j=1}^{m_1}(\Delta_j z_j)}{(1+\sum_{i=2}^{m_1} \Delta_i^1)(\sum_{j=2}^{m_1}(\Phi_j^1 z_j^1) + \sum_{j=1}^{m_1} o_j^1) - \sum_{i=2}^{m_1} \Phi_i^1 \sum_{j=1}^{m_1}(\Delta_j^1 z_j^1)}} \tag{6}$$

From Eq.(6) and Eq.(4), we find that $n_1 + 1 = \sqrt{W_1}\tau_1$. After rewriting Eq.(5), we have

$$\begin{aligned} & W_1 \frac{\sum_{j=1}^{m_1}(\Delta_j^1 z_j^1)}{(1+\sum_{i=2}^{m_1}\Delta_i^1)} + \sqrt{W_1}2\tau_1 \\ & \left(\frac{(1+\sum_{i=2}^{m_1}\Delta_i^1)(\sum_{j=2}^{m_1}(\Phi_j^1 z_j^1)+\sum_{j=1}^{m_1} o_j^1) - \sum_{i=2}^{m_1}\Phi_i^1 \sum_{j=1}^{m_1}(\Delta_j^1 z_j^1)}{1+\sum_{i=2}^{m_1}\Delta_i^1} \right) \\ & - \frac{(1+\sum_{i=2}^{m_1}\Delta_i^1)(\sum_{j=2}^{m_1}(\Phi_j^1 z_j^1)+\sum_{j=1}^{m_1} o_j^1) - \sum_{i=2}^{m_1}\Phi_i^1 \sum_{j=1}^{m_1}(\Delta_j^1 z_j^1)}{1+\sum_{i=2}^{m_1}\Delta_i^1} \\ & - \frac{(z_1^1+w_1^1)\sum_{i=2}^{m_1}\Gamma_i^1}{1+\sum_{i=2}^{m}\mathrm{E}_i^1} + (o_1^1+s_1^1) = T(W_1) \end{aligned} \tag{7}$$

Three new symbols are defined to make Eq.(7) simpler.

$$\begin{cases} F_1 = -\dfrac{(1+\sum_{i=2}^{m_1}\Delta_i^1)(\sum_{j=2}^{m_1}(\Phi_j^1 z_j^1)+\sum_{j=1}^{m_1} o_j^1)}{1+\sum_{i=2}^{m_1}\Delta_i^1} \\ \qquad + \dfrac{\sum_{i=2}^{m_1}\Phi_i^1 \sum_{j=1}^{m_1}(\Delta_j^1 z_j^1)}{1+\sum_{i=2}^{m_1}\Delta_i^1} - \dfrac{(z_1^1+w_1^1)\sum_{i=2}^{m_1}\Gamma_i^1}{1+\sum_{i=2}^{m}\mathrm{E}_i^1} + (o_1^1+s_1^1) \\ G_1 = 2\tau_1 \left(\dfrac{(1+\sum_{i=2}^{m_1}\Delta_i^1)(\sum_{j=2}^{m_1}(\Phi_j^1 z_j^1)+\sum_{j=1}^{m_1} o_j^1) - \sum_{i=2}^{m_1}\Phi_i^1 \sum_{j=1}^{m_1}(\Delta_j^1 z_j^1)}{1+\sum_{i=2}^{m_1}\Delta_i^1} \right) \\ H_1 = \dfrac{\sum_{j=1}^{m_1}(\Delta_j^1 z_j^1)}{1+\sum_{i=2}^{m_1}\Delta_i^1} \end{cases} \tag{8}$$

We can rewrite Eq.(7), the response time function for a multi-installment system, as

$$H_1 W_1 + \sqrt{W_1} G_1 + F_1 = T(W_1) \tag{9}$$

4.2 Closed-Form Formula for Two-Level Hierarchical Systems

Two-level hierarchical network topology is the simplest topology in trees family. In this topology, we have a source in level zero and some nodes in level one. We assume that each of the nodes in level one is the source of an independent group in level two. Hence, if we know the size of task for each node in level one, we can easily schedule the system by using Eq.(1) and Eq.(2).

We attempt to find the size of the data assigned to each branch so that all the branches finish their tasks simultaneously. The response time for each branch can be found by using Eq.(9). The running time for each branch is different from the running time of each source in a general multi-source system. The time of participation of each branch in the task is equal to its response time plus the time of communication between the root of the branch and the main root, as shown in Eq.(10).

$$\begin{cases} H_1 W_1 + G_1\sqrt{W_1} + F_1 = H_2 W_2 + G_2\sqrt{W_2} + F_2 + W_2 z_2^2 + o_2^2 \\ H_2 W_2 + G_2\sqrt{W_2} + F_2 = H_3 W_3 + G_3\sqrt{W_3} + F_3 + W_3 z_3^3 + o_3^3 \\ \dots \\ H_{l-1} W_{l-1} + G_{l-1}\sqrt{W_{l-1}} + F_{l-1} = H_l W_l + G_l\sqrt{W_l} + F_l + W_l z_l^l + o_l^l \end{cases} \tag{10}$$

By using these equations, we can calculate W_i based on W_1.

$$W_i = Z_i W_1, i = 2, 3, \ldots, l-1 \tag{11}$$

We know that

$$W_1 + W_2 + W_3 + \cdots + W_l = \sum_{i=1}^{l} W_i = W \tag{12}$$

Hence, $W_2 = W - (1 + \sum_{i=3}^{l} Z_i)W_1$.

To make the formulas simple, we define $\Psi_{tree} = (1 + \sum_{i=3}^{l} Z_i)$.

$$\begin{aligned} H_1 W_1 + G_1\sqrt{W_1} + F_1 =& H_2(W - \Psi_{tree}W_1) + G_2\sqrt{W - \Psi_{tree}W_1} + F_2 \\ &+ z_2^2(W - \Psi_{tree}W_1) + \sqrt{W - \Psi_{tree}W_1} + o_2^2 \end{aligned} \tag{13}$$

We solve this equation to find the value of W_1.

$$\begin{aligned} &[H_1 + (H_l + z_2^2)\Psi_{tree}]W_1 + G_1\sqrt{W_1} + [F_1 - (H_l + z_2^2)W - F_l - o_2^2] \\ &- G_l\sqrt{W - \Psi_{tree}W_1} = 0 \end{aligned} \tag{14}$$

We define two new symbols and rewrite Eq.(14).

$$\begin{cases} H_{tree} = H_1 + (H_l + z_2^2)\Psi_{tree} \\ F_{tree} = F_1 - (H_l + z_2^2)W - F_l - o_2^2 \end{cases} \tag{15}$$

$$H_{tree}W_1 + G_1\sqrt{W_1} + F_{tree} - G_l\sqrt{W - \Psi_{tree}W_1} = 0 \tag{16}$$

$$\begin{aligned} &-(H_{tree}W_1)^2 + (WG_l^2 - F_{tree}^2) + (G_1^2 - 2F_{tree}H_{tree} \\ &- \Psi G_l^2)W_1 = 2G_l G_1\sqrt{WW_1 - \Psi_{tree}W_1^2} \end{aligned} \tag{17}$$

Three new symbols are necessary to make the Eq.(17) simpler.

$$\begin{cases} P_{tree} = WG_l^2 - F_{tree}^2 \\ T_{tree} = G_l^2 - 2F_{tree}H_{tree} - \Psi_{tree}G_l^2 \\ X_{tree} = G_l G_1 \end{cases} \tag{18}$$

Eq.(17) can be written as

$$\begin{aligned} &[P_{tree} - (H_{tree}W_1)^2 + T_{tree}W_1]^2 \\ &= (2X_{tree}\sqrt{WW_1 - \Psi_{tree}W_1^2})^2 \end{aligned} \tag{19}$$

After defining five new symbols, we have a simple quartic equation, Eq.(21).

$$\begin{cases} A_{tree} = H_{tree}^4 \\ B_{tree} = -2H_{tree}^2 T \\ C_{tree} = T_{tree}^2 - 2P_{tree}H_{tree}^2 + 4X_{tree}^2\Psi_{tree} \\ D_{tree} = 2P_{tree}T_{tree} - 4X_{tree}^2 W \\ E_{tree} = P_{tree}^2 \end{cases} \tag{20}$$

$$A_{tree}W_1^4 + B_{tree}W_1^3 + C_{tree}W_1^2 + D_{tree}W_1 + E_{tree} = 0 \tag{21}$$

Eq.(21) is a quartic equation that can be solved using Ferrari's method. The root of this equation, which is larger than zero and satisfies Eq.(12) is acceptable.

5 Experiment Result and Discussion

For the experiments, we used a simulator implemented by C++, which ran with Linux-based operating system, Suse 11.1. A set of 50 processors with attributes produced randomly, was used as input for simulator. A set of jobs was used as the second input. Value of w was ten times as large as z value for all processors. This means that communication speed was much faster than computation speed.

5.1 Evaluating Quartic Method for Tree Topology

We attempted to show that the Quartic Method works correctly in tree topology, and has a response time that is the shortest. As mentioned earlier, four questions should be answered while scheduling multi-installment systems. In this section, therefore, we checked whether the Quartic Method offers the best value for each of the parameters, the proper number of processors, the proper number of installments, and the proper size of task for the internal and last installment. We did three sets of experiments to show the method gives us the best value for each parameter.

As can be seen in Fig. 2, the calculated value obtained for the size of task for each of the branches by using the Quartic Method (zero point in the graphs) is much better than the other points. In this experiment, we manually changed the size of the assigned task to each branch, and the response times for the changed sizes of tasks were calculated and shown in the graphs. We reduced 5%, 10% and 15% of the task assigned to the first branch, and added them to the second branch. We also added these percentages of the size of the tasks assigned to the first branch, and reduced the percentages in the second branch.

By using an experiment similar to the above experiment, we can evaluate the Quartic Method's efficiency in calculating the proper number of installments. Fig. 3 shows that due to the same reason as was mentioned for the calculated size of task for each processor, it is clear that the Quartic Method presents the best proper number of installments for a heterogeneous multi-installment system with a two-level hierarchical network topology.

The Quartic Method can be evaluated for its ability to find the proper number of processors with a similar experiment. The results of this experiment can be seen in Fig. 4. As shown in the graphs, we manually added one, two, three and four processors to the calculated proper number of processors, and also reduced the same number of processors. The zero points show the response times for the proper number of processors calculated using the Quartic Method.

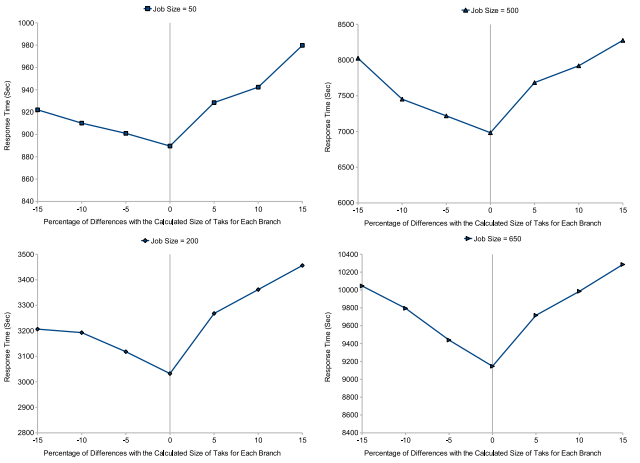


Fig. 2. Response Time vs Differences with the Calculated Size of the Assigned Task using the Quartic Method for Different Sizes of Jobs

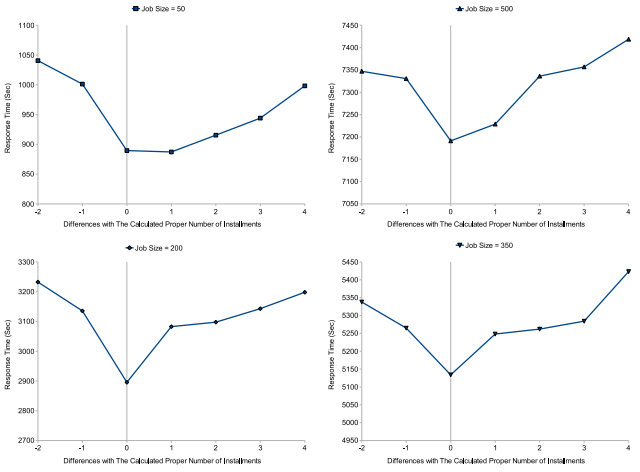


Fig. 3. Response Time vs Differences with the Calculated Proper Number of Installments using the Quartic Method for Different Sizes of Jobs

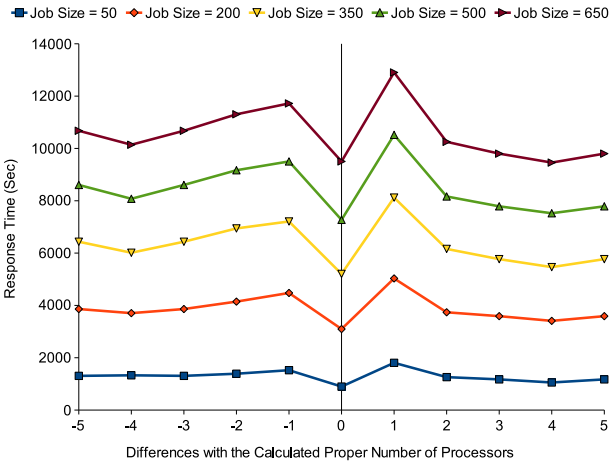


Fig. 4. Response Time vs Differences with the Calculated Proper Number of Processors using the Quartic Method for Different Sizes of Jobs

6 Conclusion

Job scheduling in a multi-installment system with two-level hierarchical topology had not been addressed before this study. In this study, we explored job scheduling in a heterogeneous multi-installment system, which include overheads. A quartic equation for finding the proper size of the task for each node in level 1 was solved by one of the known methods. After this, we used the proposed closed-form formula for job scheduling in multi-installment systems for each sub tree, independent of others. Our experiments showed the calculated size for each sub tree is the optimum size.

Acknowledgements. This work has been partially supported by the Malaysia Ministry of High Education under the Fundamental Research Grant Scheme FRGS/1/11/SG/UPM/01/1.

References

1. Robertazzi, T.: Ten reasons to use divisible load theory. Computer 36(5), 63–68 (2003)
2. Shokripour, A., Othman, M.: Categorizing DLT researches and its applications. European Journal of Scientific Research 37(3), 496–515 (2009)
3. Mingsheng, S.: Optimal algorithm for scheduling large divisible workload on heterogeneous system. Appl. Math. Model. 32, 1682–1695 (2008)
4. Shokripour, A., Othman, M., Ibrahim, H.: A New Algorithm for Divisible Load Scheduling with Different Processor Available Times. In: Nguyen, N.T., Le, M.T., Świątek, J. (eds.) ACIIDS 2010. LNCS, vol. 5990, pp. 221–230. Springer, Heidelberg (2010)

5. Shokripour, A., Othman, M., Ibrahim, H., Subramaniam, S.: A new method for job scheduling in a non-dedicated heterogeneous system. Procedia Computer Science 3, 271–275 (2011)
6. Jia, J., Veeravalli, B., Weissman, J.: Scheduling multi-source divisible loads on arbitrary networks. IEEE T. Parall. Distr. 99, 520–531 (2009)
7. Moges, M.A., Yu, D., Robertazzi, T.G.: Grid scheduling divisible loads from two sources. Comput. Math. Appl. 58(6), 1081–1092 (2009)
8. Shokripour, A., Othman, M., Ibrahim, H., Subramaniam, S.: A method for scheduling heterogeneous multi-installment systems. Future Gener. Comp. Sy. 28(8), 1205–1216 (2012)

Intelligent Water Drops Algorithm for Rough Set Feature Selection

Basem O. Alijla[1], Lim Chee Peng[2], Ahamad Tajudin Khader[1], and Mohammed Azmi Al-Betar[1,3]

[1] School of Computer Sciences, Universiti Sains Malaysia, Pinang, Malaysia
[2] Centre for Intelligent Systems Research, Deakin University, Australia
[3] Department of Computer Science, Jadara University, Irbid, Jordan

Abstract. In this article; Intelligent Water Drops (IWD) algorithm is adapted for feature selection with Rough Set (RS). Specifically, IWD is used to search for a subset of features based on RS dependency as an evaluation function. The resulting system, called IWDRSFS (Intelligent Water Drops for Rough Set Feature Selection), is evaluated with six benchmark data sets. The performance of IWDRSFS are analysed and compared with those from other methods in the literature. The outcomes indicate that IWDRSFS is able to provide competitive and comparable results. In summary, this study shows that IWD is a useful method for undertaking feature selection problems with RS.

Keywords: Feature Selection (FS), Rough Set (RS), Intelligent Water Drops (IWD).

1 Introduction

Feature Selection (FS) refers to the process of selecting the minimum subset of features that preserves the meaning of the original features [3]. An irrelevant feature is a feature that is weakly correlated to the decisional feature, which can be removed with little or no effect to the given outcomes. A redundant feature is a feature that is highly correlated with other features, and it does not carry significant knowledge when it is added to the entire set of features. If the irrelevant and redundant features can be removed, the dimension of the data set can be reduced without significantly affecting the knowledge represented by the entire features [19]. Moreover, learning and classification accuracy can be improved by simple, easy, and understandable presentation of the underlying rules, which are formulated from fewer numbers of features [10].

The main elements of an FS algorithm include subset generation, subset evaluation, and the stopping criterion [9]. Subset generation is the search technique, which is used to explore the search space. Subset evaluation then uses an evaluation approach to assess the goodness of the subset of features. The stopping criterion is used to terminate the search process. FS problem is a combinatorial NP-hard problem [19]. This is because the numbers of alternatives are proportional to the number of features in the data set. As an example, if we have a

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 356–365, 2013.

data set with N features, FS can be seen as a search process over a search space with 2^N possible subsets of features. Although exhaustive search techniques can be used to find the optimal subset of features, it is impractical in the presence of large number of features. To manage the complexity of the search process, many search strategies such as heuristic and metaheuristic methods have been proposed [1,5,8,17,18]. A detailed taxonomy and the associated algorithms of FS can be found in [9].

Rough Set Theory (RST) is a mathematical theory introduced by Pawlak in 1982 [12], which was used as a tool for analyzing incomplete or uncertain data. RST is popular for feature selection. It is characterized by its ability to evaluate features indiscernibility without needing any external information. Indeed, RST is used to analyze only the hidden information within a data set to find the minimal knowledge representation. RST has been successfully used with many search algorithms for feature selection in order to measure the goodness of the selected subsets [1,5,8,17,18].

The Intelligent Water Drops (IWD) algorithm is a meta-heuristic method introduced by Shah-Husseini [16]. It is a nature-inspired optimization algorithm. IWD imitates some of the natural phenomena of a swarm of water drops with the soil onto the river bed. Within the last 5 years, IWD has been very successful in many discrete optimization problems [4,6,11,13,16] and machine learning tasks [15]. IWD has recently been adapted for continuous optimization problems [14]. This success is partly owing to the fundamental advantages of IWD over other traditional optimization techniques [13,15,16]. IWD has a simple and easily understandable mathematical model. It can be adapted easily for many optimization problems, and is applicable to both discrete and continuous problems. It converges fast to the optimal solution. It considers the construction of solution in the population based on information given by experience (gained from the previous iteration of the search) rather than considering refinement of the existing population.

In this paper, the IWD is adapted with the rough set for feature selection. The resulting model is called Intelligent Water Drops algorithm for Rough Set Feature Selection (IWDRSFS). IWDRSFS is evaluated with 6 benchmark data sets obtained from [7]. Many of these data sets come from the UCI machine learning repository [2]. The numbers of the input features vary between 13 and 56. The results from IWDRSFS are compared with those from local search-based methods, such as hill climbing, as well as population search-based methods such as ant colony and the genetic algorithm.

The rest of the paper is organized as follows. Section 2 provides a brief introduction to RST and RS dependency for feature selection. Section 3 describes the detailed modeling and implementation of IWD for feature selection. The experiments and the associated results are presented in section 4. Conclusions and suggestions for future work are highlighted in section 5.

2 Rough Set Theory and Feature Dependency for Feature Selection

RST is an approximation approach developed to deal with incomplete knowledge Pawlak [12]. The fundamental concept of RST is the approximation of the uncertain set (knowledge) with a pair of precise sets, called the lower and upper approximations. The lower approximation is a set that describes objects that are definitely belonging to the subset of interest, while the upper approximation is a set that describes objects that are possibly belonging to the subset of interest. The pair of lower and upper approximations as a tuple is defined as a Rough Set (RS)[12].

Let *IS= (U,A$\bigcup$D)* be an information system, U is a non-empty finite set of objects (universe), A is a non empty finite set of conditional features, and D is the decisional feature. For any $S \subseteq A$ there exists an equivalence relation called the S-indiscernibility relation that can be used to group objects into classes which are called equivalence classes denoted as $[x]_s$. Each class contains the set of objects that have the same vector of features values in S. Let $X \subseteq U$. X be a target equivalence class (concept) induced by the decisional feature D. X cannot be expressed directly by $[x]_S$ because X may include an object that is not in $[x]_S$ and vice versa. RST is able to approximate this uncertainty by comparing the equivalent classes induced by the conditional features with the target equivalence class. RST defined the *lower* and *upper* approximations to find the positive region, which is a set that includes objects that can certainly be classified by a feature or subset of features.The positive region can be employed to find feature dependency, and is denoted as $\gamma_S(Q)$. $\gamma_S(Q)$ is used to measure the strength of the relation (correlation) between two set of features S, and Q. If $\gamma_S(Q) = 1$, then Q is totally dependent on S, and denoted as $(S \Rightarrow Q)$. If $\gamma_S(Q) < 1$ then Q is partially dependent on S with a degree $\gamma_S(Q)$, and is denoted as $(S \underset{\gamma_S(Q)}{\Longrightarrow} Q)$. Finally, if $\gamma_S(Q) = 0$, then Q and S are independent. The detailed information on RS can found in [9].

The main idea of FS with RS is to remove features that do not have significant effects on feature dependency. So, the FS algorithm aims to search for the minimum subset of features, S, that has feature dependency equals to the dependency of the full features C, i.e. $\gamma_S(Q) \approx \gamma_C(Q)$, where $S \subseteq C$.

FS problems require finding one subset of features that has feature dependency equals to the dependency of the full set of features. The ideal FS algorithms aim to find all subsets of features that satisfy the abovementioned condition. However, finding all subsets of features is computationally expensive. Therefore, an efficient search algorithm is required to find the optimal subset of features by considering the maximum dependency and minimum subset size.

3 Intelligent Water Drops for Feature Selection

In nature, water drops have to overcome obstacles and barriers in the environment in order to find the shortest path from its source to the destination. Water

drops prefer to follow the direction of the easy path, i.e. a path with less soil. Water drops are transferred from one point to another with a velocity. During the move, water drops carry an amount of soil gained from the bed of the path. Changes on the soil carried by the water drops, the soil in the path, and the velocity, encourage water drops to move through the shortest path that has less soil and, at the same time, to reinforce other water drops to follow the same path.

The key properties of water drops are soil and velocity. During the trip of the water drops, a certain amount of soil from the bed of the path will be carried together. The change on the soil carried by the drops is proportional to the inverse of the velocity in a nonlinear way. Specifically, during the lifetime of the water drops, the velocity will be changed with a value that is nonlinearly proportional to the inverse of the soil between two points in the path. Thus, water drops on a path with less soil become faster, and the soil on the path is decreased. Changes of the soil and velocity have an influential role on the probability of selecting the direction of flow. The probability of selecting the next path is inversely proportional to the soil of the available paths. As a result, a path with low soil has a higher probability of being the selected path. The whole process will converge when the probability of selecting the shortest path equals to 1.

The following subsections describe the detailed modeling and implementation of IWD for feature selection.

3.1 Modeling of Feature Selection as the IWD Environment

FS aims to select a subset S from the full set of features C where the knowledge represented by C is contained in S. The process of searching for the optimal subset using IWD is modeled as a complete undirected graph $G = (V, E)$, where V is number of nodes (i.e. features) connected by set of edges E. An edge represents the choice of the next feature. An edge holds an amount of soil that represents the hardness of the local path (edge between two features). A number of water drops are spreaded randomly to the set of features, where every drop is allocated with a different feature. Water drops can be used as agents that construct the solutions (population). A water drop starts to move from its source, i.e., the first allocated feature, to the next until it completes a path. A selection mechanism is required by IWD to determine the direction of the next local path, as described in section 3.2. Every water drop has a list k has a list $V_C^{IWD_k}$, which is used to record the visited features. $V_C^{IWD_k}$ is the solution k, which is constructed by the water drop IWD_k. The population is a set of solutions which are constructed by the entire water drops i.e. $T^{IWD} = \left\{ V_C^{IWD_1}, V_C^{IWD_2}, ..., V_C^{IWD_k}, ..., V_C^{IWD_{N_{IWD}}} \right\}$, where N_{IWD} is the maximum number of water drops.

In this article, RS dependency is used as the evaluation function to assess the goodness of the partial solution.

3.2 The Proposed IWDRSFS Model

In the following we present the main phases and steps of the proposed IWDRSFS model.

Initialization Phase. The initialization phase is used for initializing the static and dynamic parameters of the water drops and to spread the water drops on their sources.

i. **Initializing the static parameters**
Static parameters are parameters that assume specific initial values at the beginning of the search, and they remain unchanged during the whole process. The static parameters of the proposed IWDRSFS model are:
   - $N_{\boldsymbol{IWD}}$: a set of water drops, which represents the set of solutions.
   - **Velocity updating parameters**$(a_v, b_v\, c_v)$: set of parameters used for updating the velocity of the water drops (equation 5).
   - **Soil updating parameters** $(a_s,\ b_s,\ c_s)$: set of parameters used for computing the amount of changes in the soil of the local path (equation 6).
   - **MaxIter**: the maximum number of iterations for a water drops before terminating the IDW algorithm.
   - **initSoil**: the initial value of the local soil.

ii. **Initializing the dynamic parameters**
Dynamic parameters are parameters that are initialized at the beginning of the search, and are updated dynamically during the lifetime of search.
   - $\mathbf{V_C^{IWD_k}}$: a list of visited features for each water drop k,
   - $\mathbf{intiVel^{IWD_k}}$: the initial velocity of water drop **k** at the beginning of the search.
   - $\boldsymbol{Soil^{IWD_k}}$: the initial soil of water drop k, at the beginning of the search, where $1 \leq k \leq N_{IWD}$.

   dynamic parameters should be reset to their default initial values at the beginning of iteration.

iii. **Spread drops on their sources**
Water drops are spread randomly to the set of features, where every drop k is allocated with a different feature, which is considered as the source of water drop. $V_C^{IWD_k}$ is updated by adding the source.

Construction Phase. The main goal of the construction phase for every water drop is to complete its solution starting from the source (the first point the water drop is spread on). The construction phase is completed by the fluency of all water drops amongst the features using the following four steps:

i. **Edge selection mechanism**
A water drop k, which is resided in the current feature i can determine the next feature j, which is not in the visited list ($V_C^{IWD_k}$) using the probability

$p_i^{IWD_k}(j)$ as shown in equation (1). $V_C^{IWD_k}$) is updated by adding the selected edge.

$$p_i^{IWD_k}(j) = \frac{f(soil(i,j))}{\sum_{l \notin V_C^{IWD_k}} f(soil(i,l))} \tag{1}$$

where $f(soil(i,j)) = \frac{1}{\varepsilon + g(soil(i,j))}$, ε is a small positive number prevents the division by zero in $f(.)$

$$g(soil(i,j)) = \begin{cases} soil(i,j) & if \min_{\forall l \notin V_C^{IWD_k}} soil(i,l) \geqslant 0, \\ soil(i,j) - \min_{\forall l \notin V_C^{IWD_k}} soil(i,l) & Otherwise. \end{cases}$$

Where $soil(i,l)$ refers to the amount of soil on the local path between features i, and j. The function min (.) returns the minimum value among all available values for its argument.

ii. **Update the velocity and local soil**

The velocity of the drop k at time $t+1$ is denoted as $vel^{IWD_k}(t+1)$. It is changed every transit from feature i to feature j using equation (2).

$$vel^{IWD_k}(t+1) = vel^{IWD_k}(t) + \frac{a_v}{b_v + c_v * soil(i,j)} \tag{2}$$

where a_v, b_v, c_v are the static parameters used to represent the non-linear relationship between the velocity of a water drop k, (i.e. vel^{IWD_k}), and the inverse of soil onto the local path, (i.e. *soil(i, j)*). *soil(i, j)* and the amount of soil carried by the drop k (i.e. $soil^{IWD_k}$) are updated by $\triangle soil(i,j)$ using equations (5), (6) respectively. $\triangle soil(i,j)$ refers to the amount of soil removed from the local path and carried by the drop. $\triangle soil(i,j)$ is nonlinearly proportional to the inverse of vel^{IWD_k} as shown in equation (3).

$$\triangle soil(i,\ j) = \frac{a_s}{b_s + c_s * time(i,j : vel^{IWD_k}(t+1))} \tag{3}$$

where, a_s, b_s, c_s are the static parameters used to represent the non-linear relationship between $\triangle soil(i,j)$ and the inverse vel^{IWD_k}. $time(i,j : vel^{IWD_k}(t+1))$ refers to the time needed for a drop k to transit from feature i to feature j at time t+1. It can be calculated using equation (4).

$$time(i,j : vel^{IWD_k}(t+1)) = \frac{HUD(i,\ j)}{vel^{IWD_k}(t+1)} \tag{4}$$

where $HUD(i,j)$ is the heuristic desirability of the edge between features i and j. In this work, the RS dependency is used to evaluate the goodness of the path between two features.

$$soil(i,j) = (1-\rho_n) * soil(i,j) - \rho_n * \triangle soil(i,j) \tag{5}$$

$$soil^{IWD_k} = soil^{IWD_k} + \triangle soil(i,\ j) \tag{6}$$

where ρ_n is a small positive constant between zero and one.

Reinforcement Phase. A solution with the minimum number of features amongst T^{IWD}, called the iteration best solution (i.e T^{IB}), is selected using equation (7). For each iteration, if T^{IB} is shorter than the best solution found so far, the total best solution i.e (T^{TB}) is replaced with T^{IB}. Otherwise T^{TB} is kept unchanged. To reinforce water drops in the subsequent iterations to follow T^{TB}, , the soil of all edges (i.e. the global path soil) exist in T^{IB} is updated using equation (8).

$$T^{TB} = arg \min_{\forall l \in T^{IWB}} q(x) \tag{7}$$

where $q(.)$ is the function that is used to evaluate the quality of the solutions. In feature selection, it refers to the number of features in a solution (i.e. cardinality of the solution).

$$soil(i,\ j) = (1 + \rho_{IWD}) * soil(i,j) - \rho_{IWD} * \frac{1}{q(T^{IWB})} \tag{8}$$

where $q(T^{IB})$ is cardinality of T^{IB}, and ρ_{IWD} is a positive constant.

Termination Phase. Construction and reinforcement phases are repeated until the termination criterion (i.e. the maximum number of iterations, MaxIter) is satisfied. At any iteration, if T^{IB} is better than T^{TB}, T^{TB} is replaced by T^{IB} otherwise T^{TB} is kept unchanged, as shown in equation (9). The IWD dynamic parameters are reset to their default values at the beginning of each iteration.

$$T^{TB} = \begin{cases} T^{IB} & if q(T^{IB}) < q(T^{TB}) \\ T^{TB} & Otherwise. \end{cases} \tag{9}$$

4 Experiments and Results

The proposed IWDRSFS model was evaluated using six benchmark data sets obtained from [7], because they had been preprocessed, such as discretizing real valued features, treating the missing values, and removing outlier instances. Most of these data sets came from the UCI machine learning repository [2]. The chosen data sets had different degrees of difficulties, e.g. the numbers of features (dimensions) varied from low (13) to high (56), and the numbers of samples were small for high dimensional data sets, as shown in Table 1.

The IWDRSFS model was implemented using the Java programming language. The experiments were conducted using an Intel Pentium 4 core 2 Quad 2.66 GHz personal computer. The parameter setting of IWDRSFS is summarized in Table 2.

RS dependency was used as the evaluation function to measure the goodness of the partial solution, where a dependency of 1 was used as the stopping criterion for a complete solution.

Table 3 shows the results of IWDRSFS for the six data sets. The results of IWDRSFS are compared with those from four state-of-the-art RS methods for

Table 1. The main properties of the data sets

No.	Data sets	Abbreviations	No. of features	No. of samples
1	Artificial domains concept	M-of-N	13	1000
2	Statlog German credit data	CREDIT	20	1000
3	Letter recognition	LETTERS	25	26
4	Dermatology	DERM	34	366
5	Water Treatment Plant	WQ	38	521
6	Lung Cancer	LUNG	56	32

Table 2. IWDRSFS Parameter settings

Description	Parameters	Values
Static parameters	N_{IWD}	Number of features
	a_v, b_v, c_v	1000, 0.01, 1
	a_s, b_s, c_s	1000, 0.01, 1
	$initSoil$	100
	$MaxIter$	250
	ϵ, ρ_{IWD}, ρ_n	0.01, 0.9, 0.9
Dynamic parameters	$V_C^{IWD_k}$	Empty
	$intiVel^{IWD_k}$	4
	$soil^{IWD_k}$	0

FS, as published in [8]. They included the RS attribute reduction algorithm based on the greedy hill-climbing technique (RSAR), entropy-based data reduction (EBR), ant colony rough set attribute reduction (AntRSAR), genetic algorithm rough set attribute reduction (GenRSAR). For each data set, the experiment was repeated 20 times.

AntRSAR, GenRSAR, and IWDRSAR are multi-solution methods (i.e., every run may provide a different solution with different dimension). The results of AntRSAR, GenRSAR, and IWDRSAR in Table 3 are presented as a number with parenthesized superscript. The number refers to the dimension of the solution (i.e. a smaller subset size is better a larger subset size). The superscript refers to the number of runs that provides the corresponding dimension. On the other hand, RSAR and EBR are single solution methods, i.e. they provide the same solution even with different runs. So, the RSAR, and EBR results are presented as a single number.

As shown in Table 3, IWDRSFS outperformed RSAR and EBR in all data sets, except for CREDIT where RSAR performed better than IWDRSFS. Comparing IWDRSFS with GenRSAR; out of the six data sets; IWDRSFS found better solutions in three data sets (i.e., 4, 5, and 6), and comparable solutions in the remaining data sets (i.e. 1, 2, and 3). Comparing with AntRSAR, IWDRSFS produced comparable results for 4 data sets (i.e. 1, 3, 4, and 6).

In general; the results indicate that IDWRSFS outperforms local-based search methods (RSAR, and EBR) and are comparable with population-based search

Table 3. Results of 20 runs of IWDRSFS for six UCI datasets. The results are compared with those from four state of the art methods as published in [8].

No.	Dataset	No. of Features	Comparative methods				IWDRSAR
			RSAR	EBR	AntRSAR	GenRSAR	
1	M-of-N	13	8	6	6	$6^{(6)}7^{(12)}$	$6^{(18)}7^{(2)}$
2	CREDIT	20	9	10	$8^{(12)}9^{(4)}10^{(4)}$	$10^{(6)}11^{(14)}$	$10^{(12)}11^{(8)}$
3	LETTERS	25	9	9	8	$8^{(8)}9^{(12)}$	$8^{(16)}9^{(4)}$
4	DERM	34	7	6	$6^{(17)}7^{(3)}$	$10^{(6)}11^{(14)}$	$6^{(2)}7^{(3)}8^{(5)}9^{(5)}10^{(5)}$
5	WQ	38	14	14	$12^{(2)}13^{(7)}14^{(11)}$	16	$13^{(3)}14^{(17)}$
6	LUNG	56	4	4	4	$6^{(8)}7^{(12)}$	$4^{(6)}5^{(12)}6^{(2)}$

methods (AntRSAR and GenRSAR). The success of IWDRSFS for FS is owing to the characteristic of exploration, where the solutions from different places in the solution space are explored. Then, the search process is guided by the strategy that maintains the history of the search learned in the previous iterations.

5 Conclusions

This article has proposed a new FS method, i.e. IDWRSFS that combines the IDW algorithm with RS. IWD is used as the search procedure, and RS dependency is used as the subset evaluation function. IWDRSFS has been evaluated using six benchmark data sets. The empirical evaluation has shown that IWDRSFS is suitable for FS with RS, whereby good solutions have been produced. A performance comparative study with four RS-based FS methods has been carried out. The results of IWDRSFS are generally better than those from RSAR and EBR. Furthermore, the results of IWDRSFS are comparable with those from GenRSAR and AntRSAR.

While IWDRSFS has shown good results for FS, future work to improve the robustness of IWDRSFS by adapting a local search method and tuning the IWD parameters can be conducted. In addition, further investigations to verify the usefulness of IWDRSFS for real-valued data sets are required.

Acknowledgements. This research is supported by the grant with id no. $1001/PKOMP/817047$.

References

1. Abdullah, S., Jaddi, N.: Great deluge algorithm for rough set attribute reduction. Database Theory and Application, Bio-Science and Bio-Technology, 189–197 (2010)
2. Blake, C., Merz, C.: {UCI} repository of machine learning databases (1998), `http://www.ics.uci.edu/~mlearn/MLRepository.html`.
3. Dash, M., Liu, H.: Feature selection for classification. Intelligent data analysis 1(1-4), 131–156 (1997)

4. Duan, H., Liu, S., Wu, J.: Novel intelligent water drops optimization approach to single ucav smooth trajectory planning. Aerospace Science and Technology 13(8), 442–449 (2009)
5. Hedar, A., Wang, J., Fukushima, M.: Tabu search for attribute reduction in rough set theory. Soft Computing-A Fusion of Foundations, Methodologies and Applications 12(9), 909–918 (2008)
6. Hendrawan, Y., Murase, H.: Neural-intelligent water drops algorithm to select relevant textural features for developing precision irrigation system using machine vision. Computers and Electronics in Agriculture 77(2), 214–228 (2011)
7. Jensen, R.: A collection of datasets used in feature selection experimentation, `http://users.aber.ac.uk/rkj/test2/` (visited October 14, 2012)
8. Jensen, R., Shen, Q.: Finding rough set reducts with ant colony optimization. In: Proceedings of the 2003 UK Workshop on Computational Intelligence, vol. 1 (2003)
9. Jensen, R., Shen, Q.: Computational intelligence and feature selection: rough and fuzzy approaches, vol. 8. Wiley-IEEE Press (2008)
10. Kim, H., Howland, P., Park, H.: Dimension reduction in text classification with support vector machines. Journal of Machine Learning Research 6(1), 37 (2006)
11. Niu, S., Ong, S., Nee, A.: An improved intelligent water drops algorithm for achieving optimal job-shop scheduling solutions. International Journal of Production Research 50(15), 4192–4205 (2012)
12. Pawlak, Z.: Rough sets. International Journal of Parallel Programming 11(5), 341–356 (1982)
13. Shah-Hosseini, H.: Intelligent water drops algorithm: A new optimization method for solving the multiple knapsack problem. International Journal of Intelligent Computing and Cybernetics 1(2), 193–212 (2008)
14. Shah-Hosseini, H.: An approach to continuous optimization by the intelligent water drops algorithm. Procedia-Social and Behavioral Sciences 32, 224–229 (2012)
15. Shah-Hosseini, H.: Intelligent water drops algorithm for automatic multilevel thresholding of grey–level images using a modified otsu's criterion. International Journal of Modelling, Identification and Control 15(4), 241–249 (2012)
16. Shah-Hosseini, H.: Problem solving by intelligent water drops. In: IEEE Congress on Evolutionary Computation, CEC 2007, pp. 3226–3231. IEEE (2007)
17. Suguna, N., Thanushkodi, K.: A novel rough set reduct algorithm for medical domain based on bee colony optimization. Journal of Computing 2(6), 49–54 (2010)
18. Wang, J., Hedar, A., Zheng, G., Wang, S.: Scatter search for rough set attribute reduction. In: International Joint Conference on Computational Sciences and Optimization, CSO 2009, vol. 1, pp. 531–535. IEEE (2009)
19. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. The Journal of Machine Learning Research 5, 1205–1224 (2004)

Information-Based Scale Saliency Methods with Wavelet Sub-band Energy Density Descriptors

Anh Cat Le Ngo[1], Li-Minn Ang[2], Guoping Qiu[3], and Kah Phooi Seng[4]

[1] School of Engineering, The University of Nottingham, Malaysia Campus
[2] Centre for Communications Engineering Research, Edith Cowan University
[3] School of Computer Science, The University of Nottingham, UK Campus
[4] Department of Computer Science & Networked System, Sunway University

Abstract. Pixel-based scale saliency (PSS) work bases on information estimation of data content and structure in multiscale analysis; its theoretical aspects as well as practical implementation are discussed by Kadir *et al* [11]. Scale Saliency framework [10] does not work only for pixels but other basis-projected descriptors as well. While wavelet atoms, localization in both time and frequency domain, are possible alternative descriptors, no theoretical analysis and practical solutions have been proposed yet. Our contribution is introducing a mathematical model of utilizing wavelet-based descriptors in a correspondent Wavelet-based Scale Saliency (WSS). It treats wavelet sub-band energy density of two popular discrete wavelet transform (DWT) and dual-tree complex wavelet transform (DTCWT) as basis descriptors instead of pixel-value descriptors for saliency map estimation. Then, ROC, AUC, and NSS quantitative analysis are comparing WSS against PSS as well as other state-of-the-art saliency methods ITT [9], SUN [18], SRS [8] on N. Bruce's database [4] with human eye-tracking data as ground-truth. Furthermore, qualitative results, different saliency maps, are analyzed case by case for their pros and cons; especially their short-comings in specific situation or insensible results for human perception.

1 Introduction

After Itti *et al* proposes his multi-scale constrast-based saliency methods, several other mechanisms are suggested for modeling human visual attention (HVA) such as SUN object-based visual attention [18], Spectral Residual Saliency (SRS) [8], etc. In recent years, researchers have also used achievements of information theory and statistical image processing for HVA modeling as well. Pixel-based scale saliency (PSS) by Kadir *et al* [11] on the basic of Gilles principles [13] is among the earliest works of information-based saliency method. Then, the research direction is rapidly developed thanks to Niel Bruce's effort on An Information Maximization (AIM) theory [4], Gao *et al* 's work in Discriminative Information Saliency (DIS) [7], etc. Furthermore, the Entropy-based Saliency framework (ENT) [1,12] simplifies information estimation process and broadens applicable boundary from still images to dynamic videos with its spatio-temporal

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 366–376, 2013.

extension. Most information-based saliency approaches are motivated from assumption that information theory plays important role in HVA. In other words, human attention might be attracted to highly informative location in scenes. This informative measurement in turn could be estimated by different methods on various descriptors; intensity values in PSS [11], DCT coefficients in ENT [12], ICA components in AIM [4] and WAVELET atoms in DIS [7]. In accordance with descriptors, ENT and PSS employ histogram construction or Parzen kernel for information estimation; meanwhile, AIM estimates self-information through neural-network, and DIS uses decision theory to optimize discriminating information from classifying descriptors into center or surrounds classes. Noteworthy, PSS is so far the only approach emphasizing on spatial as well as structural information or difference of information across scales; however, it is noise-prone, rotation-variant and inefficient due to information estimation of high-dimensional data.

Projection-based atoms are recommended as alternative descriptors by Kadir *et al* [11] [10] since their sparse data representation not only boosts practical performance but also provides deeper theoretical understanding of scale saliency and multi-scale structural information. Moreover, it would make scale saliency the first-ever saliency framework capable of using both pixel-value and basis-projection descriptors. Therefore, our main contribution is proposing two scale saliency computation methods, WSS and MIS, with different wavelet sub-band energy density descriptors. Wavelet atoms are utilized because of their compact time-frequency localization; however, they depend on particular morphological shapes of mother wavelets. To justify this dependence, two wavelet descriptors based on discrete wavelet transform (DWT) and dual-tree complex wavelet transform (DTCWT) are considered in the sub-section 3. Along with wavelet-based descriptors, suitable mathematical models are derived for the proposed approaches in the sub-section 4; moreover, it also unveils inherence of WSS and MIS in the stream of information-based saliency methods. Simulations on Neil Bruce Image database [5] provide quantitative evaluation of WSS and MIS against PSS, ITT, SUN, and SRS models. In addition, judging saliency maps against normal human perception provide qualitative measurement on different sample images. These above evaluation are briefly documented in the section 5. The final section summarizes main contributions of this paper and discuss some future research directions.

2 Scale Saliency

$$Y_D(s_p, \boldsymbol{x}) = H_D(s_p, \boldsymbol{x}) W_D(s_p, \boldsymbol{x}) \tag{1}$$

$$H_D(s_p, \boldsymbol{x}) = -\int_{d \in D} p(d, s_p, \boldsymbol{x}) \log_2 (p(d, s_p, \boldsymbol{x}) d(d) \tag{2}$$

$$W_D(s_p, \boldsymbol{x}) = s \int_{d \in D} \left| \frac{\delta p(d, s_p, \boldsymbol{x})}{\delta s} \right| d(d) \tag{3}$$

$$s_p = \{s | \frac{\delta H_D(s, \boldsymbol{x})}{\delta s} = 0; \frac{\delta^2 H_D(s, \boldsymbol{x})}{\delta s^2} < 0\} \tag{4}$$

This section briefly reviews a few fundamental principles of original scales saliency which is defined as a product of maximum feature-space entropy and its inter-scale dependency across scales with above mathematical model. Feature-space saliency, (H_D) in the equation 2, is measured by its Shannon entropy of pixel-values descriptor (d) at a specific scale or sampling window size (s_p) for each image location $(\boldsymbol{x})$. Shannon entropy satisfies first four criteria of multi-scale entropy [15] filtering to estimate information from feature space, but not attend the fifth criterion about structural correlation. The entropy only concerns about uncertainty of features distribution, not their spatial arrangement. Therefore, PSS is incomplete with only feature-space information (H_D), it needs to measure feature correlation across scales (W_D) as well. Inter-scale saliency, the equation 3, actually does reveal some feature correlation due to being defined as derivatives across scales of probability distribution function from feature-space. As in any multi-scale analysis, characteristic scale has profound effect on output results. PSS chooses (s_p) such that most significant information should be contained inside the window of that scale. Beside being rich in information, significant features need to be consistent across scales as well, hence final saliency value is defined as feature-space saliency H_D weighted by W_D, the equation 1. Lets apply the above concept on a general form of signal $R = I + N$, where I is ideal noise free signal, R is the measured signal with noise N. Assumed no dependencies between random noise and the ideal signal with large features, the equation 1 can be written as.

$$Y_D(R) = (H_D(I) + H_D(N))(W_D(I) + W_D(N))$$

Since probability distribution of random noise is scale-invariant, inter-scale saliency of noise component is zero. $W_D(N) = \Delta_{s_i} H_D(N) = 0$. Meanwhile, significant features across scales has strong correlation between consecutive scales $W_D(I) = \Delta_{s_i} H_D(I) > 0$. Therefore, Saliency value of real signal Y_D can be rewritten as.

$$Y_D(R) = (H_D(I) + H_D(N))W_D(I)$$

Obviously, scale saliency purely depends on inter-scale saliency of useful signal $W_D(I)$. This briefly explains basic motivation behind scale saliency work as well as how inter-scale scale is able to describe structural correlation of image features. Further mathematical analysis and experiments results can be found in [11] [10]. The original scale saliency [11] uses pixel-value descriptors which are simple, intuitive, and straight forward interpretation of image data. Moreover, its combination with circular sampling window provides isotropic information analysis, independent of any morphological shape inside sampled regions. Nevertheless, pixel descriptors also have their drawbacks such as noise-sensitivity, high computational cost and significant bias in entropy estimation. With pixel descriptors, histogram and approximated Parzen kernel play basic roles of constructing pixel-value descriptors' PDF and estimating entropy, and their estimation bias and speed performance greatly depend on manual tuning of histogram numbers of bins or Parzen size kernel. In addition, they as well restrict extension of scale saliency to higher dimensional data. Suau [17] overcomes these problems

when bypassing PDF construction stage and using direct multivariate-data-adaptive information estimation technique [16]. However, the non-pdf approach [17] hinders computation of inter-scale saliency, defined as derivative of PDFs across scales, the equation 3. The problem can be overcame by adapting set-theory-based solution of Kadir [10] for inter-scale saliency W_D computation into kd-tree structure. Nevertheless, the solution is no longer straight-forward and unnecessarily complex; therefore, it motivates our development of (WSS), a more coherent, simple and intuitive information-based scale saliency with sub-band energy descriptors.

3 Wavelet Sub-band Descriptor

Last section has clarified basic advantages of wavelet transform in information estimation of local energy density distributions. Kadir *et al* [11,10] has actually argued possible usage of spectral distribution as saliency measure. Accordingly, a simple, flat, non-salient image region would be fully described by a single sub-band; meanwhile, complex data and structure regions would have required more sub-bands descriptors. Therefore, information amount at a spatial location would be proportional to complexity of the distribution; then, basis-projected sub-bands are potential alternative descriptors. Due to uncertainty principles of time-frequency distribution, wavelet sub-band energy descriptors need treating as discrete variables. Following available mathematical definition of PSS for discrete pixel descriptors, we roughly sketch mathematical models of WSS.

$$Y_{\mathbf{E_p}}(s_p, \boldsymbol{x}) \triangleq H_{\mathbf{E}}(s_p, \boldsymbol{x}) W_{\mathbf{E}}(s_p, \boldsymbol{x}) \tag{5}$$

$$H_{\mathbf{E}}(s, \boldsymbol{x}) \triangleq -\sum_{e \in E} p_{e,s,\boldsymbol{x}} log_2 p_{e,s,\boldsymbol{x}} \tag{6}$$

$$W_{\mathbf{E}}(s, \boldsymbol{x}) \triangleq \frac{s^2}{2s-1} \sum_{e \in E} |p_{e,s,\boldsymbol{x}} - p_{e,s-1,\boldsymbol{x}}| \tag{7}$$

$$H_{\mathbf{E}}(s_p - 1, \boldsymbol{x}) < H_{\mathbf{E}}(s_p, \boldsymbol{x}) > H_{\mathbf{E}}(s_p + 1, \boldsymbol{x})\} \tag{8}$$

where $\mathbf{E} = \{e_1, e_2, \ldots e_m\}$ is a set of sub-band energy (e) wrt location and scales parameter $(\boldsymbol{x}, s)$. A general concept of sub-band descriptor has been formulated; then, specific details of DWT and DTCWT descriptors will be discussed in the rest of this section. Lets consider wavelet coefficients ($\mathbf{w_i}$), the equation 9, generated by 2-D discrete real-wavelet transform with three sub-bands vertical (v), horizontal (h) and diagonal (d) sub-bands at each particular dyadic scale (s). Square of ($\mathbf{w_i}$) is sub-band energy density descriptors (e), the equation 10.

$$\mathbf{w_i}[f(x, y, s_j)]|_{i=\{v,h,d\}, s_j \in \mathbf{S}} = f(x, y) * \psi_{s,i}(x, y, s_j) \tag{9}$$

$$\mathbf{P_e}\{\mathbf{w_i}[f(x, y, s_j)]\}|_{i=\{v,h,d\}, s_j \in \mathbf{S}} = |\mathbf{w_i}[f(x, y, s_j)]|^2 \tag{10}$$

At each level of decomposition, DWT has three separated sub-bands. Proposed that the maximum decomposition level is 4 or 5, each image can be totally described with 12 or 15 sub-bands. DWT needs much less descriptors than normal

number of pixel values for any gray-scale image; however, it suffers from the shift-variance drawback.

$$\text{With } \exists x, y, s_j, \mathbf{w_i}, \Delta(x), \Delta(y) : \tag{11}$$

$$\mathbf{w_i} f(x, y, s_j) \neq \mathbf{w_i} f(x + \Delta(x), y + \Delta(y), s_j) \tag{12}$$

$$\mathbf{P_e}\{\mathbf{w_i} f(x, y, s_j)\} \neq P\{\mathbf{w_i} f(x + \Delta(x), y + \Delta(y), s_j)\} \tag{13}$$

Consequently, data projection on DWT basis relies on both signal values and its relative location, which contradicts to the fourth criteria for good information measurement of Starck *et al* [15]: *"Entropy must work in the same way regardless of descriptors' locations"*. Therefore, utilization of shift-variant descriptors might lead to different information estimation for the same data at two different locations. In wavelet literature, the shift-variance is solved by extension to complex wavelet methods; for instances, recently developed dual-tree complex wavelet transforms (DTCWT) [14], and its derivation Quaternion wavelet transform (QWT) [6]. Either of the above shift-invariant wavelet transform depends on a dual-tree approache in which two different wavelet filter-banks ,$\{\psi_g, \psi_h\}$, are specially designed to form analytical complex pairs.

$$(x, y, s) = \psi_g(x, y, s) + j\psi_h(x, y, s) \tag{14}$$

$$\psi_h(x, y, s) \approx H\left(\psi_g(x, y, s)\right) \tag{15}$$

The magnitudes of projected-complex coefficients are proven to be shift-invariant; therefore, its derived energy density of the sub-bands are as well shift-invariant. In this paper, we propose usage of QWT scheme [6] instead of DTCWT [14] due to its strong relation to Quaternion Fourier Transform (QFT). In fact, Chan *et al* [6] points out equivalence between local QFT and QWT which is analogous to STFT and DWT relation. Therefore, QWT descriptors can be easily integrated and explained in the proposed methods.

4 Information Measurement

As the last section introduces four possible sub-band descriptors in accordance with different wavelet transforms. Hence, PDF of energy density at each scale s_j can be computed as follows. With $i = \{v, h, d\})\}$ and $1 <= j <= m$:

$$p\{\mathbf{P_e}\} = \frac{\mathbf{P_e}\{\mathbf{w_i}[f(x, y, s_j)]\}}{\sum_j \sum_i \mathbf{P_e}\{\mathbf{w_i}[f(x, y, s_j)]\}} \tag{16}$$

Equation 16 estimates information at each location (x, y) across different sub-bands, $i = \{v, h, d\}$ (DWT, DTCWT) from the smallest scale, $j = 1$, to the currently considered scale, $j = m$. First scale possesses smallest wavelet kernels for analyzing finest image details. Then, sampling windows are doubled after each level; coarser features are generated. Noteworthy, scales are doubled in the proposed methods rather than increased by a unit in PSS. In addition, probability distribution function (PDF) has horizontal axis with increasing scales j from

level 1 (smallest wavelet atom) to level m (currently largest wavelet atom). With PDF in the equation 16 is computed a feature-space entropy as follows.

$$H\left(x, y, s_m\right) = -\sum p_i(x, y, s_j) \log p_i(x, y, s_j) \tag{17}$$

Where $p_i(x, y, s_j)$ is a short form of $p_i\left(\mathbf{P}\{\mathbf{w_i}\left[f(x, y, s_j)\right]\}\right)$ and $i = \{v, h, d\}$ and $1 <= j <= m$. The equation 17 computes feature-space entropy H_E of sub-band energy descriptors for WSS and MIS as the equation 6 does for PSS. The rest of scale saliency task is identifying computational details of inter-scale saliency W_E; in other words, determining how the equation 7 should be interpreted with wavelet-based descriptors. In the equation 7, W_D is measured as total variation in probability distribution of descriptors at two consecutive scales, and a specific pixel-value descriptor (d) can appear in both distributions. It complicates information estimation process for W_D. Meanwhile, the proposed descriptors have unique sub-bands at each level, it simplifies construction of sub-band probability distribution for different levels. However, the uniqueness makes the equation 3 inappropriate for sub-band features since it is inappropriate to find differentiation of two PDF on two different set of descriptors. Therefore, an alternative interpretation of inter-scale saliency need developing. Lets consider $p(M)$ as PDF of all sub-bands up to the current m level, sub-bands D from next level $m+1$. The D descriptors will modify the current $p(M)$ into $p(M|D)$, and distance between a prior model and a modified model can be measured by Kullback-Leibler divergence as follows.

$$K(P(M|D), P(M)) = \int_M P(M|D) \log \frac{P(M|D)}{P(M)} \tag{18}$$

Noteworthy, it is similar to Itti's Bayesian Surprise Saliency (BSS) metric [2], and the surprise model can be extended for multiple sub-band descriptors or evidences in BSS. The equation 18 becomes mutual information between the current model (M) and a set of new evidences (D). In other words, expectation of surprise for adding new sub-bands into the current model is their mutual information, shown in equation 20.

$$MI(D, M) = \int_D K(P(M|D), P(M)) \tag{19}$$

$$= \int_{D,M} P(D, M) \log \frac{P(D, M)}{P(M)P(D)} \tag{20}$$

Therefore, the mutual information in the equation 20 is chosen as inter-scale saliency for successive dyadic scales since it actually implies averaged "bayesian surprise" [2] saliency of sub-bands across scales. Furthermore, mutual information as inter-scale saliency measurement well emphasizes structural coherence of data across scales. If there are consistent features across consecutive scales such as edges or joints and, they will increase mutual information between two consecutive scales. Otherwise noises have no mutual information across scales as its self-information is zero, $I(N, N) = 0$. It is remarkable that mutual information

satisfies the fifth criterion of the good information estimation by Starck *et al* [15]. The only remaining step is identifying how the mutual should be calculated in discrete cases. Following formula shows relations between mutual information and entropy with $i = \{v, h, d\}$.

$$MI(D, M) = \quad H(D) \quad + \quad H(M) \quad - \quad H(D, M) \tag{21}$$

$$H(M) = -\sum_{\{j \leq m\}} p_i(x, y, s_j) \log p_i(x, y, s_j) \tag{22}$$

$$H(D) = -\sum_{\{j = m\}} p_i(x, y, s_j) \log p_i(x, y, s_j) \tag{23}$$

$$H(D, M) = -\sum_{\{j \leq m+1\}} p_i(x, y, s_j) \log p_i(x, y, s_j) \tag{24}$$

The mutual information can be directly calculated as difference between separated entropy of M,D ($H(D)+H(M)$) and joint entropy $H(D, M)$, the equation 21; meanwhile, $H(D), H(M), H(D, M)$ can be easily estimated by simple mathematical equations 22,23,24. The joint entropy $H(D, M)$ can be reused as $H(M)$ for a next processing scale due to uniqueness property of the descriptors. Proposed mathematical principles on wavelet-domain sub-band energy descriptors are summarized in the equation 25 as product of maximum feature-space saliency and inter-scale saliency, or product of mutual information between consecutive levels and maximum sub-band entropy.

$$\begin{aligned}
&H(M_{x,y,s_p}) = -\sum p_i(x, y, s_p) \log p_i(x, y, s_p) \\
&MI(D_{x,y,s_p}, M_{x,y,s_p-1}) = H(D) + H(M) - H(D, M) \\
&H(M_{s_p-1,x,y}) < H(M_{s_p,x,y}) \wedge H(M_{s,x,y}) > H(M_{s_p+1,x,y}) \\
&Y(M_{x,y,s_p}) = H(M_{x,y,s_p}) * MI(D_{x,y,s}, M_{x,y,s_p-1})
\end{aligned} \tag{25}$$

Where $i = \{v, h, d\} \vee i = B^2(\mathbf{w_i}), 1 <= j <= m$. A characteristic scale s_p is chosen to maximize information of model $H(M(s, x, y))$. Lets imagine that an image at a prior scale contains only noise, meanwhile at later scales it actually contains useful structures. With bias of Shannon entropy toward noise, WSS choose a prior scale which fails to enclose most useful structure. To overcome this drawback, we propose alternative approach, *MIS* in which s_p is selected so as to maximize inter-scale saliency or average "Bayesian surprise". *MIS* inherits almost all mathematical principles of WSS, equation 25 but choose a characteristic scale s_p according to the maximum value of inter-scale saliency W_D instead of feature-space saliency H_D.

$$\begin{aligned}
MI(D_{s-1}, M_{s-2}) < MI(D_s, M_{s-1}) \\
MI(D_s, M_{s-1}) > MI(D_{s+1}, M_s)
\end{aligned}$$

Experiments and results of *MIS* and *WSS* are detailed in the next section for evaluating and comparing proposed methods against state-of-the-art solutions.

5 Discussion and Results

This section leads a discussion to pros and cons of the proposed methods. WSS1, WSS2 with correspondent descriptors, DWT and DTCWT, are examined with scale selection mechanism WSS. If MIS approach is chosen instead, saliency solutions are named MIS1 and MIS2 accordingly. Both WSSs and MISs are compared against PSS [11], ITT model [9], AIM model [4], SUN model [18], and SRS model [8]. Bruce and Tsotsos [5] database is chosen mainly due to availability of their eye-tracking data which provides ground-truth for quantitative evaluation of different saliency maps and human visual performance with appropriate statistical tests (shuffled ROC, AUC, NSS). Noteworthy, all quantitative tests have scrutinized data on the basis of eliminating center-bias or other unwanted properties. While quantitative results give a general idea how each method behaves against human performance, qualitative comparisons with visual saliency maps reveal performance on individual samples and whether saliency methods give reasonable match to human perception or not.

5.1 Quantitative Results

The quantitative performance is characterized by shuffled Receiver Operating Characteristics (ROC) curves, Area Under ROC Curve (AUC), and Normalized Scanpath Saliency (NSS) as numerical results. To ensure accuracy and fairness of results analysis, we employ open-source evaluation codes for shuffled AUC and NSS [3]. Noteworthy, saliency maps are standardized around median instead mean of distributions; moreover, shuffled AUC is introduced to limit center-bias in eye-fixation evaluation. Images from Niel Bruce's database are varied from outdoor context to indoor environment, accompanied with eye-tracking data of 20 human subjects. The ground-truth data are recorded in carefully set-up psychological experiments. Furthermore, the database has been repeatedly used in other saliency methods' evaluation though it is relatively small with 120 images.

Table 1 shows performances of six different methods from top-bottom order: ITT, AIM, SRS, SUN, WSS1, WSS2, PSS. All WSS methods are better than PSS performance and comparable to ITT method. Especially, a computational time of WSS1 (1.2689s) is approximately 7 times less than that of PSS (7.1092s),

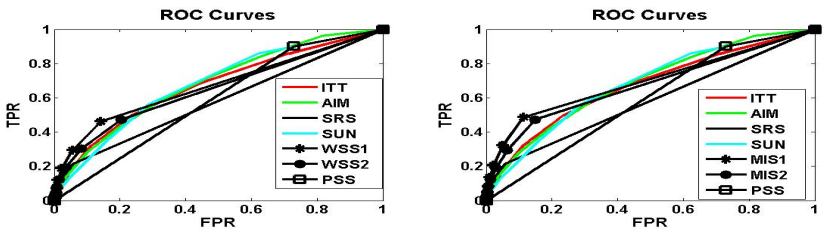


Fig. 1. ROC Curve of (a) WSS , (b) MIS methods (left-to-right order)

Table 1. AUC & NSS

MTH	ITT	AIM	SRS	SUN	WSS1	WSS2	MIS1	MIS2	PSS
AUC	0.694	0.724	0.584	0.718	**0.678**	**0.663**	**0.703**	**0.692**	0.586
NSS	0.277	0.124	0.198	0.208	**0.334**	**0.300**	**0.318**	**0.302**	-0.392

while PSS is implemented in C++, and WSSs are totally written in MATLAB. Quantitative performances of MIS methods, figure 1(b) and table 1 are even better than those of WSS. Their AUC results are comparable to other state-of-the-art methods like AIM, SRS, and SUN; meanwhile, their performances in NSS criteria surpass the others. Both MIS and WSS, which is highly sparse but quite accurate saliency maps according to eye fixation map, outperform the other methods in the low false-positive-rate region.

5.2 Qualitative Results

Qualitative samples in figures 2(a),2(b),2(c) are composed of original samples, ITT, MIS1, and MIS2 maps in left-right, top-bottom order. Saliency maps of only three methods are displayed due to lack of space and similarity of generate saliency maps. For first two samples, similar saliency maps are generated and quite matched to human perception, the figure 2(a); however, slight difference can be spotted as well. For example, ITT identifies a white button as partly salient while MIS1 and MIS2 misses it in figure 2(a). Moreover, MIS1 and MIS2 behave slightly differently as well in the figure 2(b). Finally, reasonable results can not be obtained from any methods like the figure 2(c). It usually happens if images are flooded with complex textures.

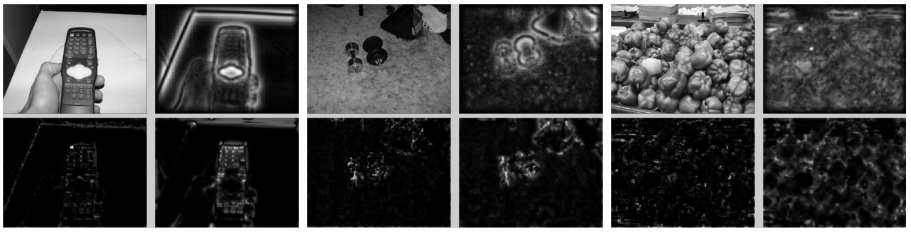


Fig. 2. Saliency Maps of (a) Image 1, Image (b) 2, (c) Image 3 (left-to-right order)

6 Conclusion

In this paper, we propose the extension of scale saliency from pixel descriptors to sub-band energy density descriptors of DWT and DTCWT wavelet transform. Comparing to pixels, the proposed descriptors are much more sparse representation with other properties such as shift-invariance, best-basis approaches; however, they are biased toward morphological shapes of mother wavelets. Along

with new descriptors are proposed, an innovative coherent information framework, and its strong relations with Bayesian Surprise Model [2] are emphasized. Beside solid theoretical development, the experimental results show competitiveness of the proposed methods against other state-of-the-art models and surpasses the original scale saliency model PSS quantitatively and qualitatively. In future research, theoretical analysis will be extended to include prior information or top-down information, perceptual grouping and other visual attention operations.

References

1. Anh Cat, L.N., Qiu, G., Geoff, U., Li-minn, A., Kah Phooi, S.: Visual Information Based on Fast nonparametric multidimensional entropy estimation. International Conference on Acoustic, Speed and Signal Processing (1) (2012)
2. Baldi, P., Itti, L.: Of bits and wows: A Bayesian theory of surprise with applications to attention. Neural Networks: The Official Journal of the International Neural Network Society 23(5), 649–666 (2010)
3. Borji, A., Sihite, D.N., Itti, L.: Quantitative Analysis of Human-Model Agreement in Visual Saliency Modeling: A Comparative Study. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society PP(99), 1–16 (2012)
4. Bruce, N.D.B., Tsotsos, J.K.: Saliency based on information maximization. In: Advances in Neural Information Processing Systems, vol. 18, p. 155 (2006)
5. Bruce, N.D.B., Tsotsos, J.K.: Saliency, attention, and visual search: An information theoretic approach. Journal of Vision 9, 1–24 (2009)
6. Chan, W.L., Choi, H., Baraniuk, R.G.: Coherent multiscale image processing using dual-tree quaternion wavelets. IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society 17(7), 1069–1082 (2008)
7. Gao, D., Mahadevan, V., Vasconcelos, N.: The discriminant center-surround hypothesis for bottom-up saliency. In: Advances in Neural Information Processing Systems, vol. 20, pp. 1–8 (2007)
8. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007). No. 800, pp. 1–8. IEEE Computer Society, Citeseer (2007)
9. Itti, L., Koch, C., Niebur, E., et al.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998)
10. Kadir, T., Boukerroui, D., Brady, M.: An analysis of the scale saliency algorithm. OUEL No: 2264 3, 1–38 (2003)
11. Kadir, T., Brady, M.: Saliency, scale and image description. International Journal of Computer Vision 45(2), 83–105 (2001)
12. Qiu, G., Gu, X., Chen, Z., Chen, Q., Wang, C.: An information theoretic model of spatiotemporal visual saliency. To Appear, International Conference on Multimedia and Expo., pp. 1806–1809. Citeseer (2007)
13. Gilles, S.: Robust Description and Matching of Images. Ph.D. thesis, University of Oxford (1998)

14. Selesnick, I., Baraniuk, R., Kingsbury, N.: The dual-tree complex wavelet transform. IEEE Signal Processing Magazine 22(6), 123–151 (2005)
15. Starck, J., Murtagh, F.: Multiscale entropy filtering. Signal Processing 76 (1999)
16. Stowell, D., Plumbley, M.D.: Fast multidimensional entropy estimation by k-d partitioning. Signal Processing 16(6), 537–540 (2009)
17. Suau, P., Escolano, F.: A New Feasible Approach to Multi-dimensional Scale Saliency. In: Blanc-Talon, J., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2009. LNCS, vol. 5807, pp. 77–88. Springer, Heidelberg (2009)
18. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: SUN: A Bayesian framework for saliency using natural statistics. Journal of Vision 8(7), 321–20 (2008)

Feature Subset Selection Using Binary Gravitational Search Algorithm for Intrusion Detection System

Amir Rajabi Behjat[1], Aida Mustapha[1], Hossein Nezamabadi – pour[2], Md. Nasir Sulaiman[1], and Norwati Mustapha[1]

[1] Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
[2] Department of Electrical Engineering, Shahid Bahonar University of Kerman
rajabi.amir6@gmail.com, {aida,nasir,norwati}@fsktm.upm.edu.my, nezam@mail.uk.ac.ir

Abstract. Due to control different infrastructures of networked computers in cyber security, intrusion detection system has been an important task essentially. Today, an effective intrusion detection system utilizes computational methods as machine learning techniques to improve detection rate with lowest false positive rate; however large number of irrelevant features as an optimization problem decrease this rate. This study using Binary Search Gravitational Algorithm (BGSA) as a feature selection method decreases irrelevant features in KDD 99 intrusion detection data set in order to improve Multi-layer perceptron performance. Results show that significant and relevant features increase performance of intrusion detection system near to 100% with lowest computational cost.

Keywords: Intrusion Detection System, KDD 99 dataset, GSA, Feature Selection, MLP.

1 Introduction

Intrusion Detection Systems (IDS) are based on a set of applications which detect attacks before computer systems are being attacked by hackers. These systems are usually installed within the network segments in strategic places. Today, the security administrators have a difficult role in managing IDS because the systems and services have grown increasingly complex while the new attacks and vulnerabilities are continuously arising. Many IDSs are using database of well-known actions to compare the normal and abnormal data or activities for sending alerts when a match is detected [1, 2]. Intrusion detection systems are divided as network based and host based. Even though the base of these systems is the same, but their operation is different. Host-based intrusion detection system monitors accessed and executed files. Normally, it controls file systems, system logs and disk resources. On the other hand, network based system checks network traffic or exchanged packets between computers. There are various attacks that would be detected by host-based detection system such as denial of service (DOS) attacks. In the data monitoring, different

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 377–386, 2013.

techniques have tried to search attack patterns. For example, anomaly detection systems apply normal behavior and identify deviations from this behavior. On the other hand, some techniques based on human input create useful models for normal behavior; however making attack signatures increase learning algorithm. The goal of learning algorithm is to divide normal and attack behaviors according to various features. One of the main issues in intrusion detection system is the large amount of data and features to decrease performance of algorithm and detection rate. Thus, selecting a set of relevant features is a serious problem in this field [3-8].

In 1999, DARPA 98 lincoln Lab dataset was collected with 41 features to provide labeled datasets for comparing different IDS systems. Recently, machine learning techniques (Decision trees, clustering, neural network and support vector machine) have been employed on KDD 99 benchmark to detect DOS and probes attacks. In [3], Information Gain as a feature selection technique selects 12 most relevant features within 41 features for classification of different attacks based on decision trees classifier. However, low accuracy decrease performance of intrusion detection system [3]. In [5], Support Vector Machine (SVM), Multivariate Adaptive Regression Splines (MARSs) and Linear Genetic Programs (LGPs) investigate the best rank of features in intrusion detection system. The result of ranking features shows that the best accuracy belongs to LGP comparing other classifiers. In [7], Genetic Algorithm (GA) is applied to select relevant features with highest rank. They selected 12 relevant features that are caused 99% accuracy of detection system without false positive rate.

In [6], Quantum particle Swarm Optimization (QPSO) and Support Vector Machine (SVM) increased performance of intrusion detection system using relevant features. In fact, the fitness of each particle is identified by SVM accuracy based on training set that uses only selected features. In [7], Genetic Quantum Particle Swarm Optimization (GQPSO) as a feature selection method decreased the number of irrelevant features and obtained the best performance comparing other feature selection methods such as QPSO and PSO algorithm. They believe that reduction of irrelevant features decrease detecting time and training time exactly.

In [9], GA optimized SVM parameters and Selected relevant features in the same time for intrusion detection system. This method minimized the number of features to increase detection rate; however uniformity of features is the main issue that should be exchanged into a new feature space.

The main objective of this study is to select a set of relevant features and evaluate the impact of these features on the computational cost, false positive, and accuracy of intrusion detection system using Binary Gravitational search Algorithm (BGSA) with an existing Multi-layer Perceptron Neural Network (MLPNN) classifier. This study attempts to prove that the high classification accuracy and low false positive are possible through feature reduction that results in lower features set dimensionality.

The paper organization keeps on as follows: Section 2 explain proposed model and applied dataset in this study; section 3 gives details on the experimental setting and findings. Finally, section 4 winds up with some suggestions for future work.

2 Proposed Model

The proposed BGSA algorithm in this paper selects a set of reliable features within extracted features in [3, 5, 6, 7, 9] that can have an effective influence on the MLP classifier accuracy. The main idea is to use the MLP accuracy over an evaluating set as the fitness function into searching the best solutions. Thus, fast classifier required to train all the possible subset of features. In addition, MLP can create more accuracy in comparison with other classifiers such as SVM. The flowchart of the proposed system is given by Figure 1.

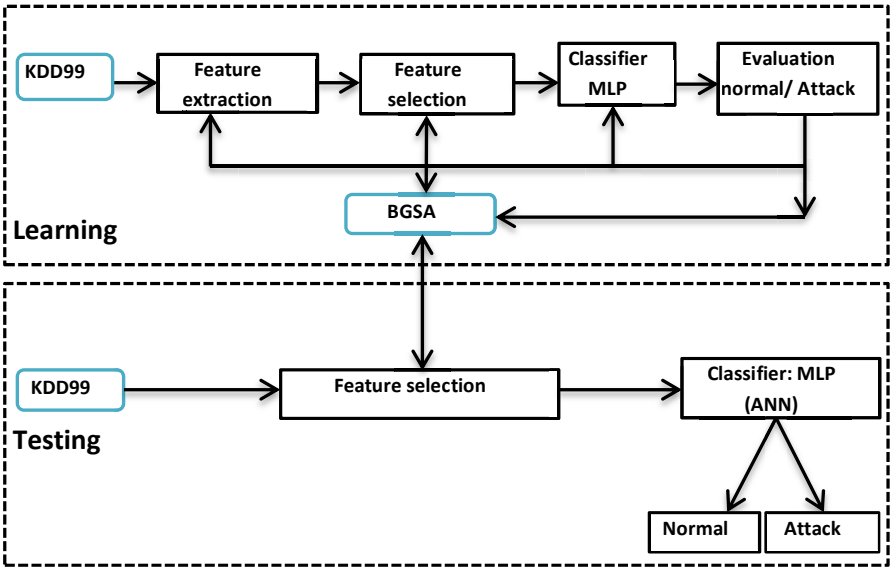


Fig. 1. The flowchart of the proposed system

The proposed system is based on C-class (C=2 for intrusion detection system) in classification process. In fact, in the n-dimensional feature space, each feature is classified into the attack or normal class. The feature selection based on BGSA in the intrusion detection system is designed into the training and performance phases. For this purpose, the system separates the learning data and the testing data according to separated datasets; so corrected KDD of data set is used in the training phase and original KDD is applied in the testing phase. During the training phase, BGSA tries to select the best feature for each class under the best fitness function in. We partition the data into the two classes of normal attack (Prob, DoS, U2Su, R2L) patterns. BGSA is initialized with N masses in n-dimension space. In fact, GSA searches one feature for each class (attack and normal); so, normally the features selected by BGSA are m-dimensional vectors that are used as the classifier inputs. The mass positions identify the candidate solutions for the problem. Then, the fitness rate of the masses is calculated for mass computing in BGSA (Equation 2). The stopping condition shows the end of finding the best feature process. After that, these best features are applied

by performance phase as initial parameters of the MLP classifier. The performance phase starts its function by entered features from the training phase. Based on the performance phase, each class label helps the classifier to find new data or features using the defined distance measure.

2.1 KDD Dataset

The KDD 99 is based on the 1998 DARPA that creates a benchmark for evaluation different methodologies. Thus, a network is content of three target machine is necessary running different operating system. These three machines spoof various IP addresses for producing traffic. At last, all network traffic is recorded by sniffer using TCP dump format. During this process, normal traffic is categorized into one category and attacks are categorized into four categories as follows.

(a) Probe
(b) DoS
(c) U2Su
(d) R2U

The KDD 99 intrusion detection benchmark is content of two characteristics, namely corrected KDD and original KDD that is detailed in Table 1. In this study, original KDD is applied for training set and corrected KDD is employed for testing set.

Table 1. Characteristics of the KDD 99 Intrusion Detection datasets

Dataset	DoS	Probe	U2Su	R2U	Normal
Corrected KDD	229853	4166	70	16347	60593
Original KDD	3883370	41102	52	1126	972780

2.2 Gravitational Search Algorithm

Gravitational Search Algorithm (GSA) is a meta-heuristic optimization tool based on stochastic population, inspired by the Newton's laws of gravity and motion [10]. The law of gravity was defined as "every massive particle in the universe attracts other massive one with a force that is directly proportional to the product of their masses and inversely proportional to the square of the distance between them" [11]. [10] believe that the classical algorithms are practically unable to represent a reasonable solution for solving optimization problems with high-dimensional search space and an exponential search space. Therefore, solving these problems is not practical by exact techniques. On the other hand, GSA-based approach provides an iterative method which simulates mass interactions and moves through a multi-dimensional search space under the influence of gravitation [13]. GSA and its binary Gravitational Search Algorithm (BGSA) have an effective efficiency in solving a set of nonlinear benchmark functions and optimization of engineering problems [12].

GSA acts in a continuous space for optimization of real-valued problems; but most of the optimization problems are set in the binary space, including feature selection, which is a binary encoded optimization problem in this paper.

Real-Valued Gravitational Search Algorithm (RGSA)

Consider a GSA system with N objects in an n-dimensional space. In the RGSA the position of objects which are candidate solutions to the problem are described as:

$$X_i=(x_{i,\dots}^1, x_{i,\dots,}^d x_{i,\dots}^n),i=1,2...N \quad (1)$$

where x_j^d indicates *j-th* agent (object) position in the *d-th* dimension.

Based on [13], when each swarm's member fitness is computed, then the mass of each agent is calculated as follows:

$$M_i(t)=\frac{fit_i(t)-worst\ (t)}{\sum_{j=1}^{N}(fit_i(t)-worst\ (t))} \quad (2)$$

where $M_i(t)$ and $fit_i(t)$ signify the mass and the fitness value of the agent *i* at *t*, and, worst (t) is defined as follows (for a maximization problem):

$$worst\ (t) = \min fit_i(t). \quad (3)$$

$j \in \{1, \dots, N\}$

Calculating the acceleration of an agent should be considered total forces of heavy objects applied to it (Equation 4) by law of gravity (Equation 5).

$$F_i^d(t)=\sum_{j\in kbest,j\neq i} rand_i\ G(t)\frac{M_j(t)\times M_i(t)}{R_{ij}(t)+\varepsilon}(x_j^d(t)-x_j^d(t)) \quad (4)$$

$$a_i^d(t)=\frac{F_i^d(t)}{M_i(t)}=\sum_{j\in kbest,j\neq i} rand_i\ G(t)\frac{M_j(t)}{R_{ij}(t)+\varepsilon}(x_j^d(t)-x_j^d(t)) \quad (5)$$

Next, Equation 6 and Equation 7, the velocity added to its acceleration and the next position is also computed respectively.

$$v_i^d(t+1)=rand_i \times v_i^d(t)+a_i^d(t) \quad (6)$$

$$x_i^d(t+1)=x_i^d(t)+v_i^d(t+1) \quad (7)$$

where *randi* and *randj* are two uniformly distributed random numbers in the interval [0, 1], ε is a small value, and $R_{ij}(t)$ is the Euclidian distance between two agents i and j defined as $R_{ij}(t)=\|X_i\ (t), X_j(t)\|_2$, kbest is the set of first K agents with the best fitness value and biggest mass. kbest is a function of time, it is initialized to K0 at the beginning and decreased with time. Here, K0 is set to N (total number of agents) and is decreased linearly to 1. The gravitational constant, G, is a decreasing Function of time where it is set to G0 at the beginning and is decreased exponentially towards zero at the last iteration as shown in Equation 8.

$$G = G_0 exp\left(-\frac{a\times t}{T}\right) \quad (8)$$

where T is the total number of iterations.

The Equations of updating force, acceleration and velocity are similar for both versions of GSA; but the BGSA is different in computation of the distance between

two agents. BGSA distance measurement is based on Hamming distance. Based on the above concepts, a proper probability function is defined in a small v_i^d that shows the probability of changing x_i^d value from 0 to 1 or contrariwise. Basically, This value must be near zero and for a large v_i^d, the probability of x_i^d must be higher than that. They defined function Sv_i^d to transfer v_i^d into a probability function. Sv_i^d should be restricted within interval [0,1] and increases with increasing v_i^d to be defined according to Equation. 9.

$$S(v_i^d(t+1))=|\tanh(v_i^d(t))| \tag{9}$$

The movement of agents is done by Equation 10.

For obtaining a better coverage rate, the velocity should be limited, $|v_i^d|<v_{max}$. where$v_{max}$is fixed to 6 [12-13].

If rand $<S(v_i^d(t+1))$ then

$$x_i^d(t+1)= \text{complement } (x_i^d(t))$$
$$\text{else}x_i^d(t+1)= (x_i^d(t)) \tag{10}$$

In BGSA G is decreased linearly with time according to Equation 11.

$$G(t)=G_0\left(1-\frac{t}{T}\right) \tag{11}$$

where T is the total number of iterations or the total age of system.

3 Experiments and Results

For classification of attack and normal classes by MLP on BGSA feature selection method, a set of settings are used. The population size and the number of iteration are identified in N=5 and T=20 respectively. Equation 9 indicates that the gravitational constant G and α 1 and 20 respectively. In addition, Classification accuracy and the number of the selected features are the two criteria used to design a fitness function. Thus, the high fitness value needs high classification accuracy and a small number of features. The aim of this paper is to solve the proposed problem by creating a single fitness function which tries to choose a low number of attacked and normal features. As defined by Equation 12, in the fitness function, weight is defined *ωF* for the size of selected feature subset that plays the most important role for fitness function and also affects the weight of classification accuracy (ωA) . Where *ωF* is the weight for the number of the selected features, *ωA* is the weight of MLP classification accuracy and *fi* is the value of feature mask —"1"represents that feature *i* is selected (fi =1) and "0" represents that feature *i* is not selected (fi =0).

$$\text{fit}_i = \text{ca}_i \times \omega\text{A} + [1-\frac{\sum_{i=1}^{p} \text{fi}}{p}] \times \omega\text{F} \tag{12}$$

$$\text{ca}_i=\frac{TP+TN}{TP+FP+FN+TN} \times 100 \tag{13}$$

For fitness evaluation, the percentage of accuracy (ca_i) is important that is defined by Eq. (13). In this equation, the number of spam e-mails which are correctly predicted

as spam (TP), the number of spams which are predicted as non-spam (FN), the number of normal e-mails which are predicted as non-spam (TN) and the number of normal e-mails which are predicted as spam (FP).

Table 2. Selected Features for MLP Classifier

Feature Description
destination bytes: number of bytes received by the source host from the destination host **dst_host_same_srv_rate:** % of connections to same service ports from a destination host **source bytes:** number of bytes sent from the host system to the destination system

Table 3. ROC report of intrusion detection system using MLP classifier on KDD 99

Method	AUC (%)	95% of CI (%)	AUC (SdErr)	P-Value	CPU Time (S)
MLP (GSA)	**100**	**100**	**0**	**<0.001**	**0.005**
GA	99				
PCA	99.23				
GQPSO	96.40				
QPSO	97.77				
IG	98				

In Equation 12, ωf is 0.2. The results of our system show that the fitness function proposed in GSA is equal to 100; for measuring performance of intrusion detection system in this paper, 10 runs are used to measure the performance. In fact, in any iteration, corrected data were used for training; while the original data are used for testing. MLPNN classifies the testing emails into the attack or normal classes.

Note that the number of nodes equals to the size of input vector. In this study, an input vector is collected from a set of "0" and "1". The nodes of hidden layer are tested from 3 to 15. Output layer followed two nodes; first node indicates attack class and second node is normal class. The transfer functions of hidden layer and output layer are 'tansig' and 'purelin', respectively. While the training function is 'trainlm' and the performance function is MSE. The network is trained for a maximum of 60 epochs to 0.01 of error goal. Additionally, the small evaluation fitness allows classifier to have reliable and better results. In addition; this value reduces errors in comparison with other algorithms. Note that table 3 shows the result of feature selection based on BGSA under MLP classifier on KDD 99 dataset that is run more than 10 times. According to this table, the BGSA and MLP results in this study are more reliable than the previous studies. Moreover, during this study, the number of

the features selected for MLP classifier reduces from 41 to 3 in Table 2; while the number of the features of another study are between 12-41 features [3, 5, 6, 7, 9].

Table 3 explains that BGSA is more precise in feature selection and has an efficient performance comparing with other classifiers and some algorithms such as GA, PCA, GQPSO and QPSO [3, 5, 6, 7, 9]. Fig. 2 and Fig. 3 show the performance of BGSA based on MLP classifier. Based on Fig 3, the proposed detection system not only increase accuracy but decrease high-dimensionality and miss rate.

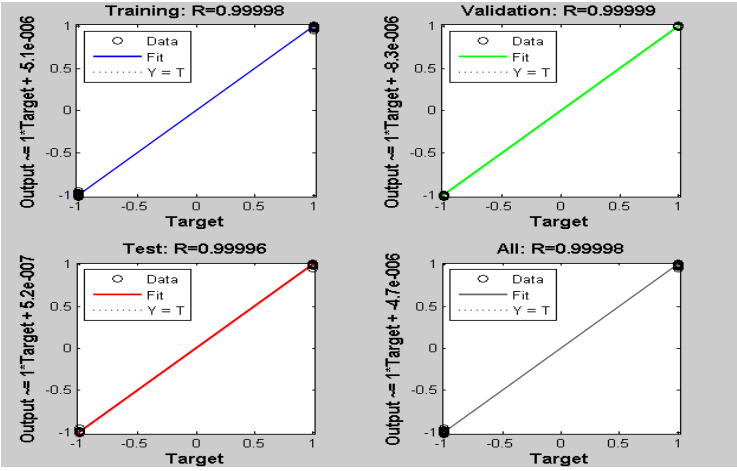


Fig. 2. Training and testing regression of MLP classifier

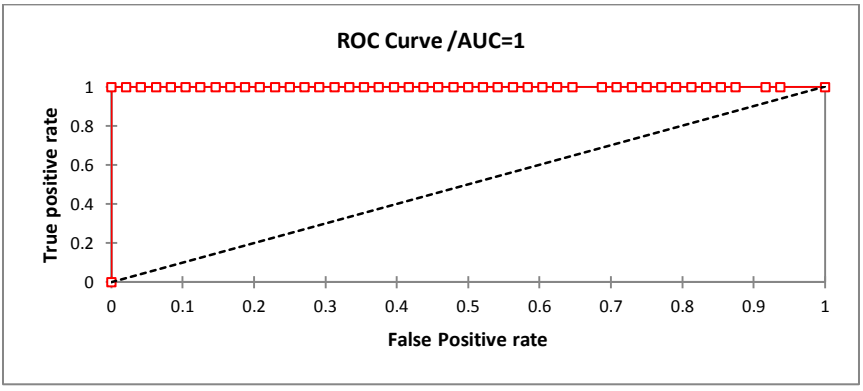


Fig. 3. ROC curve based on MLP classifier

During the testing result in the detection system, the performance measurement is a necessary requirement that evaluates the accurate of detection system as well as the CPU time. In this important aspect, the number of features can change this time. Each detection system tries to decrease this time by decreasing the number of irrelevant features. Hence, the evolution of performance is applied to consider how the

irrelevant features affect the CPU time during feature selection and detection process. Thus, The CPU time, on the opposite, conducted to consider the impact of the number of features on the performance and computational cost. The time and cost are different with changing the number of features. Table 3 shows the CPU time of classification process for BGSA using MLP classifier.

4 Conclusion

In this paper, a BGSA algorithm as a feature selection method together with MLP-based classifier in order to design an intrusion detection system that is able to decrease irrelevant features during the feature selection phase and make binary input vectors of classifier for higher rate of classification accuracy. This study compares the result of proposed system with another system, which applied PSO, GA, QPSO, GQPSO and PCA algorithm as a feature selection method, and indicates better computational performance and higher accuracy. The success of GSA in comparison with other algorithms is to explore and exploit efficiently without getting fixed in local optimum.

References

1. Miller, T.: Social Engineering: Techniques that can bypass Intrusion Detection Systems (2000)
2. Gorton, A.S., Champion, T.G.: Combining Evasion Techniques to Avoid Network Intrusion Detection Systems (2004); International Conference on Data Engineering (DSDE). pp. 169–172. IEEE (2010)
3. Kayacik, H.G., Zincir-Heywood, A.N., Heywood, M.I.: Selecting features for intrusion detection: A feature relevance analysis on KDD 99 intrusion detection datasets. In: Proceedings of the Third Annual Conference on Privacy, Security and Trust (PST 2005). Citeseer (2005)
4. McHugh, J.: Testing intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory. ACM Transactions on Information and System Security 3(4), 262–294 (2000)
5. Mukkamala, S., Sung, A.: Significant feature selection using computational intelligent techniques for intrusion detection. Advanced Methods for Knowledge Discovery from Complex Data, 285–306 (2005)
6. Zhang, H., Gao, H., Wang, X.: Quantum Particle Swarm Optimization Based Network Intrusion Feature Selection and Detection. In: Proc. of the 17th World Congress IFAC, pp. 12312–12317 (2008)
7. Gong, S., Gong, X., Bi, X.: Feature selection method for network intrusion based on GQPSO attribute reduction. In: 2011 International Conference on Multimedia Technology (ICMT), pp. 6365–6368. IEEE (2011)
8. Patcha, A., Park, J.M.: An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer Networks 51(12), 3448–3470 (2007)
9. Ahmad, I., Abdullah, A.B., Alghamdi, A.S., Hussain, M., Nafjan, K.: Features Subset Selection for Network Intrusion Detection Mechanism Using Genetic Eigen Vectors. In: Proceedings of 2011 International Conference on Telecommunication Technology and Applications (ICTTA 2011), pp. 75–79 (2011)

10. Rashedi, E., Nezamabadi-pour, H., Saryazdi, S.: GSA: a gravitational search algorithm. Information Sciences 179(13), 2232–2248 (2009)
11. Michalak, K., Kwasnicka, H.: Correlation-based feature selection strategy in neural classification. In: Sixth International Conference on Intelligent Systems Design and Applications (ISDA 2006), vol. 1, pp. 741–746. IEEE (2006)
12. Rashedi, E., Nezamabadi-Pour, H., Saryazdi, S.: BGSA: binary gravitational search algorithm. Natural Computing 9(3), 727–745 (2010)
13. Sarafrazi, S., Nezamabadi-pour, H.: Facing the classification of binary problems with a GSA-SVM hybrid system. Mathematical and Computer Modelling (2011)

Sparse Signal Recovery by Difference of Convex Functions Algorithms

Hoai An Le Thi, Bich Thuy Nguyen Thi, and Hoai Minh Le

Laboratory of Theoretical and Applied Computer Science - LITA EA 3097,
University of Lorraine, Ile de Saulcy, 57045 Metz, France
{hoai-an.le-thi,minh.le}@univ-lorraine.fr,
thi-bich-thuy.nguyen9@etu.univ-lorraine.fr

Abstract. This paper deals with the problem of signal recovery which is formulated as a l_0-minimization problem. Using two appropriate continuous approximations of $l_0 - norm$, we reformulate the problem as a DC (Difference of Convex functions) program. DCA (DC Algorithm) is then developed to solve the resulting problems. Computational experiments on several datasets show the efficiency of our methods.

Keywords: Compressed Sensing, Sparse Recovery, $l_0 - norm$, DC Programming, DCA.

1 Introduction

Compressed Sensing or Compressive Sensing (CS) which was introduced by Donoho [9] and Candes et al. [5] is an emerging area having significant interest in data analysis. It can be used for compressing higher dimensional data sets to lower dimensional ones for data analysis, signal processing and feature selection applications. Since CS was introduced, it is applied in various fields including radar imaging, signal extraction, aerial laser scanning, medical imaging, surface metrology, through wall radar imaging, space based imaging, ground penetrating radar imaging in archeology, geophysics, oil-exploration, landmine detection, forensics, civil engineering, etc.

Let us firstly give some basic definitions and notations in CS. For a complete study of CS the reader is referred to [8] and the references therein. We rely on a signal representation in a given basis $\{\psi_i\}_{i=1}^n$ for $\mathbb{R}^n$. Every signal $x \in \mathbb{R}^n$ is representable in terms of n coefficients $\{\theta_i\}_{i=1}^n$ as $x = \sum_{i=1}^{n} \psi_i \theta_i$. Arranging the ψ_i as columns into the $n \times n$ matrix Ψ and the coefficients θ_i into the $n \times 1$ coefficient vector θ, we can write that $x = \Psi\theta$, with $\theta \in \mathbb{R}^n$. In a general setting, we refer to Ψ as the sparsifying dictionary [8].

A vector $x \in \mathbb{R}^n$ is called $k - sparse$ in the basis or frame Ψ if there exists a vector $\theta \in \mathbb{R}^n$ with only $k \ll n$ nonzero entries such that $x = \Psi\theta$. The set of indices nonzero entries is called the *support* of θ and denote it by *supp(θ)*.

If a signal is not sparse itself, we can sparsify it by choosing an appropriate representation system. There exist various well known transforms to sparsify a signal, for example Fourier, Cosine, wavelet, curvelet,...

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 387–397, 2013.

The research in CS can be classified into two major contribution areas. The first one consists of the theory and applications related to finding a sensing matrix A to ensure that it preserves the information of the signal x. The second area includes reconstruction techniques for recovering the original sparse signal x from its measurement $y = Ax$ via a sensing matrix A.

In this paper, we consider the problem of sparse signal recovery. Suppose that the signal x is already sparse. The problem can be stated as follows. Given a sensing matrix $A \in \mathbb{R}^{m\times n}$ $(m << n)$ and a measurement vector $y = Ax \in \mathbb{R}^m$. We are to recover the sparse signal $x \in \mathbb{R}^n$.

Since the linear system $y = Ax$ is highly underdetermined, and has therefore infinitely many solutions, the recovery sparse signal can be seen as finding mutually the sparsest signal x being consistent with its measurement. This leads to solving the ℓ_0-minimization problem:

$$(P) \qquad \alpha := \min \{||x||_0 : Ax = y, x \in \mathbb{R}^n\}, \tag{1}$$

where $\|x\|_0 = \#\{i : x_i \neq 0, i = 1, \ldots, n\}$ is the sparsity of x.

An alternative version of (P), due to the hight underdetermination of the linear system $y = Ax$, can be formulated as follows: finding a sparse signal vector x which is as consistent with y as possible according to the square error criterion. Then the resulting optimization problem is written as

$$(P_\rho) \qquad \alpha_\rho := \min \left\{ \|Ax - y\|^2 + \rho\|x\|_0 : x \in \mathbb{R}^n \right\}. \tag{2}$$

where $\rho > 0$, called the regularized parameter, represents a tradeoff between error and sparsity.

It is well known that the problem of minimizing the zero-norm is NP-Hard ([1]). In the literature, several works in convex and nonconvex optimization approaches have been developed for solving the optimization problems dealing with ℓ_0 norm in various contexts that include sparse signal recovery and feature selection. Three of the best known strategies using convex approaches consist in discarding the ℓ_0 term and adding the ℓ_1-norm (see, e.g. [19]) or the ℓ_2- norm (see, e.g. [12]) or the both (see, e.g. [22]) to least-squares loss function. At the same time, nonconvex continuous approaches were extensively developed in the contex of feature selection in which the ℓ_0 term $\|x\|_0$ is approximated by a nonconvex function. Several approximations have been proposed. The first is concave exponential approximation developed in [2]. Lately, other very used approximations are Smoothly Clipped Absolute Deviation (SCAD) ([11]), the logarithmic approximation of Weston et al. [21], and the piecewise concave approximation proposed by Le Thi et al. in [16]. A common point of these approximations is that the resulting optimization problems are all DC (Difference of Convex functions) programs and thereby one can investigate DCA, an efficient method in nonconvex programming framework for solving them. Works in this direction can be found in [2] where the authors proposed SLA (Successive Linear Approximation) algorithm, is a special version of DCA, for solving the resulting problem with concave exponential approximation, in ([7]) where Candes et al. developed a log

penalty method using the logarithmic approximation (this algorithm is known under the name "reweighted ℓ_1 minimization") and in [16,17] where Le Thi et al. proposed a DCA scheme for the resulting optimization problem in case of piecewise concave approximation ([16]) and of SCAD approximation ([17]).

Motivated by the success of DCA in the previous works we propose to develop it for sparse signal recovering. We consider the problem (P_ρ) in which the ℓ_0 term is replaced by the piecewise concave approximation ([16]) and/or the SCAD approximation ([17]). The resulting problems are reformulated as DC programs and then solved by DCA.

The remainder of the paper is organized as follows. DC programming and DCA are briefly presented in Section 2. Section 3 deals with DCA for solving the two resulting optimization problems of sparse signal recovery. Finally, computational results are reported in the last section.

2 Outline of DC Programming and DCA

DC programming and DCA which constitute the backbone of smooth/nonsmooth nonconvex programming and global optimization were introduced by Pham Dinh Tao in a preliminary form in 1985. These tools have been extensively developed since 1994 by Le Thi Hoai An and Pham Dinh Tao (see e.g. [13,15] and the references therein) and become now classic and increasingly popular (see the list of references in [14]). They address the problem of minimizing a function f which is the difference of convex functions on the whole space $\mathbb{R}^d$ or on a convex set $C \subset \mathbb{R}^d$. Generally speaking, a DC program is an optimisation problem of the form :

$$\alpha = \inf\{f(x) := g(x) - h(x) : x \in \mathbb{R}^d\} \qquad (P_{dc})$$

where g, h are lower semi-continuous proper convex functions on $\mathbb{R}^d$. The convex constraint $x \in C$ can be incorporated in the objective function of (P_{dc}) by using the indicator function on C denoted by χ_C which is defined by $\chi_C(x) = 0$ if $x \in C$, and $+\infty$ otherwise.

2.1 Generic DCA Scheme

The main idea behind DCA is to replace in the DC program (P_{dc}), at the current point x^l of iteration l, the second component h with its affine minorization defined by

$$h_l(x) := h(x^l) + \langle x - x^l, y^l \rangle, \quad y^l \in \partial h(x^l)$$

to give birth to the convex program of the form

$$(P_l) \quad \inf\{g(x) - h_l(x) : x \in \mathbb{R}^n\} \Longleftrightarrow \inf\{g(x) - \langle x, y^l \rangle : x \in \mathbb{R}^d\}$$

whose optimal solution is taken as x^{l+1}. The generic DCA scheme is described as follows.

DCA scheme
Initialization: Let $x^0 \in \mathbb{R}^d$ be a best guess, $l = 0$.
Repeat

- Calculate $y^l \in \partial h(x^l)$
- Calculate $x^{l+1} \in \arg\min\{g(x) - h(x^l) - \langle x - x^l, y^l\rangle : x \in \mathbb{R}^d\}$ (P_l)
- $l = l + 1$

Until convergence of $\{x^l\}$.

2.2 DCA'S Convergence Properties

Convergence properties of DCA and its theoretical basis can be found in ([13,15]). For instance it is important to mention that

- DCA is a descent method without linesearch: the sequences $\{g(x^l) - h(x^l)\}$ is decreasing.
- If the optimal value α of problem (P_{dc}) is finite and the infinite sequences $\{x^l\}$ is bounded, then every limit point x^* of the sequence $\{x^l\}$is a critical point of $g - h$, i.e. $\partial h(x^*) \cap \partial g(x^*) \neq \emptyset$.
- DCA has a *linear convergence* for DC programs.

For a complete study of DC programming and DCA the reader is referred to [13,15] and the references therein.

3 Sparse Signal Recovery by DC Programming and DCA

Before developing DCA for solving the problem (P_ρ) it is important to study the relation between the two problems (P) and (P_ρ). Intuitively, if ρ decreases to zero, we attach more importance in $\|Ax - b\|_2^2$ and it seems normal that, if x^ρ is a solution of (P^ρ), $\|Ax^\rho - b\|_2^2$ decreases and x^ρ becomes a good approximation of a solution of (P). The following proposition expresses it in a rigorous way.

Proposition 1. *Assume that the linear system $Ax = y$ admits a solution. Then*

1. $\alpha_\rho \leq \rho\alpha$, $\forall \rho > 0$,
2. *there exists $\rho_0 > 0$, such that $\alpha_\rho = \rho\alpha$ and (P) and (P_ρ) have the same solution set, $\forall 0 < \rho \leq \rho_0$.*

Proof. The property 1) comes from the fact that any feasible point x of (P) is feasible for (P^ρ) and satisfies $Ax = y$.
The proof of 2) can be found in [18].

We develop now DC programming and DCA for solving (P_ρ) via the piecewise concave and/or the SCAD approximation of ℓ_0-term.

3.1 The First DCA Scheme via the Piecewise Concave Approximation

For $t \in \mathbb{R}$, let η be the function defined by

$$\eta(t, \lambda) = 1 - e^{-\lambda|t|}, \tag{3}$$

where $\lambda > 0$ and e denotes the base of the natural logarithm. In what follows, for a given λ, we will use $\eta(t)$ instead of $\eta(t, \lambda)$.

The piecewise concave approximation of ℓ_0-norm given in [16] is the following: $\|x\|_0 \simeq \sum_{i=1}^{n} \eta(x_i)$. It is easy to see that $\eta(t)$ is a DC function of the form $\eta(t) = g_1(t) - h_1(t)$, where

$$g_1(t) = \lambda|t|; \quad h_1(t) = \lambda|t| - 1 + e^{-\lambda|t|}. \tag{4}$$

Hence the resulting problem of (P_ρ) via this approximation can be written as a DC program (with $\gamma := 1/\rho$):

$$\min_{x \in \mathbb{R}^n} \left\{ F_1(x) = G_1(x) - H_1(x) := \left(\frac{\gamma}{2}\|Ax - y\|^2 + \sum_{i=1}^{n} g_1(x_i) \right) - \sum_{i=1}^{n} h_1(x_i) \right\} \tag{5}$$

$$\text{where} \quad G_1(x) := \frac{\gamma}{2}\|Ax - y\|^2 + \sum_{i=1}^{n} g_1(x_i), \text{ and} \quad H_1(x) := \sum_{i=1}^{n} h_1(x_i) \tag{6}$$

are clearly convex functions. Hence, according to the generic DCA scheme, applying DCA on (5) amounts to computing the two sequences $\{v^l\}$ and $\{x^l\}$ such that $v^l \in \partial H_1(x^l)$ and x^l is a solution to the next convex program

$$\min \left\{ \frac{\gamma}{2}\|Ax - y\|^2 + \sum_{i=1}^{n} g_1(x_i) - \langle v^l, x \rangle : x \in \mathbb{R}^n \right\}. \tag{7}$$

By introducing a new variable $u \in \mathbb{R}^n$, we reformulate (7) in an equivalent form

$$\min_{x,u} \left\{ \frac{\gamma}{2}\|Ax - y\|^2 + \sum_{i=1}^{n} u_i - \langle v^l, x \rangle : \lambda x_i \le u_i, \ -\lambda x_i \le u_i, \forall i = 1, \ldots, n. \right\} \tag{8}$$

which is in fact a linearly constrained convex quadratic program.

3.2 The Second DCA Scheme via a Modified SCAD Approximation

For $\delta > 2$ and $\lambda > 0$, an alternative SCAD approximation of ℓ_0-norm is given in [17] as following: $\|x\|_0 \simeq \sum_{i=1}^{n} \phi(x_i)$, where the function $\phi(t)$ is defined by

$$\phi(t) = \begin{cases} \lambda t \text{ if } 0 \le t \le \lambda, \ -\frac{t^2 - 2\delta\lambda t + \lambda^2}{2(\delta - 1)} \text{ if } 0 \le t \le \delta\lambda, \ \frac{(\delta+1)\lambda^2}{2} \text{ if } t \ge \delta\lambda, \\ \phi(-t) \text{ if } t < 0. \end{cases} \tag{9}$$

It is easy to see that $\phi(t)$ can be expressed as a DC function of the form $\phi(t) = g_1(t) - h_2(t)$, where the function $h_2(t)$ is given by:

$$h_2(t) = \begin{cases} 0 \text{ if } 0 \le t \le \lambda, \ -\frac{(t-\lambda)^2}{2(\delta-1)} \text{ if } 0 \le t \le \delta, \lambda t - \frac{(\delta+1)\lambda^2}{2} \text{ if } t \ge \delta\lambda, \\ h_2(-t) \text{ if } t \le 0 \end{cases} \tag{10}$$

which is clearly convex, and g_1 is already defined in (4).

So the resulting problem of (P_ρ) via this approximation is written as:

$$\min_{x \in \mathbb{R}^n} \left\{ F_2(x) := \frac{\gamma}{2}\|Ax - y\|^2 + \sum_{i=1}^{n} (g_1(x_i) - h_2(x_i)) : x \in \mathbb{R}^n \right\}. \tag{11}$$

Similar to F_1, we can express F_2 as a DC function of the form

$$F_2(x) := G_1(x) - H_2(x), \ H_2(x) := \sum_{i=1}^{n} h_2(x_i) \text{ and } G_1 \text{ is given in (6).} \tag{12}$$

Hence, applying DCA to (11)-(12) amounts to computing the two sequences $\{x^l\}$ and $\{v^l\}$ such that $v^l \in \partial H_2(x^l)$ and $\{x^l\}$ solves the convex program (8).

3.3 Algorithms

We can now describe in detail the two DCA schemes for solving the sparse signal recovery problem (P_ρ).

DCA-1: DCA applied to (5) (piesewise concave approximation of ℓ_0-norm).
Initialization Let τ be a tolerance sufficiently small, set $l = 0$. Choose $x^0 \in \mathbb{R}^n$.
Repeat

- Step 1. Compute $v^l \in \partial H_1(x^l)$ as follows:

$$v_i^l = \alpha\lambda(1 - \epsilon^{-\lambda x_i^l}) \text{ if } x_i^l \ge 0, \quad -\alpha\lambda(1 - \epsilon^{\lambda x_i^l}) \text{ if } x_i^l < 0, \forall i = 1, \ldots, n. \tag{13}$$

- Step 2. Solve the program quadratic (8) to obtain x^{l+1}.
- Set $l = l + 1$.

Until $\|x^{l+1} - x^l\| \le \tau(\|x^l\| + 1)$.

DCA-2: DCA applied to (11) (SCAD approximation of ℓ_0-norm).
Apply **DCA-1** in which Step 1 is replaced by
Step 1. Compute $v^l \in \partial H_2(x^l)$ as follows:

$$v_i^l = \begin{cases} 0 \text{ if } -\lambda \le x_i^l \le \lambda, \ \frac{x_i - \lambda}{(\delta-1)} \text{ if } \lambda \le x_i^l \le \delta\lambda, \ \frac{x_i + \lambda}{(\delta-1)} \text{ if } \delta\lambda \le x_i^l \le -\lambda, \\ \lambda \text{ if } x_i > \delta\lambda, \ -\lambda \text{ if } x_i < -\delta\lambda, \quad \forall i = 1, \ldots, n. \end{cases} \tag{14}$$

Theorem 1. *(Convergence properties of DCA)*

(i) ***DCA-1****(resp.* ***DCA-2****) generates a sequence* $\{x^l\}$ *such that the sequence* $\{F_1(x^l)\}$ *(resp.* $\{F_2(x^l)\}$*) is monotonously decreasing.*
(ii) the two algorithms ***DCA-1*** *and* ***DCA-2*** *have a linear convergence.*
(iii) The sequence $\{x^l\}$ *generated by* ***DCA-1*** *(resp.* ***DCA-2****) converges vers a critical point of* $F_1 := G_1 - H_1$ *(resp.* $F_2 := G_1 - H_2$*).*

Proof: (i) - (iii) are direct consequences of the convergence properties of general DC programs.

4 Experiments and Results

Our algorithms were developed in Visual C++ 2008, and performed on a PC Intel Core(TM)2 Quad CPU $Q9505$, 2.83 GHz and 4GB RAM. CPLEX 12.3 was used for solving the quadratic program (8). Numerical experiments were performed on 5 datasets ($Prb1 - Prb5$) taken from well known toolbox *Sparco* ([20]).

For *DCA-1* and *DCA-2*, the regularized parameter ρ is chosen in the set of values $\{0.0001; 0.0005; 0.001; 0.005; 0.01; 0.05; 0.1; 0.5; 1; 1.5; 2; 2.5; 5\}$ while the parameter of l_0 approximation λ (resp. δ) is taken from the set $\{0.5; 1; 1.5, 2; 2.5; 3; 3.5\}$ (resp. $\{4; 16; 25\}$). We stop *DCA-1*/*DCA-2* with the tolerance $\tau = 10^{-4}$.

We compare our algorithms with 4 other ones: *GPSR* (Gradient Projection for Sparse Reconstruction) ([10]) and 3 algorithms *L1eq* ([3]), *L1qc* ([4]), *L1dantzig* ([6]) of $l_1 - magic$ package. All these 4 algorithms deal with the problem of sparse signal recovery using $l_1 - norm$.

MSE (Mean Square Error) was used to compare the performances of algorithms. MSE is one of many ways to quantify the difference between values implied by an estimator and the true values of the quantity being estimated. The MSE is defined by: $MSE = ||x_0 - x||^2/n$. In Table 1, the MSE and zero-norm of reconstructed signal of each algorithm are reported (we note NA when an algorithm does not furnish a solution).

Table 1. Comparison of algorithms

		DCA-1	DCA-2	GPSR	L1dantzig	L1qc	L1eq
Prb1	$\|x\|_0$	601	617	89	393	517	NA
($\|x_0\|_0 = 4$)	MSE	**0,330**	0,522	2,069	0,699	0,421	NA
Prb2	$\|x\|_0$	71	71	67	735	71	71
($\|x_0\|_0 = 71$)	MSE	**0,002**	**0,002**	0,078	**0,002**	**0,002**	**0,002**
Prb3	$\|x\|_0$	268	252	42	290	NA	NA
($\|x_0\|_0 = 63$)	MSE	0,020	**0,015**	0,022	0,018	NA	NA
Prb4	$\|x\|_0$	521	542	468	445	NA	NA
($\|x_0\|_0 = 191$)	MSE	**0,000**	0,001	0,025	0,013	NA	NA
Prb5	$\|x\|_0$	12	14	208	1024	NA	NA
($\|x_0\|_0 = 12$)	MSE	8,4E-01	**6,7E-01**	7,7E+01	8,0E-01	NA	NA

In Figure 1 - Figure 5, we report the comparative results of all algorithms on each data ($Prb1 - Prb5$). Each figure contains: (a)-The original signal and reconstructed signal by *DCA-1*, *DCA-2*, *GPSR* and *L1Dantzig* (*L1Dantzig* usually gives best results out of 3 algorithms of $l_1 - magic$); (b)-The enlargement of the image (a) at red circle; (c)-Coefficients of original signal and reconstructed signals by *DCA-1*, *DCA-2* and (d)-Coefficients of original signal and reconstructed signals by *GPSR* and *L1Dantzig*.

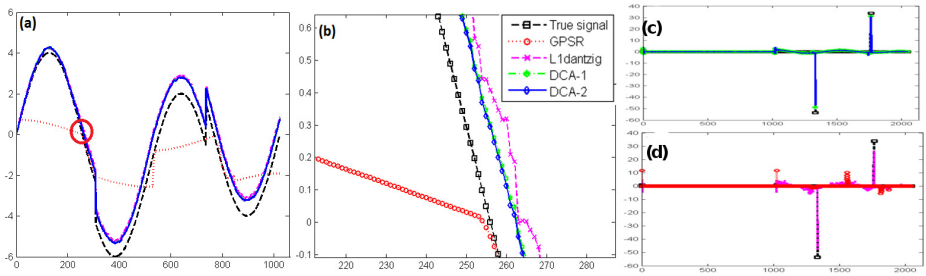


Fig. 1. Comparative result on *Prb1*

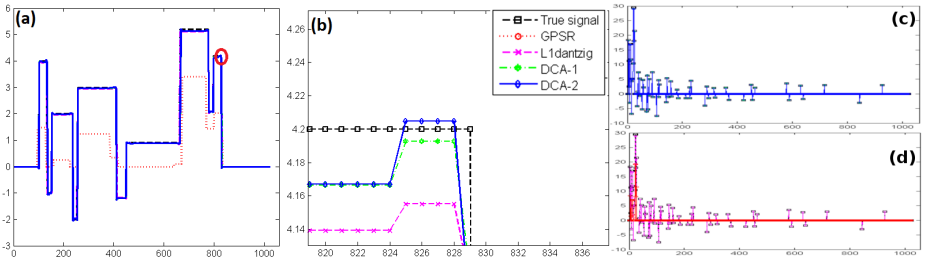


Fig. 2. Comparative result on *Prb2*

From the computational results, we observe that

- In many problems, *L1eq* and *L1qc* can not furnish a solution while our algorithms always give a solution.
- In most of case (4 out of 5), our algorithms give better results of MSE than *GPSR*, *L1eq*, *L1qc*, *L1dantzig*. For the dataset *Prb*2, we obtain the same MSE as *L1eq*, *L1qc*, *L1dantzig*.
- The recovered signal obtained with *DCA-1*/*DCA-2* is the closest one to the original signal and coefficients vector.

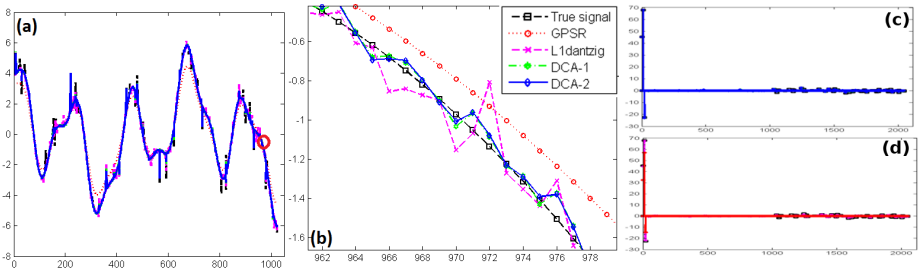


Fig. 3. Comparative result on *Prb3*

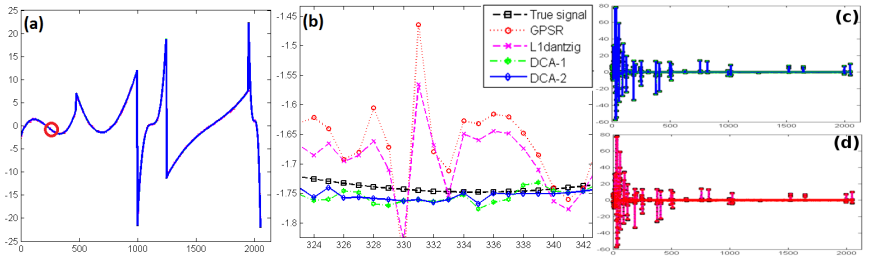


Fig. 4. Comparative result on *Prb4*

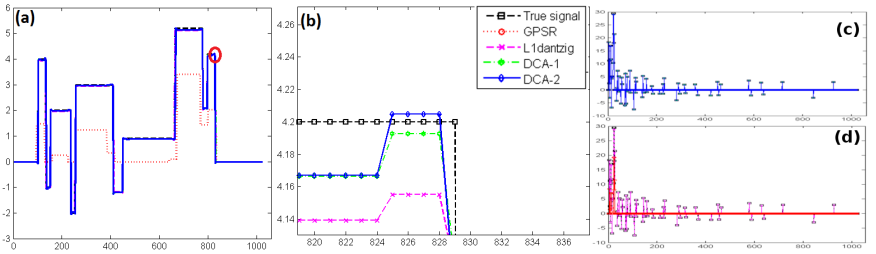


Fig. 5. Comparative result on *Prb5*

Conclusion. In this paper, we have proposed a efficient continuous nonconvex optimization approach based on DC programming and DCA for the problem of signal recovery. Using two appropriate approximation functions of zero-norm, we formulated the problem as a DC program and then developed DCA to solve the resulting programs. Numerical results on several datasets of well known toolbox *Sparco* showed the effectiveness of the DCA based schemes.

References

1. Amaldi, E., Kann, V.: On the approximability of minimizing non zero variables or unsatisfied relations in linear systems. Theoretical Computer Science 209, 237–260 (1998)
2. Bradley, P.S., Mangasarian, O.L.: Feature Selection via concave minimization and support vector machines. In: Proceeding of International Conference on Machina Learning, ICML 1998 (1998)
3. Candès, E., Tao, T.: Near optimal signal recovery from random projections: Universal encoding strategies? IEEE Trans. Inform. Theory 52(12), 5406–5425 (2006)
4. Candès, E., Romberg, J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Comm. Pure Appl. Math. 59(8), 1207–1223 (2006)
5. Candès, E., Romberg, J., Tao, T.: Robust uncertainty principles: Exact signal reconstruction from highly incomplete Fourier information. IEEE Trans. Inform. Theory 52, 489–509 (2006)
6. Candès, E., Tao, T.: The Dantzig selector: Statistical estimation when p is much larger than n. Ann. Stat. 35(6), 2313–2351 (2007)
7. Candès, E.J., Wakin, M.B., Boyd, S.P.: Enhancing Sparsity by Reweighted l_1 Minimization. J. Fourier Anal. Appl. 14, 877–905 (2008)
8. Duarte, M.F., Eldar, Y.C.: Structured Compressed Sensing: From Theory to Applications. IEEE Transactions on Signal Processing 59(9), 4053–4085 (2011)
9. Donoho, D.L.: Compressed sensing. IEEE Trans. Inform. Theory 52(4), 1289–1306 (2006)
10. Figueiredo, M.A.T., Nowak, R.D., Wright, S.J.: Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. IEEE Journal of Selected Topics in Signal Processing 1(4), 586–597 (2007)
11. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. of the American Stat. Association 96(456), 1348–1360 (2001)
12. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning. Springer, Heidelberg (2001)
13. Le Thi, H.A.: Contribution à l'optimisation non convexe et l'optimisation globale: Théorie, Algoritmes et Applications, Habilitation à Diriger des Recherches, Université de Rouen (1997)
14. Le Thi, H.A.: DC programming and DCA, `http://lita.sciences.univ-metz.fr/~lethi/english/DCA.html`
15. Le Thi, H.A., Pham Dinh, T.: The DC (difference of convex functions) Programming and DCA revisited with DC models of real world nonconvex optimization problems. Annals of Operations Research 133, 23–46 (2005)
16. Le Thi, H.A., Le Hoai, M., Van Nguyen, V., Pham Dinh, T.: A DC Programming approach for Feature Selection in Support Vector Machines learning. Journal of Advances in Data Analysis and Classification 2(3), 259–278 (2008)
17. Le Thi, H.A., Van Nguyen, V., Ouchani, S.: Gene Selection for Cancer Classification Using DCA. In: Tang, C., Ling, C.X., Zhou, X., Cercone, N.J., Li, X. (eds.) ADMA 2008. LNCS (LNAI), vol. 5139, pp. 62–72. Springer, Heidelberg (2008)
18. Thiao, M.: Approches de la programmation DC et DCA en Data mining, Thèse de doctorat à l'INSA-Rouen, France (2011)

19. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society 46, 431–439 (1996)
20. SPARCO: A toolbox for testing sparse reconstruction algorithms, http://www.cs.ubc.ca/labs/scl/sparco/
21. Weston, J., Elisseeff, A., Scholkopf, B., Tipping, M.: Use of the Zero-Norm with Linear Models and Kernel Methods. Journal of Machine Learning Research 3, 1439–1461 (2003)
22. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B 7, 301–320 (2005)

DC Programming and DCA Based Cross-Layer Optimization in Multi-hop TDMA Networks

Hoai An Le Thi[1], Quang Thuan Nguyen[2], Khoa Tran Phan[3], and Tao Pham Dinh[4]

[1] Laboratory of Theoretical and Applied Computer Science,
Paul Verlaine University of Metz, France
[2] School of Applied Mathematics and Informatics,
Hanoi University of Science and Technology, Vietnam
[3] Department of Electrical and Computer Engineering,
University of Alberta, Edmonton, AB, Canada
[4] Laboratory of Modelling, Optimization & Operations Research,
INSA-Rouen, France
lethi@univ-metz.fr, thuan.nguyenquang@hust.vn,
khoa@ece.ualberta.ca, pham@insa-rouen.fr

Abstract. Efficient design of wireless networks is a challenging task. Recently, the concept of cross-layer design in wireless networks has been investigated extensively. In this work, we present a cross-layer optimization framework, i.e., joint rate control, routing, link scheduling and power control for multi-hop time division multiple access (TDMA) networks. In particular, we study a centralized controller that coordinates the routing process and transmissions of links such that the network lifetime is maximized. We show that the aforementioned design can be formulated as a mixed integer-linear program (MILP) which has worst case exponential complexity to compute the optimal solution. Therefore, our main contribution is to propose a computationally efficient approach to solve the cross-layer design problem. Our design methodology is based on a so-called *Difference of Convex functions algorithm* (DCA) to provide either optimal or near-optimal solutions with finite convergence. The numerical results are encouraging and demonstrate the effectiveness of the proposed approach. One of the advantages of the proposed design is the capability to handle very large-scale problems which are the usual scenarios encountered in practice.

Keywords: Cross-layer optimization, TDMA, DCA.

1 Introduction

Wireless networks have recently emerged as essential means of communications to provide reliable data communication among many users. In such networks, random deployment and mobility of wireless nodes possibly result in scenarios when the traffic source nodes are far away from their corresponding destination nodes. Therefore, multi-hop transmission is necessary where nodes can forward other nodes' information. Due to interference between links, in this research, we adopt time division multiplexing (TDM) to allocate transmission rights to wireless links. It should be noted that the optimal scheduling problem in TDMA-based networks is NP-complete [18] and can be

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 398–408, 2013.

usually reformulated as some sort of vertex coloring problems in graph theory [4]. Furthermore, in a multi-hop network, power allocation, link scheduling, routing, and rate control interact with each other. Thus, not only the design of medium access control (MAC) schemes is of concern but also a cross-layer design across all layers is important (see, e.g., [12] for an overview). Such a design methodology is shown to outperform the method of designing each layer by itself. Recently, cross-layer optimization with different design objectives and constraints has received much attention from the academia [7], [5], [1], [3].

In this work, we consider a cross-layer design problem to allocate time and power resources to links in an interference-limited wireless network. Since each energy-limited node either generates or relays information that needs to be communicated to a base station, we aim at maximizing the network lifetime. The network lifetime is defined as the earliest time when the first node dies. We show that the proposed design can be formulated as a mixed integer-linear program (MILP) which is computationally intensive, especially for large-scale settings. By employing the *exact penalty method*, we first show that the proposed MILP can be equivalently recast as a *concave minimization* problem. Essentially, we show that even when binary variables are relaxed as continuous ones, optimality is still be preserved. Next, we reformulate the concave minimization problem in the form of a DC (Difference of Convex functions) program that consists of minimizing a DC function on the whole space. We propose a technique which combines *DC Algorithm* (DCA) and the traditional *branch and bound* (BnB) to solve the resulting DC problem. Generally, DCA has linear convergence and achieves near-optimal solution. One of the powerful and distinct advantage of the proposed approach compared to global optimization techniques is its ability to solve very large-scale problems. The preliminary computational results are very encouraging and demonstrate the effectiveness of the proposed method. To the best of our knowledge, the proposed mathematical approach is the first of its kind in the areas of wireless communications and networking.

Cross-layer design in wireless networks over physical and MAC layers has been well-studied. For example, [6] computes the joint link scheduling and power control to reduce the power consumption for ad hoc networks. In this research, we consider the system and interference model as in [13, 15]. [15] presents a joint link scheduling and power control scheme for TDMA-based networks. Moreover, routing is assumed to be fixed and the network throughout, i.e., sum of links' throughput instead of the network lifetime is maximized. A heuristic polynomial time algorithm to solve the proposed MILP is presented. However, its performance is inferior to that of the optimal one. Our proposed formulation can be seen as an extension to the work in [15] where we also incorporate rate control, routing and network lifetime maximization with quality-of-service (QoS) constraints on the end-to-end flows.

Optimal TDMA scheduling to maximize the average transmission rate or to minimize the cross-link interference given fixed link transmission powers is considered in [14]. Unsurprisingly, the resulting formulation is also a MILP but no efficient solution approaches are proposed. However, none of the aforementioned papers consider network lifetime or routing. In [19], the authors investigate the tradeoff between energy consumption and performance in wireless sensor networks. Specifically, the interaction between network lifetime maximization and rate allocation is studied.

Usually, the routing algorithms in wireless networks try to minimize the total energy consumption which may cause some particular nodes run out of energy quickly. Therefore, network lifetime maximization-based routing is a good candidate to prolong the network operation [2,13]. In [2], the proposed algorithm computes the routes and power levels to maximize the network lifetime. Only local information on the network condition is required, and thus, distributed implementation is possible. Similarly, in [13], a cross-layer design across physical, MAC and routing layers is proposed where each link is able to vary its transmission rate and power. Moreover, a distributed joint routing and MAC scheme is presented in [8] to maximize the lifetime of wireless sensor network.

2 System Description and Assumptions

Consider a multi-hop network with node set $\mathcal{N}$. Uplink transmission is assumed where there is one common traffic destination (not included in $\mathcal{N}$) for all the nodes which does not have any traffic to other nodes. One example of the traffic destination is the traffic sink in a wireless sensor network. Each node $n \in \mathcal{N}$ generates traffic at a rate r_n. r_n is a integer number of unit rate, and is required to be no less than a minimal value $r_n^{\min}$. Let $\mathcal{L}$ denote the set of unidirectional links; bidirectional links can be represented by two unidirectional links.

In a multi-hop network, links contend and/or interfere with each other. So it is obvious that it may not be feasible to allow all the links to transmit at the same time. In addition, since each node cannot transmit and receive simultaneously, its outgoing and incoming links cannot be active at the same time. Further, in a unicast network, a transmitter cannot transmit data to more than one receivers. In addition, any two simultaneous transmissions with a common receiver are not allowed due to collision in packet reception. In TDMA-based transmission, time is partitioned into fixed-length frames, and each frame is further divided into T time slots. The resource allocation in a frame is the same as those in other frames. In each frame, a node may need to transmit in one or more slots for its own traffic and/or relay traffic from other nodes. If a node transmits in a slot, while its transmission power can be varied from [0, $P_{\max}$], its transmission rate is fixed at a unit rate. In the TDMA-based network, a channel is specified by two elements (j, l), $j \in \mathcal{J}$, $l \in \mathcal{L}$, where $\mathcal{J} = \{1, 2, ..., J\}$. For the channel, the resource allocation is denoted by (s_j^l, P_j^l), where $s_j^l = 1$ means link l is active at slot j while $s_j^l = 0$ otherwise, and $P_j^l > 0$ denotes the transmission power of link l at slot j if $s_j^l = 1$, $P_j^l = 0$ otherwise. At each node, the difference of its outgoing traffic and its incoming traffic should be the traffic generated by itself, i.e.,

$$\sum_{l \in \mathcal{O}(n)} \sum_{j=1}^{J} s_j^l - \sum_{l \in \mathcal{I}(n)} \sum_{j=1}^{J} s_j^l = r_n, \quad n \in \mathcal{N} \tag{1}$$

where $\mathcal{O}(n)$ and $\mathcal{I}(n)$ are the set of outgoing links and incoming links at node n, respectively. The values of s_n for the non-source nodes are set to zero, or equivalently all the traffic entering such nodes must be routed.

The initial energy supply at node n is denoted as E_n. Obviously, we should have

$$\Big[\sum_{l\in\mathcal{O}(n)}\sum_{j=1}^{J}P_j^l+\sum_{l\in\mathcal{I}(n)}\sum_{j=1}^{J}\varepsilon_l s_j^l\Big]\cdot T_n\le E_n.\quad n\in\mathcal{N} \tag{2}$$

where ε_l denote the energy needed to receive a unit of traffic over link l, T_n is the lifetime (unit: number of frames) of node n. In this energy model, we assume that transmit power is the dominant source of energy consumption at the transmitter.

In this research, the QoS requirement of each node should be guaranteed during the network lifetime. Thus the network lifetime denoted T is to the moment when any node dies. Specifically, $T=\min_{n\in\mathcal{N}}\ T_n$.

Interference relations among the nodes and/or links can be modeled in various ways, for example by using contention-based model [3, 8, 19] or the signal-to-interference-plus-noise-ratio (SINR)-based model [13, 15] in which the latter is adopted in this research. Specifically, if the link $l\in\mathcal{L}$ is active at slot j (i.e., $s_j^l=1$), the following inequality should hold so as to guarantee the transmission quality of the link

$$\mathrm{SINR}_j^l=\frac{P_j^l h_{ll}}{\sum_{k\neq l}P_j^k h_{kl}+\eta_l}\ge\gamma^{\mathrm{th}} \tag{3}$$

where SINR_j^l is the SINR for link l at slot j, h_{kl} is the path gain from the transmitter of link k to the receiver of link l, η_l is the noise power at receiver of link l, and and γ^{th} is the required SINR threshold for accurate information transmission. We assume that all wireless nodes are low-mobility devices and/or the topology of the network is static or changes slowly allowing enough time for computing the new scheduler. An example of such networks is a wireless sensor network for environmental monitoring with fixed sensor locations. In this case, the need for distributed implementation is not necessary.

3 The Cross-Layer Optimization Framework

In this section, we propose a joint power allocation, link scheduling, routing and rate control framework to maximize the lifetime of the network. From the preceding discussions, the lifetime maximization cross-layer design problem can be posed as

$$\max_{P_j^l,\ s_j^l,\ T}\qquad T \tag{4}$$

$$\Big[\sum_{l\in\mathcal{O}(n)}\sum_{j=1}^{J}P_j^l+\sum_{l\in\mathcal{I}(n)}\sum_{j=1}^{J}\varepsilon_l s_j^l\Big]\cdot T\le E_n,\ n\in\mathcal{N} \tag{5}$$

$$\sum_{l\in\mathcal{O}(n)}\sum_{j=1}^{J}s_j^l-\sum_{l\in\mathcal{I}(n)}\sum_{j=1}^{J}s_j^l=r_n,\ n\in\mathcal{N} \tag{6}$$

$$r_n\ge r_n^{\min},\ n\in\mathcal{N} \tag{7}$$

$$\sum_{l\in\mathcal{I}(\hat{n})}\sum_{j=1}^{J} s_j^l = \sum_{n\in\mathcal{N}} r_n \tag{8}$$

$$\sum_{l\in\mathcal{O}(n)} s_j^l + \sum_{l\in\mathcal{I}(n)} s_j^l \leq 1,\ \forall n \in \{\mathcal{N}\cup\hat{n}\},\ j{=}1,\ldots,J \tag{9}$$

$$h_{ll}P_j^l \geq \gamma^{\text{th}}\sum_{k\neq l} P_j^k h_{kl} + \gamma^{\text{th}}\eta_l + D(s_j^l - 1), \quad \forall l \in \mathcal{L},\ j = 1,\ldots,J \tag{10}$$

$$0 \leq P_j^l \leq P_{max}s_j^l,\ \forall l \in \mathcal{L},\ j = 1,\ldots,J \tag{11}$$

$$s_j^l \in \{0,1\}, \quad \forall l \in \mathcal{L},\ j = 1,\ldots,J \tag{12}$$

where $\hat{n}$ denotes the common sink node for all data generated in the network, D is a very large positive constant. The objective function is the network lifetime. Constraints (5) require that the total energy consumed during the network lifetime for each node must be less than the available energy. [1] Constraints (6) ensure that the data generated by source nodes are routed properly. Constraints (7) guarantee that the rate for each node is no less than a minimum rate. The minimum rates are possibly different for nodes and are usually determined by the network QoS. Nodes which do not generate traffic have $r_n = r_n^{\min} = 0$. Constraint (8) is the flow conservation at the traffic destination for all the sources. Constraints (9) state that a node can not receive and transmit simultaneously in one particular time slot. Constraints (10) make sure the SINR requirement is met: if a link l is active in time slot j, then the SINR at receiver of link l must be larger than the given threshold γ^{th} which also depends on the system implementation. Constraint (10) is automatically satisfied if link l is not scheduled in time slot j. Constraint (11) states that if a link l is scheduled for time slot j, i.e., $s_j^l = 1$, then the corresponding power value P_j^l must be less than P_{max}. Otherwise, P_j^l obviously equals to zero. We also impose binary integer constraints on s_j^l. With $q = \frac{1}{T}$, the problem becomes

$$\min_{P_j^l,\, s_j^l,\, q} \quad q \tag{13a}$$

$$\sum_{l\in\mathcal{O}(n)}\sum_{j=1}^{J} P_j^l + \sum_{l\in\mathcal{I}(n)}\sum_{j=1}^{J}\varepsilon_l s_j^l \leq q.E_n,\ n \in \mathcal{N} \tag{13b}$$

$$\text{The constraints in (6)–(12).} \tag{13c}$$

It can be seen that the cross-layer optimization problem (13a)–(13c) belongs to a class of well-known mixed-integer linear programs (MILPs). The combinatorial nature of the optimization (13a)–(13c) is not surprising and it has been shown in some previous works, albeit with different objective functions and formulations [4, 13, 15]. Theoretically, MILPs are NP-hard which is clearly inviable for practical scenarios when the dimension is large. The cross-layer optimization problem (13a)–(13c) has worst case

[1] We have assumed that the destination node is endowed with unlimited-energy.

exponential complexity when BnB methods are used to compute the solution. Moreover, when modeling practical networks and depending on the number of links, nodes and time slots, problem with large sizes may arise. As a result, it is extremely difficult to schedule links optimally. Most research in literature is based on heuristic at the cost of performance degradation, for example, see [4, 6, 15]. Here, we propose a method to solve the problem(13a)–(13c) efficiently. We first apply the theory of exact penalization in DC programming [10] to reformulate the MILP as that of minimizing a DC function over a polyhedral convex set. The resulting problem is then handled by DCA. The mentioned approach has been applied successfully in several large scale problems (see [9, 11, 16, 17] and reference therein). The details are provided in the following section.

4 An Efficient Algorithm for Near-Optimal Link Scheduling

By using an exact penalty result, we can reformulate the aforementioned MILP (13a)–(13c) in the form of a concave minimization program. The exact penalty technique aims at transforming the original MILP into a more tractable equivalent problem in the DC optimization framework. Let S be the feasible set of the problem MILP (13a)–(13c). For notational simplicity, we group all the power variables and link scheduling variables in column vectors $P = [P_1^1, \ldots, P_1^J, P_2^1, \ldots, P_L^J]^\mathcal{T}$, $s = [s_1^1, \ldots, s_1^J, s_2^1, \ldots, s_L^J]^\mathcal{T}$ respectively where $\mathcal{T}$ denotes the transpose operator. We denote a new set $K := \{(P, s, q) \in S : s \in [0,1]^{LJ}\}$, and assume that K is a nonempty, bounded polyhedral convex set in $\mathbb{R}^{LJ} \times \mathbb{R}^{LJ} \times \mathbb{R}$. The cross-layer optimization problem (13a)–(13c) can be expressed in the general form:

$$(P_{\text{opt}}, s_{\text{opt}}, q_{\text{opt}}) = \arg\min\Big\{q : (P, s, q) \in S,\ s \in \{0,1\}^{LJ}\Big\}. \tag{14}$$

Consider the function $p(P, s, q)$ defined by $p(P, s, q) = \sum_{l \in \mathcal{L},\ j \in \mathcal{J}} \min\{s_j^l, 1 - s_j^l\}$. It is clear that p is concave and finite on K, $p(P, s, q) \geq 0$ for all $(P, s, q) \in K$, and $\Big\{(P, s, q) \in S : s \in \{0,1\}^{LJ}\Big\} = \Big\{(P, s, q) \in K : p \leq 0\Big\}$. Hence problem (14) can be rewritten as $(P_{\text{opt}}, s_{\text{opt}}, q_{\text{opt}}) = \arg\min\Big\{q : (P, s, q) \in K, p(P, s, q) \leq 0\Big\}$.
The following theorem is in order.

THEOREM 1: *(Theorem 1, [10]) Let K be a nonempty bounded polyhedral convex set, f be a finite concave function on K and p be a finite nonnegative concave function on K. Then there exists $\tilde{t}_0 \geq 0$ such that for $\tilde{t} > \tilde{t}_0$ the problem (P_t) $\alpha(t) = \min\{f(x) + \tilde{t}p(x) : x \in K\}$ and (P) $\alpha = \min\{f(x) : x \in K, p(x) \leq 0\}$ have the same optimal value and the same solution set.*

Furthermore, if the vertex set of K, denoted by $V(K)$, is contained in $x \in K : p(x) \leq 0$, then $\tilde{t}_0 = 0$. If $p(x) > 0$ for some x in $V(K)$, then $\tilde{t}_0 = \min\Big\{\frac{f(x) - \alpha(0)}{S_0} : x \in K, p(x) \leq 0\Big\}$, where $S_0 = \min\Big\{p(x) : x \in V(K), p(x) > 0\Big\} > 0$.

PROOF: The proof for the general case can be found in [10].

From Theorem 1 we get, for a sufficiently large number $\tilde{t}\,(\tilde{t} > \tilde{t}_0)$, the equivalent concave minimization problem to (4)

$$\min : \Big\{ q + \tilde{t}p(P,s,q) : (P,s,q) \in K \Big\} = \min : \Big\{ g(P,s,q) - h(P,s,q) \Big\} \qquad (15)$$

where $g(P,s,q) = \mathcal{X}_K(P,s,q)$, $h(P,s,q) = -q - \tilde{t}\sum_{l\in\mathcal{L},\ j\in\mathcal{J}} \min\{s_j^l, 1-s_j^l\}$, and $\mathcal{X}_K(P,s,q)$ is 0 if $(P,s,q) \in K$, otherwise $+\infty$ (the indicator function of K).

We have successfully transform an optimization with integer variables into its equivalent form with continuous variables. Now, we investigate a DC programming approach for solving (15). A DC program is that of the form

$$\alpha := \min \Big\{ f(x) := g(x) - h(x) : x \in \mathcal{R}^n \Big\} \qquad (16)$$

with g, h being lower semi-continuous proper convex functions on $\mathcal{R}^n$, and its dual is defined as

$$\alpha := \min \Big\{ h^*(y) - g^*(y) : y \in \mathcal{R}^n \Big\} \qquad (17)$$

where $g^*(y) := \max\{x^T y - g(x) : x \in \mathcal{R}^n\}$ is the conjugate function of g.

Based on local optimality conditions and duality in DC programming, the DCA consists in the construction of two sequences $\{x^k\}$ and $\{y^k\}$, candidates to be optimal solutions of primal and dual programs respectively, in such a way that $\{g(x^k) - h(x^k)\}$ and $\{h^*(y^k) - g^*(y^k)\}$ are decreasing and their limits points satisfy the local optimality conditions. The idea of DCA is simple: each iteration of DCA approximates the concave part $-h$ by its affine majorization (that corresponds to taking $y^k \in \partial h(x^k)$) and minimizes the resulting convex function.

Generic DCA scheme:
Initialization: Let $x^0 \in \mathcal{R}^n$ be a best guess, $0 \leftarrow k$.
Iteration k: Calculate $y^k \in \partial h(x^k)$ and $x^{k+1} \in \arg\min\{g(x) - h(x^k) - \langle x - x^k, y^k\rangle : x \in \mathcal{R}^n\}$ (P_k) then $k+1 \leftarrow k$.
Termination: Convergence of x^k.

Convergence properties of DCA and its theoretical basis can be found in [11, 16, 17], for instant it is important to mention that: DCA is a descent method (the sequences $\{g(x^k) - h(x^k)\}$ is decreasing) *without linesearch*; If the optimal value of problem (16) is finite and the infinite sequence $\{x^k\}$ is bounded then every limit point x^* of $\{x^k\}$ is a critical point of $g - h$; DCA has a *linear convergence* for general DC programs; DCA has a finite convergence for polyhedral DC programs ((16) is called polyhedral DC program if either g or h is polyhedral convex).

We now describe the DCA applied to the DC program (15). By the very first definition of h, a sub-gradient $(u,v,w) \in \partial h(P,s,q)$ can be chosen

$$(u,v,w) \in \partial h(P,s,q) \leftarrow u_j^l = 0; \quad v_j^l = \tilde{t} \ \text{ if } \ s_j^l \geq 0.5,$$
$$\text{otherwise } \ v_j^l = -\tilde{t}; \quad w = -1. \qquad (18)$$

Algorithm 1. (DCA applied to (15)): Let $\epsilon > 0$ and (P^0, s^0, q^0). $k = 0$, $er = 1$.
while $er > \epsilon$ **do**
-Compute $(u^k, v^k, w^k) \in \partial h(P^k, s^k, q^k)$ via (18).
-Solve the linear program: $\min\{-{v^k}^T s+q : (P, s, q) \in K\}$ to obtain $(P^{k+1}, s^{k+1}, q^{k+1})$.
-Set $er = \|(P^{k+1}, s^{k+1}, q^{k+1}) - (P^k, s^k, q^k)\|$, $k = k + 1$.
endwhile

Regarding the complexity of the proposed DCA, besides the computation of the subgradients which is trivial, the algorithm requires one linear program at each iteration and it has a finite convergence. The linear program has polynomial complexity. The convergence of Algorithm 1 can be summarized in the next theorem [16].

THEOREM 2: *(Convergence properties of Algorithm 1)*
i) Algorithm 1 generates a sequence $\{(P^k, s^k, q^k)\}$ *contained in* $V(K)$ *such that the sequence* $\{g(P^k, s^k, q^k) - h(P^k, s^k, q^k)\}$ *is decreasing.*
ii) If at iteration r we have $s^r \in \{0, 1\}^{LJ}$, *then* $s^k \in \{0, 1\}^{LJ}$ *and* $f(P^{k+1}, s^{k+1}, q^{k+1}) \leq f(P^k, s^k, q^k)$ *for all* $k \geq r$.
iii) The sequence $\{(P^k, s^k, q^k)\}$ *converges to* $\{(P^*, s^*, q^*)\} \in V(K)$ *after a finite number of iterations. The point* $\{(P^*, s^*, q^*)\}$ *is a critical point of Problem (15). Moreover such an* (P^*, s^*, q^*) *is almost always a strict local minimum of (15).*
PROOF: i) is a convergence property of general DC programs ([16, 17]) while ii) and iii) can be deduced from Proposition 2 in [9].

Since DCA works on the continuous problem (15), its solution may not be integer, i.e. not feasible to (MILP). For obtaining an integer solution we combine DCA with the branch and bound method in which a lower bound is computed by solving the corresponding relaxed linear problem. At each iteration we restart DCA from the optimal solution of the relaxed problem. We stop the combined algorithm when the solution furnished by DCA is feasible to (MILP).

Algorithm 2
Set $R_0 := [0, 1]^{L \times J}$, $k := 0$.
Solve the linear relaxation problem of (MILP) to obtain an optimal solution (P^0, s^0, q^0) and the optimal value $\beta(R_0)$.
If (P^0, x^0, q^0) is feasible of MILP **then** STOP **else** Solve (15) by DCA from (P^0, x^0, q^0) to obtain $(\overline{P}, \overline{x}, \overline{q})$.
If $(\overline{P}, \overline{x}, \overline{q})$ is feasible of MILP, **then** STOP **else** set $\Re = \{R_o\}$, goto the iteration step.
While (stop = false) **do**
-Set $k := k + 1$ and select a rectangle R_k such that $\beta(R_k) = \min\{\beta(R) : R \in \Re\}$.
-Divide R_k in to two rectangles R_{k_0} and R_{k_1} via the index $j*$ such that $s^k_{j*} = \max\{s^k_j : s^k_j \notin \{0, 1\}$: $R_{k_i} = \{s \in R_k : s_{j^*} = i, i = 0, 1\}$.
-For each $i = 0, 1$ solve the corresponding relaxed linear problem to obtain an optimal solution $(P^{k_i}, s^{k_i}, q^{k_i})$ and the optimal value $\beta(R_{k_i})$.
-Launch DCA from $(P^{k_i}, s^{k_i}, q^{k_i})$ to obtain $(\overline{P^{k_i}}, \overline{s^{k_i}}, \overline{q^{k_i}})$.
-**If** $(\overline{P^{k_i}}, \overline{s^{k_i}}, \overline{q^{k_i}})$ is feasible of MILP, **then** STOP **else** $\Re \leftarrow \Re \cup \{R_{k_i}; i = 0, 1\} \setminus R_k$
endwhile

5 Computational Experiments

In this section, we provide preliminary computational results of our approach. We have coded the **Algorithm 2** in C++ programming language and tested the instances using PC Pentium 4 3GHz, 1GB RAM. CPLEX 9.1 is used to solve the linear programs. We test several network configurations with different number of nodes, links, minimum data rate requirements, initial energy levels and so on. The maximum transmit power is taken to be equal to $P_{max} = 5$. The noise variance $\eta = -20$ dB. The SNR threshold γ^{th} equals to 10 dB. Energy consumption for receiving data ε_l is assumed to be insignificant. The gains for each link are computed using the path loss model as $h_{ij} = \frac{1}{10}[\frac{1}{d}]$ for $i \neq j$, and $h_{ii} = [\frac{1}{d}]$ where d is the Euclidean distance between nodes. The factor of $\frac{1}{10}$ can be viewed as the spreading gain in a CDMA system.

In Table 1 we report the comparative results between Algorithm 2 and the CPLEX code applied to MILP. We use the following notations: N: the number of nodes in the network; L: the number of links; J: the number of time slots; VarC: the number of continuous power variables q, $P_j^l,\ j = 1, \ldots, J,\ l = 1, \ldots, L$; VarB: the number of binary scheduling variables $s_j^l,\ j = 1, \ldots, J,\ l = 1, \ldots, L$; Con: the number of constraints in the optimization problem (4)–(12); Value: the computing objective value ($q = \frac{1}{T}$); CPU: the computing time given in seconds; iter: the number of iterations of Algorithm 2. It is shown in the table that our proposed DCA achieves optimal solutions for six testing instances and near-optimal solutions for the others. In all the cases, the computational time for DCA-BnB is much less than that of CPLEX. The ability to handle very large-scale problems make our proposed method implementable for practical networks.

Table 1. Comparative results between DCA and CPLEX code

N^o	Data						Algorithm 2			CPLEX	
	N	L	T	VarC	VarB	Con	Value	CPU	iter	Value	CPU
01	4	6	12	73	72	199	**0.014142**	0.156	04	**0.014142**	0.01
02	5	8	10	81	80	219	0.016900	0.078	01	0.012071	4865.47
03	6	9	10	91	90	251	**0.016971**	0.468	10	**0.016971**	167.45
04	6	10	10	101	100	271	**0.013553**	0.062	01	**0.013553**	903.38
05	6	10	10	101	100	271	**0.008944**	1.328	25	**0.008944**	0.01
06	7	10	10	101	100	283	0.017657	0.218	04	0.012000	4519.38
07	5	10	10	101	100	259	**0.013417**	0.546	09	**0.013417**	56.00
08	5	8	15	121	120	324	0.020241	0.390	06	0.016730	9621.15
09	6	13	10	131	130	331	**0.012000**	0.609	09	**0.012000**	850.52
10	7	14	10	141	140	363	0.017657	0.140	02	0.012000	834.22

6 Conclusion

In this paper, we have studied the cross-layer design problem in an interference-limited TDMA wireless network. We have shown that the problem can be formulated as a

mixed-integer linear program. It was then reformulated as DC program and solved by DCA. DCA is original because it gives an integer solution while it works in a continuous domain. Preliminary numerical results were encouraging and demonstrated the effectiveness of the proposed method. The superior performance in terms of both computed objective value and running time makes it feasible to implement the centralized TDMA-based wireless networks. Moreover, notice that most problem formulations arising in TDMA-based networks can be formulated as some sort of MILP problems, our proposed approach seems attractive and needs more investigation.

References

1. Bhatia, R., Kodialam, M.: On power efficient communication over multi-hop wireless networks: Joint routing, scheduling and power control. In: Proc. IEEE INFOCOM 2004, Hong Kong, pp. 1457–1466 (2004)
2. Chang, J.-H., Tassiulas, L.: Energy conserving routing in wireless ad-hoc networks. In: Proc. IEEE INFOCOM 2000, Tel-Aviv, Israel, pp. 21–31 (2000)
3. Chen, L., Low, S.H., Chiang, M., Doyle, J.C.: Cross-layer congestion control, routing and scheduling design in ad hoc wireless networks. In: Proc. IEEE INFOCOM 2006, Barcelona, Spain, pp. 1–13 (2006)
4. Commander, C.W., Pardalos, P.M.: A combinatorial algorithm for the TDMA message scheduling problem. Computational Optimization and Apps. 43, 449–463 (2009)
5. Cruz, R.L., Santhanam, A.V.: Optimal routing, link scheduling and power control in multihop wireless networks. In: Proc. IEEE INFOCOM 2003, San Francisco, USA, pp. 702–711 (2003)
6. ElBatt, T., Ephremides, A.: Joint scheduling and power control for wireless ad hoc networks. In: Proc. IEEE INFOCOM 2002, New York, USA, pp. 976–984 (2002)
7. Jiang, H., Zhuang, H., Shen, X.: Cross-layer design for resource allocation in 3G wireless networks and beyond. IEEE Communications Magazine 43(12), 120–126 (2005)
8. Kim, S.-J., Wang, X., Madihian, M.: Distributed joint routing and medium access control for lifetime maximization of wireless sensor networks. IEEE Trans. Wireless Commun. 6(7), 2669–2677 (2007)
9. Le-Thi, H.A., Pham Dinh, T.: A continuous approach for globally solving linearly constrained quadratic zero-one programming problems. Optimization 50, 93–120 (2001)
10. Le-Thi, H.A., Pham Dinh, T., Dung Le, M.: Exact penalty in DC programming. Vietnam Journal of Mathematics, 1216–1231 (2007)
11. Le-Thi, H.A., Pham Dinh, T.: The DC (Difference of Convex functions) Programming and DCA revisited with DC models of real world nonconvex optimization problem. Annals of Operations Research, 23–46 (2005)
12. Lin, X., Shroff, N., Srikant, R.: A tutorial on cross-layer optimization in wireless networks. IEEE Journal on Selected Areas in Communications 24(8), 1452–1463 (2006)
13. Madan, R., Cui, S., Lall, S., Goldsmith, A.: Cross-layer design for lifetime maximization in interference-limited wireless sensor networks. IEEE Trans. Wireless Commun. 5(11), 3142–3152 (2006)
14. Madan, R., Cui, S., Lall, S., Goldsmith, A.: Mixed integer-linear programming for link scheduling in interference-limited networks. In: Proc. of 1st Workshop on Resource Allocation in Wireless Networks, Italy (2005)

15. Tang, J., Xue, G., Chandler, C., Zhang, W.: Link scheduling with power control for throughput enhancement in multihop wireless networks. IEEE Trans. Vehicular Tech. 55(3), 733–742 (2006)
16. Pham Dinh, T., Le-Thi, H.A.: Convex analysis approach to DC programming: Theory, Algorithms and Applications, Acta Mathematica Vietnamica, dedicated to Professor Hoang Tuy on the occasion of his 70th birthday, pp. 289–355 (1997)
17. Pham Dinh, T., Le Thi, H.A.: DC optimization algorithms for solving the trust region subproblem. SIAM Journal Optimization 8, 476–505 (1998)
18. Ramanathan, S., Lloyd, E.L.: Scheduling algorithms for multi-hop radio network. IEEE ACM Trans. Networking 1(2), 166–177 (1993)
19. Zhu, J., Chen, S., Bensaou, B., Hung, K.-L.: Tradeoff between lifetime and rate allocation in wireless sensor networks: A cross-layer approach. In: Proc. IEEE INFOCOM 2007, Alaska, USA, pp. 267–275 (2007)

The Multi-flow Necessary Condition for Membership in the Pedigree Polytope Is Not Sufficient- A Counterexample

Laleh Haerian Ardekani and Tiru S. Arthanari[1]

[1] University of Auckland, Auckland, New Zealand
t.arthanari@auckland.ac.nz

Abstract. The multistage insertion formulation (MI) for the symmetric traveling salesman problem (STSP), gives rise to a combinatorial object called pedigree. Pedigrees are in one-to-one correspondence with Hamiltonian cycles. The convex hull of all the pedigrees of a problem instance is called the pedigree polytope. The MI polytope is as tight as the subtour elimination polytope when projected into its two-subscripted variable space. It is known that the complexity of solving a linear optimization problem over a polytope is polynomial if the membership problem of the polytope can be solved in polynomial time. Hence the study of membership problem of the pedigree polytope is important. A polynomially checkable necessary condition is given by Arthanari in [5]. This paper provides a counter example that shows the necessary condition is not sufficient.

1 Introduction

The traveling salesman problem (TSP) is known to be the most studied combinatorial optimization problem [14]. Many well known solution methods and heuristics such as branch and cut or simulated annealing, were first designed for or were tested on the TSP [1]. The TSP has been modeled by different formulations varying in size and in the strength of the LP relaxations [13]. The 0-1 model for the TSP given by Dantzig, Fulkerson and Johnson (DFJ) has $O(n^2)$ variables and $O(2^{n-1})$ constraints [8]. The polytope given by the LP relaxation of the DFJ model, also known as the subtour elimination polytope (SEP), is shown to give the tightest polytope for the TSP compared to other suggested models when their polytopes are projected into the DFJ variable space [15] and [16]. The multistage insertion formulation (MI) for the symmetric traveling salesman problem (STSP) has $O(n^3)$ variables and $O(n^2)$ constraints [2] and the projection of the polytope given by its LP relaxation is shown to be a subset of the SEP [7].

It is shown in [10] that a subroutine for checking membership in a polytope can be used a polynomial number of times to provide a separation algorithm for the polytope. It is also shown that such a separation algorithm could be used a polynomial number of times to solve an optimization problem over the polytope.

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 409–419, 2013.

This makes studying the complexity of the membership problem interesting. The membership problem for the pedigree polytope is studied in [4]. The pedigree polytope is an alternate polytope to study the symmetric TSP (STSP). However, the pedigree polytope has properties not found with respect to the STSP polytope. A necessary condition for membership in pedigree polytope was given in [5] and it was shown that it can be checked in polynomial time. The question whether a counterexample exists for which the necessary condition is not sufficient, was raised in [5]. In this paper such a counterexample is provided.

The remainder of this paper is structured as follows: Section 2 provides notations and preliminaries for the MI-formulation and the polytope that then are defined in Section 3. A multicommodity flow problem that is used to check a necessary condition or membership in pedigree polytope, is given in Section 4. This necessary condition and some sufficient conditions for membership in the pedigree polytope are discussed in Section 5. A counterexample for which the necessary condition is not sufficient is given in Section 6. Lastly, Section 7 includes some conclusions and future research.

2 Notations and Preliminaries

Let $K_n = (V_n, E_n)$ indicate a complete graph of $n \geq 4$ vertices, where $V_n = \{1, \ldots, n\}$ is the set of vertices labeled in some order, and $E_n = \{e = (i, j) \mid i, j \in V_n, i < j\}$ is the set of edges. The cardinality of E_n is denoted by $p_n = n(n-1)/2$, and τ_n denotes $\sum_{k=4}^{n} p_{k-1}$. Each edge $e = (i, j) \in E_n$ is assigned a label that is equal to $l_{ij} = p_{j-1} + i$. For a subset $F \subseteq E_n$ the characteristic vector of F is represented by $x_F \in \mathbb{R}^{p_n}$. Assuming that edges in E_n are ordered in increasing order of the edge labels, the characteristic vector $x_F(e)$ is equal to one if $e \in F$, and it is equal to zero otherwise.

Definition 1. *Forbidden Arcs Transportation Problem (FAT): The FAT problem can be defined as a variation of a capacitated transportation problem in a bipartite network, with some of the arcs marked as forbidden. Given positive values for the supply (demand) of each origin (destination), the FAT problem seeks to find a feasible flow from the origins to the destinations.*

Definition 2. *Rigid and Dummy Arcs: Given a FAT problem, those arcs which have the same flow in all the feasible solutions to the problem are called* rigid *arcs. The rigid arcs with zero flow are called* dummy *arcs. Finding rigid arcs can be done efficiently [5].*

3 The Multistage Insertion Formulation (MI)

The MI-formulation is based on constructing STSP tours by sequentially inserting nodes into the initial tour of three nodes 1, 2 and 3. Let T_k denote a STSP tour of nodes 1 to k, where $k \geq 3$. Given graph K_n and starting with tour $T_3 = [1, 2, 3, 1]$, nodes from 4 to n are inserted sequentially between the nodes of

this tour until a complete tour of size n is achieved. For all $1 \leq i < j \leq k-1$ and $4 \leq k \leq n$, the decision variables of the MI-formulation [1], x_{ijk}, are set equal to one if node k is inserted between nodes i and j, and zero otherwise. Let c_{ij} be the cost of an edge $(i,j) \in E_n$. The insertion of some node k between nodes i and j, would replace the edge (i,j) with two new edges of (i,k) and (j,k) in the tour that increases the total cost of the tour by $C_{ijk} = c_{ik} + c_{jk} - c_{ij}$. The objective function of the MI-formulation is to minimize the total incremental cost. The MI-formulation [2] is:

$$\min \sum_{k=4}^{n} \sum_{(1 \leq i < j \leq k-1)} C_{ijk} x_{ijk}$$

subject to:

$$\sum_{1 \leq i < j \leq k-1} x_{ijk} = 1, \qquad 4 \leq k \leq n, \tag{1}$$

$$\sum_{k=4}^{n} x_{ijk} \leq 1, \qquad 1 \leq i < j \leq 3, \tag{2}$$

$$-\sum_{r=1}^{i-1} x_{rij} - \sum_{s=i+1}^{j-1} x_{isj} + \sum_{k=j+1}^{n} x_{ijk} \leq 0, \qquad 1 \leq i < j, 4 \leq j \leq n-1, \tag{3}$$

$$x_{ijk} \in \{0,1\}, \qquad 1 \leq i < j \leq k-1, 4 \leq k \leq n. \tag{4}$$

Constraint (1) guarantees that each node from 4 to n is inserted in some edge. Constraint (2) ensures that at most one node is inserted in each of the edges of T_3. Constraint (3) makes sure that a node is inserted into an edge of the subtour only if that edge has been generated through previous insertions and is available. We obtain the MI-relaxation problem by relaxing the integer constraint from MI-formulation and also adding the following redundant constraint to the model.

$$-\sum_{r=1}^{i-1} x_{rin} - \sum_{s=i+1}^{n-1} x_{isn} \leq 0, i = 1, \ldots, n-1, \tag{5}$$

. Let the slack variables corresponding to MI-relaxation problem be denoted by u_{ij}, corresponding to the inequality for edge (i,j). It is shown in [6] that for any MI problem instance of size n, if for some $i,j \in V_n$ the slack variable $u_{ij} = 1$, then the edge (i,j) is present in the tour given by the optimal solution of the problem. The polytope of MI- relaxation problem is denoted by $P_{MI}(n)$. The affine transformation of $P_{MI}(n)$, projecting out x_{ijk} variables, is denoted by $\mathcal{U}(n)$. Arthanari and Usha [6] compared $\mathcal{U}(n)$ with the subtour elimination polytope $SEP(n)$, and proved that $\mathcal{U}(n) \subseteq SEP_n$. Thus MI formulation is as tight as the standard DFJ formulation and has only polynomially many constraints. Different formulations of TSP varying in size and in the strength of the LP relaxations have been compared in the literature [13], [15], and [16].

[1] We also use the equivalent notation $x_k(e)$ for x_{ijk} when $e = (i,j) \in E_{k-1}$.

3.1 Pedigrees and the Pedigree Polytope

Given $K_n = (V_n, A_n)$, a complete graph of size n, let H_n be the set of all Hamiltonian cycles in K_n. Given some $k \geq 3$, let $K_k = (V_k, A_k)$ be a subgraph of K_n that contains only the nodes in $V_k \subset V_n$. We use Hamiltonian cycle $T_k \in H_k$ or k-tour for a given k synonymously.

Definition 3. *Pedigree: Given some $3 \leq k \leq n$, let e_{k+1} be an edge of T_k corresponding to graph K_k. The vector $W = (e_4, ..., e_n)$ is called a pedigree if and only if there exists a Hamiltonian cycle $T_n \in H_n$ where T_n is obtained from T_3 by a sequence of insertions such that: T_k results from the insertion of k into e_k in T_{k-1},for $k \in V_n \setminus V_3$.*

Remark 1. Pedigrees are in one-to-one correspondence with Hamiltonian cycles.

Definition 4. *Characteristic Vector of a Pedigree: Consider a pedigree W, with edge $e_k \in E_{k-1}$ being its $(k-3)^{rd}$ component, for some $4 \leq k \leq n$. We define $X = (x_4, ..., x_n) \in B^{\tau_n}$ as the characteristic vector of W, where $x_k \in B^{p_{k-1}}$ is the indicator of e_k with the edge label l, such that x_k is a vector that has a 1 in l^{th} coordinate and zeros elsewhere.*

Definition 5. *The Pedigree Polytope: Let $P_n = \{X \in B^{\tau_n} | X$ be the characteristic vector of W that corresponds to some $T_n \in H_n\}$. The convex hull of P_n is called the pedigree polytope and is denoted by $conv(P_n)$.*

Definition 6. *Edge Generators: Given $e = (i, j) \in E_n$, $G(e)$ is called the set of generators of e, and it is equal to $\delta(i) \cap E_{j-1}$, if $j \geq 4$, and E_3, otherwise.*

We say a pedigree $W' \in P_{k+1}$ is an extension of a $W \in P_k$, in case there is an edge $e \in E_k$ such that $W' = (W, e)$. Notice that, given $W = (e_4, ..., e_k)$ and $e \in E_k$, (W, e) is an extension of W if and only if there exists some $4 \leq a \leq k$, such that e_a is a generator of e.

3.2 The Membership in $conv(P_n)$

Given some $k \leq n$ and given X a feasible solution to the MI-relaxation, let X/k indicate a solution from MI-relaxation including only nodes from 1 to k.

Definition 7. *The Membership Problem: Given X and k such that $X/k \in conv(P_k)$, the membership problem checks whether $X/k + 1 \in conv(P_{k+1})$.*

It is shown in [5] that for any $X \in P_{MI}$, given $X/k \in conv(P_k)$, a necessary condition for membership of $X/k+1$ in $conv(P_{k+1})$, is the existence of a flow equal to one in some layered network. The network used for checking the necessary condition is defined in [5]. The question of whether the necessary condition is sufficient was raised in that paper as well. The algorithm for checking the necessary condition and also the procedures for constructing the layered network are illustrated through an example in [12].

3.3 The FAT(λ) Problem

It is shown in [4] that given $X/k \in conv(P_k)$, if a certain FAT problem, called FAT(λ), is feasible then it is concluded that $X/k+1 \in conv(P_{k+1})$. Given some $X \in P_{MI}(n)$ and $X/k \in conv(P_k)$, let $\lambda \in R_+^{|P_k|}$ be a weight vector such that X/k can be written as a convex combination of $X^r \in P_k$, that is $\Lambda_k(X) = \{\lambda \in R_+^{|P_k|} | \sum_{r \in I(\lambda), X^r \in P_k} \lambda_r X^r = X/k, \sum_{r \in I(\lambda)} \lambda_r = 1\}$, where $I(\lambda)$ is the index set of positive coordinates of λ.

Definition 8. *The FAT(λ) Problem: Given some $X \in P_{MI}(n)$ where $X/k \in conv(P_k)$, and given a weight vector $\lambda \in \Lambda_k(X)$, the FAT(λ) problem is defined with the following components: 1) A set of origin nodes $\{X^\alpha | \alpha \in I(\lambda)\}$, with the supply of λ_α for all α. 2) A set of sink nodes $\{e_\beta \in E_k | x_{k+1}(e_\beta) > 0\}$, with demands equal to $x_{k+1}(e_\beta)$, for all β. 3) The set of arcs $\{(X^\alpha, e_\beta) \in Origins \times Sinks\ | X^\alpha$ has a generator of $e_\beta\}$.*

The feasibility of the FAT(λ) problem is equivalent to having a convex combination of pedigree extensions in the form of (W, e) that satisfies all the supply, capacity and demand restrictions imposed by $X/k+1$, where $W \in conv(P_k)$. However, if for some λ the problem is not feasible, we cannot conclude that $X/k+1$ is not a convex combination of pedigrees in P_{k+1} [4].

4 The Multicommodity Layered Network

Given $X \in P_{MI}(n)$, we can recursively construct the layered network N_k for checking the necessary condition. The set of nodes in N_k corresponds to the x_{ijk} variables in X with positive values. N_k is made up of $(k-2)$ layers. The set of nodes in the $(k-3)^{rd}$ layer of the network is $V_{[k-3]} = \{[k : i, j] | x_{ijk} > 0\}$. We start with $k = 4$; with the set of nodes as $V_{[1]} \cup V_{[2]}$ and arcs as $A_{[4]} = \{[4 : i, j], [5 : r, s] | (i, j) \in G((r, s))\}$, with capacity of the arc equal to the node capacity of starting node, yielding the network N_4.

We check whether this problem is feasible. If not, as we know from [3] $X/5 \notin conv(P_5)$ and therefore $X \notin conv(P_5)$. We stop. Otherwise we find rigid arcs in $A_{[4]}$ and identify dummy arcs and discard them and update N_4. Now we say N_4 is *well-defined*. Example 1 illustrates N_4 for a given X.

Example 1. Consider $X \in P_{MI}(10)$, with the following x_{ijk} values: $x_{124} = 1$, $x_{135} = 3/4$, $x_{145} = 1/4$, $x_{246} = 2/4$, $x_{356} = x_{456} = x_{237} = x_{157} = x_{467} = x_{567} = x_{238} = 1/4$, $x_{148} = 2/4$, $x_{468} = 1/4$, $x_{269} = 2/4$, $x_{189} = x_{689} = 1/4$, $x_{3510} = 2/4$, $x_{4710} = x_{6710} = 1/4$. In the N_4 network, we have $V_{[1]} = \{[4 : 1, 2]\}$ and $V_{[2]} = \{[5 : 1, 3], [5 : 1, 4]\}$. Since $[4 : 1, 2]$ is a generator for all the nodes in $V_{[2]}$, the set of arcs in N_4 is $A_{[4]} = \{([4 : 1, 2], [5 : 1, 3]), ([4 : 1, 2], [5 : 1, 4])\}$.

Given N_{k-1} is well-defined, we proceed to define N_k recursively. Firstly we notice that the node set of N_k is the union of nodes in N_{k-1} and $V_{[k-2]}$. Similarly the set of arcs in N_k is the union of arcs in N_k and the arcs between layer $(k-3)$

and $(k-2)$, that are newly added at this stage. Now consider the links between layers $(k-3)$ and $(k-2)$. Any link, $L = (e, e')$ with $e \in V_{[k-3]}$ and $e' \in V_{[k-2]}$ can give rise to an arc in the network N_k depending on the solution to a max flow problem defined on a sub network derived from N_{k-1} and the link L. If the maximal flow in the sub network is zero we cannot use the link (e, e').

Definition 9. *Restricted Network $N_{k-1}(L)$: Given $k \in V_{n-1} \setminus V_4$, a link $L = (e_\alpha, e_\beta) \in E_{k-1} \times E_k$ is defined where $e_\alpha = (r, s)$ and $e_\beta = (i, j)$. The network $N_{k-1}(L)$ is called the restricted network of N_k corresponding to L that is induced by the set of nodes of N_{k-1}, deleting the nodes in a set called $D(L)$ that includes the following nodes: $[l : e_\beta]$: for max $(4, j) < l < k$, $[l : e_\alpha]$: for max $(4, s) < l < k$, $[j : e]$: $e \notin G(e_\beta)$, if $e_\beta \in E_k \backslash E_3$; otherwise $[4, e_\beta]$, $[s : e]$: $e \notin G(e_\alpha)$, if $e_\alpha \in E_{k-1} \backslash E_3$; otherwise $[4, e_\alpha]$, all the nodes in $V_{[k-3]} \backslash \{[k : e_\alpha]\}$.*

4.1 The F_k Problem

After adding the new $(k-2)^{nd}$ layer to the network, for each link $L \in A_{[k]}$, the corresponding restricted network $N_{k-1}(L)$ is defined and a max flow problem is solved over the restricted network. The maximal flow through a link L is denoted as $C(L)$. We define a FAT problem called F_k in the bipartite network given by the last two layers. The set of origin nodes is $V_{[k-3]}$, with supply values equal to x_{ijk} for each node, and the set of demand nodes is $V_{[k-2]}$, with demands equal to x_{ijk} values. The arcs are $\{L \in A_{[k]} | C(L) > 0\}$, with capacities equal to $C(L)$.

If the F_k problem is feasible, the capacities of the arcs in N_k are updated. The capacities of rigid arcs are set equal to their flow in N_k, and dummy arcs are discarded from N_k. The necessity of removing dummy arcs from the network is shown in [12].

Example 2. The set of origins and destinations in F_5 are $V_{[2]}$ and $V_{[3]}$ respectively. Since $[5 : 1, 3]$ is not a generator of $[6 : 4, 5]$ and $[5 : 1, 4]$ is not a generator of $[6 : 3, 5]$, the corresponding arcs are not links and are not included in the set of the arcs of N_5. The set of arcs for F_5 is $\{([5 : 1, 3], [6 : 2, 4]), ([5 : 1, 3], [6 : 3, 5]), ([5 : 1, 4], [6 : 2, 4]), ([5 : 1, 4], [6 : 4, 5])\}$. The solution to the F_5 problem is shown in Figure 1. All the arcs are rigid. The arc $([5 : 1, 4], [6 : 2, 4])$ has zero flow and is a dummy arc and therefore is discarded from the network.

4.2 The Multicommodity Flow Problem

Given $X \in P_{MI}$ and $X/k \in conv(P_k)$ for $k > 4$, we proceed to check a necessary condition for $X/k+1 \in conv(P_{k+1})$. An enlarged network N is constructed based on N_k to solve a multicommodity flow problem. We construct this network by adding a sink node to N_k corresponding to every arc between the $(k-3)^{rd}$ and the $(k-2)^{nd}$ layers. Sink nodes are connected to the nodes of the $(k-2)^{nd}$ layer with the arcs of capacities equal to one unit. Each sink is assigned a unique commodity. We add source nodes, with supply of one for each commodity, to N_k and connect them with arcs of capacity one to first layer.

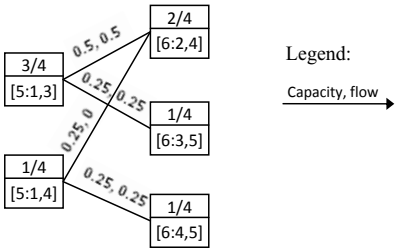


Fig. 1. The Solution to the F_5 Problem

Let S be the set of commodities corresponding to all the links between the last two layers of N_k. The set of all the arcs that are connecting the source node to the first layer, and the sink nodes to the last layer is denoted by A_{new}. Let $V(N_k)$ be the set of all the nodes in N_k and let $A(N) = A(N_k) \cup A_{new}$. Let $f_a^s \geq 0$ denote the flow of commodity $s \in S$, in an arc $a \in A(N)$. The total capacity of the arc regardless of the commodities is denoted by c_a. An upper bound for on f_a^s variables for arc $a \in A(N)$ and commodity s is set equal to one, in case $a \in A_{new}$, or c_a, in case $a \in N_{k-1}(L_s)$, or zero otherwise.

The flow capacity of the nodes is denoted as $x(v)$ which corresponds to x_{ijk} values in X. The flow demand of sink $s \in S$ is denoted as b_s and the flow of commodity s into its corresponding sink is denoted as v_s. Let S_F denote the set of arcs with rigid flow in F_k. The multicommodity flow problem in the layered network N is defined as follows [5].

$$\max z = \sum_{s \in S} v^s$$

subject to:

$$0 \leq f_a^s \leq u_a^s, \ s \in S, a \in A(N), \tag{6}$$

$$\sum_{u \ni a=(u,v)} f_a^s = \sum_{w \ni a=(v,w)} f_a^s, v \in V(N_k), s \in S, \tag{7}$$

$$\sum_{s \in S} f_a^s \leq c_a, \qquad a \in A(N), \tag{8}$$

$$\sum_{s \in S} \sum_{u \ni a=(u,v)} f_a^s \leq x(v), \qquad v \in V(N_k), \tag{9}$$

$$v^s = \sum_{u \ni a=(u,v) \in A_{new}} f_a^s, \qquad s \in S, \tag{10}$$

$$v^s = b_s, \qquad s \ni S_F. \tag{11}$$

Constraint (6) and (8) ensure that arc flows in do not exceed arc capacities for each commodity, and for the total arc capacities. Constraint (7) is a flow conservation constraint. Constraint (9) guarantees that the total flow into each

node does not exceed its capacity. Constraints (10) and (11) guarantee that the flow through the rigid arcs in the last layer is equal to their corresponding sink demands. This problem can be efficiently solved using available algorithms for the multicommodity flow problem such as the arc-chain algorithms in [9].

5 A Necessary Condition for Membership

Theorem 6.6. in [5] gives a necessary condition for membership as having a multicommodity flow equal to one in the enlarged layered network N. Given $X/k \in conv(P_k)$, if the solution to the F_k problem in N_k is not feasible, it is concluded that $X/k+1 \notin conv(P_{k+1})$. Otherwise, the algorithm for checking the necessary condition proceeds with updating arc capacities in N_k. After updating the capacities of the arcs, the multicommodity flow problem in N is solved. If the total flow is not equal to one, it is concluded that $X/k+1 \notin conv(P_{k+1})$, otherwise the necessary condition for membership is satisfied.

An algorithm for checking the necessary condition is given in [5]; and it is illustrated with an example in [12]. This algorithm includes subroutines that can all run within polynomial times. One of the subroutines of the algorithm searches for rigid arcs, given a solution to a F_k problem. An algorithm given in [11] can be used for this purpose. Another subroutine is defining the restricted networks corresponding to the links and solving a maximum flow problem in the restricted networks. It is shown in [5] that defining the networks can be done in polynomial time.

The feasibility of F_4 in N_4 is sufficient to conclude the membership of $X/5$ in $conv(P_5)$ for a given X [5]. For $k = 5$, it can be verified that the necessary condition is also sufficient.

6 A Counterexample

Consider the X given in Example 1 with $n = 10$. The sufficient condition for $X/9 \in conv(P_9)$ is satisfied, as all of the commodity flow paths in N_8 given in Figure 2 follow pedigree paths, namely: $W_1 = ((1,2),(1,3),(2,4),(1,5),(2,3),$ (2,6)), $W_2 = ((1,2),(1,3),(2,4),(4,6),(1,4),(2,6))$, $W_3 = ((1,2),(1,3),(3,5),$ $(5,6)$, $(1,4)$, $(1,8))$, $W_4 = ((1,2),(1,4),(4,5),(2,3),(4,6),(6,8))$.

To check the necessary condition for $X/10 \in conv(P_{10})$, the maximum flow problems in the restricted networks of N_9 given by the links between layers 6 and 7 are solved. The F_9 problem is also solved and the optimal solution is equal to one. None of the arcs in the network for the F_9 problem are found to be rigid. The multicommodity flow problem in N_9 is then solved and the optimal flow is found to be equal to one, as shown in Figure 3

To check the sufficient condition for $X/10 \in conv(P_{10})$, some λ vector which expresses $X/9$ as a convex combination of $X^r \in P_9$, and also makes the FAT(λ) problem feasible, needs to be found. To find such a vector, all possible pedigrees in the N_8 network are found first, using a breadth first search. We designed an LP model to find such a vector.

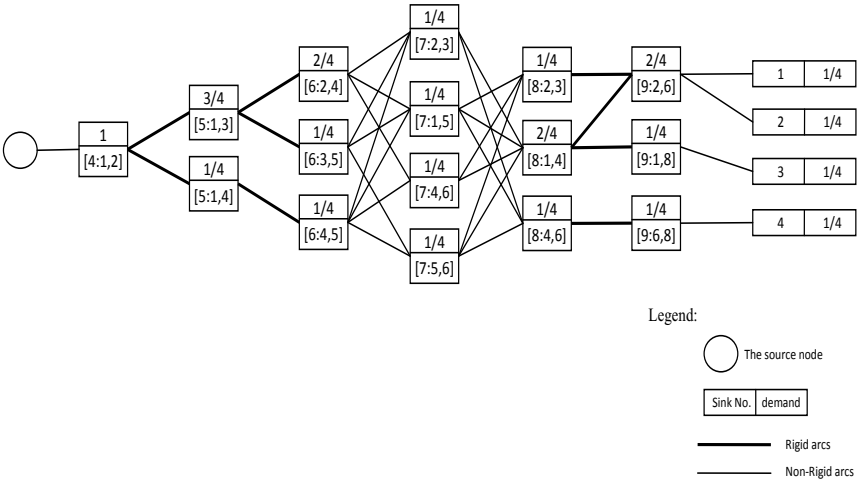


Fig. 2. The N_8 Network

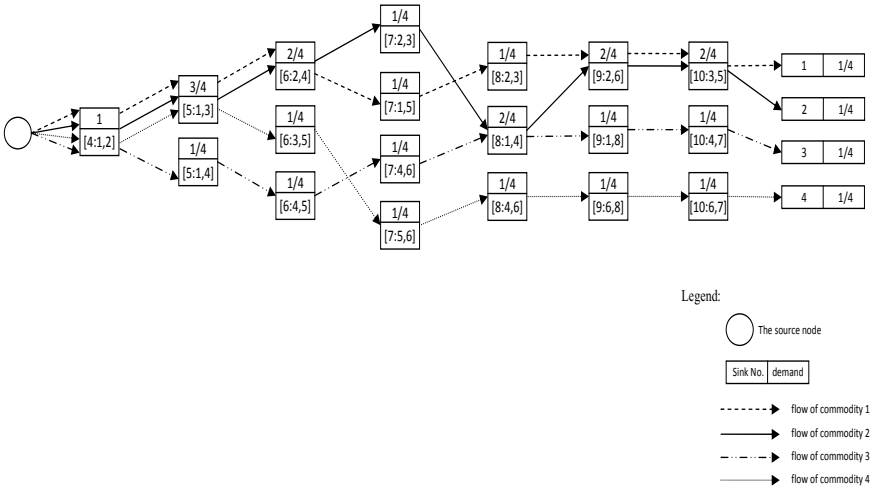


Fig. 3. The Multicommodity Flow in N_9

Given some MI-relaxation solution X/n where $X/n - 1 \in conv(P_{n-1})$, let m be the number of all the pedigrees in N_{n-2}. Let $S = \{1, .., m\}$ and let P be a 0-1 matrix with rows corresponding to pedigrees in N_{n-2}, and columns corresponding to x_{ijk} variables. Let a column vector $\lambda \in [0, 1]^m$ be a weight vector for expressing $X/n - 1$ as a convex combination of all the pedigrees in P. Let T be the set of sink nodes that correspond to the $x_{ijn} > 0$ variables in X/n, where the nodes are ordered according to their edge labels. Let $f_{st} \geq 0$ denote the flow from $s \in S$ to $t \in T$. An LP model for finding a λ vectors for which the FAT(λ) problem is feasible is given as Problem 1.

Problem 1:

$$\max \textstyle\sum_{(s,t)\in S\times T} f_{st}$$

subject to:

$$\sum_{s\in S} \lambda(s) = 1, \tag{12}$$

$$P^T\lambda = X/n - 1, \tag{13}$$

$$\sum_{s\ni G(e_t)\in P(s)} f_{st} - x_n(e_t) = 0, \qquad \forall t \in T, \tag{14}$$

$$\sum_{t\ni G(e_t)\in P(s)} f_{st} - \lambda(s) = 0, \qquad \forall s \in S, \tag{15}$$

$$\lambda(s) \geq 0, f_{st} \geq 0, \forall s \in S, \forall t \in T. \tag{16}$$

Constraint (12) ensures that the sum of $\lambda(s)$ coordinates is equal to one. Constraint (13) makes sure that the pedigree combination given by λ satisfies the availability of x_{ijk} variables in $X/n - 1$. Constraints (14) and (15) ensure that supply and demand conditions are met. When this problem is feasible, we have a λ and corresponding feasible solution to FAT(λ) problem and vice versa.

In general, though Problem 1 could be solved for checking the sufficient condition for membership in pedigree polytope, notice that there may be exponentially many pedigrees to be considered. And so Problem 1 may not be solvable in time polynomial in number of cities. However, because of the small number of pedigrees in N_8, we could use Problem 1 for checking the sufficient condition. We solved Problem 1 for N_8 and $X/10$ in Example 1 using Cplex, and Cplex reported the infeasibility of the problem. Since for no λ the FAT(λ) problem is feasible, it is concluded that $X/10 \notin P_{10}$. Thus we have a counterexample.

7 Conclusion

MI-formulation of the STSP has given rise to the combinatorial objects called pedigrees. Pedigrees are in $1-1$ correspondence with the tours of the symmetric traveling salesman problem. MI-relaxation is tight in the sense its projection is a subset of the subtour elimination polytope. The interesting properties of

the pedigree polytope are studied by Arthanari in a series of papers. In [5] we have a necessary condition for checking whether a X feasible for the MI-relaxation problem is in the pedigree polytope. In this paper we have given a counterexample that shows that the necessary condition given in [5] is not sufficient. Future research is on (1) to define cuts to be used with the MI-formulation to solve the STSP, and (2) to identify facets of the pedigree polytope.

References

1. Applegate, D., Bixby, R.E., Chvatal, V., Cook, W.J.: The traveling salesman problem: a computational study. Princton University Press (2006)
2. Arthanari, T.S.: On the traveling salesman problem. In: Mathematical Programming - The State of the Art. Springer (1983)
3. Arthanari, T.S.: Pedigree polytope is a combinatorial polytope. In: Mohan, S.R., Neogy, S.K. (eds.) Operations Research with Economic and Industrial Applications: Emerging Trends. Anamaya Publishers (2005)
4. Arthanari, T.S.: On pedigree polytopes and hamiltonian cycles. Descrete Mathematics 306(14), 1474–1492 (2006)
5. Arthanari, T.S.: On the membership problem of pedigree polytope. In: Neogy, S.K., Bapat, R.B., Das, A.K., Parthasarathy, T. (eds.) Mathematical Programming and Game Theory for Decision Making. World Scientific (2008)
6. Arthanari, T.S., Usha, M.: An alternate formulation of the symmetric traveling salesman problem and its properties. Discrete Applied Mathematics 98(3), 173–190 (2000)
7. Arthanari, T.S., Usha, M.: On the equivalence of the multistage-insertion and cycle shrink formulations of the symmetric traveling salesman problem. Operations Research Letters 29(3), 129–139 (2001)
8. Dantzig, G.B., Fulkerson, D.R., Johnson, S.M.: Solution of a large-scale traveling-salesman problem. Operations Research 2(4), 393–410 (1954)
9. Ford Jr., L.R., Fulkerson, D.R.: A suggested computation for maximal multi-commodity network flows. Management Science 50(12), 1778–1780 (2004)
10. Grötschel, M., Lovász, L., Schrijver, A.: Geometric Algorithms and Combinatorial Optimization. Springer (1988)
11. Gusfield, D.: A graph theoretic approach to statistical data security. SIAM J. Comput. 17(3), 552–571 (1988)
12. Haerian Ardekani, L., Arthanari, T.S.: Traveling Salesman Problem and Membership in Pedigree Polytope - a Numerical Illustration. In: Le Thi, H.A., Bouvry, P., Dinh, T.P. (eds.) MCO 2008. CCIS, vol. 14, pp. 145–154. Springer, Heidelberg (2008)
13. Langevin, A., Soumis, F., Desrosiers, J.: Classification of travelling salesman formulations. OR Letters 9(2), 127–132 (1990)
14. Lawler, E.L., Lenstra, J.K., Rinooy Kan, A.H.G., Shmoys, D.B.: The traveling salesman problem: A guided tour of combinatorial optimization. Wiley (1985)
15. Orman, A.J., Williams, H.P.: A survey of different integer programming formulations of the travelling salesman problem. In: Optimization, Econometrics and Financial Analysis, pp. 91–104. Springer (2007)
16. Padberg, M., Sung, T.: An analytical comparison of different formulations of the traveling salesman problem. Mathematical Programming 52(1-3), 315–357 (1991)

A Linear Integer Program to Reduce Air Traffic Delay in Enroute Airspace

Ihsen Farah[1], Adnan Yassine[1,2], and Thierry Galinho[2,3]

[1] Laboratory of Applied Mathematics of Le Havre
25 rue Philippe Lebon - B.P. 540, 76058 Le Havre - France
[2] Superior Institute for Logistics Studies
Quai Frissard - B.P. 1137, 76063 Le Havre - France
[3] Computer Science, Information Processing, and Systems Laboratory
Avenue de l'université - B.P. 8, 76801 Saint Étienne du Rouvray - France
{ihsen.farah,adnan.yassine,thierry.galinho}@univ-lehavre.fr
http://lmah.univ-lehavre.fr

Abstract. Due to fast growing of sector of air transportation, air traffic management becomes more and more complex. Therefore, the ability of systems to manage air traffic presents difficulties.

In this paper, we address the Air Traffic Flow Management (ATFM) problem, as a new Linear Integer Program (LIP). It takes into account all flights phases, i.e., taking-off, cruising and landing. The model also allows rerouting decisions. All constraints of our model and objective function are linear, differently from the model of Bertsimas et al. [2] which contains non-linear constraints.

Finally, numerical simulations are presented at the end of this article showing the effectiveness of our new formulation.

Keywords: Air Traffic Flow Management (ATFM), Air Traffic Control (ATC), Linear Integer Programming (LIP).

1 Introduction

A definition of air traffic control is :"the original intention of the air traffic control is to ensure the safety of the traffic and thus to avoid the boarding between operative aircraft in the system, then to optimize the traffic flow". The part of air traffic control aims to optimize the traffic flow is the air traffic flow management (ATFM). ATFM has an important role in order to avoid airports and sectors congestion. Air traffic is growing rapidly and is making significant progress for several years. For this reason optimize the air traffic becomes more complex to solve. Till now, most work about ATFM problem deals mainly congestion at airports. For this reason the most popular flow management approach was the assignment of "ground-holding" delays to departing flights. It is to optimize the delay at airports, by adjusting the departure time of flights. In this subject we can refer to the work of Odoni [12], who was formulating the problem in mathematical terms. From the work of Odoni, several models and algorithms have

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 420–428, 2013.

been proposed for the optimization of delay on the ground. Bertsimas and Odoni [6] , Ball et al. [1] and Hoffman et al. [9] have provided detailed relevant surveys. However, it has become clear that the problem of overloaded airspace also comes from the saturation of the enroute part of the airspace. Although the capability of enroute sectors has generally increased in recent years as a result of actions taken, the problem of capacity constraint sectors is persistent and may take at least another decade to be solved [7]. Unlike the case in which only airport congestion is considered, research literature relating simultaneously to all components of the system is rare. Helme [8] considered the ability of enroute airspace over capacity at airports to obtain a global vision of the situation. While the formulation of this model is simple and easy to understand, its computational performance is low. Lindsay et al. [10] developed a deterministic disaggregate 0-1 integer programming model for the presence of capacity constraints at airports and sectors. Their model determines the optimal temporal and spatial location of each aircraft, given a set of capacity constraints imposed by the national airspace system (NAS). Bertsimas and Patterson [5] presented a deterministic model 0-1 IP to solve a similar problem. For each flight, a predetermined set of sectors is specified. The model determines the optimal time of departure and the time of occupation of a sector for each aircraft. More recently, Lulli and Odoni [11] presented a more macroscopic model for ATFM. However, none of the models mentioned above consider rerouting or speed control. They all consider that the flight path is mentioned in advance. The first work that considers rerouting is Bertsimas and Paterson [4], which describes a dynamic, multicommodity, integer network flow model. Aggregate flows are generated using a lagrangian relaxation approach. Nevertheless, the computing performance of this model was not sufficient to solve very large scale (real problem) instances. In 2008, Bertsimas et al. [3] presented a mathematical model that overcomes this limitation. The proposed model combines the two models proposed in their previous works [4,5] to present a mathematical model to optimize for each flight, the departure time, the flight plan, the time needed to cover each sector, and the arrival time, taking into account the capacity of all components of the air traffic management system. Except, the model contains some problem on the third constraint to compute the number of flights presents in sector. In 2011, Bertsimas et al. [2] introduced a max function on a constraint to computing the number of flights in sectors correctly. However, this latest model is a non-linear model, containing non-linear constraints. In this paper, we propose a modification on the decision variable by introducing a fourth index to determine, for each flight, what is the sector for each flight comming. This change has led to the linearization of non-linear constraints to obtain finally a linear integer program.

2 The Mathematical Model

The mathematical model presented here is based upon the model proposed by Bertsimas et al. [2]. The model is provided to determine the departure time of flights and how to reroute a flight when one or more sectors in its prefered path

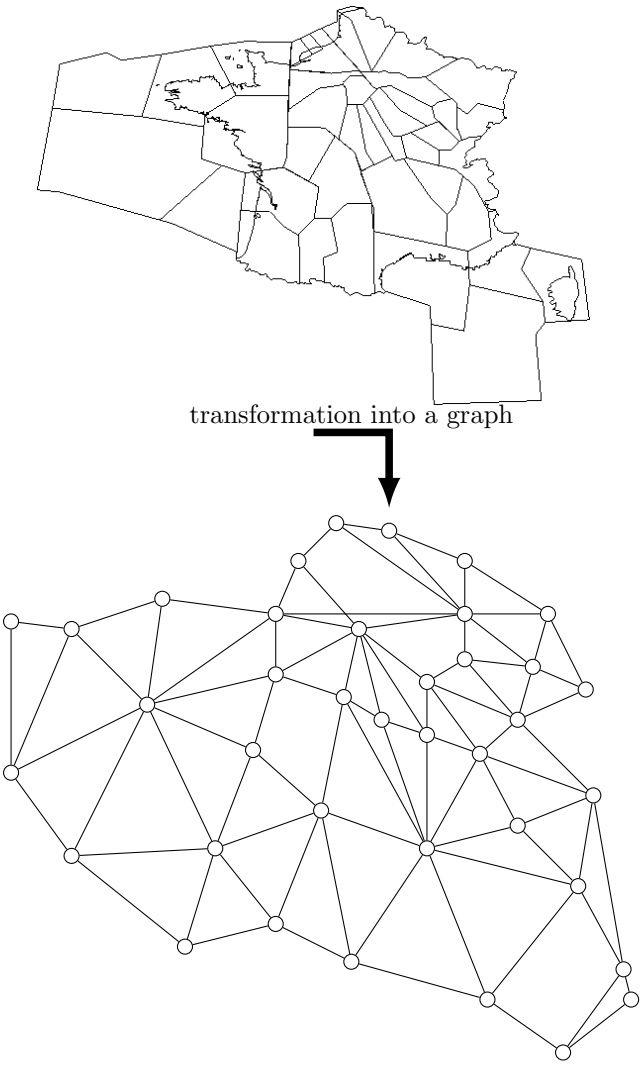


Fig. 1. Transformation into a graph of space

are saturated. The airspace is represented by a graph wher the node represents sectors and edges represents the neighborhood. Figure 1 shows an example of such a representation of airspace by a graph.

Each origin-destination route is represented by a digraph. The set of nodes of the digraph represents the set of capacitated elements of the airspace, i.e. sectors and airports. To consider rerouting in the mathematical model, we must

enlarge the set of possible sectors that might be flown by a flight f. For more details about the representation of origin-destination route by a digraph see Bertsimas et al. [2].

2.1 The Mathematical Formulation

We consider a set of planes with an origin and a destination. The objective of the problem considered here is to determine the flight plan of each aircraft by minimizing flight time. In addition, each airport has a capacity for takeoff and landing, and each sector has also an occupation capacity (the capacity of sector s is represented by the maximal number of flights that can occupy sector s per period). We consider the following notations for the model formulation:

- F : set of flights,
- K : set of airports,
- S : set of sectors,
- S_f : set of sectors that can be flown by flight f,
- T : set of time periods,
- L_j^f : set of sectors that follow sector j,
- P_j^f : set of sectors that precede sector j,
- $D_k(t)$: departure capacity of airport k at time t,
- $A_k(t)$: arrival capacity of airport k at time t,
- $S_j(t)$: capacity of sector j at time t,
- o_f : departure airport of flight f,
- d_f : arrival airport of flight f,
- l_{fj} : number of time units that flight f must spend in sector j,
- $T_j^f = [\underline{T}_j^f, \overline{T}_j^f]$: set of time periods for flight f to arrive in sector j,
- $\underline{T}_j^f$: first time period in the set T_j^f,
- $\overline{T}_j^f$: last time period in the set T_j^f.

As mentioned above, the model presented in this paper is based on the Bertsimas et al. model [2]. The decision variable of their model is as follows :

$$w_{j,t}^f = \begin{cases} 1, \text{if flight } f \text{ arrives at sector } j \text{ by time } t \\ 0, \text{otherwise.} \end{cases}$$

We introduce some modifications in the decision variable as follows : a fourth index is added. It indicates, for each flight, the previous place(Figure 2).

$$w_{i,j,t}^f = \begin{cases} 1, \text{if flight } f \text{ arrives at sector } j \text{ by time } t \text{ that come from } i \\ 0, \text{otherwise.} \end{cases}$$

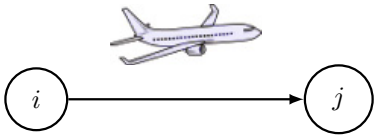


Fig. 2. Illustration of the decision variable, in this case the variable $w^f_{i,j,t}$ takes 1

$$\min \sum_{f\in F} \Big(\sum_{j\in P^f_{d_f}} \sum_{t\in T^f_d} C_{td}(w^f_{j,d_f,t} - w^f_{j,d_f,t-1}) - \sum_{t\in T^f_o} C_{gd}(w^f_{o_f,o_f,t} - w^f_{o_f,o_f,t-1})\Big)$$

$$\sum_{f\in F:o_f=k}(w^f_{k,k,t} - w^f_{k,k,t-1}) \le D_k(t) \qquad \forall\ k\in K, t\in T \quad (1)$$

$$\sum_{f\in F:d_f=k} \sum_{j\in P^f_{d_f}} (w^f_{j,k,t} - w^f_{j,k,t-1}) \le A_k(t) \qquad \forall\ k\in K, t\in T \quad (2)$$

$$\sum_{f\in F:j\in S_f}(\sum_{i\in P^f_j} w^f_{i,j,t} - \sum_{j'\in L^f_j} w^f_{j,j',t}) \le S_j(t) \quad \forall\ j\in S, t\in T \quad (3)$$

$$w^f_{o_f,o_f,\overline{T}^f_{o_f}} = 1 \qquad \forall\ f\in F \quad (4)$$

$$\sum_{i\in P^f_j} w^f_{i,j,t} \le \sum_{i\in P^f_j} \sum_{i'\in P^f_i} w^f_{i',i,t-l_{fi}} \qquad \forall\ f\in F, t\in T^f_j, j\in S_f, i\in P^f_j : j\ne o_f \quad (5)$$

$$\sum_{i\in P^f_j} w^f_{i,j,\overline{T}^f_j} = \sum_{j'\in L^f_j} w^f_{j,j',\overline{T}^f_{j'}} \qquad \forall\ f\in F, j\in S_f : j\ne d_f \quad (6)$$

$$\sum_{j'\in L^f_j} \sum_{i\in P^f_{j'}} w^f_{i,j',\overline{T}^f_{j'}} \le 1 \qquad \forall\ f\in F, j\in S_f : j\ne d_f \quad (7)$$

$$\sum_{j\in P^f_{d_f}} w^f_{j,d_f,\overline{T}^f_{d_f}} = 1 \qquad \forall\ f\in F \quad (8)$$

$$\sum_{i\in P^f_j}(w^f_{i,j,t-1} - w^f_{i,j,t}) \le 0 \qquad \forall\ f\in F, j\in S_f, t\in T^f_j \quad (9)$$

The Objective Function. The objective of the problem is to minimize the total airborne-holding delay. As the objective function of the model propose by Bertsimas et al[2], we calculate the difference between the total delays:

$$\sum_{j\in P^f_{d_f}} \sum_{t\in T^f_d} C_{td}(w^f_{j,d_f,t} - w^f_{j,d_f,t-1})$$

and ground holding delay:

$$\sum_{t\in T^f_o} C_{gd}(w^f_{o_f,o_f,t} - w^f_{o_f,o_f,t-1})$$

The Constraints. The first three constraints represent the capacity of the system components. Constraints (1) and (2) ensure that the number of flights taking off and landing in an airport k at time t does not exceed the departure and arrival capacity of airport k. Constraint (3) ensures that the number of flights

in a sector j at time t does not exceed the capacity of sector j. The first term of this constraint ($\sum_{i \in P_j^f} w_{i,j,t}^f$) takes 1 if flight f has arrived at sector j at time t. If this flight passed to one of the next sectors, the second term ($\sum_{j' \in L_j^f} w_{j,j',t}^f$) takes 1, in this case flight f will not count among the flights that are still in sector j. Otherwise ($\sum_{j' \in L_j^f} w_{j,j',t}^f$ equals 0), the difference takes 1, i.e. the flight will be one of the flights which are still in sector j. The fives following constraints (4, 5, 6, 7, and 8) model the flight plan. Constraint (4) requires that each flight f takes off from its airport of origin during the departure time window. Constraint (5) ensures that every flight must spends a minimum time l_{fs} in sector i before switching to one of its next sectors. If the first term of constraint ($\sum_{i \in P_j^f} w_{i,j,t}^f$) takes 1, the second term ($\sum_{i \in P_j^f} \sum_{i' \in P_i^f} w_{i',i,t-l_{fi}}^f$) must takes 1, i.e. flight f at time $t - l_{fi}$ was in sector i. In other words, before arriving in sector j, flight f must spends l_{fi} time units in sector i. The equality between the two terms in constraint (6) ensures that the number of flights arriving in sector j must be equal to the number of flights leaving sector j (flow conservation). Constraint (7) ensures that every flight must arrive in one of its next sectors. Constraint (8) requires that each flight f lands in its arrival airport during the arrival time window. Finally, the last constraint (9) ensures connectivity in time. In other words, if a flight f arrives at time t in sector j then for all $\overline{t} >= t$, The variable w takes 1.

The changes taken by the decision variable, compared to the formulation of Bertsimas et al [2], have changed considerably the mathematical formulation. The two most important changes are on constraints (3) and (6).

Constraint (3). To ensure that the sum of all flights which may feasibly be in sector j at time t will not exceed the capacity of sector j at time t, Bertsimas et al. introduce the following non-linear constraint:

$$\sum_{f \in F: j \in S_f} (max\{0, w_{j,t}^f - \sum_{j' \in L_j^f} w_{j',t}^f\}) \leq S_j(t)$$

In the formulation proposed in this work, the max function is not introduced. Indeed, for each sector j' we know its previous sector j. Then $\sum_{j' \in L_j^f} w_{j,j',t}^f$ can takes 1 if and only if $\sum_{i \in P_j^f} w_{i,j,t}^f$ takes 1.

$$\sum_{f \in F: j \in S_f} (\sum_{i \in P_j^f} w_{i,j,t}^f - \sum_{j' \in L_j^f} w_{j,j',t}^f) \leq S_j(t)$$

In this way, and unlike the Bertsimas et al. model [2], the term ($\sum_{i \in P_j^f} w_{i,j,t}^f - \sum_{j' \in L_j^f} w_{j,j',t}^f$) never takes -1, for this reason, the max function is not introduced. Finally, the constraint obtained is linear and thereafter the model obtained is linear.

Table 1. Computational results with ILOG CPLEX

			Bertsimas et al.[2] Formulation					New Formulation				
Instance	Capacity	Objective Function	GAP (%)	Solution Time (Sec)	Cuts			GAP (%)	Solution Time (Sec)	Cuts		
					Clique	Bound	Gomory			Clique	Bound	Gomory
3003	40	7019	0	342.80	17885	1198	-	0	119.04	19437	41	16
	50	6713	0	347.07	19162	101	11	0	101.42	19773	34	6
	100	6163	0	302.42	18566	91	16	0	83.86	18819	38	5
3140	30	11051	0	641.76	17814	1241	-	0	615,25	18219	623	5
	60	10696	0	514.46	17923	613	4	0	327.83	18342	256	4
	70	10236	0	456.14	19224	203	6	0	114.96	19456	56	3
3196	40	10821	0	867.24	19943	518	14	0	557.25	19701	116	21
	60	10345	0	645.08	20354	124	8	0	416.09	20171	68	3
	90	10068	0	371.69	19214	83	4	0	127.76	19266	11	6
3230	50	11363	0	1118.10	19984	451	63	0	767.18	20101	103	29
	70	11104	0	925.70	21738	167	18	0	651.09	21386	86	69
	100	10811	0	718.23	20819	103	11	0	428.46	20874	45	4

Constraint (6). To ensure the arrival of the flight at one of its subsequent sectors, Bertsimas et al. [2] introduce the following constraint:

$$w^f_{j,\overline{T}^f_j} \leq \sum_{j' \in L^f_j} w^f_{j',\overline{T}^f_{j'}}$$

Again, introducing a fourth index in decision variable, the constraint becomes an equality constraint.

$$\sum_{i \in P^f_j} w^f_{i,j,\overline{T}^f_j} = \sum_{j' \in L^f_j} w^f_{j,j',\overline{T}^f_{j'}}$$

3 Numerical Results

In this section, we present the computational experience on the mathematical model presented in Section 2.

To compute optimal solutions we use the ILOG CPLEX solver 12.2, implemented using OPL as modeling language on a PC Dell precision M6300 workstation 2 processors 2.50 GHz, 3.00 GB RAM with Linux Ubuntu 12.04 OS. Each instance has 15 minutes per resolution. If in this time we do not have a solution we consider that the instance does not have a feasible solution.The data used for simulation is the data used by Bertsimas et al. in [2]. We consider a five hour time horizon subdivided into 6-8 time units (the period t).

The computional results are reported in Table 1. In the first column the size of instance (number of flights) is reported. The capacity of sector and the value of objective function are reported respectively in the second and third column. The five following columns gives the numerical result for Bertsimas et al. [2] formulation and the five last column gives that of our formulation. The number of additional cuts of clique, implied bound and Gomory type, generating by CPLEX, are reported in the sixth, seventh and eighth column of Table 1 respectively for the first formulation and in eleventh, twelfth and thirteenth for our formulation. The number of cuts applied in the first formulation is greater than the number of cuts applied in our formulation which makes the formulation is faster in terms of computation time. Numerical results show the importance of modifications we provide on the formulation of Bertsimas et al. [2] and especially the modification apported to constraints (3) and (6) after which they obtained a linear integer program.

4 Conclusions

This paper presents a new formulation for air traffic flow management problem which takes account the rerouting decisions. The formulation of Bertsimas et al. [2] is taken as basic model with a modification on the decision variable in order to have a linear integer program. Computational analysis on realistic instances show the efficiency of our new formulation that is faster in terms of computation time.

For the cases of big sizes instances, CPLEX is incapable to supply optimal solutions. As perspectives, we shall propose a meta-heuristic method for resolve the cases of bigger sizes instances.

References

1. Ball, M.O., Barnhart, C., Nemhauser, G., Odoni, A.: Air transportation: Irregular operations and control. Handbooks in Operations Research and Management Science 14, 1–73 (2006)
2. Bertsimas, D., Lulli, D., Odoni, A.: An Integer Optimization Approach to Large-Scale Air Traffic Flow Management. Operations Research 59(1), 211–227 (2011)
3. Bertsimas, D., Lulli, G., Odoni, A.: The Air Traffic Flow Management Problem: An Integer Optimization Approach. In: Lodi, A., Panconesi, A., Rinaldi, G. (eds.) IPCO 2008. LNCS, vol. 5035, pp. 34–46. Springer, Heidelberg (2008)
4. Bertsimas, D., Stock, S.: The Traffic Flow Management Rerouting Problem in Air Traffic Control: A Dynamic Network Flow Approach. Transportation Science 34, 239–255 (2000)
5. Bertsimas, D., Stock, S.: The Air Traffic Management Problem with Enroute Capacities. Operations Research 46, 406–422 (1998)
6. Bertsimas, D., Odoni, A.: A critical survey of optimization models for tactical and strategic aspects of air traffic flow management. Technical report, NASA (1997)
7. EUROCONTROL Performance Review Commission, Performance Review Report, Brussels (2004)
8. Helme, M.: Reducing air traffic delay in a space-time network. In: IEEE International Conference on Systems, Man and Cybernetics, vol. 1, pp. 236–242 (1992)
9. Hoffman, R., Mukherjee, A., Vossen, T.: Air Traffic Flow Management. Working Paper (2007)
10. Lindsay, K., Boyd, E., Burlingame, R.: Traffic flow management modeling with the time assignment model. Air Traffic Control Quarterly 1, 255–276 (1993)
11. Lulli, G., Odoni, A.: The European Air Traffic Flow Management Problem. Transportation Science 41, 1–13 (2007)
12. Odoni, A.: The Flow Management Problem in Air Traffic Control. Springer, Berlin (1987)
13. Performance Review Commission. Performance review report, an assessment of air traffic management in Europe during the calendar year (2008)

Modeling the Structure of Recommending Interfaces with Adjustable Influence on Users

Jaroslaw Jankowski

Faculty of Computer Science and Information Technology
West Pomeranian University of Technology
ul. Zolnierska 49, 71-410 Szczecin, Poland
jjankowski@wi.zut.edu.pl

Abstract. Recommending interfaces are usually integrated with marketing processes and are targeted to increasing sales with the use of persuasion and influence methods to motivate users to follow recommendations. In this paper is presented an approach based on decomposition of recommending interface into elements with adjustable influence levels. A fuzzy inference model is proposed to represent the system characteristics with the ability to adjust the parameters of the interface to acquire results and increase customer satisfaction.

Keywords: recommending interfaces, online marketing, web analytics, fuzzy modeling.

1 Introduction

Recommending systems were initially used mainly in various areas of electronic commerce. In recent years several other sectors like news portals, social platforms, and entertainment websites adopted them for targeted information filtering [10][14][15]. Other direction is related to implementing recommending interfaces within B2B platforms and as a part of applications supporting intelligent supply chains [26]. One of the most explored areas of recommender systems research was data processing and algorithms related to collaborative filtering and accuracy [14]. Initially less attention was paid to the design of user interfaces of recommender systems. Research targeted to user experience apart from pure accuracy becomes indispensable for further advances [4][11][13]. Recent studies also refer to this sphere and present results related to the organization of the interface, trust in recommendation agents and explanation processes [19][24]. There is a growing area related to building recommender interfaces and topics addressed to persuasions, structures of recommending lists and emphasis on the designing of recommender interfaces [23]. Extensive usage of persuasion elements within recommendation interfaces can lead to situations where users are motivated to follow recommendations without real interest in the products or services. This can negatively affect user experience and satisfaction. In this paper, we addressed the trade-off between website operator goals so as to motivate users to follow recommendations and real interest in recommended products or services and customer satisfaction. We proposed a fuzzy

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 429–438, 2013.

model to capture the characteristics of the interface such as influence levels and built an inference system showing relations between persuasion and customer satisfaction with the goal of building compromise solutions.

2 Related Work

Earlier research in the field of recommender systems was usually related to the accuracy of recommendations and the generation of most accurate rankings and algorithms [16]. Other areas are related to trust building, case based reasoning or user-system interaction with core conditions like choice, discovery and relevance or diversity applied to a set of items [19]. Recent publications showed the importance of interfaces and apart from correct data processing at the level of filtering algorithms there is a need to build proper interfaces to attract user attention and to convince the user to select recommended products [16]. Interaction design of recommender system depends on the main goal of the system and can guide users to explore products or help to discover new areas [23]. It is emphasized that more efforts should be put into the study of recommender systems from the user centric perspective instead of concentrating mainly on recommended items and as a result recommender systems are analyzed from the human-computer interaction point of view as well [13][23]. One of the earliest research in the field of recommending interfaces was based on different layouts tested together and results showed the importance of the form of data and explanations delivered together with the recommendation [5]. The authors analyzed the differences between informative and persuasive interfaces and opened some research questions related to both approaches. Other research presented results based on exploring categories generation and text usage for titles and products description [19]. The proposed organization algorithm identified the main principles and was based on solutions where each recommendation was represented by a vector comprising of a set of attributes and trade-off pairs. Apart from these mechanisms of absorbing user attention and perception can be analyzed in relation to different layouts of recommending interfaces. Research based on list view, organization interface and quadrant arrangement was performed with eye-tracking and eye gaze patters [2]. The authors opened new areas related to investigating perceptual processes at different layouts and building predictive models of user's cognitive architecture. Interface structures and content were analyzed in relation to the importance of different types of information displayed at the recommending interface in the research presented by A. A. Ozok et. al [17]. The results indicate that price, image and names of products are identified as essential information. Other elements like product promotions, customer ratings and feedback where identified as secondary types of information. Different graphics, text, descriptions, explanations played a role in building trust in the system. The authors analyzed several aspects of the interface components and the research was based on general structures of the interface and its content. However the role of the different attributes, animations, colors was not examined at the more detailed level. Research shows that, the building interfaces are connected with problems of selecting optimal design and has high impact on the overall user satisfaction and effectiveness of the interfaces. In the area of recommender systems apart from design layer is explored and a persuasive

technology which is developed in the several areas like ambient persuasive, web persuasive and user experience based platforms [9]. Persuasive communication is any message that is intended to shape, reinforce, or change the responses. Persuasiveness of recommender interfaces takes part in recent research and can be used to change cognition, attitude and behavior. Reaction on recommender interface can be related to dual-mode processing models of persuasion, like heuristic systematic model and elaboration likelihood model and is dependent on audience characteristics. Results showed that effectiveness of recommendations in many cases is more related to persuasiveness than to correctness. The trade-off between system persuasion and satisfaction for users was analyzed by T. Nanou [16]. The research illustrates different impact of recommendations organization on persuasion and user satisfaction which is dependent on structures of the presentation layer and these results were confirmed in earlier findings [19]. Different factors related to the communication process and affecting persuasion are connected with source, message tone, argument quality and audience characteristics and it is difficult to measure all the potential relevant features because of the lack of appropriate metrics and representation [20]. The process of recommendations should include elements leading to behavior changes and model for persuasive design with principal factors like motivation, ability etc. Users with low motivation can only perform well if the task is simple enough [3]. The research showed that higher level of persuasiveness leads to higher offer acceptance rate by customers. The presented papers are related to general elements of interfaces and can be treated as a strategic level showing the approaches based on organization guidelines using heuristics to create the structure of the interfaces [1]. The complexity of the problem and many connected areas show that more detailed research should be performed in relation to different elements of recommender systems structure and multidimensional analysis of user response and use knowledge based systems like in other areas of applications [21]. The research shows that users are skeptical to marketing language, sponsored texts and high level of persuasion and instead of increased response it can negatively affect customer satisfaction [3]. In our research we assume that the ability to persuade is limited and if we move the level of persuasion above the critical point, effect will be different than assumed and can negatively affect customer satisfaction. In this paper we propose inference system based on fuzzy modeling and the usage of quantitative methods to analyze response from web users and a generation of trade-off design taking into account interaction with the interface expected by website operator and customer satisfaction rate. Our research integrates in the fuzzy inference system response from user's perspective because the goal of a website operator is the satisfaction rate which represents the user evaluation of a recommended product. The conducted research showed the difference between the interfaces with the main goal to persuade users to perform desired actions and an interface delivering content according to the user's expectations with low impact and persuasion.

3 Conceptual Framework and Experiment Design

In this section the interactive recommending interface and structure of fuzzy inference system is presented, which enables the determination of the influence levels on the

user and tracking of the response. Recommending interfaces can be identified as an interactive object integrating the components that influence users and motivates them to have interactions. For every recommending object we defined a set of available components determined by $E= \{e_1,e_2,\ldots,e_n\}$ and for every e_i there is a set of available variants $e_i = \{v_{i,1},v_{i,2}, \ldots,v_{i,cnt(i)}\}$ where $cnt_{(i)}$ describes the number of variants available for the *i*-th element. For every variant $v_{i,j}$ can be assigned an influence level $l_{i,j}$ which defines the strength of persuasion on a user. At the moment of time *t* the user interactive object with adjusted influence level on user *i* is selected from $S_{i,t} =\{sel(e_1), sel(e_2),\ldots,sel(e_n)\}$ with a function $sel(e_i)$ selecting level for each element. The selection vector $S_{i,t}$ is computed from the set of all possible design variants $D=\{d_1,d_2,\ldots,d_p\}$ where *p* is the total number of possible combinations. The main purpose of this approach is to obtain S_{opt} vector maximizing factors related to results acquired by website operators which is represented by the conversion rates CR with the positive impact on customer satisfaction represented by satisfaction rate factor SR. The measurement used for the conversion factors in the time period *t* is represented by CR_t and it determines the relation of a number of desired interactions I_t to the number of website users U_t in a given time *t*. In case of multi-dimensional monitoring and realization of advertising campaigns we can distinguish partial conversion for the chosen types of interaction and also determine the segments of audiences. Conversion CR for design variant D_k can be determined in relation to the audience segment *s* and the type of interaction *i* in the period of time *t* and can be presented according to the following formula:

$$CR_{s,i,t}[D_k] = \frac{I_{s,i,t}[D_k]}{U_{s,i,t}[D_k]} \tag{1}$$

where $I_{(s,i,t)}[D_k]$- number of interactions of type *i* generated within segment *s* of audience in the period of time *t* with delivered variant D_k, $U_{(s,i,t)}[D_k]$ - the total number of website users assigned to segment *s* in the period of time *t* receiving variant D_k of design. Apart from the conversion rates we measure customer satisfaction rate $SR_{s,i,t}[D_k]$ which can be measured using different metrics depending on areas of application like ratings, frequency or time of service usage in example based on virtual and real time dimension [8]. Our approach assumes building response surface for different levels of persuasion and measuring both conversion rate and satisfaction rate used for training inference system targeted to exploring design space and generating final variant. Influence levels can be defined by linguistic representation in a more natural way than describing them with crisp numbers. According to specifics of interactive messages different influence levels can be assigned into fuzzy sets depending on the needed accuracy and borders between levels. In the example for six influential levels we can define three fuzzy sets *Low*, *Middle* and *High* as showed in Fig. 1.

Level l_1 is assigned to *Low* set with probability p = 1, l_2 belongs to both *Low* and *Middle* sets with p =.9 and p=.1 respectively, l_3 belongs with p =.4 to *Low* set and with p=.6 to *Middle*, l_4 belongs with probability .4 to set *Middle* and with .6 to set *High*, l_5 with probabilities .1 and .9 to sets *Middle* and *High*. Level l_6 is assigned to set *High* with probability 1. Fuzzy sets theory was developed by L.A. Zadeh [25] and

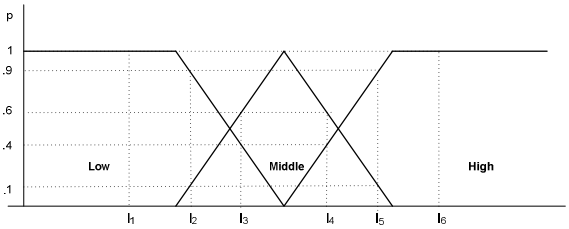


Fig. 1. Changes in influence levels assigned to fuzzy sets.

was extended in the field of decision making towards control systems and reasoning systems proposed by E.H. Mamdani et al. [12] just to name the main areas. Number of sets and shape of membership functions are dependent on the number of variables and specifics of the input data [18]. In this paper we show its application in the field of design of interactive interfaces and recommending systems based on approach presented earlier in relation to web design [7]. The influence level can be defined as fuzzy sets and covers a range of values of attributes with different probabilities and are based on membership functions. With this approach we use fuzzy sets as inputs to inference system with better ability to cover decision space. Fig. 2 shows the structure of the system with inputs $e_1,e_2,\ldots,e_n$ representing the elements of the interface and the levels of influence on users with the use of inference mechanisms based on adaptive neuro-fuzzy system introduced by R. Jang [6].

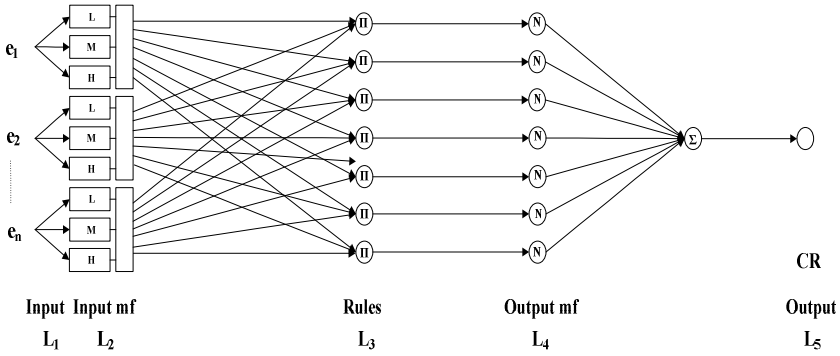


Fig. 2. Structure of adaptive neuro-fuzzy inference system [6]

The first layer is responsible for the processing of input data, a parameterization and the introduction of premises. In the next layer the support product is generated and the levels of each rule, used in a system are determined. The third layer sends normalized rules on the output. Another layer generates inference parameters, and the output of fifth layer is a deffuzificated signal. The hybrid methods, least-squares and back propagation method can be used in order to identify characteristics of the belonging function and the set of rules. Because of the many possible levels of influence and elements of interface, it would be difficult to perform full factorial experiments to get all possible combinations of the design. To reduce this problem we

assume sampling of decision space with limited number of influence level and build a response surface. At this stage we used surface sampling to get the output from the design space. Sampling methods for fuzzy inference models can be adjusted to specifics of input datasets [18]. In reference to recommending interfaces the effects can be analysed in a few dimensions. To measure both CR and SR we designed an experiment based on dynamic recommendation object with changeable influence levels. The main goal of the experiment was to observe how the intensity of the elements on recommendation interface can affect the usage of recommendations and how the intensity is related to both the measured factors. Conversion rate CR was measured by a number of users following recommendation among all users and SR was represented by the time of spent on the recommended subsite with editorial content targeted to user. We assume that users following recommendations with low persuasion were more convinced to exploring recommended content and the time was relatively longer than a result of recommendations with high intensity. The schematic structure of recommending object is presented in Fig. 3.

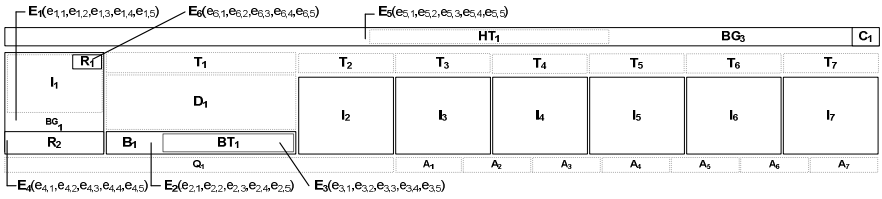


Fig. 3. Structure of experimental recommender module

The object was dynamically generated each time the website was reloaded. During the generation, seven recommended objects were selected from the database and was shown in a form of thumbnail images I_1-I_7 with size 105 x 89 pixels with titles T_1-T_7. One of the visual objects was always shown as a featured on the left side with additional description D_1 and recommendation R_1in a form of text (*Recommended* and rating R_2 info and active background BG_1). Following recommendations was possible only after clicking on the thumbnail and then clicking on button B_1 with the call to action text BT_1. The header bar with title $HT_1$was inviting users to explore recommended content and background BG_3. Web users had an option to remove recommending object by clicking the C_1 button. Changeable elements where assigned to E_1-E_6 sets of available design variants, each of them with five levels of influence. The system was selecting recommended object based on the category of the currently explored content however, this algorithm is not the scope of our research. During the experiment, the system registered 1241029 impressions, 37880 clicks on objects presented on recommending interface. With factorial ANOVA we identified that the highest impact on user interactions and decision taken to follow recommended content was related to thumbnail background e_1 (p=0,000329). A smaller but statistically significant was the impact of the call to action text on button e_3 (p=0,007635) and the recommendation text visible on thumbnail e_6 (p=0,010763). Apart from the main effects we identified interactions between elements of design e_6*e_2 with significance p = 0,027577 and e_2*e_1 with p = 0,042150 which shows that

the influence on users with two elements was easier to generate enough stimuli to motivate users to perform the expected action. The obtained data were used in the next step for fuzzy models in training and verification of the process.

4 Fuzzy Inference System Based on Adaptive Neuro-fuzzy Model

The initially conducted analysis determined dependencies between the components, their persuasion, and response levels. Based on these inputs, the structure of the fuzzy model was developed, for which measurement and data inputs were established for individual parameters. The input parameters reflected the individual components of vector E = [e_1,e_2,e_3,e_4,e_5,e_6], whose values were acquired from empirical data. On the output the response levels were obtained, which were determined for the parameters' vector. Fig. 4 shows the response from the system for changes made to the e_2 and e_1 elements. The obtained results show that changing e_1 from lowest to highest levels were influencing users in taking decision to perform any interaction and CR was grew from 2.5 up to 3. Changing e_2 while keeping e_1 on highest level was increasing the response from 3% up to 3.2% and this confirms interaction between both elements and cumulative effect as was identified earlier. Fig. 5 shows the relation between e_6 and e_2 elements. While e_2 was on lowest level changes of e_6 were not resulting in changes in response and the observed response was in the range 2.70% - 2.75% for e_6=1 and e_6=5 respectively.

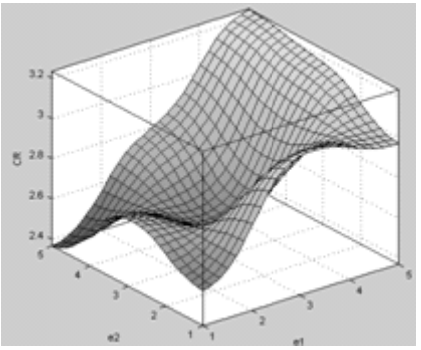


Fig. 4. Interaction between e_2 and e_1 elements in relation to conversion CR

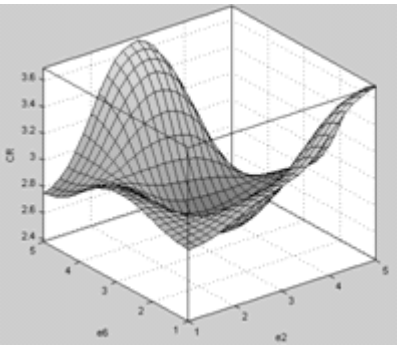


Fig. 5. Interaction between e_6 and e_2 elements in relation to conversion CR

Another situation was observed when e_2 was switched to *Middle* level and switching of e_6 to *High* level was resulted in the increase in response of CR up to 3.6%. This shows how animated button and animated *Recommended* text was resulting in cumulative influence on a user and increased the number of interactions. In the next step we analyzed how text and visual animated elements were affecting results. At the first stage we used textual call to action associated with e_3 element and results showed that it was possible to increase response using e_3 and e_1 parameters while keeping all the other parameters on the *Low* level. The highest response for e_3 was obtained at the *1-st* level while drop in response was observed at *High* level and

this is similar to results obtained from ANOVA. Recommending content to be explored with text *Recommended* with different animations was not motivating users to use this content and switching e_6 was not resulting in big changes in the response even at highly animated *5-th* level. In the next stage, we built a model for inference system showing how different levels on persuasion within recommending interface can affect user satisfaction represented by time exploring the recommended content. On Fig. 6 we showed the relation of satisfaction rate SR to changes of influence levels within element e_1 and e_3 representing the visual element and the verbal intensity of the button. While intensity was growing due to the shorter time of exploring the content, the user was motivated to follow recommendation with influence element however, satisfaction from this type of recommendation was at a relatively low level, especially when e_3 went to *5-th* level and the time of exploration went down to 210 seconds. Keeping both elements at the lowest level was resulting in a time at a high level of 450. A different situation is observed with changes of e_6 and e_5 as shown on Fig. 7. Increasing the intensity of *Recommended* sign represented by e_6 was motivating users to spend more time exploring the recommended content and the header e_5 was not strongly affecting the time.

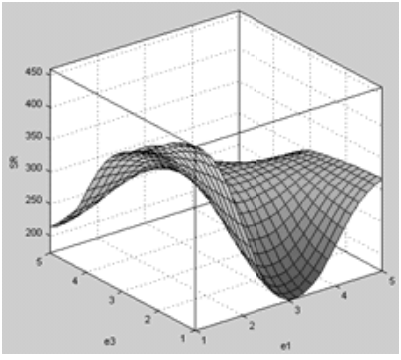


Fig. 6. Influence of elements e_1 and e_3 on SR

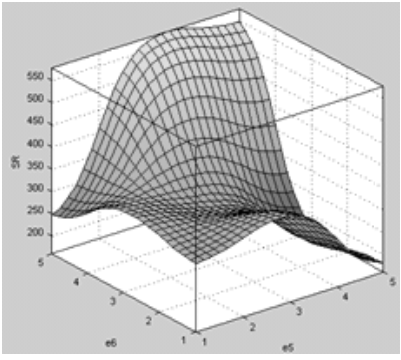


Fig. 7. Influence of elements e_6 and e_5 on SR

The overall results show that knowledge about system behavior delivered by a fuzzy inference model can be used to build interface with limited impact on user delivering both CR and SR on acceptable levels. In our model, the final design can be based on middle level of e_3 keeping e_4 on a high level. This variant of design would deliver CR on the level of 2.9 % which is higher from the lowest obtained level for control design E = [1,1,1,1,1]. At this level, the negative impact on users is limited and from both website operators goal, the user's satisfaction solution can be treated as a compromise. Our results extended earlier research related to the role of elements within recommending interfaces like price, image and names of products. They were identified as the essential information in recommending systems but a more detailed role of intensity of influence was not analyzed. The obtained model with fuzzy inputs can be generalized and this shows how this approach can be used while modeling web interfaces with parameters hard to define with crisp numbers. The obtained model showing which level is worth to increase the level of interface intrusiveness obtained higher conversion. This approach can be used in modeling web recommending

interfaces with the goal to deliver balanced solutions between high conversions and limited negative impact on users. It is especially important while marketers are using more and more intrusive techniques to get attention of users and trying to overcome the effect of dropping click through rates.

5 Summary

Recommending interfaces are important part of websites in many sectors and are integrated with ecommerce, social and entertainment pages. While the main goal is to attract users with better selection of products, marketers are trying to motivate them to use recommended items. The obtained results show interactions between levels of influence and response from web users. In our research we treated the problem as a trade-off approach where the goal was to find a balance design with the ability to increase the number of conversions without a negative impact on users resulting in a decrease of customer satisfaction. The proposed way of constructing interactive objects with the call to action intention can be used in analytical processes and searching for optimal design. The nature of the parameters which were hard to the capture the boundaries between the influence levels gave the motivation to use fuzzy inference models. The presented approach can be used for the selection of design variants and will help to avoid situations when persuasive elements with high influence on users' decisions can transform the interface into intrusive website. Elements distracting users from the main tasks within a website can result in a drop in user experience and customer satisfaction which can have a negative side effect instead of building positive relations with web users.

References

1. Barneveld, J., Setten, J.: Designing usable interfaces for TV recommender systems. In: Ardissono, L., Kobsa, A., Maybury, M. (eds.) Personalized Digital Television: Targeting Programs to Individual Viewers. Human-Computer Interaction Series, vol. 6, pp. 259–286. Kluwer, Dordrecht (2004)
2. Chen, L., Pu, P.: Eye-Tracking Study of User Behavior in Recommender Interfaces. In: De Bra, P., Kobsa, A., Chin, D. (eds.) UMAP 2010. LNCS, vol. 6075, pp. 375–380. Springer, Heidelberg (2010)
3. Fogg, B.J.: A behavior model for persuasive design. In: 4th International Conference on Persuasive Technology Persuasive 2009, pp. 40–47. ACM, New York (2009)
4. Hayes, C., Massa, P., Avesani, P., Cunningham, P.: An online evaluation framework for recommender systems. In: Workshop on Personalization and Recommendation in E-Commerce, Malaga, Spain, pp. 57–67. Springer, Heidelberg (2002)
5. Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: Proceedings of the ACM Conference on Computer Supported Cooperative, pp. 241–250. ACM Press, New York (2000)
6. Jang, J.S.R.: ANFIS: Adaptive-Network-Based Fuzzy Inference Systems. IEEE Transactions on Systems, Man and Cybernetics 23(3), 665–685 (1993)
7. Jankowski, J.: Integration of Collective Knowledge in Fuzzy Models Supporting Web Design Process. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part II. LNCS (LNAI), vol. 6923, pp. 395–404. Springer, Heidelberg (2011)

8. Jankowski, J.: Analysis of Multiplayer Platform Users Activity Based on the Virtual and Real Time Dimension. In: Datta, A., Shulman, S., Zheng, B., Lin, S.-D., Sun, A., Lim, E.-P. (eds.) SocInfo 2011. LNCS, vol. 6984, pp. 312–315. Springer, Heidelberg (2011)
9. Jawdat, A., Obeidat, Q., Aljanaby, A.: On The Design of User Experience Based Persuasive Systems. Computer and Information Science 4(4), 90–99 (2011)
10. Kazienko, P., Musiał, K., Kajdanowicz, T.: Multidimensional Social Network in the Social Recommender System. IEEE Transactions on Systems. Man, and Cybernetics - Part A: Systems and Humans 41(4), 746–759 (2011)
11. Kukla, G., Kazienko, P., Bródka, P., Filipowski, T.: SocLaKE - Social Latent Knowledge Explorator. The Computer Journal 55(3), 258–276 (2012)
12. Mamdani, E.H., Folger, T.A., Gaines, R.R.: Fuzzy reasoning and its applications. Academic Press, London (1981)
13. McNee, S.M., Riedl, J., Konstan, J.A.: Being accurate is not enough: How accuracy metrics have hurt recommender systems. In: CHI 2006 Extended Abstracts on Human Factors in Computing Systems, pp. 1097–1101. ACM, New York (2006)
14. Montaner, M., Lopez, B., De La Rosa, J.L.: A taxonomy of recommender agents on the Internet. Artificial Intelligence Review 19(4), 285–330 (2003)
15. Nageswara, R.K., Talwar, V.G.: Application domain and functional classification of recommender systems - a survey. Journal of Library & Information Technology 28(3), 17–35 (2008)
16. Nanou, T., Lekakos, G., Fouskas, K.: The effects of recommendations' presentation on persuasion and satisfaction in a movie recommender system. Multimedia Systems 16, 219–230 (2010)
17. Ozok, A.A., Fan, Q., Norcio, A.: Design guidelines for effective recommender system interfaces based on a usability criteria conceptual model. Behaviour and Information Technology 29(1), 57–83 (2010)
18. Piegat, A.: Fuzzy Modeling and Control, Verlag Heildelberg, New York (2001)
19. Pu, P., Chen, L.: Trust building with explanation interfaces. In: Proceedings of the 11th International Conference on Intelligent user Interfaces (IUI 2006), pp. 93–100. ACM, New York (2006)
20. Qinyu, L.: Empirical findings on persuasiveness of recommender systems for customer decision support in electronic commerce, PHD dissertation, Missisipi State University, USA (2005)
21. Różewski, P., Małachowski, B.: Competence Management in Knowledge-Based Organisation: Case Study Based on Higher Education Organisation. In: Karagiannis, D., Jin, Z. (eds.) KSEM 2009. LNCS (LNAI), vol. 5914, pp. 358–369. Springer, Heidelberg (2009)
22. Swearingen, K., Sinha, R.: Interaction design for recommender systems. In: Proceedings of the 4th ACM Conference on Designing Interactive Systems (DIS 2002). ACM, New York (2006)
23. Swearingen, K., Sinha, R.: Beyond algorithms: An HCI perspective on recommender systems. In: ACM Workshop on Recommender Systems SIGIR 2001. ACM, New York (2001)
24. Tintarev, N., Masthoff, J.: A Survey of Explanations in Recommender Systems. In: Proceedings of the IEEE 23rd International Conference on Data Engineering (ICDEW 2007), pp. 801–810. IEEE Computer Society, IEEE Press, Washington (2007)
25. Zadeh, L.A.: Fuzzy Sets. Information and Control 8(3), 33–353 (1965)
26. Zhang, X., Wang, H.: Recommender Systems for B2B Electronic Commerce. Communications of the IIMA 5(4), 53–62 (2005)

Increasing Website Conversions Using Content Repetitions with Different Levels of Persuasion

Jaroslaw Jankowski

Faculty of Computer Science and Information Technology
West Pomeranian University of Technology
ul. Zolnierska 49, 71-410 Szczecin, Poland
jjankowski@wi.zut.edu.pl

Abstract. A user's behavior on a website is usually based on a sequence of loaded pages and repetition of content and this is a natural part of communication with a website. Changes of persuasion on a user between repetitions can lead to the increased number of interactions expected by a website operator but can affect user experience as well. In this paper we propose the modeling of the parameters of objects used in repetitions based on the fuzzy inference system towards increased conversions with limited negative impact on a user.

Keywords: Website conversion, web analytics, fuzzy modeling, online marketing, website optimization.

1 Introduction

Communication with potential customers using interactive content is a key element of online marketing processes. Recent trends show evolution of web analytics and optimization methods toward techniques addressed to increasing the number of interactions desired by website operatorsin both retail and business to business sector [16]. A website's effectiveness is represented by conversion metrics and interactions like user clicks, signups or purchases [17][18]. Conversion maximization is often accomplished at the expense of parameters of system-usability assessment. As a result, a negative side effect of the increase of a website's intrusiveness is observed because of attempts to attract a user's attention using animations or contrasting content [15]. Searching for optimal design using experimental design based on multivariate testing usually generates a single optimal version of a website without taking into account repeated contacts with the website. The area of content repetitions is not widely discussed in research related to websites optimization but was analyzed in the field of online advertising and especially banners [1][15]. Our research extends the available approaches towards measuring response based on changes of persuasion levels between repetitions and searches for optimal level of changes. For parameters like changing levels of influence, intensity of increasing and decreasing persuasion between repetitions, we used representations based on linguistic approach and membership functions. The proposed approach based on fuzzy inference system

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 439–448, 2013.

makes it possible to process this type of data in a more natural way. The obtained model gives us knowledge about the design space with the ability to select a set of parameters which changes between repetitions and delivers conversions with limited negative impact on users.

2 Literature Review

Converting visitors into customers is one of the target stages of the process of delivering web traffic to a website from several external sources like paid advertising or organic search. While website optimization is usually addressed to the different aspects of usability, user experience or search engine optimization, it can be targeted to increasing the results obtained from web traffic as well. In this case, the main goal is increasing the number of actions performed by users like clicks, signups, software downloads, purchases and other measured factors like usage frequency or loyalty desired by a website operator [2][18][7]. Conversion in this case is defined by the number of acquired interactions in relation to the number of visitors and this is one of factors affecting a website's effectiveness. Interaction with a web system is usually based on web browsing and navigation through webpages of portal, social networking platform or ecommerce system. The term *interaction* is defined as a recursive communication with a system in which there is exchange of data or information related to a previous stage of communication [13]. In earlier research, different dimensions of interaction were analyzed. D. Hoffman and T.P. Novak combined the definitions and addressed them to the Internet, where interaction can occur both with the system as machine interactivity as well as through the system in the form of personal interactivity [4]. Motivating users to perform the desired actions during interaction with a website and initiating them is the goal of effects oriented websites and web designers, taking into account other goals of a website [17]. A. Schlosser indicates the evolution of websites in the direction of conversion-oriented systems and he also identifies the main elements connected with website design that influence the acquired results [16]. The problem of website effectiveness is multidimensional and connects several areas like marketing, psychology, usability, human factors and computer science. The process of conversion maximization is discussed in many aspects, from building, call to action messages, optimal design, toward neuro-marketing. Both researchers and practitioners usually search for optimal designs and are not dealing with switching content and ability to change user behaviors between repetitions. The area of repetitions was earlier discussed in terms of online advertising and banners. One of the discussed areas in this field is the role of frequency in communication and identification as well as how many times and in which form marketing content should be shown to achieve certain results. Research related to advertising and banners shows a relation between results and number of impressions. Among others B. Baccot showed a research based on the most appropriate presentations like textual, visual, audio, interactivity levels and sequence of banners evaluation in terms of content and time [1]. The research showed different dimensions like sequence, timing, content format and how the appropriate formats can lead to an increase in the interest of the user and how it improves the overall effectiveness of a website. P. Chatterjee showed that click response to repeated exposures to passive ads within the same visit will be negative and he discussed the passive and active

advertisement as the main two elements of online marketing communication [3]. Practitioners have identified at the early stage of online marketing that the mean click probability will decrease with each successive passive ad exposure during the same visit. The analysis of decreasing returns to repeated passive ad exposures suggests that after the third exposure the probability of interaction will drop because of the effect identified as banner burnout. But even though repeated messages are resulting in lower probabilities of interaction exposures distributed over time and the absence of perception is observed, repetitions lead to awareness, familiarity, and positive effect towards the ad stimulus under low involvement conditions [14]. Other research confirmed a non-linear effect of repeated passive ad exposures during the same visit and in the short-term during the same visit, there is a positive long-term effect which will be recorded only if the consumer is exposed to more units [3]. A research conducted by B. Baccot confirms that repetition reduces the number of interactions but it builds brands [1]. The discussed areas and research related to repetitions is addressed mainly to advertising and banner campaigns. Conversion optimization within a website has different specifics because of the characteristics of the content and the role of interactive objects which are not typical advertisements. Users after interaction are not leaving the website and their actions are usually integrated within the same website. Call to action elements are not in a form of advertisements and brand awareness is not a target of a website operator, especially if it is related to increasing conversions based on call to action like elements such as buttons or headers. Analyzing processes within a website can distinguish the main goal of visiting a website in a form of interaction with editorial content and attempts to change user behavior by a website operator to motivate a user to perform the desired action. In this phase, call to action messages and visual elements can be used and have the ability to distract a user from interacting with the main content and to move the user's attention to other areas of a website. Among other parameters, website optimization should take into account characteristics of browsing patterns in a form of series of website reloads and ability to deliver different call to action content with each repetition. In this context, it is justified to develop a model of a system which enables the determination of the influence levels and the scope to which the increased persuasion intensity reflects the acquired results. This approach is explored in our paper in the field of changing stimuli and levels of persuasion within a website dynamically during browsing activity with the ability to detect how the changes of the parameters are affecting the results. We presented a novel approach to constructing interactive object and measuring the increase or decrease of influence between repetitions. In this research, we assumed that the levels of persuasion, stimuli and intrusiveness can be represented by measurable metrics and they can be used to track changes between repetitions and this is an extension of our earlier research [8]. With our proposed approach, it was possible to perform detailed analysis of response as a result of intensity changes and to observe how the increase or decrease of influence is affecting response from users. Our approach makes it possible to use linguistic measures based on membership functions for the representation of changing states between repetitions in relation to elements of a website which are the subject of optimization.

3 Conceptual Framework for Inference Model Based on Differential Approach

Communication with web users and motivating them to perform a desired action by a website operator is usually connected with repeated messages sent to a user during the user's website content navigation and exploration. In our research we addressed this area to multi-attribute content and more detailed analysis of the changes of states. We proposed the usage of interactive objects with measurable levels of influence and repetitions which are based on changes of persuasion. In the communication process we used the interactive object $O_{u,k}$ like website header or advertisement delivered to user u in the *k-th* stage of communication consisting of n elements $O_{u,k}=\{e_1,e_2,\ldots,e_n\}$. For every element e_i we assigned a set of available variants $V_i=\{v_{i,1}, v_{i,2},\ldots, v_{i,m}\}$ where m is the number of variants for *i-th* element. For each variant $v_{i,j}$ is assigned an influence level $l_{i,j}$ which represents persuasion or intrusiveness and shows the relative strengths of level *j-th* in comparison to other levels of the element *i*. We defined the vector of changes $C[t_1,t_2]=[c_1,c_2,\ldots,c_n]$ representing changes of persuasion between time points t_2 and t_1 showing what was changed in $l_{i,j}$ calculated as $c_i=l_{i,j}(t_2)-l_{i,j}(t_1)$ between website reloads. An example of the process is shown in Fig. 1 where the system is delivering dynamic content to web system pages A,B,C,D with editorial content connected with hyperlinks. For *i-th* user and *j-th* contact with the website delivered an interactive object $O_{i,j}$ consisting of three design elements e_1,e_2,e_3 with five levels of influence each.

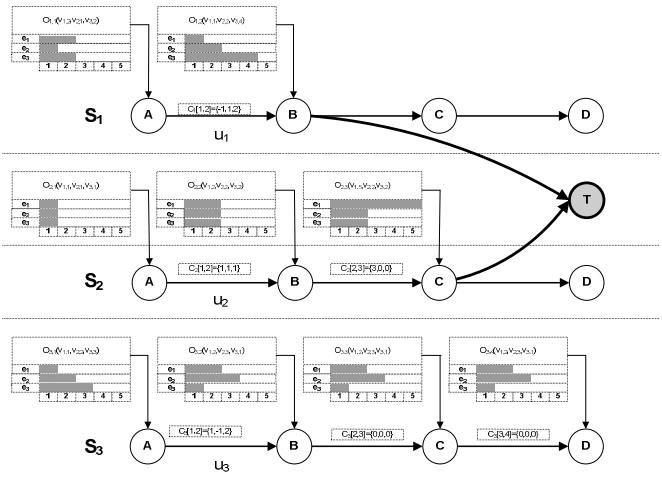


Fig. 1. Integration of adjustable influence levels with targeting system

Users u_1,u_2 and u_3 started navigation from the home page denoted as A. Within the structure which is connected by hyperlinks pages we define the transactional page and denote it as T. Redirecting users to the transactional page with call to action elements is one of goals of website operators. User u_1 started navigation and received content for first delivered object with e_1 at level 2, e_2 at level 1 and e_3 at level 2 denoted as

$O_{1,1}$=(2,1,2) and with second contact with page B received interactive object with levels for all elements adjusted to $O_{1,2}$=(1,2,4). The level for e_3 was increased from 2 to 4 and it resulted in a redirection to the target website F. The vector of changes between website loads is represented by $C_1[1,2]=\{-1,1,2\}$. User u_2 visited websites A, B, C which changes $C_2[1,2]=\{1,1,1\}$ between page A->B. The first repetition did not result in redirection to target site but the second repetition with $C_2[2,3]=\{3,0,0\}$ between B->C increased the influence enough and call to action message resulted in aninteraction. A similar communication is illustrated for user u_3 but the changes of states did not result in interactions. In our approach, we show the model of inference system which allows us to adjust persuasion levels during communication based on fuzzy inference model generated with data related to changes in stimuli in sequential communication. The bases for our approach is a generalized representation of multi attribute state changes in sequential communication where at time *t* the user *u* receives interactive object with a structure and parameters $O_t=\{e_{1,t}, e_{2,t}\ldots, e_{n,t}\}$ and at time *t+1*receives $O_{t+1}=\{e_{1,t+1}, e_{2,t+1},\ldots, e_{n,t+1}\}$. Objects are characterized by influence level vector $L_t=\{l_{1,t}, l_{2,t},\ldots, l_{n,t}\}$ and $L_{t+1}=\{l_{1,t+1}, l_{2,t},\ldots, l_{n,t+1}\}$. Having knowledge about state *t* and results from others web users, one of the goals of content management system is to assign to a user levels of component for *t+1* time. To evaluate the effectiveness of the design variants together with exposition we observed the conversion. The measurement used for such actions is the conversion rate CR_t, which determines the relation between the number of desired interactions I_t to the number of website users U_t receiving the interactive object. We assume that the conversion rate at *t+1* as a function of influence level changes between time points *t* and *t+1* denoted as $CR_{t+1}(c_1,c_2,\ldots,c_n)$ where $c_1=l_{1(t+1)}-l_{(1,t)}$, $c_2=l_{2(t+1)}-l_{2(t)},\ldots$, $c_n=l_{n,(t+1)}-l_{n,(t)}$. Conversion based on state changes represented with values c_1, $c_2,\ldots$, c_n can be determined in a multidimensional way in relation to audience segment *s*, type of measured interactions *i* in the period of time *t* according to the following formula:

$$CR_{s,i,t}[c_1,c_2,...,c_n]=\frac{I_{s,i,t}[c_1,c_2,...,c_n]}{U_{s,i,t}[c_1,c_2,...,c_n]} \quad (1)$$

Where $I_{s,i,t}$ represents the number of interactions of type *i*, generated within audience segment *s* in a period of time *t* are after changes in influence elements $e_1,e_2,\ldots,e_n$ and is equal to $c_1,c_2,\ldots,c_n$ respectively. Together with the number of interactions, a system can gather information about the total number of users $U_{s,i,t}$ acquired in time *t* in segment *s* who received interactive content with changes represented by $c_1,c_2,\ldots,c_n$. Building a model on whole state changes creates a big decision space. We use *m* elements and *n* states each time we obtain the set of states changes for each element from values (1-n) for maximal reduction from level *n* to level 1 and maximal increase (n-1) from 1 to highest value n. In this case, for each element we obtain |1-n|+n possible changes and it makes a combinatorial number of possible situations. With limited resources it would be difficult to build inference system for all possible combinations because of the decision space which is not covered. Characteristics of changes of influence levels can generate similarities and the possibility to reduce search space using linguistic representation for ranges of changes like *Low*, *Neutral* and *High* change of persuasion. It is a more natural way than using crisp numbers to define values which are subjective and not precise. Our approach uses fuzzy sets to represent values with membership function based on theory introduced by L. Zadeh,

where fuzzy set is defined as $F=\{(\mu_F^*(l_i),l_i)\}, \forall l_i \in L$ where μ_F^* is membership function of fuzzy set F which is assigning to each element $\forall l_i \in L$ level of belonging to the L set $\mu_F^*(l_i)$ where $\mu_F(l_i)$ [0,1][19]. Fuzzy sets theory was extended in the field of decision making towards control systems, linguistic approach and multi objective decisions making, just to name the main areas. Other areas of application are discussed in relation to online environments [6]. Recent emerging applications and research are conducted in the area of behavior adaptation [10]. In this paper is showed its applications in the area which is not yet explored. Levels of influence can be assigned as members of fuzzy sets with different probabilities according to specifics of interactive content and needed resolution.

4 Experiment Design and Empirical Research

In the next phase, we designed an experiment and conducted empirical research in the real environment. The experimental object was designed in a form of recommending interface with the main goal to motivate users using both textual and visual elements to click within the interface and follow the recommendations. The sequence of users' behaviors was related to exploring several pages with content during a single session. The experimental interactive object O was defined as a set of components $\{e_1, e_2, e_3, e_4, e_5, e_6\}$ with different role in communication process. Each component was assigned to different elements of design as follows: e_1-background color, e_2-background of button, e_3 - call to action text on button, e_4-users rating info, e_5-header background, e_6 - recommendation sign. During the experiment, the object was shown 1287704 times and it received 20546 clicks. For the acquired interactions where the assigned state changes and the example changes between page views PV_m and PV_n for user u_j can be computed as follows $C_{n,m}(u_j)=PV_n(6,3,4,4,5,6)$ - $PV_m(1,2,3,2,1,2)= \{5,1,1,2,4,2\}$. Using this approach, it was possible to track all sequences, the number of impressions registered for each state changes and the interactions of each type. An example of the subset of data is shown in Table 1. For each level of change denoted as c_1-c_6 we registered the number of impressions and the interactions and the conversion factor was computed.

Table 1. Selected set of state changes resulting conversions

c_1	c_2	c_3	c_4	c_5	c_6	Views	Interaction	CR
0	0	0	0	-4	0	407	3	0,2457
0	-2	-2	0	2	0	287	2	0,3484
-2	0	2	0	0	-2	270	2	0,3704
-2	0	0	-4	0	0	270	3	0,3704
-2	0	0	2	-2	0	269	2	0,3717
-2	0	0	-1	0	0	259	2	0,3861
0	2	0	-2	-2	0	256	2	0,3906
0	0	0	-2	-2	2	255	2	0,3922
0	0	0	-2	0	-4	254	3	0,3937
-4	0	0	0	0	-2	252	3	0,3968

For each factor, we generated changes from a set of value {-4,-3,-2,-1,0,1,2,3,4}. During the experiment a total of 3411variants of changes of states were registered for conversion CR rates with a total 8917 interactions and 593910 page views. The average conversion rate was 1.51%. In the next part, the goal was to build inference system trained on the obtained results showing how changes of states are affecting response from users.

5 Fuzzy Inference System

In this part, we applied the inference model based on fuzzy sets generated from input data and changes of states. During the research, we used neuro-fuzzy reasoning system based on model defined by E. Mamdani et al. [14] and extended by J. Jang towards neuro-fuzzy system [7]. This model is presented in the literature and discussed among others by A. Piegat [15]. Each rule determines one fuzzy point in the area of Cartesian product X x Y and the rules take a form of $x \in A \Rightarrow y \in BS$. In the proposed solution we assumed the integration of inference subsystem with real system and built the inference model based on inputs from experimental environment. To explore these possibilities, we built a model based on aggregated data, the structure and ANFIS fuzzy models were designed for the conversion analysis CR and customer satisfaction rate SR. For each input three fuzzy sets with Gaussian membership function were applied. For each item we identified fuzzy representations for the differences in influence levels. The shape of membership function was defined to distinguish three sets for each input with *Low*, *Neutral* and *High* influence on a user after comparison with previous state. Influence change with value -4 representing drop from level 5 to level 1 belongs to set *Low* with a probability of 1. Influence with change -2 belongs to both *Low* and *Neutral* sets with a probability of 0.5 while in a situation when a user receives an object with no changes in states where $c_i=0$ it is the central part of *Neutral* set with membership function mf=1 and mf=0.75 belonging to *Neutral* set for changes at levels of -1 and 1. -1 and 1 values belong to sets *Low* and *High* respectively with membership functions at .25 levels. After defining the input sets, we performed model training with hybrid optimization method which is a combination of least-squares and back propagation gradient descent method. Using inference model, we can obtain knowledge about the relation between inputs and output and the fuzzy approach helps to cover decision space using limited empirical data. In Fig. 2, we can observe the relation between the selected parameters setting and conversions.

Results show that if the difference between *t* and *t-1*drops to a minimal value in range of -4 for both c_2 and c_1 we obtain a drop in conversion rates to the level of 1%.These changes where represented by changing the thumbnail image background from animated to light blue and text version from persuasive, were less influential on users. If together with the background, changes were performed in the text we obtained an increase in the conversion from 1% up to 1.7%and this is represented by maximal changes for c_1 and c_2 from 1 to 5. The other relation is shown on Fig.3, where we changed the intensity of rating info represented by c_4 value and thumbnail background

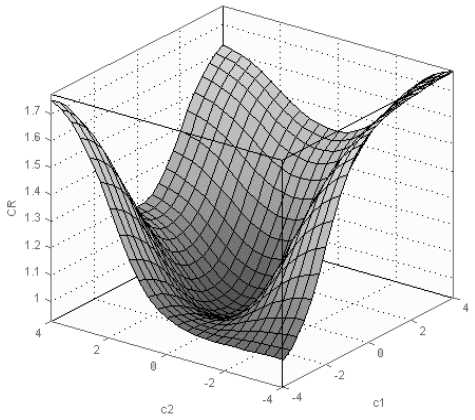


Fig. 2. Response from the inference system with changes of states for c1 and c2 inputs

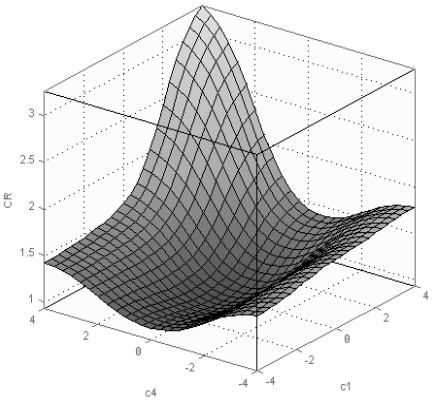


Fig. 3. Response from the inference system with changes of states for c4 and c1 inputs

changes represented by c_1. Response from inference model shows that if we change states from -1 to 4 for both parameters separately, conversion rate growth stabilizes at the region of 1.5% however increasing intensity of both elements from 1 to 5 results in conversions growth up to 2.5%. Changes of c_3 parameter shownin Fig. 4 have no big influence on results and even with neutral changes CR factor achieves level of 1.5% while keeping the neutral level of $c_4$leads to a drop of conversions when we compare it with big changes where c_4 is represented by high values in the range of 4.

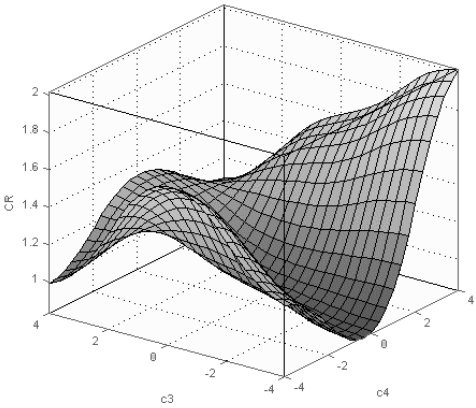


Fig. 4. Response from the inference system with changes of states for c_3 and c_4 inputs

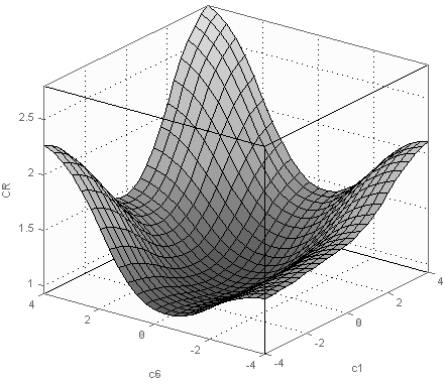


Fig. 5. Response from the inference system with changes of states for c_6 and c_1 inputs

In Fig. 5, the output from inference system shows that changes registered for c_6 and c_1 separately are not resulting in high growth of CR, even with a change at the level of 4 for c_1 is represented by growth from 1.5% to 2% but can result in a drop in

user experience. Response from the model illustrates that increase of response can be obtained by increasing influence of factor e_1 and e_6 together without changing other factors. This variant resulted in the increase of CR up to 2.5% and can be used as variant with accepted level of interactions and limited negative impact on users because of low levels associated to other parameters. The results show that the proposed approach identified the relation between output and changes of states. It is important to identify this relations especially in environments where it is difficult to motivate web users to perform the desired actions and especially when the probability of actions drops after first the impression. Construction of fuzzy model, which reflects the effects of changes depending on the levels of impacts in various stages of communication in an interactive system, makes it possible to determine the levels of influence and the role of individual elements with their impact on the number of interactions. The results indicate the existence of links between changes in levels of impact and the obtained response. The use of fuzzy models can reflect the phenomena occurring in the interactive environment and can contribute to a better understanding of existing relationships. The presented solution can be used to design components of the advertising or websites. Further work maybe aimed at studying the impact of the message sequence and not only the state changes between the change-over. Conversion to fuzzy sets as inputs for inference system can be scalable and we can perform data clustering and for values of each change, we can use fuzzy sets and the dimensionality of the problem will be reduced and less resources will be required.

6 Summary

Website design requires identification characteristics of the audience and the selection of impact in such a way as to ensure a certain level of conversions. If taken into account the communication process and sequence of interactions with content repetitions, there are complex relationships between components of the site, which affect the obtained results. Components selection, adjusting parameters and testing requires the use of analytical methods that enables us to design solutions to fit an existing condition. The resulting inference system allows us to determine the levels of the system response for the specified components represented by fuzzy numbers and takes into account the imprecise inputs data. The presented solution shows an alternative way of interactive media design, which provides the ability to test the levels of interaction with the use of fuzzy inference methods and taking into account the effects of repetitions addressed to the area of website conversion maximization. As demonstrated by the results obtained in this paper, the relationships between the components of the site may affect the results. Changes of states delivering conversions at an acceptable level using optimized variants of the design can be generated in an automated manner towards real time website optimization.

References

1. Baccot, D., Choudary, O., Grigoras, R., Charvillat, V.: On the impact of sequence and time in rich media advertising. In: Proceedings of the 17th ACM International Conference on Multimedia (MM 2009), pp. 849–852. ACM, New York (2009)

2. Barnes, S.J., Vidgen, R.: An Integrative Approach to the Assessment of E-Commerce Quality. Journal of Electronic Commerce Research 3(3), 114–127 (2002)
3. Chatterjee, P., Hoffman, D.L., Novak, T.P.: Modeling the Clickstream: Implications for Web-Based Advertising Efforts. Marketing Science 22(4), 520–541 (2003)
4. Hoffman, D.L., Novak, T.P.: Marketing in Hypermedia Computer-Mediated Environments: Conceptual Foundations. Journal of Marketing 60, 50–68 (1996)
5. Jang, J.S.R.: ANFIS: Adaptive-Network-Based Fuzzy Inference Systems. IEEE Transactions on Systems, Man and Cybernetics 23(3), 665–685 (1993)
6. Jankowski, J.: Integration of Collective Knowledge in Fuzzy Models Supporting Web Design Process. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part II. LNCS (LNAI), vol. 6923, pp. 395–404. Springer, Heidelberg (2011)
7. Jankowski, J.: Analysis of Multiplayer Platform Users Activity Based on the Virtual and Real Time Dimension. In: Datta, A., Shulman, S., Zheng, B., Lin, S.-D., Sun, A., Lim, E.-P. (eds.) SocInfo 2011. LNCS, vol. 6984, pp. 312–315. Springer, Heidelberg (2011)
8. Kazienko, P.: Multi-agent System for Web Advertising. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3682, pp. 507–513. Springer, Heidelberg (2005)
9. Klir, G.J., Yuan, B.: Fuzzy Sets and Fuzzy Logic. Prentice Hall, New York (1995)
10. Kolomvatsos, K., Hadjiefthymiades, S.: Buyer behavior adaptation based on a fuzzy logic controller and prediction techniques. Fuzzy Sets and Systems 189(10), 30–52 (2011)
11. Mamdani, E.H., Folger, T.A., Gaines, R.R.: Fuzzy reasoning and its applications. Academic Press, London (1981)
12. Piegat, A.: Fuzzy Modeling and Control. Physica Verlag, Heildelberg (2001)
13. Rafaeli, S.: Interactivity: From new media to communication. Sage Annual Review of Communication Research: Advancing Communication Science 16, 110–134 (1988)
14. Briggs, R., Hollis, N.: Advertising on the Web: Is There Response before Click- Through? Journal of Advertising Research 37(2), 33–45 (1997)
15. Rohrer, C., Boyd, J.: The rise of intrusive online advertising and the response of user experience research at Yahoo! In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1085–1086. ACM Press, New York (2004)
16. Schlosser, A.: Converting Web Site Visitors into Buyers: How Web Site Investment Increases Consumer Trusting Beliefs and Online Purchase Intentions. Journal of Marketing 70(2), 33–148 (2006)
17. Tarafdar, M., Zhang, J.: Analysis of Critical Website Characteristics: A Cross-Category Study of Successful Websites. The Journal of Computer Information Systems 46(2), 14–24 (2005)
18. Udo, G.J., Marquis, G.P.: Factors affecting e-commerce Web site effectiveness. The Journal of Computer Information Systems 42(2), 10–16 (2001)
19. Zadeh, L.A.: Fuzzy Sets. Information and Control 8(3), 33–53 (1965)

Virtual Collaboration in the Supply Chains – T-Scale Platform Case Study

Marcin Hajdul

Institute of Logistics and Warehousing, Estkowskiego 6, 61-755 Poznań, Poland
marcin.hajdul@ilim.poznan.pl

Abstract. Current model of organization of supply chains results in inefficient use of transport resources, high transport costs, increasing congestions and CO_2 emission. This effect has been demonstrated by research conducted by the author as well as by the European Environmental Agency. This situation can be change by development of alternative business model for collaboration in organisation of the transport processes within the supply chains. The aim of this paper is to present practical implementation of the developed by the author T-Scale platform that enables collaboration between independent transport users and transport service providers. Moreover, an overview of existing communication platform with its major functionalities are presented. The work is summarized by the major benefits of collaboration achieved by the group of companies operating in the FMCG sector in Poland.

Keywords: virtual collaboration, sharing supply chains, communication platforms, load factor, empty runs.

1 Introduction

The European economy has been experiencing some radical changes in the last few years. The analysis of the data of the European Statistical Office shows a 5% increase in the sales and turnover in wholesale and retail trade in European Union states. The effects of the global economic recession appeared in 2009, causing the slowdown of the economic progress. Still, companies have remained active and have been adjusting their strategies to the changing market conditions [9]. Mergers of companies take place, new process management concepts are introduced. At the same time, competition gets stiffer and consumers' expectations grow. It should be also noted that regardless of the economic growth rate, the transportation of goods by road increased in the last four years. As an example, on the basis of the latest data made available by the Main Statistical Office (GUS), the share of road transportation in goods shipping in Poland was 84.4% in tons, and 70.4% in ton-kilometers, [6].

These changes forced companies who not only wish to survive, but also to develop and bring the expected profits, to introduce changes to their operation. Hence, it was necessary to search sets of activities, most often completed in sequences, that would allow to make a product or provide a service whose value is specified and acceptable to the customer.

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 449–457, 2013.

Therefore, the paper's objective is to present a implementation of a model for collaboration in the supply chains that enable increase of efficiency and effectiveness of executed processes. The implementation has been carrying out within members of ECR Poland.

ECR Poland, member of ECR Europe - a non-profit association focused on optimising value chains in order to deliver better value for consumers/shoppers. ECR Mission is working together to fulfil consumer/shopper needs – better, faster and at less cost in a sustainable way. ECR Poland gathers large, medium and small companies representing:

- retailers and wholesalers,
- manufacturers (mostly supplying all Europe)
- service providers (including logistics and IT services).

2 Utilisation of Available Transport Resources in Europe

The effects of the currently applied approach to transport organization within existing supply chains lead to heavier traffic, reduced travel safety and increased emission of harmful substances. The growing congestion lowers the average technical speed of vehicles, ultimately increasing delivery time and possibly impacting customer dissatisfaction, which may even cause a part of orders to be cancelled. Hence, in the long run the companies unwittingly work towards worse financial results and reduced competitiveness.

The above situation is confirmed by the research of the European Environment Agency. The research shows that the utilization of the available load capacity of transportation means is poor across UE states. In case of the most popular type of transportation, namely road transportation, the average utilization of the available load capacity of trucks for delivery or distribution purposes is at 54% [10]. Naturally, the situation varies among specific countries.

This results were confirmed by research conducted by the European Statistical Office and Professor Alan McKinnon of the Heriot-Watt University. According to their analyses the EU average percentage share of empty runs, as a total number of covered kilometers, for road transportation is at 25% [5]. Unfortunately, it often happens that truck owners cannot find return loads and their truck come back empty or only carrying minor loads.

The analysis of presented results leads to a conclusion that transportation resources are used uneconomically, simply speaking are wasted [4]. These activities not only apply to improper use of the available resources, but also confirm that possibilities of completing given actions with reduced outlays are either omitted or intentionally ignored [10].

Taking into consideration presented above information, together with 30 production companies from ECR Poland a detailed measurement was carried out. The aim of this action was to identified what was the load factor (utilisation of the truck space) while cooperation with small and medium transport companies (partial and FTL transports described in previous chapter). The utilization of the available load capacity of trucks for delivery or distribution purposes was at 57% [9].

3 Pros and Cons of Existing Web-Based Solutions Supporting Organization Transport Processes

Numerous possible cooperation modes within supply chain provoke interest in easy and quick exchange of electronic data among potential transport providers and transport users [8].

The disadvantages of the current approach to the organization of processes within supply chains, as described in the previous chapter, may be eliminated through implementing of new model for collaboration of independent companies, either associated in clusters or functioning in close proximity. The collaboration should apply to common organization of transport processes within supply chains and their proper coordination in order to achieve the effect of synergy [4, 9]. However, collaboration requires secure, reliable and dynamic data exchange.

Dynamic development of the Internet caused development of web-based solutions for communication in the supply chains. However, based on the information from the industry most of them support classical approach to organisation of transports within supply chains. Detailed research of eleven newly developed communications were carried out. Based on the first analysis, seventeen major functionalities were selected that are essential both for model with and without collaboration of independent transport users and transport service providers. Results of this analysis is presented in table 1. All verified platforms are web-based platforms, provides on-line notification about all events within the supply chains, reporting and management of transport orders. Some of the verified platforms are typical transport freight exchanges, whereas others offer more possibilities to its users.

Table 1. Summary of the platforms supporting analysed functionalities

No.	Functionality	Number of products supporting analysed functionality	% share of products supporting analysed functionality
1.	Web-based communication platform	11	100,0%
2.	On-line notification	11	100,0%
3.	Reporting	11	100,0%
4.	Transport orders management at company level	11	100,0%
5.	Monitoring of the performed task	10	90,9%
6.	Digital map	10	90,9%
7.	Route optimisation	10	90,9%
8.	Invoicing	10	90,9%
9.	Fleet management	8	72,7%
10.	Freight exchange	6	54,5%

Table 1. (*continued*)

11.	Real-time monitoring of the performed tasks	5	45,5%
12.	Transport orders management at group of independent companies level	4	36,4%
13.	Verification of business partners	4	36,4%
14.	Coordination of transport orders and transport resources from independent companies	2	18,2%
15.	Optimisation of truck loading process	1	9,1%
16.	Share of savings among group of cooperating companies	1	9,1%
17.	Support of existing communication standards	1	9,1%

What is important only two of eleven platforms supports collaboration in joint organisation of transport processes between independent transport users and transport service providers. These platforms are T-Scale and TRI-VISOR. As T-Scale was created in order to support collaboration it covers sixteen out of seventeen functionalities which were verified during this analysis. T-Scale only does not support truck loading process. Table two presents summary of the products and their coverage of analysed functionalities.

Table 2. Summary of the functionalities supported by the analyzed communication platforms

No.	Product	Producer	Number of functionalities supported by analysed products	% share of functionalities supported by analysed products
1.	T-Scale	ILiM	16	94,1%
2.	TRI - VIZOR	WaterFront Research Park	12	70,6%
3.	TRANSPOREON	TRANSPOREON GmbH	12	70,6%
4.	LOG INTEGRA	Vesper Software	11	64,7%
5.	TIMOCOM	TimoCom Soft- und Hardware GmbH	11	64,7%
6.	Wtransnet	Wtransnet	11	64,7%
7.	InterLan	InterLan	10	58,8%
8.	Sky Logic	Benson Consultans	10	58,8%
9.	ORTEC	ORTEC	10	58,8%
10.	RAMCO	RAMCO	9	52,9%
11.	Pooling France	DIAGMA, ECR France, IPS Europe	5	29,4%

4 Virtual Collaboration in Practice – T-Scale Platform Case Study

The idea of developing new bussines model for joint transport processes organisation within member companies of ECR Poland began in 2010. The whole 2010 was spent on the development of theoretical model and discussion among production companies about way of possible cooperation. Based on that steps which allows implementation of the solution were carried out:

- cost and value analyses for transport users, as well as service providers,
- development of practical web-accessible tool (T-Scale) enabling automated information exchange between involved parties within the whole supply chain in order to start vertical cooperation between companies to reduce transport cost,
- guidelines on information sharing based on the unified communication standards,
- some possible pre-defined scenarios, based on:
    - product categories,
    - current distribution network set-up,
    - geographies,
    - scale economies,
- guidelines on ordering processes optimisation within the supply chains,
- other changes in transport processes organisation to present operations that are envisaged,
- key performance indicators.

Developed T-Scale platform plays the crucial part in the virtual collaboration in transport organisation within the supply chains. T-Scale allows real time exchange of information among companies participating in the realization of transportation processes. It enables to form temporary cooperation network (virtual supply chains). There are four key roles applied:

- The transport users define transportation needs.
- The transport service providers offer their services.
- The planning of deliveries and generating of consolidated transportation orders are made by the transportation coordinator, who also acts as an intermediary between group of independent producers and carriers.
- 4th party role responsible for auditing of all companies, verifying if the agreed conditions for cooperation are obeyed and providing technical solutions. The Institute of Logistics and Warehousing acts as technical and content-wise coordinator. The Institute oversees the technical aspect of operation of the platform. Moreover, ILiM carries out monthly impartial audits of effectiveness of planning of transportation and ensures stability and safety of the solution.

The principal advantage of the discussed solution (T-Scale platform) is the complete coordination of cooperation among different companies involved in the common transport planning and scheduling process in order to use the available transportation resources in a balanced manner. Furthermore, T-Scale is based on agreed global communication standards.

The following transport communication standards (GS1 standards for transport and logistics) were selected and agreed to be used in the new bussines model for joint transport processes organisation:

- Global Location Number (GLN) - is the GS1 ID Key used to identify locations and legal entities. Using a GLN rather than a proprietary internal numbering system for locations gives a company significant advantages, because it provides a standardised way to uniquely identify entities and locations throughout the supply chain [3].
- Serial Shipping Container Code (SSCC) - the GS1 ID Key used to identify individual logistic units. A logistic unit is defined in T-Scale as combination of units put together on a truck/container, where the specific unit load needs to be managed through the supply chain [3].
- format for naming point of origin and destination,
- type of products groups and its transport susceptibility,
- type of transport units and its equipment,
- transport request,
- transport order,
- transport service description [2],
- format of the transport pricelists,
- common process for placeing of the transport requests and orders,
- common process for sharing of the savings in the transport costs.

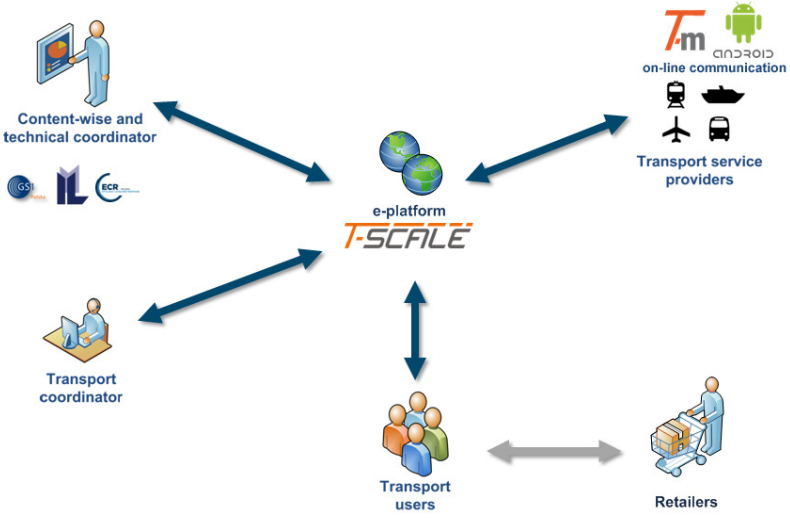


Fig. 1. Roles on the T-Scale platform

The T-Scale platform improves communication between virtual supply chain participants for purposes of joint organization of deliveries, which translates to a number of benefits from the cooperation between companies, such as:

- optimization of transportation costs due to the achieved scale effect,
- improved availability of cargo space,
- better utilization of the load capacity of trucks,
- elimination of "empty runs",
- reduction of road traffic intensity.

On the selected seventeen routes which belong to ten producers, T-Scale in June 2012, was able to significantly improve effectiveness and efficiency of the transport processes. At the beginning of the pilot implementation the following options for collaboration were defined:

- First option: Cooperation within FTL transports in order to find partner which allows two or more producers to close the whole route (e.g. from point A to point B and from point B to point A). Therefore, total transport rate is going to be calculated on the basis of the number of run kilometres. It means production company is paying for transporting the route from point A to point B and from B to A.
- Second option: Cooperation within joint organisation of partial transports (above 10 pallets) and LTL in order to increase utilisation of truck and reduction of transport unit costs.

Table 3. Results of the T-Scale operations in June 2012

Parameter	Without T-Scale and communication standards	With T-Scale and communication standards
Pallets carried out [pcs]	18202	18202
Total number of transport routes [pcs]	821	649
Total number of kilometres [km]	198682	157058
Total transport costs [euro]	168880	119757
Total savings in transport costs [euro]	-	49123
Average savings in transport costs per company [%]	-	15%
Average share of empty runs in total number of kilometres [%]	data not available	7,7%

To sum up, sharing of resources and cooperation in transport organisation according to agreed communication standards is of multi-dimensional nature. It positively impacts both companies that use transport services and the ones that provide such services. However, it is still a challenge to change companies attitude and approach with respect to the business processes organisation and being more open for cooperation in the field of logistics. Moreover, presented case concerns only road transport cooperation, there is a great challenge to implement similar business model for intermodal transport, where more actors are involved in the process.

5 Conclusions

The intensive development of Business Intelligence and Competitive Intelligence tools, access to information from multi-dimensional data analysis [1] aggregated from various enterprise IT systems sources (usually in case of heterogeneous environments) has been significantly facilitated in recent years. However, vast majority of available tools supports only classical approach to organisation of transport processes within supply chains. To improve effectiveness and efficiency of transport processes a new approach is a must is near future [7]. Companies needs to collaborate based on the agreed data standards within secure and reliable virtual supply chains.

Virtual collaboration allows sharing of resources and joint cooperation of transports which is of multi-dimensional nature. It positively impacts both companies that use transport services and the ones that provide such services. Furthermore, these companies are closely connected to the environment in which they operate. In many cases the main objectives of companies and the society are not identical. The proposed solution makes it possible to organize logistics process while taking into account economic, social and environmental aspects.

Additionally, the positive reception of the solution by the leading manufacturers and distributors in Poland allows to hope that the solution will soon be accepted and employed in business activity.

This hope is also fortified with the growing awareness the companies have of their impact on the environment. It can now be observed that companies exhibiting advanced social awareness often shape their activities not only with their own strategies in mind, but also taking into account the objectives and values of the society. Corporate social responsibility is a method of creating generally understood benefits, both for companies, as profits, and for its environment. Hence, in can said that a company following the principle of sustainable development can achieve a balance between its profitability and effectiveness, and social interests.

References

[1] Sołtysik-Piorunkiewicz, A.: Controlling in Organisation and Management. Computerisation concept. Humanitas Publishing House, Sosnowiec (2009)
[2] Pedersen, T.J., Paganelli, P., Knoors, F.: One Common Framework for Information and Communication Systems in Transport and Logistics. DiSCwise project deliverable, Brussels (2010)
[3] GS1 (2010) GS1 standards in transport and logistics, GS1 Global Office, Brussels (2010)
[4] Hajdul, M.: Model of coordination of transport processes according to the concept of sustainable development. LogForum 3(21), 45–55 (2010)
[5] McKinnon, A.: European Freight Transport Statistics: Limitations, Misinterpretations and Aspirations. Report Prepared for the 15th ACEA Scientific Advisory Group Meeting. Heriot-Watt University, Edinburgh (2010)
[6] Golinska, P., Hajdul, M.: Multi-agent Coordination Mechanism of Virtual Supply Chain. In: O'Shea, J., Nguyen, N.T., Crockett, K., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2011. LNCS, vol. 6682, pp. 620–629. Springer, Heidelberg (2011)

[7] Golinska, P., Hajdul, M.: European Union Policy for sustainable transport system – challenges and limitations. In: Golinska, P., Hajdul, M. (eds.) Sustainable Transport, pp. 3–20. Springer, Heidelberg (2012)

[8] Sliwczynski, B., Hajdul, M., Golinska, P.: Standards for Transport Data Exchange in the Supply Chain – Pilot Studies. In: Jezic, G., Kusek, M., Nguyen, N.-T., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2012. LNCS (LNAI), vol. 7327, pp. 586–594. Springer, Heidelberg (2012)

[9] Golinska, P., Hajdul, M.: Virtual Logistics Clusters – IT Support for Integration. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ACIIDS 2012, Part I. LNCS, vol. 7196, pp. 449–458. Springer, Heidelberg (2012)

[10] European Enviromental Agency Road freight load factors (during the laden trips) (2012), http://www.eea.europa.eu/data-and-maps/figures/road-freight-load-factors-during (access: September 14, 2012)

Cooperation between Logistics Service Providers Based on Cloud Computing

Arkadiusz Kawa and Milena Ratajczak-Mrozek

Poznań University of Economics,
al. Niepodległości 10, 61-875 Poznań, Poland
{arkadiusz.kawa,milena.ratajczak}@ue.poznan.pl

Abstract. The paper describes the use of cloud computing in logistics, especially the creation of the multi-modal platform designed for cooperating logistics service providers and their customers. The research conducted within the EU project is presented. The article focuses primarily on the findings of its initial phase – the analysis of information requirements needed for cloud computing platform. The processes maps and use case are proposed.

Keywords: cloud computing, companies cooperation, logistics service providers.

1 Logistics Service Providers

A characteristic trait of the modern enterprise is the growing importance of its relationships, interactions and interdependencies with other entities from its environment. Finding themselves under increasing competitive pressure, companies are ceasing to treat these relationships and cooperation with partners as a solution that can help all parties achieve substantial benefits. At the same time they look for solutions facilitating this cooperation.

Logistics service providers are companies belonging to the so-called transport, forwarding and logistics industry. This sector covers activities of companies of different sizes, multiplicity of services and global range. It includes very large but also small firms, offering a range of services – from simple transport services, through service forwarding, warehousing, palletizing, packing, packaging, to full service of supply chains. Their range of activities may comprise a region (e.g. a province), country, continent or the whole world [6]. To ensure a fast and correct flow of information between individual entities of the operating system, a logistics service provider has to use appropriate information technologies [7]. Logistics service providers, in general, apply information technology in order to increase efficiency and automate their work. A particularly important goal, however, is to meet the expectations of potential and existing customers. Unfortunately, customers using the services of various logistics service providers always have to adapt to their tools. The logistics services industry lacks solutions that would integrate the services of different operators in one place.

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 458–467, 2013.

A more serious problem arises in the case of small and medium (SME) logistics service providers. These companies are generally more flexible but often lacking the needed resources for growth. Additionally, SME logistics companies have limited or no IT-competence and investments.

The next problem of the logistics industry is a lack of formal semantics which prevents automated data integration. There are not any universal solutions which let smaller companies work together in the changing conditions and access resources, software and information provided to computers and other devices on demand [7].

The solution to aforementioned problems may be cloud computing.

2 Cloud Computing Conception

According to the National Institute of Standards and Technology cloud computing is "a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (for example networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [8]. In other words, cloud computing is a pay-per-use consumption and delivery model that enables real-time delivery of configurable computing resources. Typically, these resources are delivered over the Internet to multiple companies, which pay only for what they use [2]. One may also say that cloud computing is a large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualized, dynamically-scalable, managed computing power, storage, platforms, and services are delivered on demand to external customers over the Internet [5].

There are many advantages resulting for companies from the use of cloud computing. Generally speaking, it allows enterprises to manage the building blocks of IT, provided by other people in the same way they would their in-house infrastructure, but without the challenges that such complex architecture would normally produce [10]. The most important advantages are cost reduction and more flexible use of resources and operations.

Cloud computing represents a fundamental shift in how organizations pay for and access IT services. It is a model for provisioning and consuming IT capabilities on a need and pay by use basis ("pay as you go" method) [10]. It eliminates the need to purchase licenses and pay for software installation and administration. The IT service is consumed according to the actual user demands. It enables externalization of IT costs in the logistics sector by a change of fixed costs to variable costs [1] and shifting the cost structure from capital expenditure to operating expenditure and also helps the IT systems to be more agile [10].

With cloud computing, there is a paradigm shift to an asset-free provision of technological resources [10]. Moreover thanks to these changes and among other things thru, effecting economy of scales and pooling together of knowledge and experience in diverse industries, cloud computing enables expanded products (service) sophistication, allows for more end-user simplicity, increases product (service) relevance and drives potential new businesses [2]. The enterprise can get

the advantage of developing and marketing the product or service earlier to the market, ahead of its competitors [10].

Companies are not only relying on cloud services to enhance internal efficiencies (cost reduction, more flexible use of resources), but also to target more strategic business capabilities. Research has confirmed that number-one objective for adopting cloud services is an external capability – that of increased collaboration with external partners [2].

One must underline that due to the aforementioned advantages cloud computing is a very good solution for small and medium sized companies. Thanks to the use of cloud computing they themselves do not have to invest in the entire infrastructure (equipment, facilities and staff). All they need to do is rent a server from a professional company.

3 Logistics Project Based on Cloud Computing

Logical project is implemented through the Central Europe Programme. The project is promoted by six infrastructure providers that are supported by economic development agents and logistics cluster associations. The partnership is completed by research institutes and universities with enormous capacities to deliver innovative ICT solutions for the logistics sector [11].

The objective of the project is "to enhance the interoperability of logistics businesses of different sizes, to improve the competitiveness of Central European logistics hubs through a decrease of transaction costs (better access to systems of global players), and to promote collective (sustainable) modes of transport (multi-modal co-operation)" [11].

The main tasks of the Logical project are to identify the need for cloud services including logistics companies, create a platform for cloud computing and its implementation in selected logistics centers in Central Europe and to develop patterns of cooperation of future users of the system.

The basis of the platform is easy and cheap access to information about the current demand for logistics services and possibilities of their implementation. Assumed key benefits for users, resulting from the Logical project, should be: more efficient flow of information by providing access to global information systems and institutional players (such as infrastructure providers) and a balanced and optimal use of transport [12].

The Logical project began in 2011 and will last until 2014. In the first stage of the project, a survey among small and medium sized enterprises based in Poland, Germany, the Czech Republic, Hungary, Slovenia and Italy from the logistics industry was conducted. During the study, respondents were asked questions concerning the scope of logistic services, the use of solutions from the IT industry, the main problems with the use of the systems they owned. They were asked to express an opinion on the solutions based on data processing systems in the cloud [12].

As a result of the survey interviews and their implicit instructive elements, about 59% of the interviewed logistics service providers stated that they were planning to make use of cloud computing in the future provided that suitable software tools are available.

When analyzing, the companies were asked about their expectations of the clouds. The most important expectations for cloud data of the surveyed companies were:

- improve / simplify communications with our business partners,
- improve / simplify communications with our customers,
- achieve higher quality in logistical services (greater reliability, better supplier loyalty),
- achieve greater transparency in handling data / better information flows,
- improve integration in the supply chain / transport links.

At the same time, it was noted that Polish companies were far less enthusiastic about the new information technologies than German companies, which saw a lot more benefits in them. This may be due to the fact that Polish companies from the SME sector have less experience in implementing IT solutions over the Internet [12].

In the next step of the project, the information requirements analysis was individually performed with prospective users of the system. This involved, inter alia, profiling the enterprises on the basis of questionnaires, identification of processes and an analysis of the logistics processes.

Below, the results of the research carried out in one of the companies (Trans Logistics[1]) are presented.

4 Exemplification of Cloud Computing Idea

The overall business aim of the Trans Logistics (Poland) is to acquire an established position in the market of logistics services in the SME segment. The key activity of the company is to provide services related to road freight forwarding (95% of activity, the remaining relates to handling and storage). Trans Logistics (Poland) intends to develop the newly opened office in Poland, which will provide comprehensive logistics services in the field of Trans Logistics orders from the headquarters in Denmark and gain new customers in the Polish market. All the tactical and operational decisions are taken by a board member from the Polish office of Trans Logistics and their contributors. However, every day, Trans Logistics (Poland) works closely with the Danish office which gives direction for the company.

The mega process of the Trans Logistics company is Forwarding. It is the coordination and organization of the transport process. An important part of this process is conceptual work, which includes exchange of information between the participants of the transport process as well as the commercial process and arranging all kinds of formalities, such as filling in the transport and commercial documents.

The mega process „forwarding" is divided into the following processes (see fig. 1): Acquisition of clients, Calculation of transport rate, Acquisition of carriers, Shipment realisation order, Shipment coordination, Settlement and turnover of documentation.

[1] The company name has been changed.

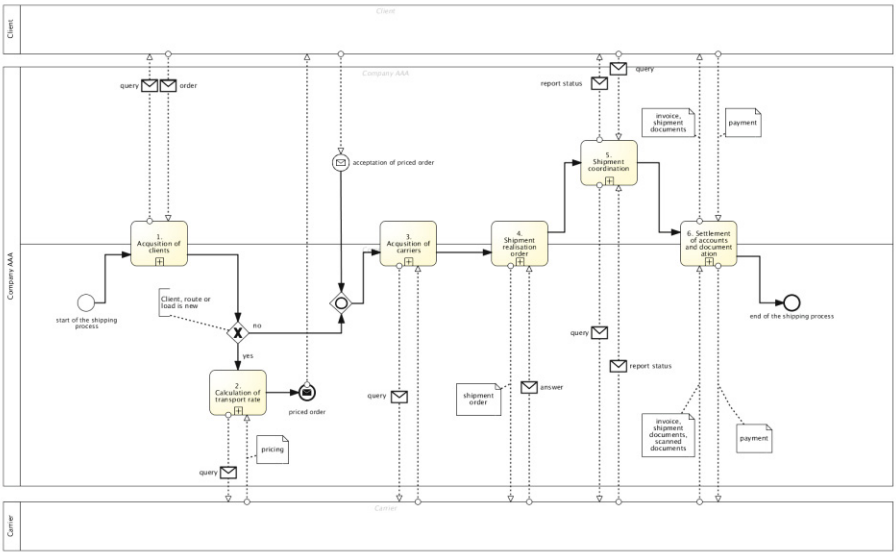


Fig. 1. Main map – forwarding process of the Trans Logistics company

After analyzing the business processes of the studied companies, a question arose which of the processes could be moved to the cloud. After a lot of careful examination, we came to the following conclusion: it is best to start with the processes that are performed most frequently and that are strongly linked with the environment. Examples are processes associated with obtaining the order and the carrier, the process of coordination of transport. Some of these processes can be placed entirely in the cloud, and some only in part.

With the exception of the process "Calculation of transport rate", all of listed processes have been transferred (entirely or partly) to the cloud.

In this paper the authors present only one example of the process. Below, the process of settlement of accounts and documentation is described. It is one of the processes, which is the most complex one in the Trans Logistics.

Settlement of Accounts and Documentation Process

Settlement of accounts and documentation starts when the carrier puts the documents into the cloud-based database. When the data is uploaded, the forwarding agent verifies the documents. If the documentation is incomplete or incorrect, the forwarding agent puts appropriate information into the system. The carrier is required to correct the data in the system.

Upon receipt of correct and complete documents, the groups of processes are executed in parallel. These are: (1) sharing the electronic version of the invoice with the client and (2) invoice preparation and invoicing.

Within invoice preparation and invoicing the forwarding agent validates the prepared invoice and invoicing is done by the financial module. In the case of discrepancies, correction is made manually. Once the invoice is prepared properly and invoiced, the forwarding agent orders the payment for the carrier. Simultaneously, the system sends an electronic copy of the invoice to the client and waits for notification about the payment. After notification, the payment is booked automatically.

The forwarding process finalizes the realization of both groups of the processes (see fig. 2).

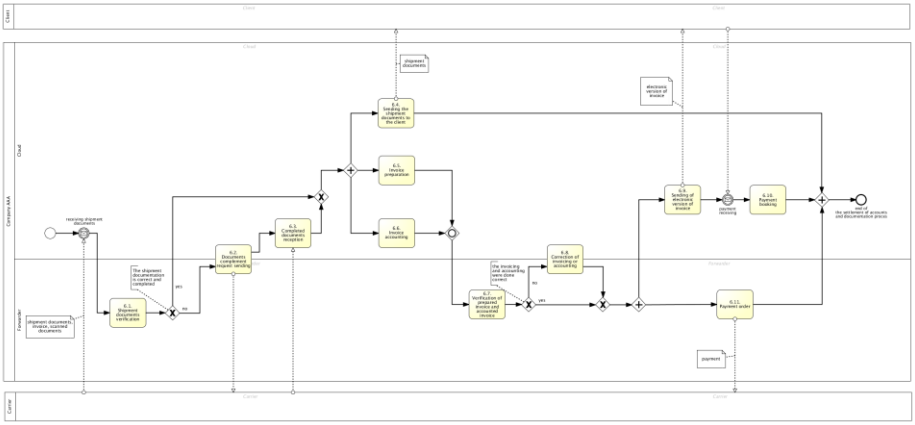


Fig. 2. Settlement and turnover documentation process

To present the process, use cases have been designed which are perfectly helpful to describe the system requirements. An use case shows the interaction between the actor (user system) initiating the event and the system itself. Furthermore, properly designed use cases allow the development of future system design, and an affordable and comprehensive platform for collaboration and communication system developers, investors and owners of the company.

An use case represents a basic course of operations, the so-called "basic flow" or "happy flow". In fig. 3 and 4 there are graphical use cases of the actors and the relationships between them.

In order to facilitate the presentation of the use cases, the process of Settlement of accounts and documentation is divided into two parts:

- Settlement of accounts with customer,
- Settlement of accounts with carrier.

The process of Settlement of accounts with customer consists of preparation, sending and accounting of invoices (see fig. 3).

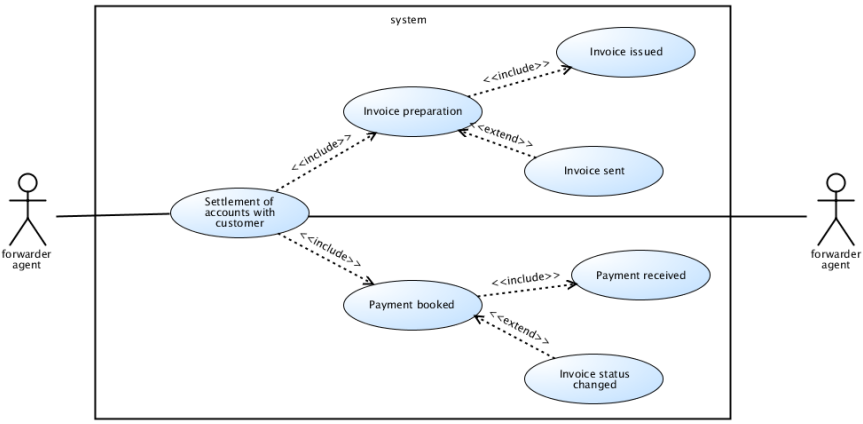


Fig. 3. Settlement of accounts with customer

In table 1, the main issues of this process are described (assumptions, preconditions, postconditions, and steps).

Table 1. Main issues of settlement of accounts with customer process

Assumptions	1. Trans Logistics and the customer use Electronic Data Interchange (EDI) 2. The client does not require paper version of invoices 3. Accounting system is able to put down the invoice automatically 4. The invoice can be generated from the data collected by the system in previous steps 5. Information from the bank's IT system can be sent to the system of Trans Logistics
Preconditions	Receiving completed and correct documentation from the carrier
Postconditions	Invoice accounted
Steps	1. Invoice preparation – system fills in the form with data collected in previous steps 2. Giving the invoice a status "unsettled" 3. Sending the invoice to the client electronically 4. Accounting the payment after receiving information from the bank's IT system 5. Changing the invoice status to "settled"

The process of Settlement of accounts with carrier is about receiving and accounting invoices and settling liabilities (see fig. 4).

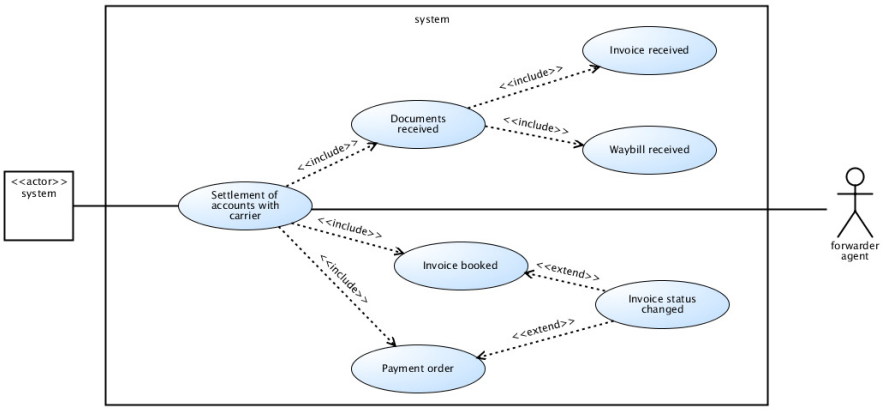


Fig. 4. Settlement of accounts with carrier

In table 2 the main issues of this process are described (assumptions, preconditions, postconditions, and steps).

Table 2. Main issues of settlement of accounts with carrier process

Assumptions	1. Trans Logistics and the carrier use Electronic Data Interchange (EDI) 2. Trans Logistics does not require paper version of invoices 3. Accounting system is able to put down the invoice automatically 4. The system of Trans Logistics is able to send information to the bank IT system, with payment order
Preconditions	Receiving the documentation from carrier
Postconditions	Settlement of liabilities with carrier
Steps	1. Receiving invoice from carrier electronically 2. Invoice accounting 3. Generating payment order to the bank's IT system

5 Conclusion

The continous development of new technologies, emergence of network interdependencies mean that companies no longer must invest in self-development of complex and sophisticated IT solutions. Instead they can use the ready-made solutions, that yet at the same time meet their requirements. Such an option for logistics service providers may be cloud computing and the platform, proposed within the Logical project.

The proposed platform should allow to: improve the process of acquisition and transmission of orders (including the process of settlement of accounts and documentation), reduce or completely eliminate the reporting process, reduce time processes by facilitating access to information, make it possible to easily evaluate the

processes – thanks to information from all stakeholders gathered in one place, and enable to create a virtual call to foster cooperation.

Of course during the implementation phase of the proposed platform many potential problems may occur. Many enterprises still remain very sceptical about the idea of cooperation and network relationships for fear of leaking information and knowledge – especially where these determine a company's competitive advantage [9]. And in the case of cloud computing, data is held "outside the company", which only intensifies aforementioned concerns. There is a lot of concern about the security and privacy of the data. Security is a great concern for most organizations. Many managers are not comfortable about their data located in a data centre in a foreign country. And true standards for how to control applications that are in a vendor's cloud have not yet been established. The current challenges must be addressed including developing acceptable compliance and security policies, reducing the risk by developing robust infrastructure for reliability and high availability along with performance guarantee [10]. Possible source of difficulties related to cooperation within the platform is the fear of losing independence and control over the company. Research has confirmed that companies are reluctant to engage in partnership relations if they fear becoming dependent on the other entity [3]. Moreover poor information flow between cooperating parties may damage relations, cause conflict and result in an "information deficiency" that threatens the well-being of projects. Research has confirmed that information processing requirements and information processing capability affect intention to adopt cloud computing [4]. Above-mentioned conclusions raise an important problem concerning the negative attitudes of companies and their managers. Many of them still view their companies as isolated units and are reluctant to engage in any form of cooperation, including cloud computing.

References

1. Arnold, U., Oberländer, J., Schwarzbach, B.: LOGICAL - Development of Cloud Computing Platforms and Tools for Logistics Hubs and Communities. In: Proceedings of the Federated Conference on Computer Science and Information Systems, Wrocław, pp. 1083–1090 (2012)
2. Berman, S.J., Kesterson-Townes, L., Marshall, A., Srivathsa, R.: How cloud computing enables process and business model innovation. Strategy & Leadership 40(4), 27–35 (2012)
3. Biong, H., Wathne, K., Parvatiyar, A.: Why do some Companies not Want to Engage in Partnering Relationships? In: Gemünden, H.G., Ritter, T., Walter, A. (eds.) Relationships and Networks in International Markets, red. Elsevier Science, Oxford (1997)
4. Cegielski, C.G., Jones-Farmer, L.A., Wu, Y., Hazen, B.T.: Adoption of cloud computing technologies in supply chains: An organizational information processing theory approach. The International Journal of Logistics Management 23(2), 184–211 (2012)
5. Foster, I., Zhao, Y., Raicu, I., Lu, S.: Cloud Computing and Grid Computing 360-Degree Compared. In: Grid Computing Environments Workshop, GCE 2008, Austin, Texas, USA, pp. 1–10 (2008)

6. Jeszka, A.M.: Problem redefinicji branży na przykładzie przesyłek ekspresowych. Gospodarka Materiałowa i Logistyka 7, 10–16 (2003)
7. Kawa, A.: SMART Logistics Chain. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ACIIDS 2012, Part I. LNCS (LNAI), vol. 7196, pp. 432–438. Springer, Heidelberg (2012)
8. NIST, Final Version of NIST Cloud Computing Definition Published (2012), http://www.nist.gov/itl/csd/cloud-102511.cfm
9. Ratajczak-Mrozek, M.: The premises for establishing business networks in the internationalisation process (research project 2010-2012). In: Fonfara, K. (ed.) The Development of Business Networks in the Company Internationalisation Process, pp. 71–81. Poznań University of Economics Press, Poznań (2012)
10. Subhankar, D.: From outsourcing to Cloud computing: evolution of IT services. Management Research Review 35(8), 664–675 (2012)
11. http://www.project-logical.eu/
12. http://www.zpds.com.pl/

Discovering Missing Links in Large-Scale Linked Data

Nam Hau[1], Ryutaro Ichise[2], and Bac Le[3]

[1] University of Technology, Ho Chi Minh City, Viet Nam
namhnt@uit.edu.vn
[2] National Institute of Informatics, Tokyo, Japan
ichise@nii.ac.jp
[3] National Science University, Ho Chi Minh City, Viet Nam
lhbac@fit.hcmus.edu.vn

Abstract. The explosion of linked data is creating sparse connection networks, primarily because more and more missing links among difference data sources are resulting from asynchronous and independent database development. DHR was proposed in other research to discover these links. However, DHR has limitations in a distributed environment. For example, while deploying on a distributed SPARQL server, the data transfer usually causes overhead on the network. Therefore, we propose a new method of detecting a missing link based on DHR. The method consists of two stages: finding the frequent graph and matching the similarity. In this paper, we enhance some features in the two stages to reduce the data flow before querying. We conduct an experiment using geographic data sources with a large number of triples to discover the missing links and compare the accuracy of our proposed matching method with DHR and the primitive mix similarity method. The experimental results show that our method can reduce a large amount of data flow on a network and increase the accuracy of discovering missing links.

Keywords: graph mining, linked data, link prediction, distributed RDF data.

1 Introduction

Linked data uses the Resource Description Framework (RDF) to represent a structure of data having different formats such as RDF-XML, N-Triples, and N3. A triple of the RDF is composed of a subject, a predicate, and an object. The predicate describes the relationship between a subject and an object. Using this mechanism, most entities in data sources can link to other entities in another data source. For example, when one is looking for information about a place, an artist, or a song, information can be obtained from a single data source to other data sources using predicates such as seeAlso, sameAs, and redirect. However, since all data sets are published independently and methods of finding the missing link between two entities from different data sources are still under development. Resolving this issue is essential for the future use of linked data.

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 468–477, 2013.

Many data points are created and all entities on each data points can be linked together with the RDF mechanism. Finding links among all entities in a group of data sources can be done by using similar link patterns. To find those patterns, the frequent graph mining method was applied in previous work [1]. The method proposed in this paper uses a comparable approach to discover the missing link between two entities in different data sources. To find a frequent graph using data from all data sources, we must query all triples that are linked together by a predicate called the link type. This process can incur overhead in a network because much data is loaded in batch. We propose a two-step method that can solve this problem. In the first step, we query triples on demand. In the second step, we just have to focus on a frequent pattern, which is useful while consuming linked data. After we obtain the required data, we apply Apriori-based graph mining to find all the needed frequent graphs. In other words, our method for discovering the missing link not only uses the frequent graph method but also the matching similarity method. We introduce a dataflow-efficient way to match data entities, because of the current context of linked data development.

In the next section we describe related works. In Section 3, the main section, we introduce our method in detail. In Section 4, we report experiments on the amount of data we can reduce and compare the accuracy of the proposed method EMSMatching with the accuracy of other methods. In the last section, we present a discussion about the proposed method, its contribution, and future work.

2 Related Work

Several studies have proposed methods for discovering links between different data sources in linked data. For music data, Raimond et al. developed several different methods based on looking up a literal to match a data entity [2]. When they tried to interlink an artist with a record and tracks in two distributed music data sets, Jamendo and Musicbrainz, their method encountered some drawbacks because of ambiguous data, which reduced the accuracy of linking entities in the two data sets. Silk Framework [3] is an efficient tool often used to generate RDF links among data entities based on user-defined link specifications. The specifications to create the link between two entities are expressed using the Silk Link Specification Language. However, the Silk Framework is only applicable for a pair of data sources and must have the configuration of the matching template, which involves very sophisticated work. In the Silk Framework, the more properties we use to match, the more data we need to query and the more time we need to calculate the similarity. Problems occur if data sources are enormous. Silk Server [4] is a tool to add missing links from an entity to all related datasets, when it has already been linked to one entity of a data set**.** Silk Server helps to generate many links to make more and more links to others. We can say that Silk Server is a tool to enrich information on web or linked data. The LInk discovery framework for MEtric Spaces (LIMES) [5] introduces a time-efficient approach to interlink data based on the Silk Framework using the mathematical features

of a metric. LIMES can filter a large number of instance pairs before mapping. As a result, the time used for matching the similarity is reduced. Yet on the first step of LIMES, it also needs to query all data on the target dataset to find a set of exemplars. For geographical data, DHR [1] is a novel method to discover missing relations by finding a closed frequent graph in many geography data sources. To calculate the similarity between two entities, it uses statistics for the appearance of words of useful attributes. However, it needs many data properties to calculate the similarity. In addition, we have to deploy all data in the local data set**.** This means that this method is not well suited to a distributed environment, where linked data is missing. Although most of the above methods focus on the accuracy of discovering missing links, we focus on the problem of data transfer on a network for discovering missing links.

3 Discovering Missing Links

3.1 Term Definition

First, we give the definitions of terms used in this paper.

Dataset *A,...,Z*: each dataset represents a linked data set, such as DBpedia, and GeoNames.

Entity $a_1,a_2,...,a_m$: entities in Dataset *A*. where *m* is the number of entities in *A*.

Linked-dataset graph (LDG): directed graphs for *Dataset*. A vertex represents a dataset and an edge represents the relationship between two datasets.

Linked-entity graph (LEG): directed graphs for *Entity*. A vertex represents entities and an edge represents the links between two entities.

Pair frequent matching graph (PFMG): pair of frequent LDGs consisting of two frequent graphs, *Parent Graph* and *Child Graph*. The *Child Graph* is constructed by eliminating one edge in the *Parent Graph*.

3.2 General Idea of Our Approach

Our method consists of two main stages, "finding the frequent LDG" and "matching data entities using the extend mix similarity metric" for discovering the missing links. In the first stage, Let us describe in detail by the example shown in Figure 1. In this example, three datasets (*A*, *B*, *C*) are given. When some entities are connected with the owl:sameAs relationship, we can create an LEG in the middle. If the LEG graph is frequent, then we can get the LDG from the frequent LEG. The LDG created from the LEG is shown in the upper left LDG. Then, we create PFMGs from the LDG. It consists of a *Parent Graph* (upper left), and a *Child Graph* (upper middle). If we find the LEG in the same pattern of the *Child Graph*, as with the LEG shown in lower right, then we can expect that c_k may have a link to an entity in *Dataset B*. After we create the candidates of missing links, we identify links with matching data entities using the extend mix similarity metric in the next stage.

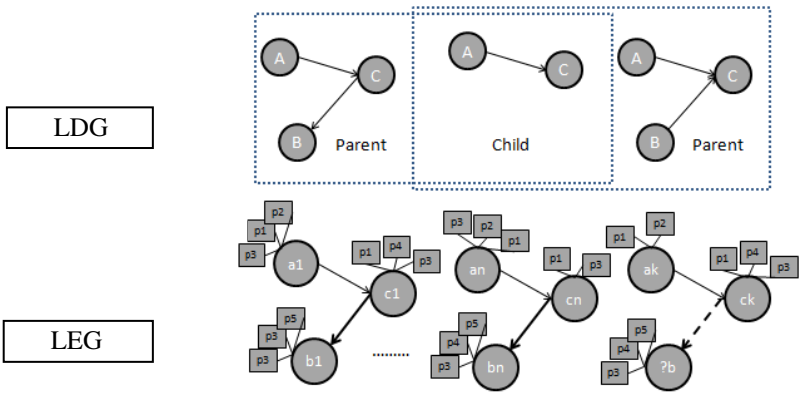


Fig. 1. Discovering missing links

3.3 Finding the Frequent LDG

Here, we go into details about the first stage of our method. First, we focus on finding the frequent LDG. In previous work [1], the process had to accomplish the following steps:

- Query all links related together from all data sources in which we are interested.
- Make a connected graph from these links, then generate a large amount of linked-entity graphs.
- Apply cSpan graph mining to find the frequent LDGs.

This method seems to be useful and easy to deploy if all datasets are stored in a local system. However, we cannot deploy in a real SPARQL server system, because it involves heavy network traffic. Instead of querying a batch of links to make graphs, we propose to get links on demand. Moreover, many frequent LDGs are obtained after the frequent pattern mining process, so we generate a large number of PFMGs. If we can omit some or choose one of them with many criteria, we can reduce many links to form frequent graphs before query. When we analyze some aspects of the linked data, not all graphs are useful for discovering missing links. In this case, we can omit them according to the following conditions:

First, we make assumption that the longer graph we have, the more confident missing link we can find out and the more information a user can obtain. So, long LEGs are needed to maintain the missing link.

Second, based on the user habit, most users normally focus on big data sources to start searching data or finding information, such as Dbpedia and DBLP. We call such data as an important data source (IDS). So, the most relation we are interested in is the one containing a link from an IDS.

Last, because of the feature of our method, the PFMG must be a closed frequent linked-dataset graph. For example, we assume that we have two frequent LDGs $G \to D, D \to U, D \to G$ and $G \to D, D \to U, D \to G, U \to G$ with the same number of

LEGs. In case, we can easily recognize that the pair is a PFMG, but the entity does not need to check because there is no possibility to find new links for the PFMG.

In this situation, if we try to apply the method in previous work with above conditions, we also have to query all links, although many of them are redundant. To solve this problem, we proposed to use the Apriori graph mining method. It not only helps us to optimize query data, but also this method has the following advantages:

- Base on Apriori theory, we count the number of links as a candidate on the item set before query the data. If the number is lower than threshold, we can remove the links. In addition to it, we can omit all links that are neither linked to any entities on important data nor linked to more than two datasources because this link is not useful for detecting the missing link.
- Generate parallel linked entity graphs while querying all needed links.
- Query data entities on demand.

Apriori graph mining requires choosing a support threshold for the frequency. However, it is difficult to choose a fixed threshold for data sources having a different number of links. When a huge dataset is connected to a small dataset, the following situation may occur: with a fixed threshold, a graph created from a huge data source tends to be very frequent because of the dense connections comparing with a small data source. Hence, important graphs may be dropped due to the fixed threshold. So depending on all linked data sources, the threshold for each graph is determined by the percentage of links in the smallest data source. Whenever a new item-set candidate is created, we need to identify the dynamic threshold again. This is an extension of the original Apriori graph mining algorithm in our system.

Now, we find all frequent LDGs at the same time using query data from the SPARQL server and satisfying the three conditions above by Apriori graph mining. Then we generate all PFMGs used in the next stage of discovering the missing link. For graph g_1 and g_2 in the PFMGs, g_1 is a *Parent Graph* of PFGM and g_2 is a *Child Graph*, which is a subgraph of g_1, which eliminates one edge from g_1.

3.4 Matching Data Entities Using the Extend Mix Similarity Metric

We perform the second stage of detecting the missing link. After obtaining all the PFMGs, we have to find out which link is missing for a given entity by a matching similarity. Before we go into the details of our matching method, we discuss our concept "extend mix similarity metric".

3.4.1 Extend Mix Similarity Metric

Although many kinds of similarity measures[7] are available, we propose a novel method to calculate the distance between two entities. The major idea of our method is the extension of the previously proposed mix similarity metric [6]. The difference between the extend mix similarity metric and the original one is to apply some similarity metrics for each attribute to calculate the distance between two entities with a

suitable data type. Given a data type, we can apply many similarity metrics because of the different characteristics, such as the string similarity with Jaro, JaroWinkler, Levenshtein, Jaccard, TFIDF, *n*-gram or the coordinate similarity with the Cosine and Euclidean distance. Not all metrics is good for all. They are depending to the feature of dataset[8][9]. Therefore, we propose the extend mix similarity metric method, which is a combination of each calculated similarity metric on a data type to achieve higher accuracy with the matching method:

$$d(i,j) = \frac{\sum_{k=1}^{m} \delta_k^{(i,j)} \frac{\sum_{l=1}^{n} \gamma_{k,l}^{(i,j)} d_{k,l}^{(i,j)}}{\sum_{l=1}^{n} \gamma_{k,l}^{(i,j)}}}{\sum_{k=1}^{m} \delta_k^{(i,j)}}$$

m: The number of an entity's attributes in a data source.
$\delta_k^{(i,j)}$: The weight of the distance against attribute k for two entities i, j.
n: The number of similarity metrics against an attribute of the two entities i, j.
$\gamma_{k,l}^{(i,j)}$: The weight of the distance at similarity metric l against attribute k for two entities i, j.
$d_{k,l}^{(i,j)}$: The distance between two entities i, j against an attribute. The distance can be calculated by the value of the attribute and the corresponding similarity metric l.

3.4.2 Matching Data Entities

Now, we go into details of how to apply similarity metric for discovering the missing link. Large data flows on a network are a problem in the matching process. So we propose many techniques to reduce data flow on a network.

The first one is the Top *k* entity. If we want to check whether entity $\boldsymbol{a_k} \in \boldsymbol{A}$ can be linked with entity in $\boldsymbol{B}$, we have to query $|\boldsymbol{B}| * \boldsymbol{m}$ values in the target data source $\boldsymbol{B}$ to find the best suitable entity ($\boldsymbol{m}$ refers the number of attributes needed for matching). In a small dataset, this is not a problem, but if a dataset is enormous, data transfer on the network becomes a serious issue. In this case, the solution for Top *k* is started by indexing data source $\boldsymbol{B}$ by a useful attribute, called "index property", which can help us to normally discriminate two objects. With the feature and the kind of data, we can choose suitable properties to serve as the index property. This choice also influences the accuracy of discovering the missing link. After that, we calculate the similarity indexed for each entity in data source $\boldsymbol{B}$, such as the label with the corresponding property in a_k, and the name. We just get the top $\boldsymbol{k}$ entities with the highest value, and then apply the matching method on the $\boldsymbol{m}$ attribute with the $\boldsymbol{k}$ entities. As a result, the total cost converges to $|\boldsymbol{B}|$ if $|\boldsymbol{B}| \gg \boldsymbol{k}$, and the data flow can be reduced to approximately the value $(\boldsymbol{m} - \boldsymbol{1}) * |\boldsymbol{B}|$. Moreover, given n is number of datasource will be processed in our system, we can refer that $\boldsymbol{n} - \boldsymbol{1}$ is maximum matching condition (MC) in each datasource with all remain ones. If we choose index property which differs from matching properties in each MCs, we would choose another suitable index property to calculate similarity for achieving Top *k* entities. So

Table 1. The number of triples queried on the SPARQL server for each data set

	Dbpedia	Geonames	Uscensus	Factbook	Total
DHR_cSpan	74313	88506	16372	0	179251
PFMG	74313	61399	7063	0	142775

the maximum index property we choose for one dataset is $\boldsymbol{n-1}$, and minimum index property is one. As a result, total property values $\boldsymbol{S}$ are queried:

$$|\boldsymbol{B}| \leq \boldsymbol{S} \leq (\boldsymbol{n-1})|\boldsymbol{B}| \text{ if } |\boldsymbol{B}| \gg \boldsymbol{k}$$

The second technique uses as few matching properties as possible. Assuming that the number of properties $\boldsymbol{m}$ and the top $\boldsymbol{k}$ entities used for matching are sufficiently big, it means possibly we can query almost data on a target datasource $\boldsymbol{T}$. The number of pair matching property belongs to the feature of data and the candidate of pairs can be determined by some metrics. With the solution of using the top $\boldsymbol{k}$ and choosing the least possible matching, the accuracy for finding the missing link can be reduced. So, we apply the extend mix similarity metric to improve the accuracy of our method. Now the system is ready to find the missing links. We implement this algorithm in our system, called EMSMatching. Moreover, in the matching process, we have to normalize the query data, because, for some reason, the accuracy is reduced when calculating the similarity.

4 Experiment

We conducted an experiment to test our idea. In this experiment, we only retrieve the LDG graph, which has more than two edges. We use a support threshold of 20% for our experiment. The data sets used for our experiment are Geonames[10], Dbpedia[11], Uscensus[12], and Factbook[13] with Dbpedia is important data source because of its popularity for all linked data.

First, we evaluate our method from the aspect of data reduction for querying. We compare our proposed method PFMG with DHR_cSpan used in [1] on the same condition for retrieving frequent graph patterns. The proposed method can reduce approximately 20% of the triples compared with the DHR_cSpan method with the same constrains against frequent LDGs (graph length more than two), as shown in Table 1. Because DBpedia is an important data source, the number of triples of this *link* does not change depending on the method. The Uscensus data contain 16372 entities linked to Geonames, but only 7603 entities can be connected by DBpedia. The number of links in the World Factbook is zero because it does not have links to other data. It is only linked by Dbpedia.

Next, we analyze the PFMGs that can be used for discovering the missing links. We obtain 19 PFMGs, as shown in Table 2. This table shows the relationship between the *Parent Graphs* and the *Child Graphs* and the number of graph instances for each. We consider the difference in the numbers as the candidates for missing links.

Table 2. PFMGs for Geonames, Dbpedia, Uscensus, and Factbook

Index	Data Point Prediction	Parent Graph	Number of Parent Graphs	Child Graph	Number of Child Graphs
Db1	**Dbpedia**	D-G U-G	4143	U-G	7063
Db2		D-U U-G	6921	U-G	7063
Db3		D-G G-D	58410	D-G	66881
Db4		U-G G-D D-G	4097	D-G U-G	4143
Db5		U-G G-D D-U	6825	D-U U-G	6921
Db6		U-G G-D D-U D-G	3962	U-G D-U D-G	4003
Us1	**UsCensus**	D-G U-G	4143	D-G	66881
Us2		D-U U-G	6921	D-U	7262
Us3		U-G G-D D-G	4097	D-G G-D	58410
Us4		U-G D-U D-G	4003	U-G D-G	4143
Us5		U-G G-D D-U D-G	3962	U-G G-D D-G	4097
Geo1	**GeoNames**	D-G G-D	58410	G-D	61399
Geo2		D-G D-F	83	D-F	168
Geo3		D-F G-D	142	D-F	168
Geo4		U-G D-U D-G	4003	D-U D-G	8011
Geo5		U-G D-U D-G	4003	D-U U-G	6921
Geo6		U-G G-D D-U D-G	3962	U-G G-D D-U	6825
Fact1	**Factbook**	D-G D-F	83	D-G	66881
Fact2		D-F G-D	142	G-D	61399

For example, the PFMG in Geo4 has 4003 *Parent Graphs* and 8011 *Child Graphs*. As a result, 4008 (=8011-4003) entities for *Child Graphs* may contain missing links. This implies that we can double the links in the data using our method. Finally, we conduct an experiment to evaluate our matching method. EMSMatching is our proposed method with full features. MSMatching is based on our method, but it uses the mix similarity metric [6] instead of the extend mix similarity metric and not applying many techniques for reducing data flow in matching. DHR_DE is the method used in [1]. In EMSMatching, we use the JaroWinkler (string-based similarity) and the Jaccard (token-based similarity) metric for matching the label of the name, title, and the Cosine and Euclidean metric for the geography coordinate. However, we only use the Jaro Winkler metric and the Cosine metric for MSMatching. The index property we choose is not different from matching properties in matching conditions. On matching method, δ weight for string matching is set 0.3 and coordinate matching equals 0.7 because our dataset related to geography. With γ weight, average value is set to all metrics applied for each data type.

The main focus is about transactions. We greatly reduce amount of data in our method. DHR_DE method use almost all data for their matching method because this method need to find all useful attributes to match entities, so we investigate the dataflow reduced for MSMatching and EMSMatching. In expreriment, for example, when conducting matching in Geonames, we obtain 6 PFMGs with approximately 60,000

	EMSMatching	MSMatching	DHL_DE
Geo1	99.3	78.8	98
Geo2	80	80	30
Geo3	42	40	80
Geo4	99.1	99	75
Geo5	**99.3**	**94**	**75**
Geo6	99.1	96	98
Fact1	**100**	**100**	**35**
Fact2	**100**	**100**	**78**
Db1	86	73	78
Db2	88.3	83.4	78
Db3	83.1	78	89
Db4	86.1	73.4	80
Db5	88	70.2	80
Db6	89	73.9	86
Us1	**100**	**99**	**75**
Us2	**99**	**97**	**75**
Us3	**100**	**99**	**78**
Us4	**98.6**	**94**	**75**
Us5	**98.7**	**94.5**	**75**
Average	***91.3***	***85.4***	***75.7***

Fig. 2. Accuracy for detecting missing links of each PFGMs

LEGs and the number of entities in Geonames are approximately 2,600,000. Because the Top Item $k = 30 \ll$ "$number\ of\ entities\ in\ Geonames$"and one index property can be used for all matching conditons to get Top k entities, 2,600,000 data values (value of each property) are queried for EMSMatching and 7,800,000 data values for MSMatching. In other words, the data flow for our method shows a great reduction of query data on the network (approximately 5,200,000 data values). For all the matching conditions between the two data sources, we use only two properties in our method (For coordinate matching on GeoNames, we need to query two properties instead of one as DBpedia).

In the last experiment, we evaluate our method from the aspect of accuracy of discovering missing links. To create correct answers for detection, we search an edge in the *Parent Graph* that does not exist in the *Child Graph*, and then we scan it in all the LEGs and finally remove the links for creating the test data. We compare three methods in this experiment. The calculation of accuracy is based on the exact entity found and all entities used for predicting.

Figure 2 illustrates the results of the experiment with the first column denotes the graph ID used in Table 2. The results indicate that our method EMSMatching obtains higher accuracy than do MSMatching and DHR_DE in most cases with the average of accuracy 91.3 %. We can say that the extend mix similarity metrics help us to

improve the accuracy in matching. From all above, we conclude that our method is effective for discovering missing links.

5 Conclusion

In this paper, we present a method to discover missing links on large-scale data-sources. The experimental results show that our method performs better than previous approach. However, more work is needed in the future, such as using an independent matching approach instead of the user configuring a matching template. One issue is that the accuracy is not high in some cases, but we cannot solve this problem completely because of the noise and ambiguous data. Finally, if the links in all datasets are too dense, the method to reduce the data triples in the first stage is not efficient.

References

1. Le, N.-T., Ichise, R., Le, H.-B.: Detecting Missing Relations in Geographic Data. In: Proceedings of the 4th International Conference on Advances in Semantic Processing, pp. 61–68 (2010)
2. Raimond, Y., Sutton, C., Sandler, M.: Automatic Interlinking of Music Dataset on Semantic Web. In: Proceedings of Linked Data on the Web (2008)
3. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Discovering and Maintaining Links on the Web of Data. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 650–665. Springer, Heidelberg (2009)
4. Isele, R., Jeentzech, A., Bizer, C.: Silk Server - Adding Missing Links While Consuming Linked Data. In: Proceedings of the 9th International Semantic Web Conference, pp. 650–665 (2010)
5. Ngomo, N., Auer, S.: LIMES — A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, pp. 2312–2317 (2011)
6. Nguyen, N.B., Ho, T.-B.: A Mixed Similarity Measure in Near-Linear Computational Complexity for Distance-Based Methods. In: Zighed, D.A., Komorowski, J., Żytkow, J.M. (eds.) PKDD 2000. LNCS (LNAI), vol. 1910, pp. 211–220. Springer, Heidelberg (2000)
7. Ichise, R.: An Analysis of Multiple Similarity Measures for Ontology Mapping Problem. International Journal of Semantic Computing 4(1), 103–122 (2010)
8. Cohen, W.W., Ravikumar, P.D., Fienberg, S.E.: A Comparison of String Distance Metrics for Name-Matching Tasks. In: Proceedings of IJCAI 2003 Workshop on Information Integration on the Web, pp. 73–78 (2003)
9. Christen, P.: A Comparison of Personal Name Matching: Techniques and Practical Issues. In: Proceedings of the 6th IEEE International Conference on Data Mining, pp. 290–294 (2006)
10. Wick, M.: The GeoNames geographical database, `http://www.geonames.org/`
11. DBpedia Team, The DBpedia database (2009), `http://wiki.dbpedia.org/`
12. Tauberer, J.: The U.S. census data, `http://www.rdfabout.com/`
13. CIA Factbook D2R Server, The World Factbook database, `http://www4.wiwiss.fu-berlin.de/factbook/`

Effective Hotspot Removal System Using Neural Network Predictor

Sangyoon Oh[1], Mun-Young Kang[1], and Sanggil Kang[2]

[1] Department of Computer Engineering, Ajou University, Korea
{syoh,hanamy}@ajou.ac.kr
[2] Department of Computer and Information Engineering, Inha University, Korea
sgkang@inha.ac.kr

Abstract. Monitoring and prediction of resource usage are two major methods to manage distributed computing environments such as cluster, grid computing, and most recent cloud computing. In this paper, we propose a novel hotspot removal system using a neural network predictor. The proposed system detects and removes hotspots with resource specific removal algorithm. The system also improves neural network predictor by introducing prediction confidence. The effectiveness of our proposed system is verified with empirical examples, and evaluation results show that our system outperforms a popular hotspot removal system in hotspot predication and hotspot removal.

Keywords: Neural Networks, Virtual Machine, Hotspot Removal, Prediction.

1 Introduction

Workload and resource management has been a popular research topic in distributed computing for a long time [1-2]. Currently, managing workload and resource in a large-scale infrastructure is becoming easier in terms of efficiency and manageability as virtualization technology becomes more mature.

Large public cloud is the distinct trend in last few years. However, private cloud will grow faster because it provides more secure and trustworthy computing infrastructure. However, good architecture design and effective resource management scheme are needed more than public cloud or cluster computing because handling workload fluctuation in private cloud is harder. Dynamic workload fluctuation, which is caused by incremental growth, time-of-day effects, and flash crowds [4-5], tends to generate a hotspot. Applications in datacenter are not able to operate above the performance level specified in service level agreement (SLA) [5]. In addition to this complexity of resource management, effective management of private cloud that is dedicated to a single organization is more challenging, since it is mostly built by leveraging existing IT infrastructure and personnel and easily suffers from resource constraints.

Conventionally, resource management task is involved with classic issues like resource discovery, acquisition, and location transparency. For example, Condor-G

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 478–488, 2013.

[3] is one of most popular software frameworks that support features listed above for grid resource management. However, with the help of virtual machines and hypervisors, we are able to focus on essential components of management to manage distributed resource dynamically; those are monitoring which is the act of collecting status information of resources of interest, load-balancing, handling fault events such as hotspots, and prediction of resource usage.

Among them, monitoring and prediction of resource usage are two major tools to manage resource dynamically and automatically. Monitoring represents the approach that remaps resources based on the current usage. Using information that is gathered through monitoring, we are able to 1) deploy workload as balanced among physical machine nodes, 2) remap workload for either higher utilization or stable use of resource, and 3) detect over-use of resource (i.e. hotspot).

On the other hand, prediction of resource usage [6] is another approach to manage resource. We achieve many of noted tasks by using prediction of usage and it is a powerful approach for workload provisioning because of its proactiveness. However, monitoring and prediction are mix-used in most cases because it is hardly possible to detect fault events such as hotspots without monitoring. While, prediction is powerful tool for workload provisioning.

2 Related Works

Sandpiper is a popular resource provisioning and management system in cloud computing that automates the task of monitoring and detecting hotspots, redistributing physical to virtual resources, and initiating any necessary migrations [4]. To predict future values, it employs time-series prediction techniques [7]. Specifically, Sandpiper relies on the auto-regressive family of predictors, where the *n*th order predictor *AR(n)* uses *n* prior observations in conjunction with other statistics of the time series to make a prediction. For example, consider a sequence of observations: $u_1, u_2, \ldots, u_k$. Given this time series, the 1st order predicator *AR(1)* makes a predication in *(k + 1)*th interval using the previous value u_k , the mean of the time series values μ and the parameter φ which captures the variations in the time series [7]. The prediction $\hat{u}_{k+1}$ is given by Equation (1):

$$\hat{u}_{k+1} = \mu + \varphi(u_k - \mu) \tag{1}$$

However, there is room to improve hotspot prediction in Sandpiper's mechanism; since the applied autoregression technique is a simple and linear statistic method (i.e. the sum of the mean and the variations of the time series). In this case, hotspot patterns are overwhelmed by normal patterns in time series. Thus, it does not provide reliable prediction result.

Another interesting example of prediction in resource management is a fault prediction mechanism based on artificial neural network (ANN) in a high performance computing domain. Charoenpornwattana et al. [8] provide their

experimental results that are conducted and analyzed online via the Intelligent Platform Management Interface (IPMI) [9] for faults prediction. Data is collected from the sensors that are installed on nodes in a cluster periodically (10 sec. interval in their experiments). The approach is interesting since hardware based ANN application for fault prediction is not popular. However, their research publication (i.e. Ref. [8]) does not provide detail formulas and mechanisms of ANN in detail. Thus it is hard to learn lessons from their experiments.

To improve Sandpiper's prediction method, we apply neural network predictor (NNP). In this paper, the predictor is trained with equal number of hotspot and normal patterns. We also consider variations of prediction errors as prediction confidence, to prevent false prediction confidence for the final prediction. The effectiveness of our approach is presented with empirical examples in Section 4.

3 Effective Hotspot Removal System Using Neural Network Predictor

In this section, we describe our architecture proposal of dynamic resource management system. The overview of the system architecture is depicted in **Figure 1**.

3.1 Motivation of Research and Design Overview

In our proposal, we address the workload fluctuation with automatic resource usage monitoring and virtual machine (VM) live migration algorithm based on resource usage information predicted by a NNP. Monitoring and prediction of resource usage are critical counter parts for each other in our design. To be effective on hotspot detection and VM migration, both monitoring and predication should collaborate together as well as monitoring module provides state information of PMs and VMs to predictor (i.e. neural network model with prediction confidence). We assume that the target system of our proposal is a cluster of virtualized server nodes in large scale and configurations of physical hardware are already known. On each PM node, applications are operated over VMs and the workloads of applications are all different.

To achieve our goal, we suggest three key features in our scheme and system design as follows:

Resource usage monitoring (Monitoring Module) – collecting state information of resources of all PM nodes and VMs through host OS and virtual machine monitor (VMM or Hypervisor). Collecting process is periodical with interval *p*. Information is provided as an input to Analysis Module for hotspot detection and prediction.

Analyzing monitoring information (Analysis Module) – receiving data from Monitoring Module. The module analyzes received monitoring data to detect hotspots and predict future occurrence of hotspots using neural network model. It detects hotspot occurrence by analyzing monitoring information and requests its removal to Migration Module. Also it runs the neural network model for hotspot occurrence

prediction and uses this prediction result to select a destination PM node for VM migration for hotspot removal

Hotspot removal (Migration Module) – executing hotspot removal operation. Migration Module determines the destination PM for migration based on the result of hotspot removal algorithm which is provided by Analysis Module. Migration Module live-migrates the overloaded VM to the destination PM.

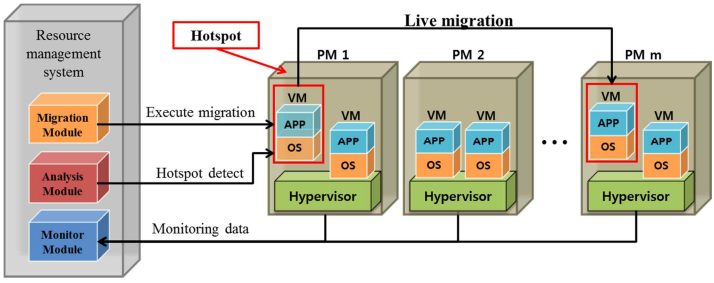


Fig. 1. Overview of Hotspot Removal System

3.2 Prediction of Resource Usage Based on Neural Networks Using Hotspot Patterns

Figure 2 is our one-step ahead hotspot predictor modeled by using a neural network (NN) based on resource usage in time series. To predict the occurrence of hotspot for each physical machine (PM), we first need to collect various hotspot patterns and normal usage patterns in time series which can be monitored on each PM. Then, we train an NN using delayed (or past) resources from current time on each PM. For convenience of training, we set '1' to the target resource usage for hotspot patterns and '0' for the normal patterns. For the i^{th} PM, denoted as PM_i in the figure, one-step ahead prediction at current time t, denoted as $\hat{R}_{i,t+1}$ can be formularized as seen in Equation (2).

$$\hat{R}_{i,t+1} = f(R_{i,t}, R_{i,t-1}, \ldots, R_{i,t-j}, \ldots, R_{i,t-d-1}) \tag{2}$$

where, the optimal f is a nonlinear function of trained NN, $R_{i,t-j}$ is the jth delay resource usage, and d is delay index. The optimal f is determined from the updating process of the weights in the NN to the direction in which mean square prediction error is minimized.

In general, reliability of the prediction depends on workload types on PMs that are monitored. For example, as seen in **Figure 3 (a)** in Section 4, monitored history variance is high while **Figure 3 (b)** in Section 4 has low variance. The workload types of **Figure 3 (a)** can be observed from many of Internet applications that have high fluctuation of workload and applications that have stable workload patters such as e-Science services shows workload type of **Figure 3 (b)**. If reliability of the prediction is varying, we have to expect that the quality of the prediction is bad and the hotspot prevention is hard as a consequence.

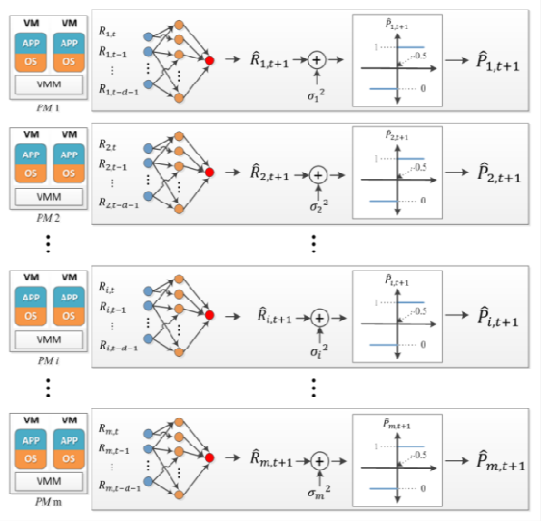


Fig. 2. Neural Network Predictors

To solve this problem, we consider prediction confidence for each PM. The prediction confidence is computed with the assumption that the prediction errors for each PM are normally distributed. $\Pr(e_i)$ is the probability of the prediction error e_i, defined as $\Pr(e_i) = \frac{1}{\sigma_1 2\pi} e^{-\frac{1}{2}\left(\frac{e_i-\mu_i}{\sigma_i}\right)}$. Also, σ_i, and μ_i are standard deviation of prediction errors and mean of errors, respectively. In this case, the prediction confidence of PM_i can be considered as σ_i^2 which is the variance of past prediction errors. Our prediction of resource usage of PM_i, denoted as $\hat{P}_{i,t+1}$, is defined as Equation (3).

$$\hat{P}_{i,t+1} = \hat{R}_{i,t+1} + \sigma_i^2 \tag{3}$$

The output $\hat{P}_{i,t+1}$ is converted to 0 or 1 according to its value as below as a final result whether hotspot will occur or not at time t+1.

$$\hat{P}_{i,t+1} = \begin{Bmatrix} 1, & if\ \hat{P}_{i,t+1} > 0.5 \\ 0, & otherwise \end{Bmatrix} \tag{4}$$

In the evaluation section, we show the accuracy of our prediction technique according to different patterns of resource usages.

3.3 Resource Management Scheme and Hotspot Removal Algorithm

Our system executes live-migration operation when there is threshold violation (i.e. hotspot) on a specific PM. To remove the violation, it migrates a source of the

violation (i.e. a VM) from the current PM to a destination PM. This is a NP-hard multidimensional bin-packing problem of matching available resource on PM nodes and required resource of VMs [5].

System Models

In this subsection, we propose resource usage model for a virtualized cluster. Suppose there are *M* nodes in the cluster and *K* virtual machines. The resource usage of PM_i, $R_{i,}$, is defined as $R_i = \{R_{i,c}, R_{i,m,} R_{i,n}\}$, where $R_{i,c}$ is the CPU utilization, $R_{i,m}$ is the memory utilization, and $R_{i,n}$ is the network utilization. Likewise, the resource usage of VM *j*, denoted as $V_j = \{V_{j,c}, V_{j,m,} V_{j,n}\}$, where, $V_{j,c}$ is the CPU utilization ratio of the current usage versus usage in service level agreement (SLA), $V_{j,m}$ is the memory utilization ratio of the current usage versus usage in service level agreement (SLA), and $V_{j,n}$ is the network utilization ratio of the current usage versus usage in service level agreement (SLA). R is a set of resource usage state of the PM cluster, denoted as $R = (R_1, R_{2,} \ldots, R_i, \ldots, R_M)$. Likewise, V is a set of resource usage state of the PM cluster, denoted as $V = (V_1, V_{2,} \ldots, V_j, \ldots, V_K)$.

Algorithm	HotspotRemoval(R, V, $\hat{P}$)
Input:	a node H where a hotspot occurred, a set of resource usage state of the PM cluster (R), a set of resource usage state of the VMs (V), and predicted resource usage ($\hat{P}$)
Output:	a success flag of hotspot removal (S)
1:	(V_h that is the main cause of hotspot, T that is the specific resource that cause the hotspot) ← *findViolateResource(H)*
2:	Remove H from set R
3:	*sortDescendingOrder(T, R)*
	// sort R in terms of specific resource T
4:	**for** i= 1 **to** (M -1)
5:	**if** (V_h + R_i < Threshold)
6:	**if** (*isMigrateSafe*(V_h , R_i , $\hat{P}$)) **then**
	// *isMigrateSafe* tests the safeness of live-migration
7:	$R_d \leftarrow R_i$
	// R_d is a destination PM node for live migration
8:	*liveMigrate*(R_d , V_h)
9:	**end if**
10:	**end if**
11:	**breakfor**
12:	**if** ($R_d \neq null$) **then**
13:	**return yes**
14:	**else**
15:	**return no**

Virtual Machine Deployment Algorithm

In our system, we execute a hotspot removal algorithm to remove hotspot from the cluster and re-balance the workload. In this algorithm we use the predicted resource usage for determining the destination PM node to prevent future hotspots. The basis of our algorithm is derived from the first fit decreasing (FFD) algorithm [10].

In our algorithm, we find the specific resource that causes a hotspot in node H (*findViolateResource*) by traversing the set of resource usage of VMs in node H. For example, if the memory usage ratio of the 2nd VM in PM node H is the biggest among the resources of all VMs in node H, V_h is V_2 and T is $V_{2,m}$ for this example.

Then, after removing node H from R, we sort nodes in R in descending order in terms of T (*sortDescendingOrder*) to apply the first fit decreasing algorithm. At the line 4, we start looping the R to find the largest available resource for the V_h and check whether the migration will generate a new hotspot in the near future by considering the prediction of resource usage of the candidate node (*isMigrateSafe*). If there is enough room and it will not make a new hotspot, we process live-migration of V_h to R_d (*liveMigrate*). Finally, the algorithm returns the flag of success.

Table 1. virtual machine instance types

Instance ID	# of cores	Size of memory (GB)	Network (Mbps)
1	1	2	100
2	4	8	100
3	8	16	200
4	1	1	100
5	8	16	100
6	4	2	100

Table 2. Simulation Environment Setup

CPU type	Intel i5-2320
Number of cores	4
CPU clock	3.00 GHz
RAM	4GB
OS	Window 7
Development Software	Java 1.7.0
DBMS	MySQL 6.0
IDE	Eclipse Indigo

4 Evaluation

To verify the effectiveness of our hotspot removal algorithm and neural network predictor, we conduct simulation experiments with empirical data. We set up the experimental environment by developing the workloads on VMs. First, we reference 6 types of VM instances from Amazon EC2 documents [11] that are shown in **Table 1**. Then, we collect workload patterns from various references: 8 patterns of CPU workloads, which are defined by Ranganathan, et al. [12] about a CPU utility in an enterprise data center, 4 patterns of memory workloads, which are studied by Lee, et al. [13], and 4 patterns of network workloads, which are studied by G. Wang, et al. [14]. The network patterns are identified from Amazon EC2 data center's usage. Thus, we are

able to define 128 various workload types that is a combination of VM types and workload patterns of each resource. **Table 2** shows our simulation environment.

For our simulation, we create 12 PMs and each PM is equipped with 40 CPU cores, 40GB memories, and 1Gbps network bandwidths. On these PMs, we simulate workloads, lifecycle, and type of instances to VMs to create 1,800~2,000 VMs. The simulation is equivalent to 3,000 hours of operations.

4.1 Evaluation of Neural Network Predictor

For this comparison, we collected 1,000 patterns (500 each for hotspot and normal cases) and divided them into training and test with 50% each. Our neural network is trained with 100 epochs and 5 neurons in the hidden layer of the NN. **Table 3** shows the result of the hotspot prediction of our method and Sandpiper as varying the delay index in time series. As seen in the table, Sandpiper provides the best performance for d=6,7 and the worst performance for d=2. While, our method provides the best performance for d=2 and the worst for d=6, 7.

Table 3. Accuracy of Hotspot Prediction (%)

Method	Delay (d)							
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**
SandPiper	**64%**	**68%**	**72%**	**72%**	**76%**	**76%**	**68%**	**68%**
Our	**76%**	**76%**	**80%**	**80%**	**84%**	**84%**	**80%**	**80%**

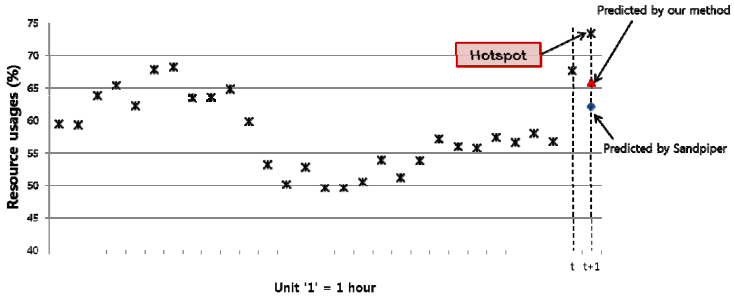


Fig. 3. (a) Workload Type with High Variance

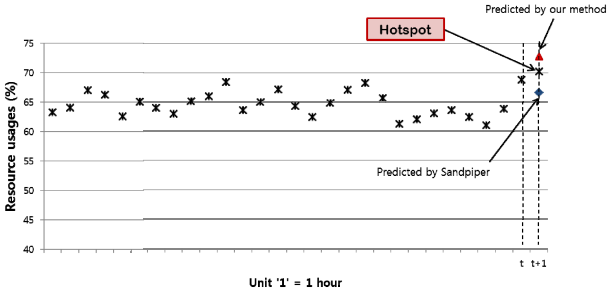


Fig. 3. (b) Workload Type with High Variance

Table 4. Hotspot Removal Ratio (%)

Simulation	Our Hotspot Removal System		Sandpiper	
	Hotspot Occurrences	Removed Hotspots	Hotspot Occurrences	Removed Hotspots
A (1912 VMs)	191	165 (86.2)	299	201(67.2)
B (1813 VMs)	94	65(69.1)	115	79(68.7)
C (1845 VMs)	205	145(70.7)	217	139(64.1)

To our analysis, Sandpiper's prediction method is not appropriate for the patterns changing dynamically (see **Figure 3 (a)** and **Figure 3 (b)**), because it utilizes the statistics of the hotspot occurrences. Our method can learn various types of hotspot patterns in a nonlinear fashion because of the generalization performance of NN. However, our method is not able to predict the hotspot in noisy patterns in which hotspot occurs abruptly as seen in **Figure 3 (a)**.

As seen in the table, overall performance is better than Sandpiper. It is because the performance of Sandpiper depends on data size more than our method. In other words, Sandpiper needs huge number of various patterns to provide better performance.

4.2 Evaluation of Hotspot Removal Algorithm

To verify the effectiveness of our hotspot removal algorithm, we measure the number of hotspots detected by our approach and Sandpiper's approach and compare them. We run simulations using the combination of VMs' types and workload patterns to measure the hotspot removal ratio of our system and that of Sandpiper. **Table 4** shows the number of hotspots occurred during 3,000 hours equivalent simulation and number of removed hotspots. We conduct three simulations with different data sets that are generated randomly with given types and patterns.

Even though the ratio improvement are varying, it is clear that with the new destination PM node selection method (i.e. resource specific first fit decreasing algorithm) and hotspot predictor modeled based on resource usage in time series, it is clear that our proposed system is removing hotspots as well as preventing hotspots more effectively.

5 Conclusion

Workload and resource management is one of key research issues in distributed computing. Currently, handling fault events such as hotspots is getting important since achieving high utilization of resource by using virtualization technology as well as classic workload fluctuation problem makes the problems of operating environment more complex and challenging. In this paper, we proposed a novel hotspot removal system with the neural network predictor to improve hotspot removal performance (i.e. success ratio of hotspot removal).

In order to verify the effectiveness of our approach, we conducted the evaluation of 1) the performance of our neural network predictor and 2) success ratio and

performed the comparison of our system with Sandpiper using empirical workload data. The results showed that our approach performs better in both evaluations.

As our future work, we are planning to run our system on mid-size private cloud to evaluate the system with workloads that are generated by real applications.

To verify the effectiveness of our hybrid approach, we conduct empirical experiment with a RDF management system based on our approach. The evaluation results and the analysis show that we can expect notable performance gains with our hybrid approach. Even though there is one strong assumption we made and it can be a limitation of our approach, we confirm that our approach is effective especially when there is a high number of requests from the users and the query repetition ratio is high. Since the query translation processor is not included in our design, the keyword query is only effective to the query which is asked previously. The query translation issue should be studied further.

There is one strong assumption we made and it can be a limitation of our approach. Since the query translation processor is not included in our design, the keyword query is only effective to the query which is asked previously. The query translation issue should be studied further.

Acknowledgement. This work was jointly supported by the MKE, Korea, under the ITRC support program supervised by the NIPA (NIPA-2012-(H0301-12-2003)) and the Basic Science Research Program through the NRF of Korea (No. 2012-0003198).

References

1. Hwang, K., Fox, G.C., Dongarra, J.J.: Distributed and Cloud Computing: From Parallel Processing to the Internet of Things. Morgan Kaufmann, Waltham (2011)
2. Buyya, R., Broberg, J., Goscinski, A.: CLOUD COMPUTING: Principles and Paradighms. John Wiley & Sons, Hoboken (2011)
3. Frey, J., Tannebaum, T., Livny, M.: Condor-G: A Computation Management Agent for Multi-Institutional Grids. Cluster Computing 2, 237–246 (2002)
4. Appleby, K., Fakhouri, S., Fong, L., Goldszmidt, M., Krishnakumar, S., Pazel, D., Pershing, J., Rochwerger, B.: Oceano-SLA-based management of a computing utility. In: Proc. of the IFIP/IEEE Symposium on Integrated Network Management (May 2001)
5. Wood, T., Shenoy, P., Venkataramani, A., Yousif, M.: Sandpiper: Black-box and gray-box resource management for virtual machines. Computer Networks 53, 2923–2938 (2009)
6. Zhang, Q., Cherkasova, L., Mi, N., Smirni, E.: A regression-based analytic model for capacity planning of multi-tier applications. Cluster Computing 11, 197–211 (2008)
7. Box, G.P., Jenkins, G.M., Reinsel, G.C.: Time Series Analysis Forecasting and Control, 3rd edn. Prentice Hall (1994)
8. Charoenpornwattana, K., Leangsuksun, C., Tikotekar, A., Vallée, G.R., Scott, S.L.: A Neural Networks Approach for Intelligent Fault Prediction in HPC Environments. In: Proc. of the High Availability and Performance Computing Workshop, Denver, Colorado (2008)
9. OpenIPMITool, `http://ipmitool.sourceforge.net`
10. Verma, A., Ahuja, P., Neogi, A.: pMapper: Power-aware dynamic placement of HPC applications. In: Proc of. 22nd Supercomputing Conference, pp. 175–184. ACM, New York (2008)

11. Amazon EC2, http://aws.amazon.com/ec2/
12. Ranganathan, P., Jouppi, N.P.: Enterprise IT Trends and Implications on System Architecture Research. In: Proc. of the High-Performance Computer Architecture, pp. 253–256. IEEE CS Press (2005)
13. Lee, K., Park, S.: A Dynamic Allocation Scheme for Improving Memory Utilization in Xen. Journal of KIISE: Computer Systems and Theory 37(3) (2010)
14. Wang, G., Ng, T.S.E.: The impact of virtualization on network performance of Amazon EC2 data center. In: Proc. of IEEE INFOCOM, San Diego, CA (2010)

A Case Study on Trust-Based Automated Blog Recommendation Making

Nurul Akhmal Mohd Zulkefli[1], Hai Trong Duong[2], and Baharum Baharudin[1]

[1] Faculty of Computer and Information Technology, Universiti Teknologi Petronas, Malaysia
skygur85@gmail.com, baharbh@petronas.com.my
[2] IESLAB, INHA University, South Korea
haiduongtrong@gmail.com

Abstract. We are presented with a situation in which a visitor wants to travel to Malaysia. Several questions arise at this point: Should the visitor believe the information provided in the Malaysian official tourism website? Or should the visitor refer to some other "unofficial" sources like blogs which contain the blogger's own experiences? In the travel domain, almost all information shared in blogs naturally derives from blogger's experiences. Positive correlation might exist between blogger and information. This correlation must point to the fact that users tend to be attracted towards finding information through blogs. To support this crucial issue, a survey on the actual people's opinions in finding the relationship between a person and his/her blog information has been done in the travel blog's domain. Results have shown that user usually prefers to refer to the information mentioned by people they trust or, more accurately, friends rather than other sources. In addition, the previous works on trust and blogs also share an agreement that the positive correlation between a blogger and his/her information should affect the trust value. This survey has created an inspiration for the recommendation systems based on the trust exerted on blog information.

Keywords: Trustworthiness, Blog, Travel, Recommendation-Making.

1 Introduction

In the blog circle, the information is kept updated- a condition known as real-time information. Although the information shared in the blog has been increasing exponentially, and a wide range of search engines or plug-ins are created, they are quite unsuitable for information-searching in blogs because the information in blogs cannot completely be trusted. The information in blogs is not merely about the contents of the entries but also about the people who support the information by submitting the information and about those contributing the information. This survey aims to examine bloggers' and non-bloggers' opinions about the correlation of "trust" with a blogger and the information that he or she includes in their blogs for blog recommendation-making. A blogger is considered as trustworthy, if the following criteria are met: when he or she has many followers, whether they have many positive comments in their posts and if they could collect as many "likes" as possible.

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 489–498, 2013.

The issue of 'trust' takes into account the distribution of followers, comments, ratings, and contents which are similar to those of other bloggers. Bloggers may share their ideas and experiences about something in the blog, for example their trips to interesting places, their fine dining and wonderful gourmet experiences, as well as experiences in using public accommodations and public transports. All these information are shared freely and enthusiastically in the blog. Bloggers usually include detailed information on every experience and every story they wish to share in their blogs. For example Alice who comes from somewhere, had visited Kuala Lumpur, Malaysia for three days and two nights. She may include these details in her blog:

- Which airline she took to fly to Kuala Lumpur?
- The mode of transport taken in Kuala Lumpur (public bus, rented car, taxi etc)
- Her accommodation arrangement and the costs involved.
- If the accommodation was a hotel, how much did she spend and how comfortable was the hotel?
- The places she visited and the costs involved, if any.
- Where did Alice have breakfast, lunch and dinner and the respective costs involved?

The growing blog fraternity has indirectly facilitated Malaysians to make well-informed decisions in important matters like travelling especially involving overseas travel. In matters pertaining to food, most Malaysian also like to share recipes in their blogs and provide other people with various recipes. Comments left on the bloggers' websites provide an indication of the level of interaction that occurs between the visitors and the blogger himself or herself.

A common situation emerges when a user finds it difficult to trust the recommendation systems. From the Google search engine, user may receive many suggested blogs, websites, images and videos in response to what they are looking for. Bhuiyan T. [1] indicates that recommendations may be received through a chain of friend's networks and the problem for the user is to be able to evaluate various types of opinions and recommendations. User may be unable to choose and evaluate, as well as decide which site should be trusted more. For example from the user's query or keyword for "Kuala Lumpur", the search engine will display many results, whether from official or unofficial websites. However, nowadays, the search engine is no longer categorized, making the results less trustable but it categorizes the results by keyword and user profile. Despite the fact that many collaborative recommendation systems are present, their performance is regarded as poor especially when previous information or experience is not made available [2]; a phenomenon known as the cold start problem. Therefore, to overcome this problem, the trust-based approach has been introduced to the recommendation system [3, 4, 5]. This survey is done to identify the person's trust characteristics and use the survey results for improving the recommendation system based on user trust.

In the following sections, we have described briefly the current status of the online social networks and the issue of trust and similar issues that arise in online environment. In this paper we have surveyed and analyzed online users'opinions about the relationship between trust and similar interests and the findings presented could be

useful in the research areas of trust-based automated recommender systems. The rest of the paper is organized as such- in section 2, we have discussed the fundamental tenets of trust by formally defining trust and listing its characteristics. Section 3 presents a brief summary of online social network evolution. Section 4 describes an analysis of the current research work on trust and similarity of interests. Section 5 elaborates on the survey method in detail and provides discussion of the survey results and findings. The paper is concluded in section 6.

2 Trust Fundamentals

2.1 Overall Trust

Gambetta [6] defines "*trust (or, contrastingly, distrust) as a particular level of subjective probability with which an agent assesses another agent or group of agents who will perform a particular action, both before he can monitor such action (or independ-ently of his capacity to ever be able to monitor it) and in a context in which it affects his own action*". Donovan Artz [7] defines that trust stands as an integral component to many angels of human interactions. By trust, people can act in uncertain situations and face the risk of having to suffer from negative consequences. In the semantic web, trust becomes a central component [8-9]. [9-10] include a trust layer that serves to assimilate the ontology, rules, logic, and proof layers. Donovan Artz [7] asserts that trust is used to verify who the source claims to be. All the information should be checked and verified as proof that the information can be trusted. Hussain and Chang [11] further explain that there is no correct definition of trust with regards to the contexts of dependence, time dependence and the dynamic nature. Moreover, Deustch [12] elaborates that when a person perceives an ambiguous path, the trusting behaviour will tend to exist. Whether the result following the path is good or otherwise, it will contingent on the action of another person [12].

In the meantime, two researchers Golbeck and Hendler [13] define trust in a person as "a commitment to an action based on belief that the future actions of that person will lead to a good outcome". This definition supplies a meaning where a positive outcome is always placed on the first place, and vice versa. Trust can be complex to one person and it is not dependent on only one situation or domain. The interpretation of trust can differ according to motivations and goals. Jøsang et al., [14] define two general trust; reliability trust *(the term "evaluation trust" is more widely used by the other researchers, therefore this term is more commonly used)* and decision trust.

Our main purpose is to provide a survey of trust in view of the natural content of travel blog and the user behaviour. Previous works relevant to the classification of trust in computer science include a Survey of Trust and Reputation Systems for Online Service Provision [20], a Survey of Trust in Computer Science and the Semantic Web [7], a Survey of Trust in Internet Applications [21] and the latest Survey of Trust in Workflows and Relevant Contexts [22]. Although there are many aspects of the definition of trust with multiple uses, there is yet to be produced, a

proposal which highlights the aspect of trust in natural content, i.e in blog, where nowadays, it is for a fact that most Internet users tend to prefer to find travel information through natural content (or what has been defined as 'blog'). Natural content in non-official websites such as blog establishes important guidelines to users who are looking for certain information based on user's own personal experiences.

2.2 Characteristics

Golbeck [15] proposes three main properties of trust in the web-based social environment which is (i) Asymmetric, (ii) transitive, and (iii) personalised. Below are the scenarios of properties that have been implemented in our comtext.

Trust is Asymmetric: Between two parties, the trust level is not balanced. Bob shares ample information on "Kuala Lumpur" in his blog. In this case, Alice may trust Bob 100%, but Bob may not necessarily trust Alice at all and may not want to share details or information with her. Bob may only trust Alice at the rate of 50%.
Arguably, trust can be transitive: Let's say Alice and Bob know each other very well; in fact, they are best friends. Bob has a friend named Carlo whom Alice has not met before. However, since Alice knows Bob so well and trust Bob's choices in making recommendation, Alice may trust Carlo to a certain extent even though they have never met. Now let's say Carlo has a friend named Daniel whom neither Alice nor Bob knows well, and understandably Alice finds it hard to trust Daniel. Hence, some argue that as the link between nodes grows longer, trust level has become more and more decreased . Previous researches [16, 17] disagree with the statement that trust is transitive. Moreover, [18] defines that if one has a good friend whom he or she trusts dearly, who also trusts that the president would not lie, does that mean that he or she would therefore trust that the president would not lie either?
Trust is personalised: Trust is a subjective point of view; two parties can have very different opinions about the trustworthiness of the same information. For example, a person in a group may like to visit the island but another member of the group does not share this preference. So there can be two different reasons that very much depend on the way a person behaves.

2.3 Blogging in Malaysia

According to *thestar.com*[1], blogging activities in Malaysia rank among the highest in the world. Malaysians start to develop the trend of sharing information, experiences and knowledge through a modern medium known as blog. Through the statistics from *Sysomos.com*[2], Malaysia is at the 14th place among the top 15 countries in the ranking of growing blogs. Furthermore, *Utusan Malaysia* had reported that in 2008 there were

[1] Thestar.com: http://thestar.com.my/
[2] Sysomos.com: www.sysomos.com/

500,000 active blogs in Malaysia. In 2012, MyEventsInternational[3] with the endorsement by the Ministry of Information, Communications and Culture in Malaysia had successfully organized the "World Bloggers and Social Media Awards 2012". In this event, a total of 803 blogs were nominated where 80% of judging is based on public voting and 20% on judges' decision on several aspects of the blog. From the evaluation by the judges, 80% come from public voting or the Internet users. This means that users have the power to evaluate whether the blog content is useful, trustworthy, interesting or others. A large number of votes imply that these users have demonstrated what they have decided to be their favorite information. If they believe in the content, and like the information provided by the blogger, they will vote for the blogger in question. Otherwise, the blog can only be satisfied with low ratings, and an indication that the blog has a low level of trust. From this event, the winner can improve their rankings and his or her position to the world's view. For example, the Best Travel Blog is won by PlacesandFoods.com. This blog has developed rather well, once it has become well-known by outside organisations such as Thailand Tourism and Australia Tourism. The power of blogs lies in its appeal or attraction to the readers or users, and such fascination can be created merely by writing down everything that has been experienced and thought to be worth sharing with other readers in the virtual world.

3 Material and Methods

3.1 Study Objective

The objective of this survey is to collect information about the users (especially bloggers)- and how they view the characteristics regarding the relationship between a blog audience and his/her information in the travel blog they produce. This survey is based on Touhid B. [19]'s study in which questionnaire is supplemented to support and improve the results. In addition, we maintain the objective to identify the relationship between a person (blogger) and his/her supplies of information.The subject is categorized in several sub-topics, as shown below:

- User and blog's profile
- Perceptions about the Recommendation System
- Relationship between person and information

3.2 Study Design

We have chosen the Google Drive (Survey) *drive.google.com* to host the survey as it is convenient for large participants, that it is very flexible and the data gathering process is easy to administer. This survey was distributed in July, 2012 and ended in the first week of September, 2012. Invitations to take part in the survey were sent out via Facebook.com, email, at the workplace and places of study.

[3] MyEventsInternational- myevents.com.my is top organization in Malaysia which mission to provide their clients with a first-class, cost effective service, organized by professional staff, for an inspirational approach to total event management.

3.3 Questionnaire Participants

In total, 65 people had taken part in the online questionnaire. Among these, 74% were female and 26% were male. Most of them are 25 to 35 years old, and 47% of respondents are undergraduates, 37% have master or professional degree and 8% are Ph.D students. Most survey participants have some Information Technology (IT) background with 36% claiming to be Internet users, 34% bloggers where they had identified their blog's link in the survey answer and 27% admitted that they are Facebook.com users. None of them answered "I am not an Internet user", so we believe that in the real world, all of them are Internet and Facebook.com users. Fig. 1 shows some results from the Demography category.

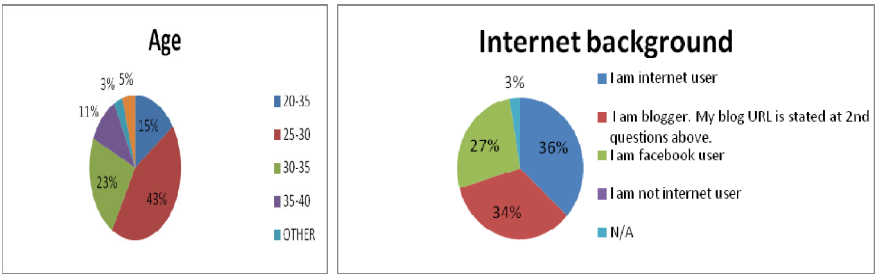


Fig. 1. A Survey of Online Social Networks Participant Background

4 Discussion on Results and Findings

The results are categorized in two different sections which shall be discussed in the following sub-sections below:

4.1 Category One : Blog and Blogger

There were 18 respondents who do not have a blog and most of them had listed the URLs blogs that they always visit. Most of the respondents which is 92%, characterize their blogs as a personal blog, 60% regard their blog as a pastime and 25% viewed their blogs as Academic Brainstorm (Fig. 2). We separate the personal life and pastime because the latter can be characterized as a personal life but personal life cannot be characterized as pastime. For example, traveling with family to Kuala Lumpur can be both under 'personal life' and 'hobby', but expressing trouble faced in Kuala Lumpur cannot be characterized as a hobby, but rather as an experience in one's personal life. Moreover, most respondents clarified their blog as moderate (32%) and slightly (34%) private, as illustrated in Fig. 3. It means many bloggers are willing to share their information to the public and it is consistent with the aspect of blog

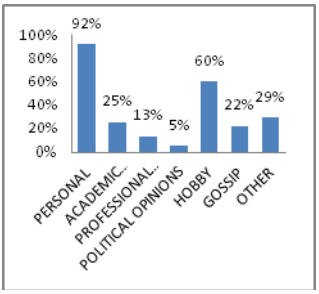


Fig. 2. Blog Characterization

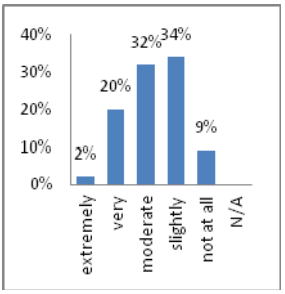


Fig. 3. How private are the things you write about on your blog?

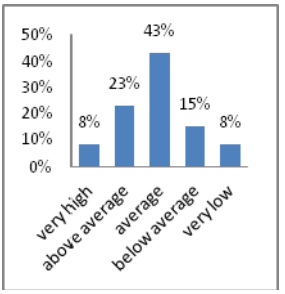


Fig. 4. How well do you feel you know your blog's visitor?

characterization where respondents had chosen "Personal Blog" and "Hobby". We asked the question of what the blogger would feel if they knew who their visitors are in Fig. 4, 43% respondents chose "average" and 23% picked "above average". In addition, 35% respondents had stated "very careful" followed by 25% who chose "moderately careful" in question 14; If you were aware of all the people who read your blog, how likely is it that you would become more careful about what you write? In conclusion, it shows that once the respondent knows his/her visitor and he/she can keep track which post that the visitors always read, he or she will try to share the same element of information in the blog. For example, when Bob knows that Alice and Carlo always visit his blog to read information about travelling, then Bob will keep sharing about most of his travelling experiences and information in his blog as compared to news about academics or sports.

4.2 Category Two: Recommendation System

In the second category, we decided to collect user's views on behalf of getting the travel information via Internet sources and how they trust people and the information from the Internet. In the first question, we asked whether user prefers to have automated recommendation for travel information or not, and 68% answered "Yes" and 31% stated otherwise, as can be seen in Fig. 5.

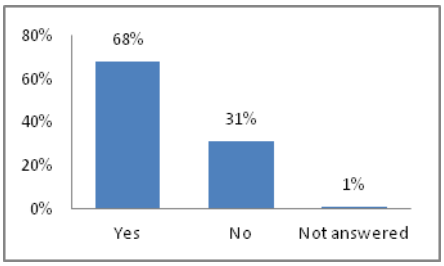


Fig. 5. "Do you prefer to have the automated recommendation for travel information or not?

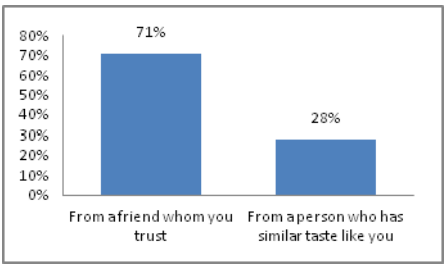


Fig. 6. "Which recommendation do you prefer most?"

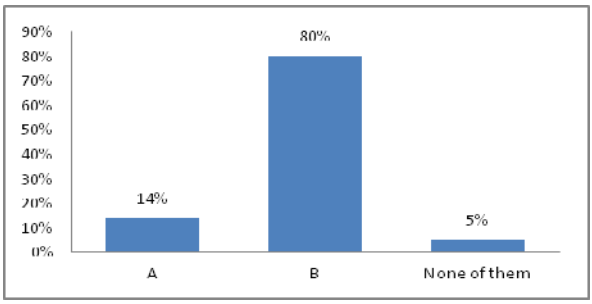


Fig. 7. "Assume that an unknown traveler expert A and one of your friends B who is an expert about travelling is available for recommendation as you are going to travel / visit some places. Which recommendation will you prefer?"

It means the users prefer to have travel recommendation as a main source for obtaining the travel information they need. We asked users *Which recommendation do you prefer most?,* as seen in Fig. 6. From the result, 71% chose "From a friend whom you trust" compared to 28% who selected "From a person who has similar taste like you". From the result, users mostly prefer information from their trusted friends, compared to friends who have similar taste and this is strongly supported by other results, where 80% users choose his/her friends who have the expertise about travelling to relay the information on travelling compared to any expert traveler that they are not familiar with. This result is shown in Fig. 7. Since many users prefer their trusted friends, we then asked them whether they consider people they have met online as a friend or not. A large number had said 'yes' and the rest did not consider them as friends, at least not yet. It contrasts with our previous findings where most users define 'friend' as someone they trusted, in comparison with a person who only has similar taste with them in terms of the recommendations, as illustrated in Fig. 8. Thus, it is a wonder why they consider their online contacts as friends when the person they actually believe most is their own friends. For example, if Bob has 50 friends in his blog (or rather, 50 followers), it does not mean that Bob happens to really know all these people on the same level. Some of them might be new acquaintances in his blogosphere, whereas some are his close friends. To answer this question, we asked users what would be a recommendation that they would trust, as in Fig. 8. and 34% answered "A recommendation from a person who is competent in the area of recommendation", 38% chose 'from a person who is believed by them', and 15% 'from a person who has good reputation' while the rest claimed that they believed in a recommendation through their confidence. From the result, it shows that the user has the confidence to believe in a person in different situations. For example, 3 of the 50 friends of Bob might be excellent at sharing travel information and 5 of 47 friends are excellent at sharing about historical places that can be visited. Bob's fifty friends can have the expertise in some things and may not have the expertise in other things too. However, what proves to be important in our survey is that we want to know that users have their own rules in choosing people in the Internet as their friends. To clarify the previous questions, the next question arises in other to find the best describe their opinion and 51% said "From people I know (including friends)". Fig. 9 shows

the result where 25% chose "Automated recommendation generated by the expert system" and 22% chose "From my family". This means that users prefer to choose "People/friends" they "know and trust" compared to resorting to other recommendation systems.

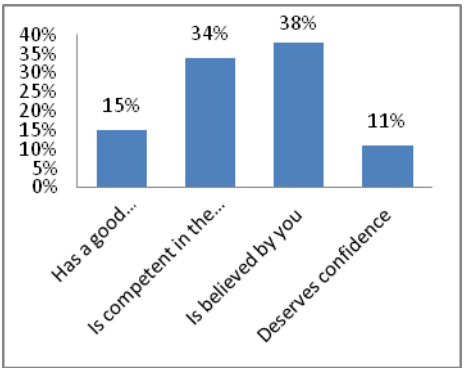


Fig. 8. "Which one is more important to you? A recommendation from a person who..."

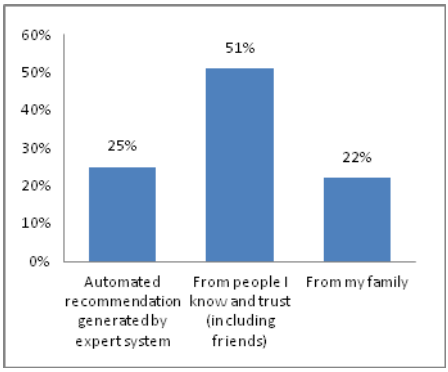


Fig. 9. "In terms of making recommendations,, which one best describes your opinion?"

5 Conclusions

Although trust has been studied and has become an increasingly popular principal in the recommendation system, we have begun to figure out the user behavior on trust determined in their findings. The hypothesis is derived, in order to validate the fact that people and the information that they have play an important role in gaining trust from users or their followers. Thus, we have done a survey to show that this hypothesis is approved and that our results have strongly supported the hypothesis. User preferred to choose the people or friends that they know and whom they feel can be trusted more compared to other factors. Besides, users also strongly choose the recommendation that comes from a person they believe, compared to the existing recommendation system which by means a person they choose is a blogger. In future works, we will use our survey data for our next development of the blog recommendation system. We do believe that our survey can offer a helping hand to other researchers in finding the pattern of behavior of users in this particular field of blogging.

References

1. Sarwar, B., Karypis, G., Konstan, J., Reidl, J.: Application of Dimensionality Reduction in Recommender Systems. In: ACM Workshop on Web Mining for E-Commerce Challenges and Opportunities (WebKDD), Boston, USA (2000)

2. Jamali, M., Ester, M.: TrustWalker: A Random Walk Model for Combining Trust-based and Item-based Recommendation. In: KDD, Paris, France (2009)
3. O'Donovan, J., Smyth, B.: Trust in Recommender Systems. In: Proceedings of the 10th International Conference on Intelligent User Interfaces, pp. 167–174 (2005)
4. Massa, P., Avesani, P.: Trust-aware Recommender Systems. In: Proceedings of the ACM Conference on Recommender Systems, Minneapolis, USA, pp. 17–24 (2007)
5. Andersen, R., Borgs, C., Chayes, J., Feige, U., Flaxman, A., Kala, A., Mrrokni, V., Tennenholtz, M.: Trust-based Recommendation Systems: An Axiomatic Approach. In: Proceeding of the 17th International Conference on WWW, Beijing, Chine, pp. 199–208 (2008)
6. Gambetta, D. (ed.): Can We Trust Trust?, vol. 13. University of Oxford, Oxford (2000)
7. Donovan, A., Yolanda, G.: A survey of trust in computer science and the Semantic Web. Web Semantics: Science, Services and Agents on the WWW 5(2), 58–71 (2007)
8. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Sci. Am. (2001)
9. Berners-Lee, T., Hall, W., Hendler, J., O'Hara, K., Shadbolt, N., Weitzner, D.: A Framework for Web Science, Found. Trends Web Sci. 1(1) (2006)
10. Berners-Lee, T.: Semantic Web on XML. Presentation at XML (2000), http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html
11. Hussain, F.K., Chang, E.: An Overview of the Interpretations of Trust and Reputation. In: 3rd Advanced International Conference on Telecommunications, Mauritius (2007)
12. Deutsch, M.: Distributive Justice: A Social Physiological Perspective. Yale University Press, USA (2004)
13. Golbeck, J., Hendler, J.: Inferring binary trust relationships in Web-based social networks. ACM Transactions on Internet Technology 6(4), 497–529 (2006)
14. Jøsang, A.: Probabilistic Logic Under Uncertainty. In: The Proceedings of Computing: The Australian Theory Symposium (CATS 2007), CRPIT, Ballarat, Australia, vol. 65 (2007)
15. Golbeck, J.: Combining Provenance with Trust in Social Networks for Semantic Web Content Filtering. In: Moreau, L., Foster, I. (eds.) IPAW 2006. LNCS, vol. 4145, pp. 101–108. Springer, Heidelberg (2006)
16. Grandison, T.: Trust Management for Internet Applications, Ph.D. Thesis, University of London, UK (2003)
17. Abdul-Rahman, A.: A Framework for Decentralised Trust Reasoning, Ph.D Thesis, University of London, UK (2004)
18. Zimmermann, P.: PGP(tm) User's Guide (1994)
19. Touhid, B.: A Survey on the Relationship Between Trust and Interest Similarity in Online Social Networks. Emerging Technologies in Web Intelligence 2(4), 291–299 (2010)
20. Jøsang, A., Ismail, R., Byod, C.: A Survey of trust and reputation systems for online service provision. Decision Support Systems 43(2), 618–644 (2007)
21. Grandison, T., Sloman, M.: A Survey of trust in internet applications. IEEE Communications Surveys and Tutorials 3(4), 2–16 (2000)
22. Viriyasitavat, W., Martin, A.: A Survey of Trust in Workflows and Relevant Contexts. IEEE Communications Surveys and Tutorials 14, 1–30 (2012)

Consensus for Collaborative Ontology-Based Vietnamese WordNet Building

Tuong Le[1], Trong Hai Duong[2], Bay Vo[3], and Sanggil Kang[4]

[1] University of Food Industry, Hochiminh City, Vietnam
[2] Faculty of Mathematics and Informatics, Quang Binh University, Vietnam
[3] Information Technology College, Hochiminh City, Vietnam
[4] Department of Computer and Information Engineering, Inha University, Korea
{tuonglecung,haiduongtrong,bayvodinh}@gmail.com,
sgkang@inha.ac.kr

Abstract. Ontology-based Vietnamese WordNet (OVW) has an extremely important role for most of areas relating to Vietnamese language processing. In this paper, we supplement some structural changes to enrich the structure of Ontology-based WordNet and use it to develop the OVW. A consensus-based collaboration method with reliability measurement is proposed for collaborative OVW building. The knowledge contributed through collaborative processes by participants is considered as in consistent data for our consensus method to make a reconciled version. In experiment, OVW is automatically initialized by using Vietnamese word list. Participants collaborate to improve this initial version via our system. To evaluate our method, we compare the accuracy rate of OVW and Vietnamese WordNet using Asian WordNet's approach.

Keywords: Ontology-based Vietnamese WordNet.

1 Introduction

WordNet[1] is a lexical database that groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations among these synsets. The purposes of WordNet are to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications.

In 2006, Assem et al. presented a conversion of WordNet 2.0 to RDF/OWL [2] for direct use by Semantic Web application developers. One year later, Huang et al. [9] described a data representation for WordNet 2.1 that can be used to enrich the work in progress of standard conversion of WordNet to the RDF/OWL representation [2] at W3C. In 2008, Suchanek et al. [13] proposed YAGO, a large ontology with high coverage and precision that has been automatically derived from Wikipedia and WordNet. It comprises entities and relations, and currently contains more than 1.7 million entities and 15 million facts that have been extracted from the category system and the infoboxes of Wikipedia and have been combined with taxonomic relations from

[1] http://wordnet.princeton.edu

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 499–508, 2013.

WordNet. In addition, the author also proposed a powerful query model to facilitate access to YAGO's data. Simultaneously, Duong et al. [5] in 2008 introduced a method for integration of WordNet-based Ontologies using distance measures. This paper applied the meaning of concepts of upper ontologies to an ontology integration process by providing semantic network called OnConceptSNet. In 2010, Assem, one of the authors of [2], continuously proposed the RDF-based WordNet3.0 [1].

For Vietnamese, WordNet is also extremely important in Natural Language Processing, Machine Translation, Information Retrieval, Semantic Web, etc. Therefore, building Vietnamese WordNet becomes imperative. In 2003, Ho et al. [8] proposed an approach to build Vietnamese WordNet focused mainly on creating the noun tree by using the Vietnamese dictionary. Recently, Isahara and Sornlertlamvanich in [10] built the AsianWordNet (AWN) project in which Vietnamese WordNet was translated from WordNet 3.0 with collaboration of Luong in 2007. Unfortunately, the structure and content of this Vietnamese WordNet have not been published, so we cannot inherit from this project. The progress of this project is quite slow, 10.44% in five years and the due time cannot be determined. Furthermore, their method of collaborative translation does not seem appropriate in English and Vietnamese. For example, in English, *fall* has many senses: *autumn*, *waterfall, etc*. However, when the system translates this word from English to Vietnamese, *mùa_thu (fall)* does not have sense *thác (waterfall)*. This is the biggest obstacle of building Vietnamese WordNet by translating from WordNet.

Nowadays, the initial approaches to create the Vietnamese WordNet-based Ontology have several publications. National key project [4] headed by Cao, created a non-publicly knowledge based on Ontology that is a set of well-known named entities in Vietnam. Besides, Nguyen built an open Ontology for Vietnamese Language[2] having 2,543 classes, 10,024 individuals, 312 relations and 87 properties. In this paper, we propose a structure of OVW improved from WordNet-based Ontology. From this structure, we supplement words or phrases in Vietnamese word list [7] to this in order to create an initial OVW. However, the initial OVW does not have senses and relations. Therefore, researching to reach a collaborative approach and conflict processing accordance with this problem is necessary in OVW building.

Collaborating and resolving inconsistency of knowledge in Ontology building were great attractive [3, 6, 11, 12]. Bao et al. introduced Wiki@nt [3], an ontology building environment, which supports collaborative ontology development included knowledge integration and knowledge reconciliation. In 2008, Nguyen et al. [11] published the book including a set of methods for resolving inconsistency of knowledge, in which consensus method was presented as an effective solution. After that, Duong et al. [6] used consensus method in collaborative ontology building. In 2012, Nguyen et al. [12] proposed a relatively new approach of consensus method to solve conflicted issues in collaborative knowledge through social network. However, this approach mainly focused on the importance of users in social network, not exploits users' intellectual contributions. In this paper, we propose a new approach of consensus-based collaboration method with reliability measurement to build OVW.

[2] `http://ovl-open.sourceforge.net`

2 Collaborative Ontology-Based Vietnamese WordNet Building

2.1 Consensus Methods for Collaborative Ontology Building

Engineering-oriented method [14] is used to develop most of ontologies. This method has a small group of engineers carefully build and maintain a representation of their view. Maintaining in engineering-oriented manner is a highly complex process: participants need to regularly merge and reconcile their modifications to ensure that the ontology captures a consistent and unified view of the domain. Stanford University uses Protégé [14] for knowledge acquisition that provides a graphical and interactive ontology design and knowledge base development environment.

Consensus is a collaborative process in which participants work together to solve a problem. Consensus is not to find the best idea or the most correct idea, but is to find the consensus idea of all participants. Currently, consensus has two common methods: Nominal Group Technique (NGT) and Delphi. NGT is a method of voting all ideas to make the final decision with the most support from participants. The drawback of this method is that the experts have to work directly with each other. Therefore, Delphi method was presented that uses normal discussion such as email instead of complex communication among participants. The facilitator makes the results by analyzing the feedbacks from participants.

The most important issue of consensus method is to solve the conflict profile that has a set of different versions of knowledge explaining the same goal. Nguyen et al. [11] proposed a function consensus for solving conflicts in participant opinions. There are two cases can occur: the solution is independent or dependent from the opinions of the conflict participants. According to this author [11], consensus method is an effective approach that can be used to solve the conflict profile.

Some consensus methods have a leader while for other methods all participants have the same role [11] in the collaborative ontology building. In this paper, to save money, time and effort, we propose a consensus-based collaboration method with reliability measurement of participants determined based on the agreements of other participants for previous contributions.

2.2 Ontology-Based Vietnamese WordNet

About the content, WordNet only includes four main types: Noun, Verb, Adjective and Adverb. They are organized into synsets, which describe and represent a basic content, are connected by different kinds of relationships. About the structure, Ontology-based Wordnet [2] has three main classes: Synset, Word and WordSense. Synset and WordSense have subclasses based on the distinction of lexical groups. For Synset this means subclasses NounSynset, VerbSynset, AdjectiveSynset and AdverbSynset. For WordSense this means subclasses NounWordSense, VerbWordSense, etc. Word has a subclass Collocation used to represent words that have hyphens or underscores in them. To develop OVW, we add a main class VWord (see Fig.1) that has a subclass VCollacation to OVW in order to store words or phrases in Vietnamese.

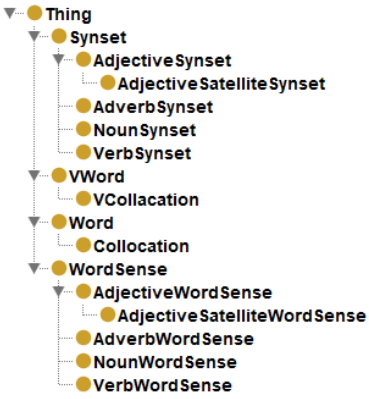


Fig. 1. The class hierarchy of Ontology-based Vietnamese WordNet (OVW)

In Table 1, we list the properties and its significance in OVW. We propose three relations to enrich the structure of OVW: *partOf*, *originalSenseOf*, *vietEng*. (1) the *partOf* relation shows relation of geographic among places. For example, Chua Mot Cot *(Chùa_Một_Cột)* is a famous place in Ha Noi *(Hà_Nội)*. Therefore, synset of *{Chùa_Một_Cột, Chùa_Mật, Nhất_Trụ_Tháp,...}* has a *partOf* relation to synset of *{Hà_Nội, Thăng_Long, Thủ_đô_Việt_Nam,...}*. (2) The *originalSenseOf* relation shows original sense of a word or phrase in Vietnamese. Because Vietnamese borrows so many words from other languages, especially Chinese, the *originalSenseOf* relation is more clear. For example, a compound word *khắc_cốt_ghi_tâm* in Vietnamese has the *originalSenseOf* relation to synset of *{in_sâu_vào,...}* because *khắc* in the compound word *khắc_cốt_ghi_tâm* has means *in_sâu_vào* in Vietnamese. However, a compound word *thời_khắc* has *originalSenseOf* relation to the synset *{thời_gian,...}* because *khắc* in *thời_khắc* means a quarter of an hour. (3) the *vietEng* relation shows the equivalent sense of a Vietnamese synset and an English synset. For example, synset {*sinh_viên,..}* in Vietnamese has a *vietEng* relationto the synset *{student,...}* in English.

Table 1. The relations in Ontology-based Vietnamese WordNet

Property	Domain	Range	Target	Meaning of Property (A: the first, B: the second)
hyponymOf	synset	synset	Nouns , Verbs	A is a hyponym of B.
entails	Synset	Synset	Verbs	B is an entailment of A.
similarTo	Synset	Synset	Adjectives	B is a satellite A.
memberMeronymOf	Synset	Synset	Nouns	B is a member meronym of A.
substanceMeronymOf	Synset	Synset	Nouns	B is a substance meronym of A.
partMeronymOf	Synset	Synset	Nouns	B is a part meronym of A.

Table 1. (*continued*)

classifiedByTopic	Synset	Synset	Nouns, Adjectives, Verbs	A has been classified as a member of the class represented by B.
classifiedByUsage	Synset	Synset	Nouns, Adjectives, Verbs	
classifiedByRegion	Synset	Synset	Nouns, Adjectives, Verbs	
causes	Synset	Synset	Verbs	B is a cause of A.
sameVerbGroupAs	Synset	Synset	Verbs	Verb synset grouped together are similar in meaning.
attribute	Synset	Synset	Nouns to Adjectives	The adjective synset is a value of the noun synset.
derivationallyRelated	WordSense	WordSense	Nouns, Verbs, Adjectives, Adverbs	A is derived from B by means of a morphological affix.
antonymOf	WordSense	WordSense	Nouns, Verbs, Adjectives, Adverbs	This operator specifies antonymous words.
seeAlso	WordSense	WordSense	Verbs, Adjectives	Additional information about A can be obtained by seeing B.
participleOf	WordSense	WordSense	Adjectives to Verbs	The adjective synset is a participle of the verb synset.
adjectivePertainsTo	Synset	Synset	Adjectives to Nouns or Adjectives	An adjective synset pertains to either the noun or adjective.
adverbPertainsTo	Synset	Synset	Adverbs to Adjectives	An adverb synset is derived from the adjective.
gloss	WordSense	xsd:string	Synset and Sentence	The gloss for a synset.
frame	Verb-WordSense	xsd:string	Synset and a verb construction pattern	A generic sentence frame for one or all words in a synset.
partOf	Synset	Synset	Nouns	A is an area of B.
originalSenseOf	Synset	Synset	Nouns, Verbs	A is original sense of B
vietEng	Synset	Synset	Nouns, Verbs, Adjectives, Adverbs	An English word B corresponds to a Vietnamese word A.

2.3 Consensus for Collaborative OVW Building

The Reliability Measurement

Each participant has a value of trust denoted by μ in [0,1]. Where: $\mu = 0$ means that the system cannot trust this participant and $\mu = 1$ means that the system absolutely trusts this participant. When a new participant joins to the system, his/her trust value

is $Const$ specified by the system. We denote U is a set of participants. The trust function t is defined as follows:

$$t: U \rightarrow [0, 1] \tag{1}$$

To solve a problem Pr in OVW building, each participant gives knowledge called a profile $P = \{(e_i, \delta_{e_i})\}$ consisting of many pairs of element and its value that expresses the strength of this element. In the specific case of this paper, P is the set of senses and relations of one word or phrase in OVW. Element $e_i \in P$ is a specific sense or specific relation to other ones of this word or phrase. Other participants have a right to express their agreement for each element by giving a value from 0 to 1 denoted by λ for each element $e_i \in P$. Each participant gives only one agreement value, but he/she can change this value later. The agreement of each participant for each element is defined as follows:

$$a: U \times P \rightarrow [0,1] \tag{2}$$

The function $a(u_j, e_i)$ returns the value of participant u_j give for an element e_i denoted by λ. For example, we assume that an element e_1 in profile P can have four agreement values from four participants $\{u_1, u_2, u_3, u_4\}$ respectively 0.4, 0.1, 1.0 and 0.9. Agreement function can be written as: $a(u_1, e_1) = 0.4$, $a(u_2, e_1) = 0.1$, $a(u_3, e_1) = 1.0$, $a(u_4, e_1) = 0.9$.

Let U_{e_i}be the set of participant expressing the agreement for the element e_i and $u_{max} \in U_{e_i}$ be the participant whose trust value $t(u_{max})$ is the largest. We propose formula (3) to determine the strength of each element.

$$\delta_{e_i} = a(u_{max}, e_i) + \frac{\sum_{j=1}^{|U_{e_i}|} \Delta a(u_j, e_i) * t(u_j)}{\sum_{j=1}^{|U_{e_i}|} t(u_j)} \tag{3}$$

with:

$$\Delta a(u_j, e_i) = a(u_j, e_i) - a(u_{max}, e_i) \tag{4}$$

where:

- δ_{e_i} is the strength of element $e_i \in P$
- U_{e_i} is the set of participants expressing the agreement for element e_i
- $|U_{e_i}|$ is the number of participants in U_{e_i}
- $t(u_j)$ is the trust of participant $u_j \in U_{e_i}$
- $a(u_{max}, e_i)$ is the agreement of the participant u_{max} for the element e_i.
- $a(u_j, e_i)$ is the agreement of the participant u_j for the element e_i.

Example 1. Assume that participants express their agreement for e_1 and e_2 (see Table 2).

Table 2. Participants' agreement for e_1 and e_2

e_1	a	t	$\Delta a(u_j, e_i)$
u_1	0.9	0.8	0
u_2	0.5	0.2	-0.4
u_3	0.4	0.6	-0.5
u_4	0.5	0.7	-0.4

(a)

e_2	a	t	$\Delta a(u_j, e_i)$
u_1	0.9	0.9	0
u_2	0.9	0.7	0
u_3	1	0.8	0.1
u_4	1	0.9	0.1

(b)

In the example of element e_1 (see Table 2 (a)), applying formula (3) we have:

$$\delta_{e_1} = 0.9 + \frac{0 * 0.8 + (-0.4) * 0.2 + (-0.5) * 0.6 + (-0.4) * 0.7}{0.8 + 0.2 + 0.6 + 0.7} = 0.61$$

The same to the example of element e_2 (see Table 2 (b)), we have:

$$\delta_{e_2} = 0.9 + \frac{0 * 0.9 + 0 * 0.7 + 0.1 * 0.8 + 0.1 * 0.9}{0.9 + 0.7 + 0.8 + 0.9} = 0.96$$

According to the formula (3), the strength of an element e_i depend on not only the values of the participants expressing their agreement for this element but also the reliability of these participants.

Let E_{u_i} be the set of elements in all profiles that participant u_i contributed to the system. When finding the consensus knowledge of a word or phrase, the system will update the trust of participant proposed this element by using the following formula:

$$t(u_i) = \frac{\sum_{k=1}^{|E_{u_i}|} \delta_{e_k}}{|E_{u_i}|} \quad (5)$$

where:

- $t(u_i)$ is the trust of participant u_i.
- $|E_{u_i}|$ is the number of elements in E_{u_i}.
- δ_{e_k} is the strength of element $e_k \in E_{u_i}$.

Solving Conflict Profile Algorithm

When there are many profiles to solve a problem Pr, an element e_i can appear in some profiles. It will lead to the conflict profiles. An example of conflict profiles for solving a problem: $Pr = \{P_1 = \{(e_1, 0.6), (e_2, 0.3)\}, P_2 = \{(e_1, 0.9), (e_2, 0.3)\}, P_3 = \{(e_1, 0), (e_3, 1)\}\}$. In the above example, element e_1 appears three times with the strength values determined by formula (3) respectively 0.6, 0.9, and 0; element e_2 appears two times with the strength values 0.3 and 0.3 respectively; element e_3 appears only one time with the strength value 1; We propose an algorithm for giving consensus knowledge from profiles in the state of conflict profile based on the strength of each element.

Table 3. Solving conflict profile algorithm

Input: Given n profiles: $P = \{P_1, P_2, \ldots P_n\}$
Output: P^* the best represents to solve the problem
1. For each $P_k \in P$
2. For each $e_i \in E_k$ /* E_k: the set of elements of P_k */
3. Determine the $u_{max} \in U_{e_i}$ with $t(u_{max})$ is the largest /* U_{e_i} is the set of participants expressing the agreement for e_i */
4. $\delta_{e_i} = a(u_{max}, e_i) + \frac{\sum_{j=1}^{|U_{e_i}|} \Delta a(u_j, e_i) * t(u_j)}{\sum_{j=1}^{|U_{e_i}|} t(u_j)}$
5. Let $E = \bigcup_{i=1}^{n} E_i$
6. For each $e_k \in E$
7. Determine $\delta_{max} = \max(\delta_{e_k})\, in\, P$
8. If δ_{max} < *consensus_threshold* then Remove e_i from E
9. Create $P^* \leftarrow e_i \in E$
10. Return P^*

The Consensus Methodology

In this section, we present a consensus methodology for collaborative OVW building from the initial OVW. There are four phases:

- **Phase 1** – Preparation: (1) We build an initial OVW based on Ontology-based WordNet [2] automatically by adding words or phrases in Vietnamese word list [7] to *UnknownWord,* the list of words or phrases that should be contributed. (2) Our system automatically invites participants to contribute information of k words or phrases in *UnknownWord* everyday via social networks.
- **Phase 2** – Contribution: (1) Information contribution: Participants provide information about senses and relations of one word or phrase. Each sense or relation is considered as an element of this word or phrase. A set of elements of this word or phrase contributed by a participant is regarded to a Profile. (2) Voting contribution: the participants express the agreement for each element of one profile by giving a value in [0,1].
- **Phase 3** – Solving the conflict profiles: after receiving the profiles from participants and the values of agreement from other participants, our system uses the algorithm in **Table 3** to integrate information of those words or phrases.
- **Phase 4** – Controlled feedback: (1) If no consensus is reached [11], our system will not save this words or phrases into OVW. Our system backs to Phase 1 until finding the consensus version. (2) If the system finds the consensus knowledge of a word or phrase, the system updates information of these words or phrases and updates the trust values of participants who contribute information by using formula number 5.

3 Experiments

Firstly, we use Jena to build an initial OVW based on Ontology-based WordNet [2] automatically by adding words or phrases in Vietnamese word list, a large Vietnamese list with about 74,000 words and phrases [7] to *UnknownWord* in OVW.

Secondly, we build a system for collaborative OVW building that automatically invites participants to contribute the knowledge of word or phrase everyday via social networks. The participants, who know the senses or relations of this word or phrase, go to the system to enter its senses or relations. Other participants have the right to express their agreement to each element. The system determines the strength of each element based on the agreement values and the reliability measurement. After that, the system creates consensus version of the words or phrases. This process will repeat until there is no word or phrase in *UnknownWord.*

We perform the collaborative Ontology-based Vietnamese WordNet Building for 1000 random words or phrases in *UnknownWord* with five participants. We compare the accuracy rate of senses of these words or phrases and 1,000 random words or phrases in the Vietnamese WordNet [10] of AWN (see Fig. 2). The result of comparison shows our approach is an effective approach for collaborative Ontology-based Vietnamese WordNet building.

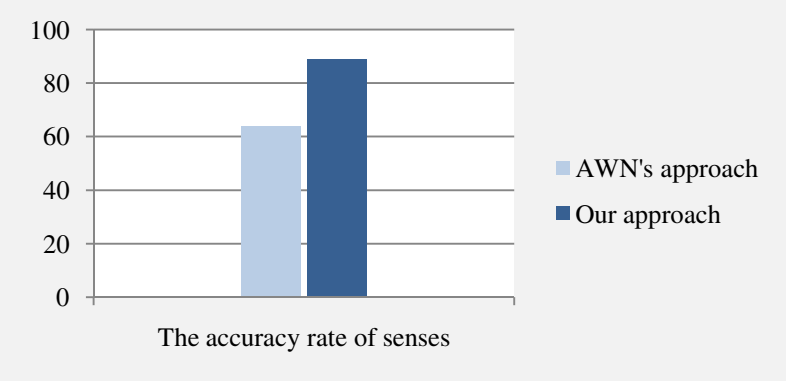


Fig. 2. The comparison of the accuracy rate of senses of 1,000 random words or phrases between our approach and AWN's approach

4 Conclusions

We propose some structural changes to enrich the structure of Ontology-based WordNet and use it to develop the OVW. Besides, we propose the collaborative OVW building and the reliability measurement of participants. Consensus method based on the reliability measurement is effective process creating the quality results. OVW has high accuracy because it is consensus knowledge of the participants who have expertise in what they contribute. However, if OVW is built automatically

before applying our method, time of this process will be greatly reduced and the accuracy of the OVW will increase. Even if this work is done, this collaborative OVW building still required a lot of time and the enthusiasm of participants. Therefore, this process needs the support from experts, society and government.

Acknowledgement. This work was partially supported by Business for Cooperative R&D between Industry, Academy, and Research Institute funded Korea Small and Medium Business Administration in 2012.

This research also was partially supported by Quang Binh University under grant number CS.2.2012.

References

1. Assem, M.V.: Wordnet 3.0 in RDF, http://semanticweb.cs.vu.nl/lod/wn30/
2. Assem, M.V., Gangemi, A., Schreiber, G.: RDF/OWL Representation of WordNet, http://www.w3.org/TR/wordnet-rdf/
3. Bao, J., Honavar, V.: Collaborative Ontology Building with Wiki@nt, A multi-agent based ontology building environment. In: Proceedings of the 3rd International Workshop on Evaluation of Ontology-based Tools, pp. 1–10 (2004)
4. Cao, H.T.: Vietnamese Knowledge and Information Management, http://www.cse.hcmut.edu.vn/~tru/VN-KIM/index.htm
5. Duong, T.H., Nguyen, N.T., Jo, G.-S.: A Method for Integration of WordNet-Based Ontologies Using Distance Measures. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part I. LNCS (LNAI), vol. 5177, pp. 210–219. Springer, Heidelberg (2008)
6. Duong, T.H., Jo, G.S.: Collaborative Ontology Building by Reaching Consensus among Participants. Information-An International Interdisciplinary Journal, 1557–1569 (2010)
7. Ho, N.D.: Vietnamese word list, http://www.informatik.uni-leipzig.de/~duc/software/misc/wordlist.html
8. Ho, N.D., Nguyen, T.T.: Towards Building a WordNet for Vietnamese. In: First International Workshop for Computer, Information and Communication Technologies, Hanoi, Vietnam (2003)
9. Huang, X., Zhou, C.: An OWL-based WordNet lexical ontology. Journal of Zhejiang University SCIENCE A 8(6), 864–870 (2007)
10. Isahara, H., Sornlertlamvanich, V.: Vietnamese WordNet, http://vi.asianwordnet.org
11. Nguyen, N.T.: Advanced Methods for Inconsistent Knowledge Management. Springer, London (2008)
12. Nguyen, Q.U., Duong, T.H., Kang, S.: Solving Conflict on Collaborative Knowledge via Social Networking Using Consensus Choice. In: Nguyen, N.-T., Hoang, K., Jędrzejowicz, P. (eds.) ICCCI 2012, Part I. LNCS, vol. 7653, pp. 21–30. Springer, Heidelberg (2012)
13. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: A Large Ontology from Wikipedia and WordNet. Journal Web Semantics: Science, Services and Agents on the World Wide Web 6(3), 203–217 (2008)
14. The Protégé, http://protege.stanford.edu/

An Ontological Context-Aware Approach for a Dynamic Business Process Formulation

Hanh Huu Hoang

Hue University
3 Le Loi Street, Hue City, Vietnam
hhhanh@hueuni.edu.vn

Abstract. Ontologies are used in Business Process Management (BPM) to reduce the gap between the business world and information systems, especially in the context of the cross enterprise collaboration. For a dynamic collaboration, virtual enterprises need to establish collaborative process with appropriate matching levels of tasks. However, the problem of solving the semantics mismatching is still not tackled or even harder in the case of querying space between different enterprise profiles as ontologies. This paper proposes an approach based on the ontological and context-awareness during the integration and matching task for forming collaborative processes in the problem of the cross enterprise collaboration.

Keywords: BPM, Semantic BPM, B2B integration, Ontology, Cross-enterprise collaboration, context-aware.

1 Introduction

Cross enterprise collaboration has become on of the main course of Semantic Business Process Management (BPM) research recently [1]. There are several approaches for this issue with different perspective which have been surveyed in [2]. State-of-the-art research trends have been focused on two issues for this problem: 1) forming collaborative business process (CBP) dynamically using ontologies or existing BPM standards; 2) Solving the semantics conflicts or mismatching during the process integration and mapping into the execution level. The both issues have been studied in our previous work [1, 3], in which, we choose to move our focal research into the second issue by proposing an conceptual architecture based on business processes ontologies. Our recent work [3] has more focused on the mapping into the execution level with the semantic web services composition approach based on an ontological hierarchical task networking (O-HTN) for the CBP formulation.

In the other hand, the second issue has not been discussed in more details in recent research efforts. Approaches often ignore this crucial issue in this challenging problem of semantic B2B. The main point, according to our survey in [2], is the heterogeneity of used ontologies in several different forms and domains. This makes the querying space for needed processes become huge and the matching process faces a real challenge for finding matched patterns.

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 509–518, 2013.

The context-aware approach proposed in this paper is about introducing a solution to scope down the querying space for processes in the matching process for CBP formulation. This paper uses BizKB Ontology (BO) as "context ontology" (or the context in short) with our O-HTN solution to achieve the goal. The remainder of the paper is organised as follows: the related work to my research is discussed in Section 2. Section 3 describes the ontological BizKB model. Main points of the paper are introduced in Section 4. Section 5 and 6 will deliver discussions and the conclusion.

2 Related Work

Since the failure of the non-semantic approaches as mentioned above, research efforts have been emerged from the motivation of knowledge management and applying Semantic Web technologies into BPM researches to bring the administrative side and IT side together.

Jenz's BPM Ontology approach [4] argued that the third generation business process management is different in that it provides an integrated view on business processes. According Jenz's, the business-oriented view has a counter piece in the form of the IT view, and both must be on an equal footing. The business view can be segmented into three layers: core business ontology layer; industry-specific ontology layer; and organization-specific ontology layer. The IT view is not segmented into layers and is completely organization-specific.

SUPER [5] addresses the ever enduring need of new weaponry in struggle for survival in optimistic business environment where profit margins dramatically drop while competitiveness reaches the new sky high limits. The major objective of the SUPER project is to raise BPM to the business level, where it belongs, from the IT level where it mostly resides now [5]. This objective requires that BPM is accessible at the level of semantics of business experts. SUPER's approach has tried to transform existing BPMN and BPEL standards into a semantics-enriched form, respectively called sBPMN (so-called BPMO – Business Process Modeling Ontology) and sBPEL [6, 7] in the attempt to realize their goals.

In the same line, the SemBiz project[1] aims at bridging the gap between the business level perspective and the technical implementation level in Business Process Management (BPM) by semantic descriptions of business processes along with respective tool support. This approach takes emerging frameworks for Semantic Web Services, namely the Web Service Modeling Ontology (WSMO)[2] as a basis for defining an exhaustive semantic description framework for business processes. On basis of this, novel functionalities for BPM on the business level can be supported by inference-based techniques that work on semantic process descriptions.

Haller in [8] extended the *multi metamodel process ontology* (m3po) introduced with concepts for a full formalisation of the meta-model of XPDL. In the context of their approach, to deal with collaborative processes (choreographies) these internal

[1] SemBiz Project, `http://www.sembiz.org/`
[2] Web Service Modeling Ontology, `http://www.wsmo.org/`

workflow models are aligned to the external behaviour advertised through web services interfaces. The *m3po* ontology presented explicitly models the complete semantics of XPDL. The integrated *m3po* is used as shared representation to perform the integration. The advantage of this approach is that authors use a web ontology language to formalise proposed model into linked data with established business document standards.

One of recent efforts in cross-enterprise collaboration research is Genesis approach based on its ontology called Business-OWL (BOWL) [9]. The core of the approach is about BOWL that is a hierarchical task networking (HTN) modelled in OWL describing the hierarchical relations between tasks of collaborative business processes consist of compound tasks, primitive tasks and task decomposition methods. HTN keeps hierarchical relations of compound and primitive tasks, however, HTN's typical techniques store the knowledge and the specification domain in text files and they could not be processed in the Web environment and not suitable for current dynamic e-commerce today. Therefore, the knowledge described by HTN needs to be modelled in forms of OWL ontologies proposed in this approach.

Through the evaluation and comparison of these approaches, we can see that the fusion of BPM and the Semantic Web or ontology-based techniques becoming a promising research direction in the domain. This research approach can bring new opportunities, new prospects and useful tools for e-business and B2B integration especially. The effort follows this line is Jung's work [10] which focus on basic problems of applying ontology aligning for business process integration. However, there is still room for the two issues mentioned above. Thus, our approach is BizKB-based by combining ontological profiles with the context in the information retrieval [11].

3 Ontological BizKB

3.1 Ontological Enterprise Profiles

According to [12], the Enterprise Architecture refers to a comprehensive description of all of the key elements and relationships that make up an organization. Through the Enterprise Architecture, enterprises can implement enterprise integration to cope with dynamically changing business environment.

Existing Enterprise Architectures, however, lack of semantics for humans and systems to understand them exactly and commonly, which causes communication problems between humans or between systems or between human and system. These communication problems keep enterprises from implementing integration and collaborating with other enterprises.

In order to solve this problem, an ontology-based Enterprise Architecture is proposed [12] and depicted in Fig. 1. The Enterprise Architecture ontology is composed of ontologies in three levels. Ontologies of business terms are in the first level, ontologies of Enterprise Architecture components are in the second level, and ontologies of relationships among Enterprise Architecture components are in the top level (Fig. 1).

In the scope of this paper, we focus on ontologies for business processes that are used for the CBP formulation in the B2B integration problem (level 1 and 2-processes). We call those are enterprise profiles which are modelled and stored in BizKB framework as a knowledge base. As depicted in Fig. 2, the overall conceptual architecture of the BizKB framework consists of two main parts: the BizKB and the Process Formulator. The output of BizKB framework is CBP with semantic web services profiles attached to the CBP. BizKB is the core part of BizKB Framework containing the business knowledge in the form of BPMO-based[3] collaborative business processes with different levels of the abstraction [1].

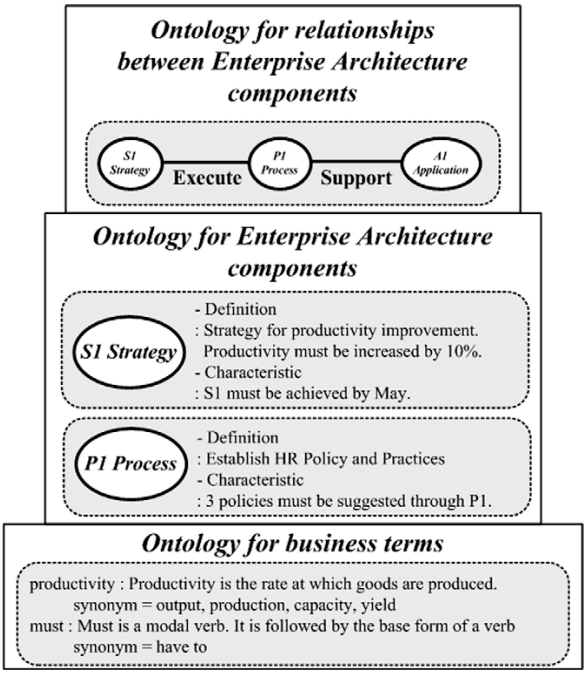


Fig. 1. Levels of Enterprise Architecture ontology

3.2 Ontology Matching for CBP Formulation

3.2.1 CBP Formulation

In order to formulate these BPMO-based processes to store in the BizKB, the BP analysts are required as an important human factor of the system. Based on the analysis on the BPs, the found CBP patterns, level of the abstraction and associate business rules are also extracted and realised.

As described in Fig. 2, extracted artifacts of BPs are modelled using BPMO according to specific domains and kept in the BizKB. This repository is considered as the process feeder for the later stage of the CBP pattern discovery and CBPs formulation.

[3] Business Process Modeling Ontology.

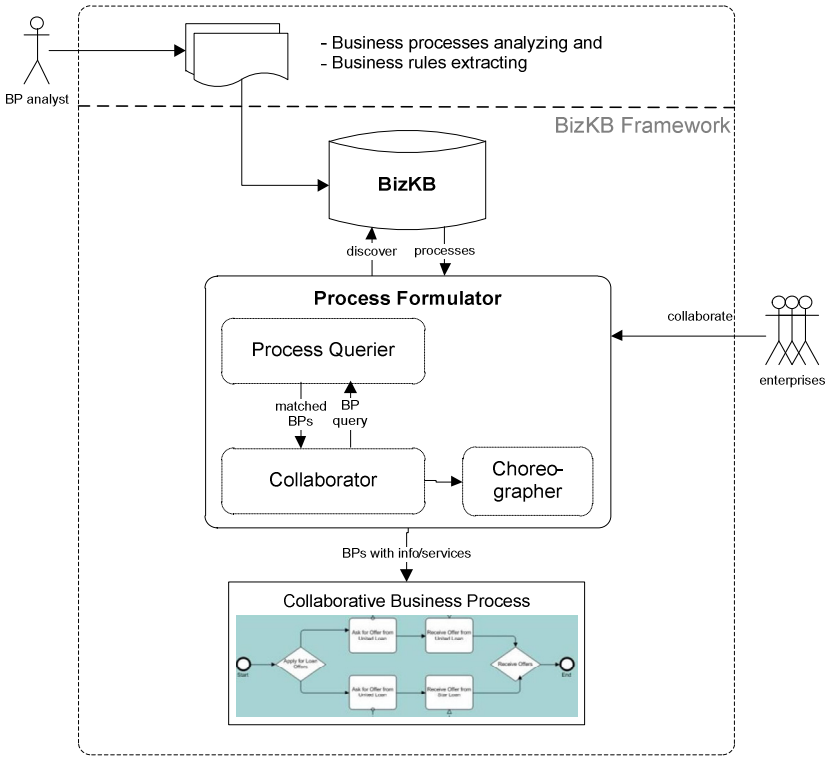


Fig. 2. BizKB architecture

Establishing a complete reference collection as a knowledge base beforehand is very unlikely due to the number of standards, their evolution speed and the cost a complete analysis would create, if it were at all possible. Thus the knowledge base has to be flexible, in the sense that its evolutionary growth is not only possible but also a substantial building criterion. Clearly, an approach that does not start with a fully developed knowledge base shows weaknesses in the starting phase. Due to its initially small knowledge base, references supplied by the system might be erroneous and incomplete. But with the growth of the knowledge base, quality improvement occurs quickly.

From B2B collaboration phases, a comprehensive list of CBP tasks can be modelled in BizKB Ontology (BO). First, the sequences and hierarchies of granular tasks were synthesised into the three B2B collaboration phases.

BO is a set of ordered compound or primitive task and methods. Compound tasks have one more "*hasMethod*" property since they can be decomposed into primitive tasks that can be performed directly using O-HTN. Each method has a prescription for how to decompose some task into a set of subtasks, with different restrictions that must be satisfied in order for method to be applicable and also various constraints of the subtask and relationship among them.

3.2.2 Ontology Matching for CBP Formulation

In BizKB, we do not focus on research for new approaches for ontology matching algorithm. We use existing ontology matching and alignment algorithms mentioned in [10] and [13] to build an ontology matching framework by integrating matching techniques to create a new effective matching results [13]. The following framework describes the matching mechanism for CBP:

Matching Repository is the repository of ontology matching (OM) artifacts that could be reused and metadata describing their properties. *Ontology Repository* is used to manage input data of the OM process described by ontology metadata. *Rule Repository* is considered as associations of ontologies and matching properties, and used to identify appropriate OM rules for input ontologies.

Matching Engine is responsible of the selection (through rules) and the execution of the OM algorithms according specific input ontologies.

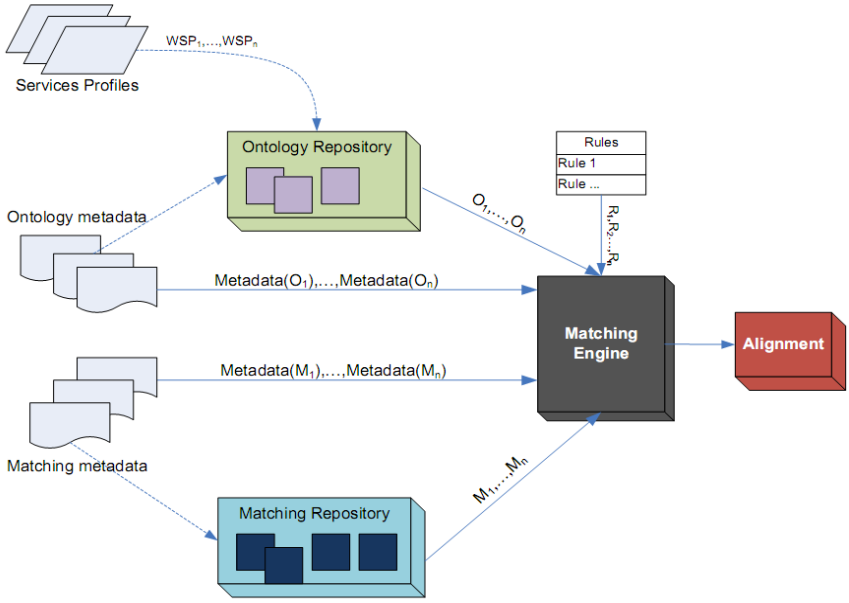


Fig. 3. Ontology matching framework

Metadata (Matching metadata, Ontology metadata) are used to represent the semantics of OM algorithms' properties and ontologies. Based on these metadata, the Matching Engine will automatically compare input value's metadata to constraints of given algorithms along with rule sets built by experts that eliminate the applying inappropriate OM algorithms, and the algorithms not satisfied with attributes of ontologies to be applied for the OM.

For BizKB framework, the Ontology Repository is the BizKB that contains the ontological enterprise profiles are modelled in BPMO. However, the OM-based querying mechanism for CBP formulation in a dynamic manner could have very large querying space, especially when we finding web services to be fitted into the CBP

with service profiles. We cannot limit the scope of domain in on-the-fly CBP formulation attached with web services. The context-aware could help narrow down the scope for querying relevant concepts according to the application domain.

4 Context-Aware Semantic Web Services Discovery

The formed CBP with service profiles has it own semantics described by BizKB artifacts. Concepts for a new CBP generated from BizKB are organised as an ontology. The next step is the discovery phase for appropriate web services that match CBP's service profiles. In order to do so, we have to do a mapping from different ontologies into the CBP ontology-called context ontology. We call this process is the contextualisation of web services into the CBP's conceptual space.

4.1 Concept Contextualisation

Definition 1. A *concept contextualization*, *Con*, in BizKB is a mapping of concept (class) *C* of service ontology O_1, to the context ontology O_2. The relationship between *C* and other concepts in O_2 will be reformed.

$$Con(C):\langle C,O_1\rangle \mapsto \langle C,O_2\rangle$$

The concepts in the CBP context ontology ('context ontology' in short) is still associated to the BizKB artifacts. The contextualisation is realised by applying the mapping mechanism mentioned above.

4.2 Context-Aware O-HTN Service Discovery

Definition 2. The *BizKB Enterprise Context* (BizKB-EC) is a set of concepts from the context ontology linked to the enterprise profiles and associated resources, as well was the properties are in querying action. Let call *U* is a BizKB-EC, we have:

$$U=\langle C,R,P\rangle$$

where *C* is the set of the underlying concepts, *R* is a set of associated resources, and *P* is a set of queried properties.

The O-HTN based architecture [3] for the Process Formulator is described in Fig. 4. User' request is presented in WSMO ontologies and a WSMO Goal. The context-aware mapping process for the goal *G* with *n* enterprise profiles is described as follows:

$$G=OntoMap(Con(P_i,U),U),\quad i=\overline{1,n}$$

where *OntoMap* is the used ontology matching algorithm.

4.3 Context-Aware Service Discovery Framework

Based on [3], in this framework (Fig. 4), the WSMX component uses the discovery component to find web services profiles which have semantic descriptions registered

through their capabilities and interfaces. A set of properties strictly belonging to a goal is defined as non-functional properties of a WSMO goal. A goal may be defined by reusing one or several already-existing goals by means of goal mediators.

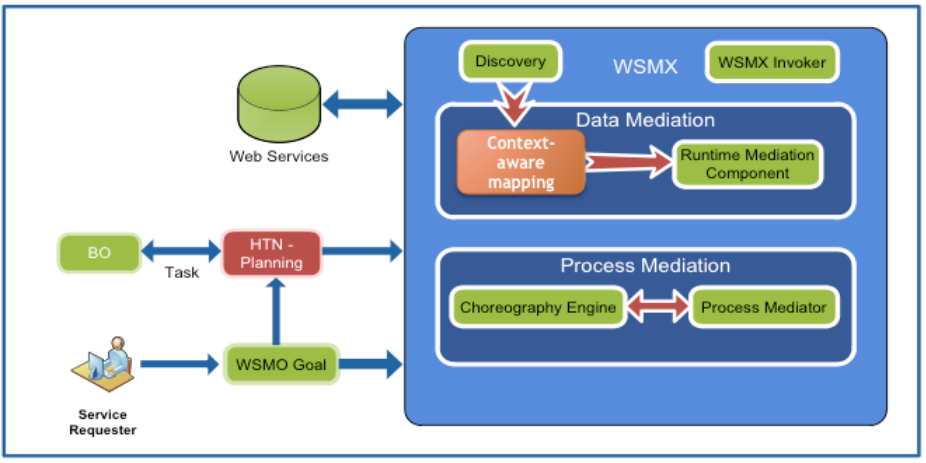


Fig. 4. The O-HTN-based Context- aware process formulator

During the discovery process the users' goal and the web services description may use different ontologies. If this occurs, *Data Mediation* is needed to resolve heterogeneity issues. Once these mappings are registered with WSMX, the runtime data Mediation component can perform automatic mediation between the two ontologies. We apply the *contextualisation process* here to make the service matching more efficient and reduce the mapping and matching spaces according to the enterprise's description model in its profiles. The context-aware approach is the matching process with the target for comparison is the context ontology, that is CBP's service profiles and ontological enterprise profiles.

Every Semantic Web service has a specific choreography that describes the way in which the user should interact with it. This choreography describes semantically the control and data flow of messages the Web Service can exchange. In cases where the choreography of the user and the choreography of the Web Service do not match, process mediation is required. The process Mediation component is WSMX is responsible for resolving mismatches between the choreographies of the user and web service. If there is no single web service that satisfies the request then the request will be offered to the planner.

The planner then tries to combine existing Semantic Web services and generate the process model. In the proposed framework, the process generator is based on HTN-planning with ontological context-awareness. The process generator to tackle the problems of heterogeneous ontologies and choreography uses discovery component of WSMX. Thus via this component, the process generator will be able to discover the appropriate semantic web services for the dynamic cross-enterprise collaboration. Finally the process model with matched services will be transferred to the WSMX for its execution. The stages for execution of Web services as a process model are like as single web services.

5 Conclusion and Outlook

In this paper we have proposed an ontological context-aware approach using Ontological-HTN, "context" ontology and WSMO for forming collaborative business processes in the dynamic cross-enterprise collaboration and service discovery in the process enactment. The approach is motivated by the semantic web approach in efforts of bridging business perspective and IT world together, and provides an architecture that supports the dynamic semantics-based collaborative business process management in a new e-business environment.

This new approach reduces the querying space and helps discover the most appropriate services according the formed CBP and enterprises' profiles in BizKB framework. For the future work, we plan to improve the algorithm and implement with some experiments for benchmarking, especially for the web services discovery with new approach to ontology mapping mechanism and carry out experiments with mapping of attached web services into the execution level with practical examples.

Acknowledgement. This work was generously sponsored by Hue University and Vietnam's National Foundation for Science and Technology Development (NAFOSTED) in the framework of the Grant 102.02-2010.14.

References

1. Hoang, H.H., Le, T.M.: BizKB: A Conceptual Framework for Dynamic Cross-Enterprise Collaboration. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) ICCCI 2009. LNCS, vol. 5796, pp. 401–412. Springer, Heidelberg (2009)
2. Hoang, H.H., Tran, P.-C.T., Le, T.M.: State of the Art of Semantic Business Process Management: An Investigation on Approaches for Business-to-Business Integration. In: Nguyen, N.T., Le, M.T., Świątek, J. (eds.) ACIIDS 2010, Part II. LNCS, vol. 5991, pp. 154–165. Springer, Heidelberg (2010)
3. Hoang, V.M., Hoang, H.H.: An Ontological Approach for Dynamic Cross-Enterprise Collaboration. In: WAINA 2012, pp. 1355–1360. IEEE Computer Society, Fukuoka (2012)
4. Jenz, D.E.: Ontology-Based Business Process Management: The Vision Statement. Jenz & Partner GmbH (2003)
5. Born, M., Drumm, C., Markovic, I., Weber, I.: SUPER - Raising Business Process Management Back to the Business Level. ERCIM News 70, 43–44 (2007)
6. Dimitrov, M., Simov, A., Stein, S., Konstantinov, M.: A BPMO Based Semantic Business Process Modelling Environment. In: Workshop on Semantic Business Process and Product Lifecycle Management (SBPM 2007), CEUR-WS, Innsbruck, Austria, vol. 251 (2007)
7. Yan, Z., Cimpian, E., Zaremba, M., Mazzara, M.A.M.M.: BPMO: Semantic Business Process Modeling and WSMO Extension. In: Cimpian, E. (ed.) IEEE International Conference on Web Services, ICWS 2007, pp. 1185–1186 (2007)
8. Haller, A., Marmolowski, M., Gaaloul, W., Oren, E., Sapkota, B., Hauswirth, M.: From Workflow Models to Executable Web Service Interfaces, pp. 131–140. IEEE Computer Society (2009)

9. Ko, R., Jusuf, A., Lee, S.: Genesis Dynamic Collaborative Business Process Formulation Based on Business Goals and Criteria. In: World Conference on Web Services - I, pp. 123–129. IEEE Computer Society (2009)
10. Jung, J.J.: Semantic business process integration based on ontology alignment. Expert Systems with Applications 36, 11013–11020 (2009)
11. Solskinnsbakk, G., Gulla, J.A.: Combining ontological profiles with context in information retrieval. Data & Knowledge Engineering 69, 251–260 (2010)
12. Kang, D., Lee, J., Choi, S., Kim, K.: An ontology-based Enterprise Architecture. Expert Systems with Applications 37, 1456–1464 (2010)
13. Quoc Hoan Mau Nguyen, H.H.H.: Ontology matching appraoches for B2B integration problem. Journal of Science, Hue University 58, 61–75 (2009)

SMAC - Dataflow and Storage Modeling for Remote Personnel Identification in Restricted Areas

Piotr Czekalski and Krzysztof Tokarz

Silesian University of Technology, Gliwice, Poland
{piotr.czekalski,krzysztof.tokarz}@polsl.pl

Abstract. Automated identification of persons staying and working within closed and restricted areas is required for many guarded centers and restricted areas including airports, power plants, railway and sea container terminals, military training grounds, etc. This process is essential to ensure security and support identification of threats over the area to prevent terrorism acts and ensure adequate protection level. In case of large areas and many employers involved, it is necessary to introduce automated identification methods to support or even replace traditional security forces that are usually human guards using optical and infrared vision, also enabling operation in the darkness and heavy weather conditions. This paper represents general assumption for the integrated solution using Global Positioning System (GPS) devices, range radars, communication and software, also contains in-depth description of the database-related part of the system, including dataflow model and underlying Database Management System (DBMS) design as a part of the integrated "Friend" or "Foe" (IFF) identification solution.

Keywords: security, identification, threats, databases, GPS, radar, restricted areas, access control, DBMS, IFF.

1 Introduction

An increasing insistence to the security has been observed over last two decades. Growing terrorism risk on one hand and requirement of protecting private and public property against vandalism on the other leads to increased demand for automated and semi-automated protection, identification and authorization techniques and systems, particularly for restricted areas. While it is common that some sensitive zones are protected by guards and systems that forbid enemies to enter area at all, in case of large, outdoor expanses it is difficult to ensure limited access. It is also common that some humans are supposed to operate there thus the problem of authorization of the persons is essential here. The examples of such areas are (among others) airports, sea container terminals, military training grounds, power plants and industrial systems. It is also common that presence of unauthorized personnel, whether having designs against on someone or accidentally entered restricted area, may lead to large scale hazards. Complex hardware and software solution may be helpful here to provide authorization systems that help or even replace human guards and human based

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 519–528, 2013.

monitoring and security systems, with means of remote identification and location using integrated system constituted of personal devices with GPS receivers and radio transmission (UAIs), microwave radar and software solution, altogether. Such system is capable to operate in heavy weather conditions and in the darkness.

This paper is intended to provide partial description, related to the data layer and database oriented part of the project. Presentation of the full SMAC integrated solution is out of scope of this paper because of its limited volume and SMAC system complexity. A brief overview has been presented in the following chapter.

The SMAC system has been invented and implemented by the team of researchers working in Silesian University of Technology, Gliwice, Poland, Faculty of Automatic Control, Electronics and Computer Science and Faculty of Mechanical Engineering. The research leading to these results has received funding National Centre for Research and Development in the frame of Project Contract No 0133/R/T00/2010/12.

2 SMAC System Overview

The SMAC system is using microwave radars to find, enumerate and localize objects (possible threats) that are visible within radars range (objects have to be moving) and provide their longitude, latitude and bearing to the data storage. Every member of the authorized personnel carries GPS receiver with communication device (UAI) that is able to communicate with the system on-demand of the base station (BS) and provide its location specified by longitude, latitude, altitude and unique identifier. Once knowing possible threats identified by radar scans, the SMAC system is requesting positions from the UAIs then juxtaposes them to the enumerated objects, thus identifying them as enemies or authorized ones, so called “Foe” or “Friend”. This simple approach has many disadvantages, however. Modern microwave, general purpose radar devices are capable to deliver enumerated objects up to 1Hz [6],[12] as well as GPS receivers [10]. That leads to frequent requests sent to UAIs, communication bottlenecks for crowded areas and high power consumption in mobile UAI devices. This is a point when modern, heuristics based software comes in hand. Using stored data and movement prediction, implemented software that constitutes part of the SMAC system is capable to identify objects and track them without necessity of UAIs requests every radar notification. Regular GPS receiver chips are able to locate with 3m or better accuracy (RMS) [8],[9],[10],[11], so radar devices can, even up to 0.25m [6],[12]. A time series organization of the location data is required with separated data streams to enable synchronization among sources. As conjunction of the data streams of many origins, coming from the data sources, have to be synchronized over time, every item within each data stream should contain timestamp to enable reliable correlation among data sources with operating error no greater than 1s. It is achieved by design, because both UAIs and radar devices are equipped with GPS receivers capable to set RTCs of the devices or at least with NTP enabled network interfaces, accurate enough to provide location data that requires very accurate RTC synchronization [11]. The exact consideration on time-related data modeling is presented in chapter 3.

SMAC system is constituted of the following components (see **Fig. 1**):

- BSs including radar devices and proxy communication modules for UAI devices, as required for restricted area coverage.
- UAI gateways enumerating entrances / exits of the personnel to / from the restricted area.
- Graphical clients with threat identification and appropriate software capable to identify authorized and enemy objects.
- SQL relational database and data storage along with database management software and configuration tools.

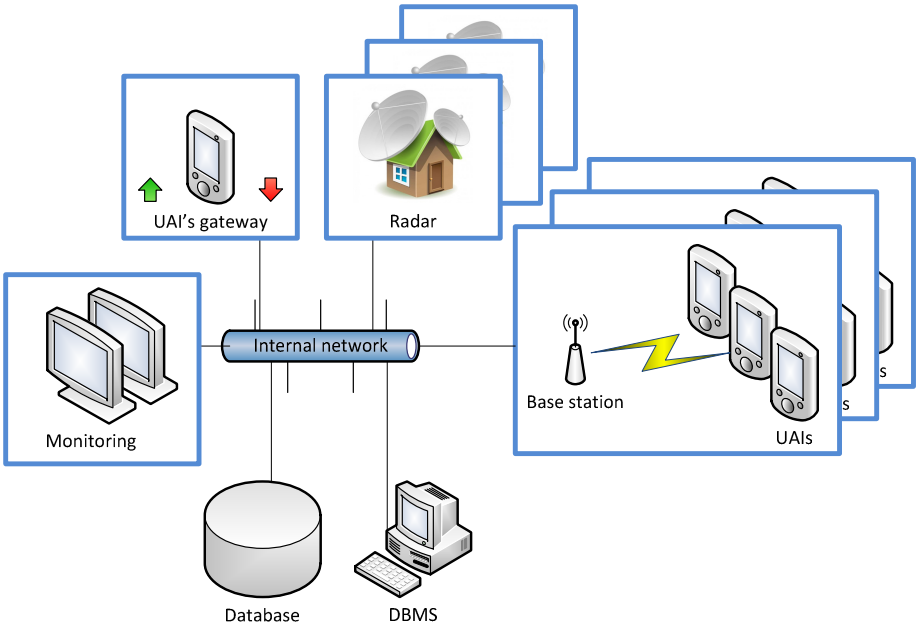


Fig. 1. SMAC system components

This paper describes in depth the fourth on the list above, with respect of the whole SMAC system dataflow. Data-oriented features of the SMAC system that played key role in database and management software design are presented below:

- enabling software solution to perform tracking and enemy recognition of objects identified by surveillance radars and UAIs,
- possibility to playback tracks of the objects (historical and current) by storing large volume of location data (UAIs tracks),
- UAIs management and attachment of the device to the personnel,
- identification of active and inactive devices i.e. corresponding to shifts, using automated gate for entrance / exit to the restricted area,

- technical configuration of the system,
- users authorization,
- integrated security solution that minimizes possibility of database compromise even, if end user credentials are compromised.

The requirements above lead to selection of the relational database as an underlying for the data and configuration storage. In case of the SMAC system, the MS SQL Server 2010 database was chosen, however almost any relational database engine may be applied, as system was designed to be aware of specific server-side features using regular DDL, DCL and DML[1], thus being compatible with SQL:2003 standard [5]. A front side management application was implemented using Microsoft .NET Framework 4.0 and WPF thick client model. As central database engine is located on the physically and logically separate OSs, all software modules constituting SMAC system connect to database server using TCP/IP protocol, secured by SSL. The detailed description of dataflow is presented in chapter 3, while security related issues are described in chapter 5.

3 SMAC Dataflow Model

As general purpose of the application is to identify authorized and enemy persons and objects (Friends and Foes), the main data sources delivering large volumes of the data are radar devices and UAIs. Those devices bring XML data on objects found by the radar echo every 1s. This data is sent from the radar devices o the tracking software as XML file of proprietary format, similar to ICD-0100 communication standards [10], [6]. Every transferred file contains a header, list of objects detected by radar device, their location given by absolute and relative coordinates as well as heading. On the other hand tracking application is following objects and classifying them as Friends and Foes by juxtaposing object location given by radar devices and information from the UAIs location delivered through BS. To construct such system that enables possibility to efficiently juxtaposing locations of two sources in appropriate time fashion, two approaches were subject of research:

1. database centric approach, where each data source loads information into the SQL database, while tracking application is reading those data from the database,
2. memory centric approach, where current and about 5 second history of each data source is stored in temporary memory cache, FIFO organized, then asynchronously stored in the SQL database.

A storage stress tests were performed to estimate possibility of database centric approach as considered to be simpler in implementation and more flexible. Assuming typical scenario where about 50 objects is identified as Friends and 10 as Foes, giving total of about 60 objects in an XML file and storage interval is 1s, the table below represent achieved results in one, two and three radar data source configurations (typical case), loading data parallel to the MS SQL 2010 database:

[1] DDL, DCL and DML stand for SQL Data Definition Language, Data Control Language and Data Manipulation Language.

Table 1. Storage stress test results for 60 objects pack, 1s interval

	1 radar device	**2 radar devices**	**3 radar devices**
Avg. storage time	71 ms	72 ms	76 ms
Std. dev. of above	47 ms	23 ms	48 ms

Stress tests performed indicate that storing radar data directly into the SQL database for feature processing showed unacceptably large standard deviation that may lead to the bottlenecks on database operation (see **Table 1** for details). This leaded to memory centric approach (see **Fig. 2**) that was chosen to implement SMAC system, with separated SQL storage channel using dedicated FIFO queue and special, efficient form of data storage (see chapter 4 for reference)[1]. SQL database also contains the list of devices and their attachment to the personnel as well as configuration information used by software modules and hardware components, including IP addresses of BSs and radar devices. The following figure present conceptual dataflow model for the SMAC system using UML dataflow diagram [7]:

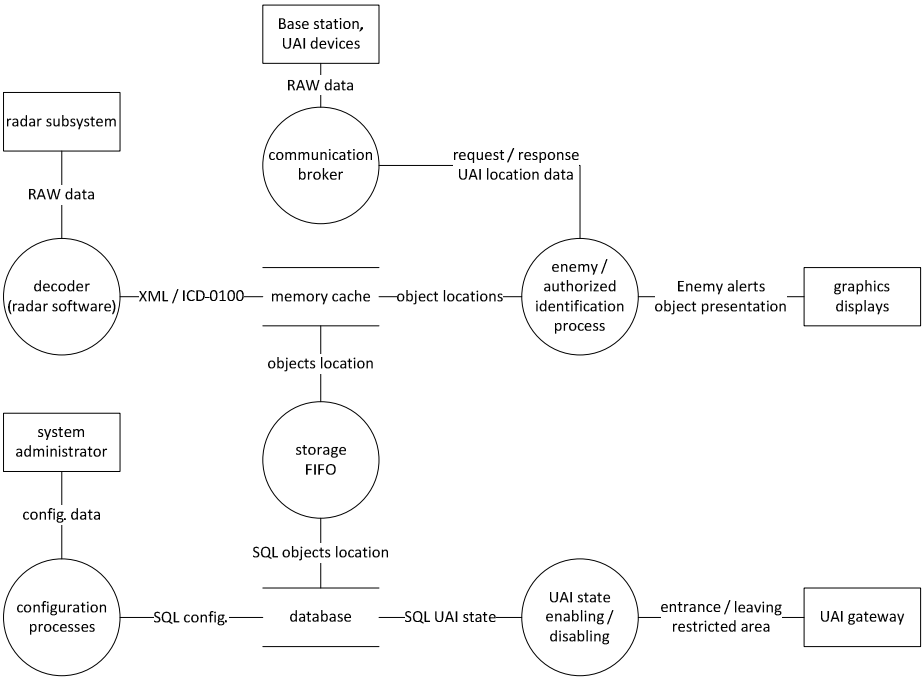


Fig. 2. SMAC system dataflow

The final enemy / authorized state is presented on graphical terminals. The exact presentation of the Friend and Foe classifier algorithm is above scope of this paper.

4 Data Structures

Underlying database contains four main groups of tables and dedicated indexes (see **Fig. 3** for physical data model) to provide rapid searching capabilities [3]:

- devices configuration and dictionaries,
- current active configuration (UAIs, on and off) including workers,
- trace storage,
- security model.

The device data configuration and data dictionaries constitute regular, relational database tables, so current active configuration related tables does. `Devices` table is updated by entrance and exit of workers carrying UAIs to/from the restricted area. This constitute the state machine, presenting current authorized UAIs set. This state can be modified using administrator's application as well. `Workers` table contains regular staff identification data and photos of personnel, enabling presentation of avatars on graphical terminals. Workers are grouped into categories by `Groups` table.

Security model tables are implemented similarly to the .NET WEB provider security model [2] and constituted by two entities: `Users` and `Roles`. More information on security system applied one can find in chapter 5. The special attention was paid to the trace storage system as it should be capable to store and process large volumes of data. Data volume has been measured for MS SQL 2010 Standard Server. Assuming 50 UAIs, 10 enemy objects and 1s update rate, SMAC system requires about 700MB of storage space for trace logs (including both data and indexes) every day. Trace data is stored up to 1 month, constituting rotating registry. After reaching last day of the month it is necessary to clean data for first day of the next month. Deleting relatively large amount of data causes extra load on database. This usually leads to bottlenecks on storage system [3] when performing `delete` on records limited by `where` clause. On the other hand using `truncate` operation is much faster, but requires separate entities one for each day. During the design of trace storage system approaches where considered:

- one table and high efficiency storage channel,
- one table, partitioned, one partition for each day,
- 31 tables, one for each day, stored in separate files.

The first approach was abandoned because of bottlenecks on `delete` operation, inability to use `truncate` operation and high cost of storage devices. The second approach was abandoned as causing bottlenecks on `delete`, non-standard DDL syntax and inability to `truncate` table. Moreover, deleting data using `delete` SQL command frequently causes file system fragmentation that lowers performance of the storage channel over the time. Last model was chosen for implementation as brings flexibility, lack of bottlenecks, as every table and indexes related are using

separate file store, ensures DDL standards and enables usage of `truncate` operation on table level. Those tables are `HSRSTrace1` through `HSRSTrace31`. `Truncate` operation is executed using stored procedure, scheduled for daily run. Its purpose is to clean trace data in advance for feature re-usage of space.

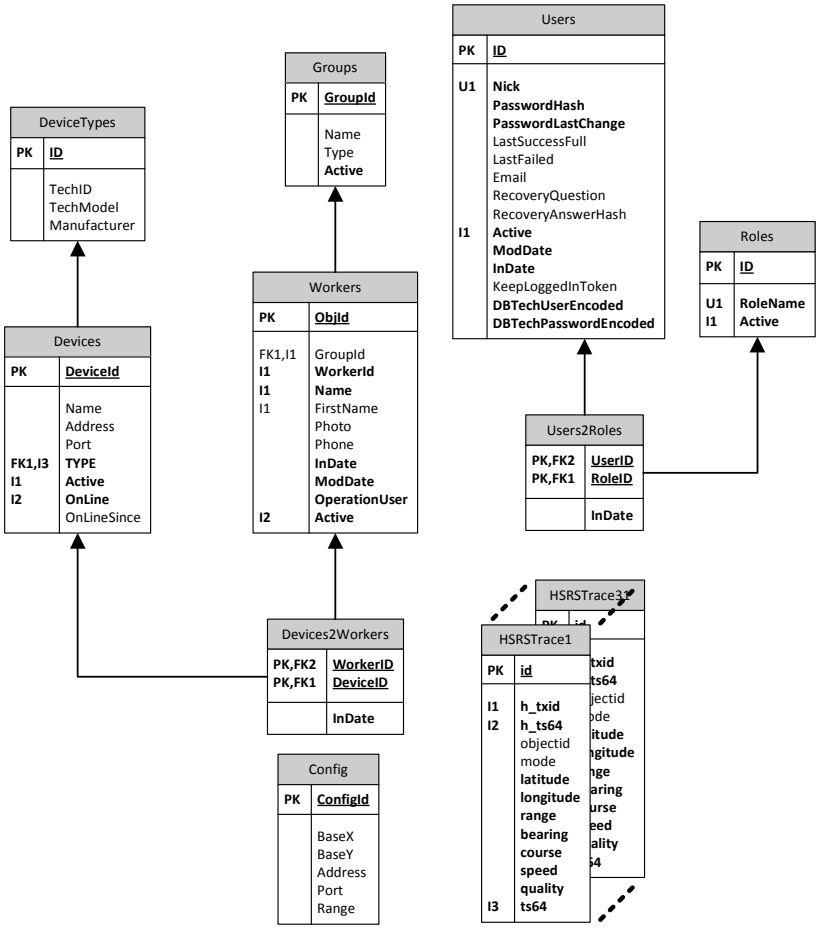


Fig. 3. Database physical model

5 Security

Whenever sensitive data processing and storage is performed, special level of the security has to be introduced. In case of SMAC system there is a bunch of applications connecting to the database. Compromising of the database access may lead to unidentified threats i.e. by enabling invalid UAI that should not appear in the restricted area at a moment or that was stolen or canceled from the authorized devices list. It is common that secure communication almost ensures inability to compromise

logon credentials. The most insecure part of the IT systems are operators. Most thick client applications require providing credentials that are used directly for underlying database connection as connection credentials [2]. This is insecure approach, because once compromising user credentials one may gain access to the SQL database as well, usually with administrators or at least with wide privileges. On the other hand most WEB based systems are secured using security provider model [2], where user provides some sort of login and password that is verified in the security provider's store, while using some other technical credentials to connect to the underlying database. It is secure, because in case of WEB applications user is usually physically and logically separated from the application server and the underlying SQL database server as well, thus has no access to the technical connection credential settings.

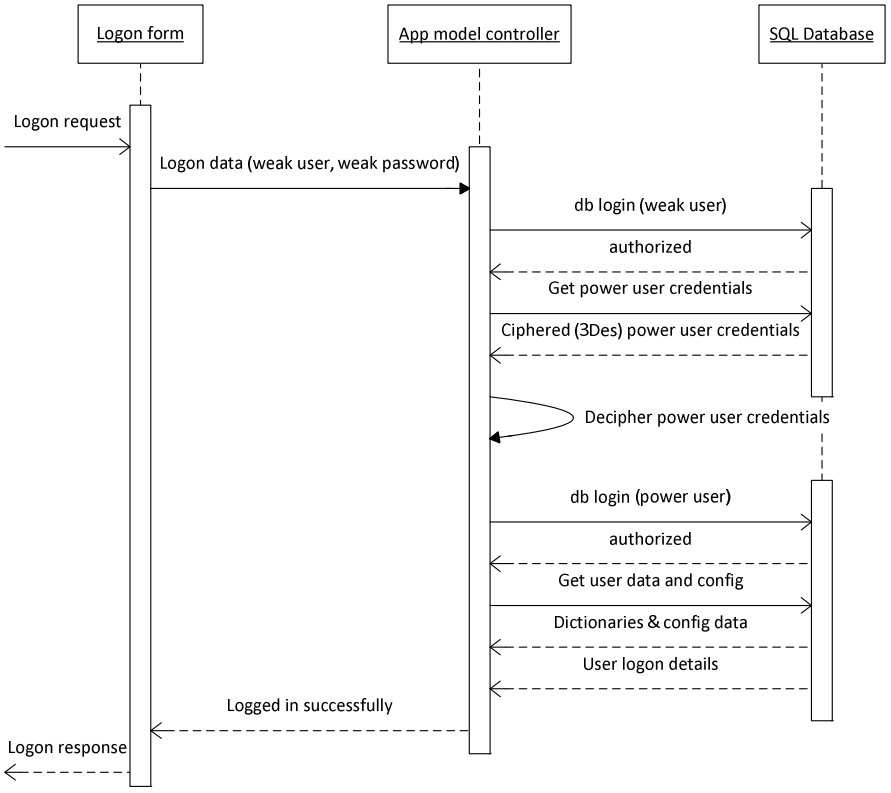


Fig. 4. SMAC security – logon model using mixed approach

A wide research on security model was provided when designing SMAC system. The WEB model approach, where user logs on to the application using logon credentials that do not have wide and high privileges on the database side (or has no connection privilege at all) is a good choice, however leads to the necessity of using another, higher privileges credentials to operate connection to the SQL server. Including fixed,

technical credentials within the software (embedded or stored as configuration parameter) as in case of the WEB application server based solutions is considered to be insecure as well in case of SMAC system, because those credentials are stored along with thick client application that leads to compromise opportunity. SMAC system is using smart mixture of both – classical and WEB-based approach. User provides WU (Weak User) logon credentials that enable application to verify credentials along with database user's table (see chapter 4 for structure). Those credentials are also used to preliminarily connect to the SQL database with lowest possible privileges, particularly reading only `Users` table in a limited range. The next step is performed to obtain ciphered, high privilege PU (Power User) credentials from the `Users` table. Application then reconnects to obtain full database access using those credentials and works using them until logout or timeout occurs. The advantage of this solution is that user does not even know the high privilege technical PU credential details, thus being unable to compromise them. As WEB based model also brings roles table along with users (using role provider [2]), it is not necessary to apply database related role model. The PU is granted operations on all tables. Separate role provider is limiting user access to the function on the code level. This simplifies database exception handling as special security exceptions does not need to be handled and explained to the user. WU has read privileges on only one table. This way it may obtain no more than ciphered credentials for PU (i.e. 3DES ciphering). This data, particularly password is subject of automated generation thus is unknown for users and typical cracking methods i.e. those using dictionary approach won't be helpful here. Moreover, password is 32 characters long that discourages potential compromising trials. This settle the approach to be more secure than most of the recommended approaches for thick client applications.

6 Summary

During development of SMAC system some problems connected with flow and storage of large amount of data and ensuring the database efficiency and security had to be resolved. Research has been done for selection of proper approach for storing track data from radars and UAIs into the database. Method of storing the data in separate files for tables of every day of the month has been selected as ensuring flexibility and eliminating the bottleneck of SQL `delete` method as well as leading to standard approach to the creation and management of the data structures. To ensure security of database the thick client application that is executed for CU is allowed to read the data from database only in very limited range. If user is allowed to have full access to the database its PU credentials are transmitted from database in ciphered form, using re-logon procedure, ensuring inability of compromise credential data and simplify database operation.

References

1. Agrawal, S., Narasayya, V., Yang, B.: Integrating vertical and horizontal partitioning into automated physical database design. In: Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data (SIGMOD 2004), pp. 359–370. ACM (2004)
2. Baier, D.: Developing More-Secure ASP.NET 2.0 Applications (Pro Developer). Microsoft Press (2006)
3. Delaney, K.: Inside Microsoft SQLServer 2005: query tuning and optimization. Microsoft Press (2007)
4. Department of Defense: Joint Gatekeeper System (DRAFT). Supporting communications between Commercial-off-the-Shelf Access Control Systems and ENABLER for identity management data from Department of Defense Authoritative Databases. Interface Control Document. SEIWG, https://www.siaonline.org/content.aspx?id=5938
5. Eisenberg, A., Kulkarni, K., Melton, J., Michels, J.E., Zemke, F.: SQL:2003 Has Been Published. SIGMOD Record 33(1) (2004)
6. FLIR Systems Inc.: Mid-Range Perimeter Surveillance Radars, http://gs.flir.com/uploads/file/products/brochures/midrangeradars.pdf
7. Gane, C., Sarson, T.: Structured systems analysis: tools and techniques.Prentice-Hall (1979)
8. Guo-Shing, H.: Application of the Vehicle Navigation via GPS Carrier Phase. In: Proc. of the 6th WSEAS Int. Conf. on Robotics, Control and Manufacturing Technology, Hangzhou, China, pp. 218–223 (2006)
9. Guo-Shing, H.: Control the Vehicle Flow via GPS Monitor Center. In: Proc. of the 6th WSEAS Int. Conf. on Signal Processing, Computational Geometry & Artificial Vision, Elounda, Greece, pp. 174–181 (2006)
10. GlobalTop Technology Inc.: FGPMMOPA6C GPS Standalone Module Data SheetRevision: V0A, http://download.maritex.com.pl/pdfs/wi/FGPMMOPA6C.pdf
11. Herb, L., Paduch, J., Tokarz, K.: Influence of Receiver Parameters on GPS Navigation Accuracy. In: Czachórski, T., Kozielski, S., Stańczyk, U. (eds.) Man-Machine Interactions 2. AISC, vol. 103, pp. 85–93. Springer, Heidelberg (2011)
12. Navtech Radar Ltd, AdvanceGuard W500 and W500-X Features, http://www.navtechradar.com/Documents/SecuritySystems/W500Datasheet.pdf

Infrastructure vs. Access Competition in NGNs

João Paulo Ribeiro Pereira

School of Technology and Management, Polytechnic Institute of Bragança (IPB), Portugal
jprp@ipb.pt

Abstract. With the introduction of NGNs, operators need to upgrade their access networks because in several cases, existing access networks can no longer meet increasing customer expectations. Evolving consumer expectations will require changes to the existing access network – next generation access. However, existing technologies faces some difficulties and are not ready for large-scale roll-out yet. For example, in the case of DSL technologies, the great majority of operators with copper networks are improving their networks, making investments to deploy fiber optics closer to customers and offering higher-speed access, which is required for new emerging services (reducing the distance between fiber and the users.). The entry of new competitors can be based on the resale of services from the incumbent, on building up their own infrastructures, on renting unbundled infrastructure from incumbents, or, on the combination of the above elements. Then, is important create the right incentive for operators to make an efficient build/buy choice and define the appropriate pricing principles.

Keywords: NGNs, Broadband Access Networks, Telecommunication network operators, policy and regulation.

1 Introduction

The advent of NGN (new network technologies, access infrastructures, and even services) has changed the concept of telecommunication networks and has profound implications for operators and regulators. The definition of policies and regulations for competition in the access networks constitute one of the most debated issues in telecommunications today. The regulation of telecommunications networks and services is seen as a necessary requirement in most countries to meet government objectives and to ensure public interest. Regulation is fundamental to generate positive welfare effects where markets alone would not tend to perfect competition.

But, as referred by [1], the major problem is how to measure these welfare effects, as they can occur as consumer surplus, producer surplus, societal gains (e.g., increased tax income, better working conditions, etc.). Their empirical study uses price situation to examine the welfare effects measured by the state of competition. They assumed that the increase of competition reduces prices in the market and that competition can also increase consumer welfare without reducing prices (achieved by

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 529–538, 2013.

innovation). Public policies should promote an efficient investment and competition in all markets (see fig. 1).

The two main economic reasons that have been used to justify interventions in access networks are the beliefs that access networks constitute a natural monopoly for which competition is not feasible in principle and that regulation is, therefore, necessary to control monopoly power and to achieve universal service in which all (or most) users have the opportunity to affordably access the services of the network. The challenge of telecom operators to provide a profitable deployment of broadband services depends if is a high or a low competition area. In areas with high competition already exists competition between broadband network operators, and the main question is know the market share of all intervenient. However, in low competition areas high investments cost must be incurred to promote broadband. [2] argued that national or regional policy concerns can also affect NGA roll out. Without some type of intervention, there is the risk for a new digital divide, with urban customers on short loops being able to receive IPTV/multi-media services and HDTV while those in rural areas might not be able to receive such services. Therefore, the access network poses serious challenges to the regulator [3].

The question then becomes whether it is more important to stimulate investment or to ensure competition. Investment in network quality is important for consumers because it provides access to both better quality and speed to services, such as Web browsing and email, and services that require more bandwidth, such as video. Investment in network quality also improves the service value for consumers and attracts new consumers to the market. Therefore, there are two major options for access regulation [2]: temporary or permanent deregulation (i.e., the removal of sector-specific rules and regulations) or mandated access (i.e., the obligation to grant access to bottleneck facilities at a regulated price and quality). Deregulation increases investment incentives, as it overcomes the "truncating problem" and allows above-normal profits. However, in the absence of alternative infrastructures or in areas of low population density under limited competition or the threat of entry into the upstream market, an integrated incumbent might leverage its market power to competitive downstream segments.

For NRAs, one request of decisive importance is if they must foster service-based competition in the first phase of liberalization or to focus on infrastructure-based competition. This decision (infrastructure or service-based) would lead to lower prices, more differentiated and innovative products and improved services for consumers. When access is available at different levels of the incumbent's network, new entrants will be able invest in the infrastructure gradually as sufficient economies of scale became achievable - This concept is the ladder of infrastructure competition. This concept defends that new entrants (or access seekers) may enter the market offering broadband access by reselling the wholesale services of the incumbent operator (requires least investment) where they only cover minor elements of the value chain (Figure 2). When the number of customer grows and financial means become available, the operator move on to higher rungs of the ladder [1, 4]. Next, new entrants need to building their own infrastructure and acquiring only the residual infrastructure from the incumbent's wholesale department. This includes a move for the operators from service to infrastructure-based competition.

The migration to NGAN has raised a range of issues related to building wiring and infrastructure sharing. The deployment strategies for operators and entrants are completely different. In addition, parameters, such as existent infrastructure, geographical characteristics, infrastructure renting costs, and consumer willingness to pay, influence the definition of the strategy. So, telecommunication operators can select among a set of deployment strategies that are characterized by path dependency and diminishing usage of the legacy copper loop. The range of the selection space is based upon how much of the copper they use and, consequently, how far toward the customer they deploy new fiber. In the final step, operators replace all of the copper with FTTH. Within that scenario, FTTH can be implemented as either active Ethernet or passive optical networks, although most incumbent operators tend to select PON.

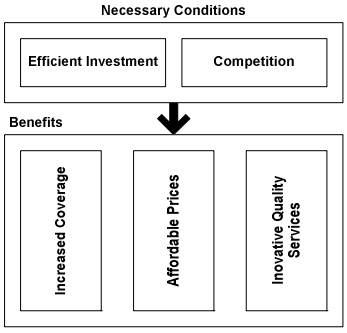


Fig. 1. Policies effects [5]

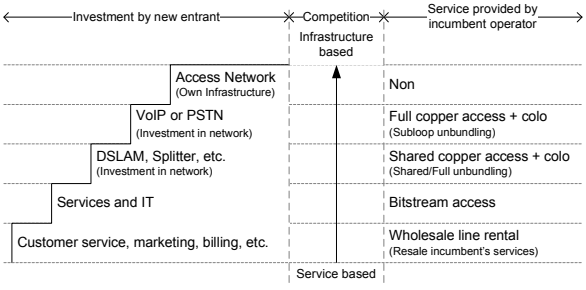


Fig. 2. Ladder of investment [6, 7]

The strategy of a new entrant in an access network that does not have an infrastructure can be one of the following three alternatives (Fig. 3): (1) Renting infrastructure (i.e. conduit, cable, equipment, ...) from other operators and offering only services (infrastructure sharing); (2) Deploying a new infrastructure; or (3) Not participating at all.

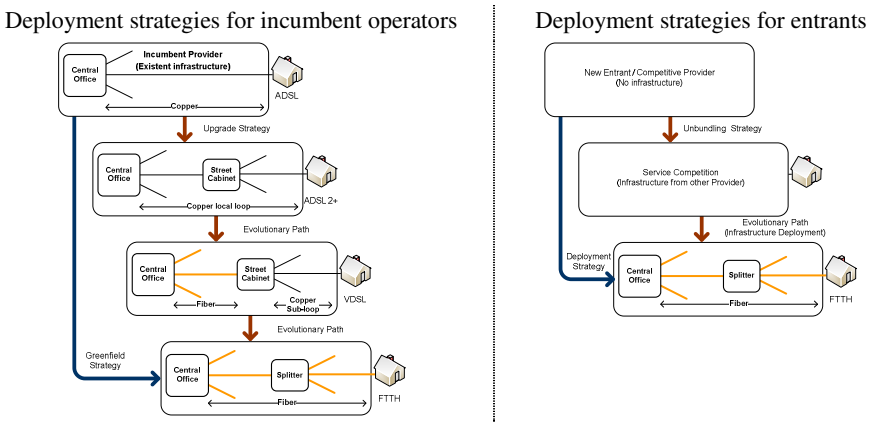


Fig. 3. Deployment strategies for incumbent operators and new entrants [8, 9]

Regulators must decide whether to promote competition on the basis of a single infrastructure with regulated access (service competition) or to encourage the build-up of competing, parallel infrastructures (infrastructure competition) [10]. Then, is important create the right incentive for operators to make an efficient build/buy choice and define the appropriate pricing principles. To obtain economic efficiency, a regulator should [6]: (1) Encourage the use of existing infrastructure of the incumbent operator where this is economically desirable, avoiding inefficient duplication of infrastructure costs by new entrants (incentive to buy); and (2) Encourage investment in new infrastructure where this is economically justified by (a) new entrants investing in competing infrastructure, and (b) the incumbent operator upgrading and expanding its networks (incentive to build).

In this context, the cost models are fundamental in the determination of the access price that can be used by regulators in the definition of wholesale prices.

2 Infrastructure-Based Competition

European Commission argues that infrastructure-based competition is the best and fastest way for broadband development. The arguments are that infrastructure based competition provides efficiency incentives to operators, reduces prices, increase penetration, stimulates innovation, etc. The empirical study deployed by [1] concludes that prices are lower and penetration rates are higher in those countries with predominantly infrastructure-based competition. However, broadband development and infrastructure-based competition has higher costs (the cost of laying out these infrastructures and operation inefficiencies of duplicating/redundant infrastructures).

The entry of competitors in the infrastructure-based market is dependent on the cost of the alternative technology. An efficient market entry is based in retail rates and access prices (reflect the cost of using the incumbent's network). So, infrastructure-based entrants can offer differentiated services at equal (or lower) costs (and thereby increase consumer welfare), when the cost of providing broadband access by an alternative network is lower (or equal) than the incumbent's cost of providing broadband access services [11]. The share of LLU is increasing in favor of bitstream

access. This could be understood that the competition in the lowest possible layers allows more degrees of freedom to differentiate. However, competition in this layers could not be not technically feasible (e.g., constraints in ducts or sewers implies that fiber cables cannot be installed) or economically feasible (e.g., in rural areas where the business case for FTTH isn't attractive).

2.1 Build or Upgrade Infrastructure

In NGNs, infrastructure-based entry into the local loop can occur in two ways: by constructing new networks (greenfield approach) or by upgrading existing networks [12]. Upgrading existing networks can be driven by the increase in number of subscribers, the introduction of new services, the conversion to broadband access infrastructure, or modernization of existing access solutions using different access technology than previously installed.

With the exception of wireless networks, which constitute a low-cost investment, the construction of parallel infrastructures that are similar in costs and capabilities remains unlikely in NGNs. However, providers can enter the market successfully by deploying a superior network with lower costs and/or higher quality alongside an existing infrastructure. In many European countries, next generation access will most commonly use FTTCabinet architecture. The replacement of the traditional copper-based access with new fiber-based access potentially both enables a significant increase in the capacity and ability to support new services and lowers operating and maintenance costs as compared to copper [13]. However, the initial cost of deploying it is substantial. Alternative operators require access to civil engineering. For an operator rolling out a fiber optical network, access to existing civil engineering changes the economic equation considerably. Therefore, all operators are not on equal footing. Alternative operators have begun to introduce optical fiber only in large markets. In addition, incumbent operators are rolling out optical fiber in their civil engineering ducts, which they inherited from the former monopoly.

NGA fiber rollout requires substantial investment, and incumbents are better positioned than new entrants to make these investments on a large scale because of the associated lower costs of infrastructure usage, investment savings by dismantling MDFs, better use of passive infrastructure, and larger subscriber base [5]. The technologies used by most, but not all, network operators are FTTC/VDSL and FTTH/B (P2P and PON). The FTTC/VDSL solution uses copper access line from the network operator's MDF in the CO, to the street cabinet is replaced by a fiber optic line. In FTTH PtP networks, the copper access line from the network operator's CO to the end user's residence is replaced by a fiber optic line that is effectively dedicated to a single customer. In FTTH (PON) the copper access line from the network operator's CO to the end user's residence is replaced by a fiber optic line.

3 Service-Based Competition

The large investments required to build/upgrade a network capable of supporting broadband access services could be a barrier to enter into the market. [14] argues that service-based competition is only viable with regulatory intervention in the market. National Regulatory Authorities (NRA) intervention is important for regulation costly

and to offer the right incentives for innovation and cost-minimization. For that, the regulators need correct information on technologies, costs, market, etc.

By NRA definition, incumbent networks were opened to facilitate competition. With this policy, new entrants could rent the elements of the incumbents' networks with stipulated prices, which enabled competitors to utilize last-mile incumbent network facilities at cost-oriented prices under the following arrangements [15-16]: Bitstream access, Local loop unbundling (LLU), and shared access. The report from the ERG [7] identified LLU and wholesale broadband access (Bitstream) as key areas where harmonization might significantly help deliver benefits of a single electronic communications market.

Mandated access to the bitstream or the unbundled line reduces uncertainty and protects competition in the downstream market, although the effects on investment depend upon the allowed margin. Further, regulated access to cable ducts can help competitors to deploy, restructure, or upgrade their access infrastructure. In NGNs, physical unbundling becomes increasingly difficult with the rollout of FTTH deployments, as current points of interconnection, such as the MDF or the street cabinets, become obsolete and are phased out. In the case of FTTH, investments by competitors to interconnect physical access points in the local loop might ultimately be stranded.

4 Wholesale Access Services

Some operators have decided to share their networks under several options in order to reduce their CAPEX and OPEX (e.g., ducts, fiber, site sharing, etc.). For new entrants, is possible without higher investments in local loops, reduce the risks of entry, which leads in the long run to more investment in alternative access network infrastructures [17]. Table 1 presents the types of wholesale access.

Table 1. Types of wholesale access (Summary)

Type	Description
Local loop unbundling (LLU) or unbundled local loop (ULL)	Is the recommendation by which incumbent operators are obliged to open their copper-based line access networks other operators (new entrants). In the case of full unbundling, a copper pair is rented to a third party for its exclusive use. This access method uses the incumbent's copper access line from the CO to the customer premises, but the LLU operator provides its own electronics at the CO to enable the local loop for broadband [18].
Shared access lines	Shared access lines supplied by the incumbent to other operators (new entrants). A fundamental feature of shared access is that it is provided over a subset of the full frequency spectrum of the line (copper, coax, fiber, …). Cable is a separate access network generally not owned by the incumbent operator, except in Portugal [189]. The cost of the line is shared between the shared access services.
Bitstream access	It refers to the situation where the incumbent installs a high-speed access link to the customer premises, and makes this access link available to new entrants, to enable them to provide high-speed services to customers [16]. Bitstream access is a wholesale access product which allows alternative operators to offer BB Internet access to the final costumer without having an own access line [19].

4.1 Wholesale Pricing

One of the variables that can influence the level of competitiveness is the access price definition. Access prices are essentially the wholesale prices that network owners (normally incumbent operators) charge to the other operators (competitors) for access to infrastructure or services provided by the incumbent network infrastructure.

Where the incumbent's network is opened to competitors at more than one level (e.g. LLU, wholesale broadband access and wholesale line rental), regulators have to project correctly the relative prices of the different options in relation to one another and in relation to the retail prices prevailing in the market [20]. An incorrect definition of the access price can affect the level of competition and welfare loss. For example, a lower price will reduce the incentive of the infrastructure owners to invest in the infrastructure because they do not receive an adequate return on their investment. Further, a higher price will reduce the capacity of competitors to compete - a low price for access encourages use but discourages investment and vice versa.

To be able to make the ladder of investment operational it is necessary that prices in wholesale markets are consistent for the different products [1]. The definition of access pricing is critical when network access is a vital input to deploying services to end consumers. In a market where the network owner also operates at the retail market, competing with other firms the access pricing definition is a key question and is frequently considered to be an economic regulation problem. To assure a legal competition, the NRAs have to control the price scheme of the wholesale regimen for the bitstream access services or even determine the upper price levels. The EU has to provide the directives to promote homogeneity in the price regulation scheme.

Without access price regulation, network owners may be tempted to exercise market power and block access to the network. On the other hand, significant control of the access price can discourage the realization of investments in the quality of the network. Consequently, a lower access price can result in inefficient input, whereas a higher access price can lead to inefficient investments with the objective of reducing dependence upon the incumbent operator. Then, the regulator can use the access pricing definition to influence the decisions of new firms' inputs in the retail segment.

In Portugal, a recent European Commission report [21] shows that LLU prices have not been modified and continue to be just above the EU average. The monthly average total cost was €10.05 for full unbundling and €3.57 for shared access in October 2010 (compared to EU averages of €9.61 and €3.29 respectively).

4.2 Wholesale Access Market

For new entrants, LLU (fully unbundled lines and shared access) is the main wholesale access with 76.2% (in January 2011) of DSL lines, up from 65.2% in July 2008 (Source: EC). New entrants' use of bitstream access for local loop unbundling in the provision of broadband services went up by 1 percentage point since July 2008.

The share of resale, which represents a type of access for low-investment intensive new entrants, has shrunk by 7.6 % during the last year. Previous figure shows the evolution since 2007 of DSL lines by type of access.

Table 2 presents the types of access of the new entrants for Portugal and EU. In Portugal, LLU continues to be the preferred wholesale option for alternative operators, with 83.4% (2010) of lines. However, the number of fully unbundled lines (shared access is not used) for the provision of broadband services has decreased from 269.066 in January 2010 to 229.098 in January 2011 [21]. Table 2 shows that there is a bigger difference in the use of shared access between Portugal and EU average.

Table 2. New entrants` DSL lines by type of access (EU and Portugal), 2009/11 (Source: EC)

Type of access	EU level			Portugal	
	2009	2010	Jan. 2011	2009	2010
Own network	0,9%	0,9%	1.2%	0,0%	0,0%
Full ULL	54,4%	61,5%	63.9%	86,6%	83,4%
Shared access	16,9%	13,3%	12.3%	0,0%	0,0%
Bitstream	17,0%	15,4%	14.5%	13,4%	16,4%
Resale	10,7%	8,9%	8.1%	0,0%	0,2%

In Portugal, competition is strong in urban areas, where unbundling is highly developed, but in rural areas, incumbent operator remains largely dominant. In 2009, the % of DSL connections was different by type of provider: a) Urban area - ULL:41,0%, Bitstream/Resale:4,0% and Incumbent retail:55,0%; b) Rural area - ULL: 8,5%, Bitstream / Resale: 5,5% and Incumbent retail: 86,0%. [6] defends that an effective and sustainable infrastructure competition is superior to service competition, as it allows for head-to-head competition between operators and requires a minimal need for regulatory intervention with competitors not being reliant on the incumbent infrastructure. So, operators, especially new entrants, will have a choice as to whether they should invest in their own infrastructure (i.e. build) in order to provide services to end-users, or to seek access (buy) from an existing provider (normally the incumbent).

5 Conclusion

The analysis of the broadband market suggests that where infrastructure competition exists, as in DSL and wireless broadband, service providers will more aggressively price their offerings, driving down the access price for consumers. However, in the case of limited infrastructure competition, broadband access price remains high for consumers. Infrastructure competition between DSL, Cable and wireless solution, had a significant and positive impact on the broadband penetration. We verify that opening access networks (and network elements) to competitive forces increases investment and the speed of development. Despite increasing competition,

incumbents are maintaining their dominant position. More than 60% of all broadband subscriptions make use of incumbent's broadband access infrastructure. In countries/regions where alternative technological platforms are not developed, the deployment of the DSL technology depends on the use of the networks infrastructures that are propriety of incumbent operators. To facilitate market entry of new competitors and develop competition in the access market, the regulatory authorities are focused on unbundled access to the local loop (fully unbundled local loop and shared access to the local loop) and on different forms of network access (bitstream and resale).

References

1. Kittl, J., Lundborg, M., Ruhle, E.-O.: Infrastructure-based versus service-based competition in telecommunications. Communications & Strategies, 67 (2006)
2. Collins, H.: Next Generation Networks - Creating a Dedicated Cost Model. InterConnect Communications Ltd., United Kingdom (2009)
3. Marcus, J.S., Elixmann, D.: Regulatory Approaches to NGNs: An International Comparison. Communications & Strategies 69, 21 (2008)
4. Xavier, P.: Geographically Segmented Regulation for Telecommunications. In: OECD 2010 (2010)
5. Oliveira, A.: Next Generation Access Networks: A key factor of development. In: NGON Seminar – ISCTE, Lisbon, pp. 1–33 (2009)
6. Andersen Int., Pricing Shared Access in Sweden, Post- of Telestyrelsen., Sweden (2004)
7. ERG, Report on ERG Best Practices on Regulatory Regimes in Wholesale Unbundled Access and Bitstream Access. European Regulators Group (2008)
8. Pereira, J.P., Ferreira, P.: Game Theoretic Modeling of NGANs: Impact of retail and wholesale services price variation. Journal of Communications (2012)
9. Pereira, J.P., Ferreira, P.: Infrastructure Sharing as an Opportunity to Promote Competition in Local Access Networks. Journal of Computer Networks and Communications, 11 (2012)
10. Höffler, F.: Cost and benefits from infrastructure competition. Estimating welfare effects from BB access competition. Telecommunications Policy 31, 401–418 (2007)
11. Vanberg, M.A.: Competition in the German BB Access Market. ZEW Mannheim (2004)
12. Kirsch, F., Hirschhausen, C.V.: Regulation of Next Generation Networks: Structural Separation, Access Regulation, or no Regulation at all? In: NFRA, Rotterdam, The Netherlands, pp. 1–8 (2008)
13. Jaag, C., Lutzenberger, M.: Approaches to FTTH-Regulation: An International Comparison. In: Second Annual Conference on Competition and Regulation in Network Industries, Brussels, Belgium, p. 23 (2009)
14. Laffont, J.-J., Tirole, J.: Competition in Telecommunications. The MIT Press, Cambridge (2001)
15. Marcus, J.S., Elixmann, D., Wernick, C.: Next Generation Networks (NGNs). European Parliament, Brussels (2009)
16. European Commission, Electronic Communications Market Indicators. EC (2011)
17. Cas, J.: Alternative local loop technologies - impact on regulation and competition. Presented at the ITS European Conference, Turin, Italy (1999)

18. Cadman, R.: Inconsistent Regulation, Market Structure and Broadband Adoption in the EU: a Dynamic Model. Strategy & Policy Consultants Network Ltd and ESRC Centre for Competition Policy (2008)
19. Cardona, M., Schwarz, A., Yurtoglu, B., Zulehner, C.: Demand estimation and market definition for broadband Internet services. Journal of Regulatory Economics 35, 70–95 (2009)
20. OECD, Geographically Segmented Regulation for Telecommunications. OECD Digital Economy Papers 173, 78 (2010)
21. European Commission, Main market developments. European Commission (2011)

Modeling and Verifying DML Triggers Using Event-B

Hong Anh Le[1] and Ninh Thuan Truong[2]

[1] Hanoi University of Mining and Geology
Dong Ngac, Tu Liem, Hanoi
[2] VNU - University of Engineering and Technology
144 Xuan Thuy, Cau Giay, Hanoi
{anhlh.di10,thuantn}@vnu.edu.vn

Abstract. Database trigger is a block code that automatically executes in response to changes of table or view in the database system. The correctness of a trigger usually can be verified when it is executed. It is apparently useful if we can detect the trigger system's errors in the design phase. In this paper, we introduce an approach to model and verify data manipulation language (DML) triggers in the database system by a formal method. In the first phase, we formalize a database trigger system by an Event-B model. After that, we use the Rodin tool to verify some properties of the system such as termination, preservation of constraint rules. We also run an example to illustrate the approach in detail.

1 Introduction

Triggers are active rules of some commercial database systems such as Oracle, SyBase, etc.. which are formed in ECA (Event - Condition - Action) structure. Triggers are widely and commonly used in database systems of many applications to implement automatic tasks and ensure integrity constraints. In some commercial databases, triggers have two kinds: data manipulation language (DML) and system triggers. The former are fired whenever events such as DELETING, UPDATING, INSERTING occur, while the latter are executed in case that system or data definition language (DDL) events occur. A trigger is made of a block of code, for example in Oracle, a trigger is similar to a stored procedure containing blocks of PL/SQL code. These codes are human readable and without any formal semantic. Therefore, we can only verify that if a trigger terminates or conflicts to integrity constraints after executing it or with human inspection step by step. It is important if we can show that triggers execution is correct at the design time. Several works have attempted to solve this question by using termination detection algorithms or model checking [13] [14] [10] [7] [16]. However, in our thought, most of results focused on the termination property, while few of them addressed to both termination and integrity constraints of the database system. Furthermore, these approaches seem such complicated that we can not apply them in the database development.

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 539–548, 2013.

The B method [1] is a formal software development method, originally created by J.-R. Abrial. The B notations are based on the set theory, generalized substitutions and the first order logic. Event-B [2] is an evolution of the B method that is more suitable for developing large reactive and distributed systems. Software development in Event-B begins by abstractly specifying the requirements of the whole system and then refining them through several steps to reach a description of the system in such a detail that can be translated into code. The consistency of each model and the relationship between an abstract model and its refinements are obtained by formal proofs. Support tools have been provided for Event-B specification and proof in the Rodin platform.

In this paper, we propose an approach to formalize database triggers system by a proof-based method, e.g Event-B. The main idea of the approach comes from the similarity between structures of Event-B EVENT and ECA. We first translate a database system to an Event-B model. In the next step, we bring this model to more practical approach by using the Rodin platform to verify some properties such as termination and constraint preservation based on its proof obligation engine. The advantage of our approach is that a real database system including triggers and constraints can be modeled easily by logic expression phrases in Event-B such as INVARIANTS and EVENTS. Therefore, the correctness of the entire system can be achieved by formal proofs. It is valuable especially for database developers since they are able to ensure that the trigger systems avoid the critical issues at the design time. Furthermore, the approach is such practical that we can implement a tool following its main idea to transform a database model to an Event-B model in Rodin platform automatically (or partly). It makes sense as we can bring the formal verification to database implementation. It also overcomes one of disadvantages that makes formal methods absent in the database development process because of the complexity of modeling.

The remainder of this paper is structured as follows. Section 2 gives a brief introduction of Event-B and background of database triggers. Next, in Section 3, we introduce some transformation rules between a database triggers system to an Event-B model. To show the approach in detail, we model a specific trigger system in an example in Section 4. Followed by Section 5, we give some information and adjustment of related works so far. We conclude our contribution and present the future works in Section 6.

2 Backgrounds

In this section, we briefly introduce the overview of relational database triggers and basic knowledge of Event-B.

2.1 Database Triggers

Database trigger is a block code that is automatically fired in response to an defined event in the database. The event is related to a specific data manipulation

of the database such as inserting, deleting or updating a row of a table. Triggers are commonly used in some cases: to audit the process, to automatically perform an action, to implement complex business rules.

The structure of a trigger is followed EAC structure, hence it takes the following form: *rule_name:: Event(e) IF condition DO action.*

It means that whenever Event(e) occurs and the *condition* is met then the database system performs *actions*. Users of some relational database systems such as Oracle, MySQL, SyBase are familiar with triggers which are represented in SQL:1999 format (the former is SQL-3 standard). Database triggers can be mainly classified by two kind: DML and Data Definition Language (DDL) trigger. The former is executed when data is manipulated, while in some database systems, the letter is fired in response to DDL events such as creating table or events such as login, commit, rollback..

2.2 Event-B

Event-B is a kind of formal method which combines mathematical techniques from the set theory and the first order logic. It is used as a notation and method for the formal development of discrete systems. Event- B is an evolution of others formal method notations like B-method (also know as classical B), Z and Action Systems. It is considered as an evolution because it simplifies the B machine notations, is easy to learn and more suitable for parallel and distributed reactive system development. Another advantage Event-B is the tool support for system modeling. The basic structure of an Event B model consists of a MACHINE and a CONTEXT.

Contexts form the static part of the model while machines form the dynamic part. Contexts can extend (or be extended by) other context and are referred (seen) by machines. The machine contains the dynamic part of the model. It describes the system state, the operations to interact with the environment together with the properties, conditions and constraints on the model. A Machine is defined by a set of clauses which is able to refine another Machine. We briefly introduce main concepts in Event-B as follows:

- Variables: represents the state variables of the model of the specification.
- Invariants: describes by first order logic expressions, the properties of the attributes defined in the variable clauses. Typing information, functional and safety properties are described in this clause. These properties are true in the whole model. Invariants need to be preserved by events clauses.
- Events: defines all the events that occur in a given model. Each event is characterized by its guard (i.e. a first order logic expression involving variables). An event is fired when its guard evaluates to true. If several guards evaluate to true, only one is fired with a non deterministic choice.

A Context consists of the following items:

- Sets: describes a set of abstract and enumerated types.
- Constants: represents the constants used by the model.

– Axioms: describes with first order logic expressions, the properties of the attributes defined in the CONSTANTS clause. Types and constraints are described in this clause.

After having the system modeled in Event-B, we need to reason about the model to understand it. To reason about a model, we use its proof obligation which show its soundness or verify some properties. As we mention in the first part of this Subsection, behaviors of the system are represented by machines. Variables v of a machine defines state of a machine which are constrained by invariants $I(v)$. Events E_m which describe possible changes of state consisting of guards $G_m(v)$ and actions $S_m(v, v')$ and they are denoted by

$$\text{when } G_m(v) \text{ then } v :| \ S_m(v, v') \text{ end}$$

Properties of an Event-B model are proved by using proof obligations (PO) which are generated automatically by the proof obligation generator of Rodin platform. The outcome of the proof obligation generator are transmitted to the prover of the Rodin tool performing automatic or interactive proofs.

3 Modeling and Verifying Database Triggers System

In this section, we present an approach to model a database systems including triggers. The main idea is mapping between entities of the database systems and Event-B elements in which we emphasize on modeling triggers by Event-B Events. After the transformation, we are able to verify some properties based on achieved Event-B model.

3.1 Modeling Database Systems

A database system is normally designed by several elements such as tables (or views) with integrity constraints and triggers. Whenever users modify the database table contents, e.g Insert, Delete and Update actions, the modification should be conformed to constraints and it also can fire the corresponding triggers. Before modeling the trigger system by Event-B, we introduce some definitions related to Event-B specification, they are useful in the modeling process.

Definition 1. A database trigger is modeled by a 3-tuple $db = <T, C, G>$ where T is a set of table, C states system constraints, G indicates a set of triggers.

Definition 2. For each $t \in T$, denoted by a tuple $t = <row_1, .., row_m>$where m is the number of table row, row_i, $(i \in 1..m)$ is a set indicating the i-th row of the table. A row is stated by a tuple $row_i = <field_1, .., field_n>$

Definition 3. Each trigger g of the system is presented as a 3-tuple such as $g \in G, g = <e, c, a>$ where, e is the corresponding event of the trigger, c is condition of the trigger, a is the action of the trigger.

We model a database system by mapping these definitions to Event-B concepts in Table 1. These rules are described in detail as follows:

Table 1. Transformation between database system and Event-B concepts

	Database definitions	**Event-B concepts**
Rule 1.	$db = <T, C, G>$ $T = \{t_1, .., t_m>$	$db_B = \{S_T \leftrightarrow I \leftrightarrow E\}$ $S_T = \{t_1, .., t_m\}$
Rule 2	$t = <r_1, .., r_m>$	$t_B = \{r_{B1}, .., r_{Bm}\}$.
Rule 3	$r_i = <f_{i1}, .., f_{in}>$	$r_{Bi} = \{1 \mapsto f_{Bi1}, .., m \mapsto f_{Bin}\}$

- Rule 1. Where set of tables T is mapped to set S_T, constraints C is translated to a set of invariant I, triggers set G is transformed to a set of events E
- Rule 2. A table is translated to a set of rows.
- Rule 3. A row of a table is transformed to an ordered set of fields, where m is a number of columns of the table and f_{Bij} is the value of column j at row i, where $i \in 1..m, j \in 1..n$

In the next subsection, we present in detail how to formalize database triggers.

3.2 Formalizing Triggers

As illustrated in Table 2, a trigger is formalized as an Event-B event where trigger's type and its condition is the guard of the event. Action of a trigger is transformed to the body part of an Event-B event.

Table 2. Modeling a trigger by an Event-B Event

IF (*type*) ON (*condition*)	WHEN (*type* $\wedge$ *condition*)
ACTION (act)	THEN (act) END

To show our approach, we simplify by considering the case that the Action part of a trigger contains a single action, though it can be a sequence of actions. It is clear that we are able to model such sequence of actions using Event-B if we can formalize a single Action. An Action of a trigger body is Insert, Update or Delete statement. In case of Update and Delete statements, the action contains a condition that shows which rows are affected. Therefore, we combine statement and trigger condition into guard of transformed event. Specifically, mapping rules of each kind of statements are presented in Table 3.

3.3 Verifying System Properties

After the transformation, taking advantages of Event-B method and its support tool, we are able to verify some properties of the database system model as follows:

Table 3. Translating SQL statements to Event-B events

UPDATE table_name SET column1=value, column2=value2 WHERE some_column=some_value	WHEN *update_condition* THEN $r := \{1 \mapsto value1, 2 \mapsto value2\}$
DELETE FROM table_name WHERE some_column=some_value	WHEN *delete_condition* $table\_name := table\_name - \{col_1 \mapsto val_1, .., col_n \mapsto val_n\}$
INSERT INTO table_name VALUES (value1,..,valuen)	WHEN *insert_condition* $table\_name := table\_name \cup \{col_1 \mapsto val_1, .., col_n \mapsto val_n\}$

- Termination: Since a trigger can fire the other triggers, hence it probably leads to infinite loop. The termination of a trigger is able to be verified by the deadlock property of the Event-B model. This situation occurs when after a sequence of events, state of the system does change. This property is proved by proof obligations which state that the disjunction of the event guards always hold under the properties of the constant and the invariant. The deadlock freedom rule is stated as $I(v) \vdash G_1(v) \vee ... \vee G_n(v)$, where v is variable, $I(v)$ denotes invariant, $G_i(v)$ presents guard of the event. At the moment, the deadlock freeness PO is not generated automatically by the Rodin tool yet. However, we can generate it ourself by as a theorem saying the disjunction of guards.
- Constraint preservation: Since these properties already are modeled by Event-B INVARIANTS as the approach illustrated in Subsection 3.1, hence we can prove them by using invariant PO rules.

4 An Example

In order to make our approach more clear, in this section, we take an example to present it in detail. We first describe the example, after that we model it by an Event-B machine and verify its properties.

4.1 Example Description

Let assume that we have a database system including two tables EMPLOYEES and BONUS structured in Table 4.

Table 4. Table EMPLOYEES and BONUS

EMPLOYEES		BONUS	
E_Id	level	E_Id	amount
0911	2	0911	2
0912	2	0912	2
0913	4	0913	4

The database system has a constraint: *The bonus of an employee with a level greater than 5 is at least 20.*

It includes two triggers that do the following tasks:
Trigger 1. Whenever the level of employee is updated, his bonus is increased by 10 if the level is even
Trigger 2. If the employee's bonus amount is updated, then his level is increased by 1.

These two triggers are rewritten in the format of PL/SQL as follows:

```
CREATE TRIGGER Trigger_1 BEFORE UPDATE
    OF level ON employees
    FOR EACH ROW
    BEGIN
          IF MOD(employees.level,2)=0 THEN
             UPDATE bonus SET bonus.amount
                    =bonus.amount  + 10
       WHERE bonus.E_id = employees.E_id;
    END IF;
END
```

```
CREATE TRIGGER Trigger_2 BEFORE UPDATE
   OF amount ON bonus
   FOR EACH ROW
   BEGIN
    UPDATE employees SET
       employees.level = employees.level+1
    WHERE bonus.E_id = employees.E_id;
 END
```

4.2 Modeling an Example

Followed the approach presented in Section 3, we formalize two tables which are involved in the trigger statements by two variables such as *empl* and *bonus*. Variables *bonus_rec* and *empl_rec* present a row of the table Bonus and Employees respectively.

$\texttt{inv7}: \texttt{bonus} \in \mathbb{P}((\mathbb{N}_1 \times \mathbb{N}_1) \times (\mathbb{N}_1 \times \mathbb{N}_1))$
$\texttt{inv11}: \texttt{empl} \in \mathbb{P}((\mathbb{N}_1 \times \mathbb{N}_1) \times (\mathbb{N}_1 \times \mathbb{N}_1))$
$\texttt{inv16}: \texttt{bonus\_rec} \in \texttt{bonus}$
$\texttt{inv17}: \texttt{empl\_rec} \in \texttt{empl}$
$\texttt{inv5}: \texttt{trigger\_type} \in \mathbb{P}(\texttt{TIME}) \leftrightarrow \mathbb{P}(\texttt{COMMAND})$
$\texttt{inv20}: \texttt{active\_field} \in \mathbb{P}(\texttt{TABLE\_NAMES}) \leftrightarrow \mathbb{P}(\texttt{FIELD\_NAMES})$

The constraint of the database system is also formalized by an INVARIANT

INVARIANTS
$\texttt{inv21}: \texttt{empl\_level} < 5 \lor \texttt{bonus\_amount} \geq 20$

We next formalize two triggers of the system as the approach presented in 3.2. Since DML actions are performed on the table, we model the table involved in triggers by an Event-B VARIABLE *table* such that *table* is the identifier of the table.

Event *trigger1* $\widehat{=}$
  when
    grd1 : $\texttt{trigger\_type} = \{\texttt{AFTER} \mapsto \texttt{UPDATE}\}$
    grd3 : $\texttt{empl\_levelmod2} = 0$
    grd5 : $\texttt{empl\_level} \in \texttt{ran}(\texttt{ran}(\texttt{empl}))$
    grd6 : $\texttt{empl\_id} \in \texttt{ran}(\texttt{dom}(\texttt{empl}))$
    grd8 : $\texttt{ran}(\texttt{dom}(\{\texttt{bonus\_rec}\})) = \{\texttt{empl\_id}\}$
    grd9 : $\texttt{bonus\_amount} \in \texttt{dom}(\texttt{dom}(\{\texttt{bonus\_rec}\}))$
    grd10 : $\texttt{active\_field} = \{\texttt{EMPLOYEES} \mapsto \texttt{EMP\_LEVEL}\}$
  then
    act1 : $\texttt{trigger\_type} := \{\texttt{AFTER} \mapsto \texttt{UPDATE}\}$
    act3 : $\texttt{bonus\_amount} := \texttt{bonus\_amount} + 10$
    act5 : $\texttt{bonus\_rec} := (1 \mapsto \texttt{empl\_id}) \mapsto (2 \mapsto \texttt{bonus\_amount})$
    act6 : $\texttt{active\_field} := \{\texttt{BONUS} \mapsto \texttt{BONUS\_AMOUNT}\}$
  end
Event *trigger2* $\widehat{=}$
  when
    grd1 : $\texttt{trigger\_type} = \{\texttt{AFTER} \mapsto \texttt{UPDATE}\}$
    grd2 : $\texttt{active\_field} = \{\texttt{BONUS} \mapsto \texttt{BONUS\_AMOUNT}\}$
    grd3 : $\texttt{ran}(\texttt{ran}(\{\texttt{empl\_rec}\})) = \{\texttt{empl\_id}\}$
    grd4 : $\texttt{empl\_level} \in \texttt{ran}(\texttt{ran}(\{\texttt{empl\_rec}\}))$
  then
    act1 : $\texttt{trigger\_type} := \{\texttt{AFTER} \mapsto \texttt{UPDATE}\}$
    act2 : $\texttt{empl\_level} := \texttt{empl\_level} + 1$
    act3 : $\texttt{empl\_rec} := (1 \mapsto \texttt{empl\_id}) \mapsto (2 \mapsto \texttt{empl\_level})$
    act4 : $\texttt{active\_field} := \{\texttt{EMPLOYEES} \mapsto \texttt{EMP\_LEVEL}\}$
  end

4.3 Checking Properties

- Termination: To verify the termination property in the Rodin tool, we generate an invariant clause which is the disjuntion of two events' guards. Using PO engine of the Rodin tool, we can prove that the system is not deadlock free, i.e the system is terminated.
- Constraint violation: Since the constraint property of the system is modeled by INVARIANT *inv*21, hence it is also proved by invariant preservation rules. The invariant is proved to be failed through events of the model, hence the triggers execution violates the system constraint.

5 Related Works

From the beginning, the previous works focused on the termination of the triggers by using static analysis, e.g. checking set of triggers is acyclic with triggering

graph. In [13] and [14], Sin-Yeung Lee and Tok-Wang introduced algorithms to detect the correctness of updating triggers. However, this approach is not extended apparently for general triggers and it is presented as their future work.

E.Baralis *et al* performed the dynamic analysis to check active rules at run time to see if a state of the database system is repeated.

L. Chavarria and Xiaoou Li proposed a method verifying active rules by using conditional colored Petri nets [7]. Since Petri nets are mainly used in modeling transitions, hence it is quite elaborated when normalizing rules. The approach has to classify rules by the logic condition of these rules to check if they involve disjunction or conjunction operators. In our opinion, if the number of these operators are enormous then the transition states can be exploded.

Some works applied model checking for active database rule analysis [12][10]. In [12], T. S. Ghazi and M. Huth presented an abstract modeling framework for active database management systems and implemented a prototype of a Promela code generator. However, they did not describe how to model data and data actions for evaluation.

Eun-Hye CHOI *et al* [10] proposed a general framework for modeling active database systems and rules. The framework is feasible by using a model checking tool, e.g SPIN, however, constructing a model in order to verify the termination and safety properties is not a simple step and can not be done automatically.

More recently, R. Manicka Chezian and T.Devi [17] introduced a new algorithm which does not pose any limitation on the number of rules but it only emphasizes on algorithms detecting the termination of the system.

6 Conclusion and Future Works

Most of the researches to date that have worked on verifying and modeling database active rules or triggers mainly focuses on the termination property. A few works presented methods to model a database system and verify some properties of the system. However, in our opinion, these results are complex to bring them to software development and are not feasible to be performed automatically without human analysis. In this paper, we propose an approach to formalize and verify the database system with constraints and triggers by using Event-B. Our main contribution is that we perform the mapping between elements of the database such as tables, triggers, constraints to Event-B clauses. We also reuse the obligation engines and tool supported by Event-B to prove the correctness of the system. Moreover, the transformation is also simple and clear such that it is feasible to formalize the database system by an Event-B model automatically. Therefore, it makes sense if we want to bring the formal verification to software development.

Besides the advantages, the paper still has some limitation such that we do not address how to model a more complex case study with more complicated triggers. These issues, along with development of a tool which takes into account to translate a database system to an Event-B model in the format of Rodin platform, are our future works.

Acknowledgments. This work is supported by the project no. 102.02–2010.06 granted by Vietnam National Foundation for Science and Technology Development (Nafosted).

References

1. B method web site (2012), `http://www.bmethod.com`
2. Event-b and the rodin platform (2012), `http://www.event-b.org`
3. Abrial, J.-R.: Modeling in Event-B - System and Software Engineering. Cambridge University Press (2010)
4. Ait-Sadoune, I., Ait-Ameur, Y.: From bpel to event-b. In: IM FMT 2009 Conference, Dsseldorf Germany, Fevruary (2009)
5. Baralis, E.: Rule analysis. In: Active Rules in Database Systems, pp. 51–67. Springer, New York (1999)
6. Baralis, E., Widom, J.: An algebraic approach to static analysis of active database rules. ACM Trans. Database Syst. 25(3), 269–332 (2000)
7. Chavarría-Báez, L., Li, X.: Verification of active rule base via conditional colored petri nets. In: SMC, pp. 343–348 (2007)
8. Chavarría-Báez, L., Li, X.: Ecapnver: A software tool to verify active rule bases. In: ICTAI (2), pp. 138–141 (2010)
9. Chavarría-Báez, L., Li, X.: A petri net-based metric for active rule validation. In: ICTAI, pp. 922–923 (2011)
10. Choi, E.-H., Tsuchiya, T., Kikuno, T.: Model checking active database rules. Technical report, AIST CVS, Osaka University, Japan (2006)
11. Choi, E.-H., Tsuchiya, T., Kikuno, T.: Model checking active database rules under various rule processing strategies. IPSJ Digital Courier 2, 826–839 (2006)
12. Ghazi, T., Huth, M.: An Abstraction-Based Analysis of Rule Systems for Active Database Management Systems. Technical report, Kansas State University, Technical Report KSU-CIS-98-6, p.15 (April 1998)
13. Lee, S.-Y., Ling, T.-W.: Are your trigger rules correct? In: Proceedings of the 9th International Workshop on Database and Expert Systems Applications, DEXA 1998, p. 837. IEEE Computer Society, Washington, DC (1998)
14. Lee, S.-Y., Ling, T.-W.: Verify Updating Trigger Correctness. In: Bench-Capon, T.J.M., Soda, G., Tjoa, A.M. (eds.) DEXA 1999. LNCS, vol. 1677, pp. 382–391. Springer, Heidelberg (1999)
15. Li, X., Medina Marín, J., Chapa, S.V.: A Structural Model of ECA Rules in Active Database. In: Coello Coello, C.A., de Albornoz, Á., Sucar, L.E., Battistutti, O.C. (eds.) MICAI 2002. LNCS (LNAI), vol. 2313, pp. 486–493. Springer, Heidelberg (2002)
16. Ray, I., Ray, I.: Detecting termination of active database rules using symbolic model checking. In: Caplinskas, A., Eder, J. (eds.) ADBIS 2001. LNCS, vol. 2151, pp. 266–279. Springer, Heidelberg (2001)
17. Manicka chezian, T.R.: A new algorithm to detect the non-termination of triggers in active databases. International Journal of Advanced Networking and Applications 3(2), 1098–1104 (2011)

A Conceptual Multi-agent Framework Using Ant Colony Optimization and Fuzzy Algorithms for Learning Style Detection

Ghusoon Salim Basheer, Mohd Sharifuddin Ahmad, and Alicia Y.C. Tang

Universiti Tenaga Nasional, Jalan IKRAM-UNITEN,
43000 Kajang, Selangor, Malaysia
noora2003@yahoo.com, {sharif,aliciat}@uniten.edu.my

Abstract. This paper examines the progress of researches that exploit multi-agent systems for detecting learning styles and adapting educational processes in e-Learning systems. In a summarized survey of the literature, we review and compile the recent trends of researches that applied and implemented multi-agent systems in educational assessment. We discuss both agent and multi-agent systems and focus on the implications of the theory of detecting learning styles that constitutes behaviors of learners when using online learning systems, learner's profile, and the structure of multi-agent learning systems. We propose a new dimension to detect learning styles, which involves the individuals of learners' social surrounding such as friends, parents, and teachers in developing a novel agent-based framework. The multi-agent system applies ant colony optimization and fuzzy logic search algorithms as tools to detecting learning styles. Ultimately, a working prototype will be developed to validate the framework using ant colony optimization and fuzzy logic.

Keywords: Multi-agent System, Learning Style, e-Learning, Learner Modeling, VARK Learning Style.

1 Introduction

Recently, more attention has been given to the use of multi-agent systems (MAS) in many distributed applications. Studies in multi-agent systems include the inquiry for rational, autonomous and flexible behaviour of entities, and their interaction and co-ordination in different areas [1]. The foundation of multi-agent systems play a significant role in the growth of teaching systems, because the basic issues of teaching and learning could be easily resolved by multi-agent systems [2].

An agent is a software entity that has the ability to execute flexible autonomous actions in an intelligent manner to carry out tasks that meet its goals. Depending on the functions of agents in their environments, their abilities may differ [3].

The quest for effective learning strategies has instigated educators and researchers to continuously exploit the information and communication technology in developing better and improved learning systems. Such development has produced a diversity of learning systems that claim to improve the learning process. The literature provides

A. Selamat et al. (Eds.): ACIIDS 2013, Part II, LNAI 7803, pp. 549–558, 2013.

ample evidence that suggests a spectrum of learning dimensions that are examined and incorporated in these learning systems. These include the modes with which the systems are used, the specific learners' domains and learning stages, the quality of the subject matter and the ease of using the systems. We believe that the development of these systems considers issues on the way people learn in these online environments, the way the resources are accessed and utilized and the learners' preferences in using the resources. These considerations manifest adaptable and flexible learning systems that are arguably compatible with the learners' behaviors and profiles. While we are not disputing the findings and development of such systems, we propose yet another alternative framework that not only considers the common dimensions used for learning styles detection but also introduces a novel proposition of learners' social surrounding and their conversational attributes, e.g., the words they used, as other possible dimensions to investigate.

In a learning environment, the learning style defines each learner's preferences in the way he/she approaches the learning process. Educators and researchers consider the learning style as an important factor that contributes to the learning process [4]. In the last decade, many researchers investigated the issues of learning style and proposed the most important feature of e-learning systems, which is personalized learning. Since each learner has his/her learning preferences, enhancing the learning systems to have the ability of classifying differences in the skills and preferences of the learner is of vital importance [5].

This paper examines recent educational systems that use multi-agent approach to improve the learning process and using such approach to detect learners' learning styles. The first part of this paper contains a review of the previous papers, their approaches, and current methods. The review discusses the applications of multi-agent approach to learning styles detection followed by a deliberation on learning style, its theory, use of multi-agent systems, and their properties, both theoretically and practically. We then propose our method, in which we infer a learners' learning style based on using the current methods that depends on the learners' behaviors and profiles and a novel method that considers the learners' social surrounding.

We organize the rest of the paper as follows: Section 2 reviews the related work of this research. In Section 3, we highlight the limitation of existing applications. Section 4 discusses our proposed framework and Section 5 concludes the paper.

2 Related Work

2.1 Aspects of Learning: Single-Agent and Multi-agent Learning

In order to understand the issues in detecting learners' learning styles using multi-agent systems; we analyze the existing theories and applications. We identify the limitations of the existing methods and show the need to offer a framework that enables discovery of learners' styles by multi-agent systems.

Current agent-based approaches that detect learning styles in adapting the learning process are offered in many different ways. A review of the literature reveals that most agent-based learning systems are developed based on a single agent but the effectiveness of the systems is doubtful since single agent approaches only improve its individual skills.

An agent-based online learning system developed by Schiaffino et al. [6] uses a single intelligent software agent called "e-Teacher" that supports learners by analyzing their online behaviors. The agent then creates a profile for each learner with which it recommends the learner with personalized learning strategies to help him/her progresses. The profile contains important information about the learners, such as their behaviors and preferences (e.g. examination results, exercises done, and objects studied). The system utilizes the Bayesian networks to detect a learner's learning style, which the e-Teacher agent uses to suggest appropriate course of actions.

Boff et al. [2] presented a method which uses the BDI and Bayesian networks with one intelligent agent (called Social Agent) to act in an educational environment. In this system, a learner's actions are stored in his/her model. The model is used when the agent looks for learners to join a workgroup. The workgroups are created by social agents based on learners' recommendations.

Alonso et al. [1] opined that single agent application in learning systems might not always produce optimal performance. They proposed that there are domains in which multi-agent learning is coordinated to improve the learning effectiveness.

In learning theory, "a learning object is a self-standing, reusable, discrete piece of content that fulfills an instructional objective". The decomposition of educational content into learning objects permits a discrete learning object to be exploited in many different educational courses. There is a special agent called Object Agent which is responsible for collaborating between learning style and the learning objects [3]. There are three parts that implement this strategy:

- Defining learner's learning style scheme.
- Classifying learning objects depending on the learning style.
- Suggesting appropriate learning objects.

For practical implementation of multi-agent systems where learning objects have to be taken into account, [7], [8] independently proved the importance of classifying each learning object and suggest appropriate object to each learner.

The researchers in [8] presented a multi-agent system for adapting distance learning on the Web. The system classifies the learners according to their skills. It uses intelligent agents to motivate the learners by letting them study what they want and test all possible chances to develop the teaching process.

Joy [3] presented a new approach that merges two learning style theories. They designed, implemented and evaluated the educational research contribution. Their system has five agents: Object Agent, Record Agent, Student Agent, Modeling Agent and Evaluation Agent. Each of these agents is developed to process all requirements that are compatible with the learning system. Their system has interleaving Bayesian Network and Fuzzy Logic techniques to implement a novel approach that develops architecture of learners' skills, by which it creates a complete model which represents the learners' requirements and knowledge.

Pham and Florea [4] proposed a method that automatically detects learner's learning style using the Felder-Silverman Learning Style (FSLS) theory. They developed a multi-agent adaptive learning system that compares the expected time spent on each learning object with the time a learner actually spent on it. If the recorded times are similar to the ones calculated for each learning style labeled for the learning object, then the learning style is correct for that learning object.

Rabbat [7], in his PhD thesis, developed a new model using the Bayesian network to detect each learner's learning style. A multi-agent tutoring system used this learning style to suggest learning materials that match the learner's style.

2.2 Learner Modeling Methods

In 2012, Pham and Florea [4] presented the concept to classify the modeling methods for learners as Explicit Modeling and Implicit Modeling Methods.

Explicit modeling is a simple and straightforward method for inferring learners' learning styles. In this method, learners are asked to answer a dedicated psychological questionnaire from which a preference towards one or more of the learning style dimensions can be inferred. Some disadvantages of this model are:

- The measuring instruments used may not be reflective of the way a particular learner learns.
- The learners may not be aware of the importance of the questionnaires as such they may tend to choose answers randomly.
- The learner model is rather static since once a learner model is created, it is not subjected to change or update.

There is another method which uses an implicit and/or dynamic modeling method. Three approaches are identified:

- Use the learner's performance as index of his/her style through evaluation tests. A higher percentage of some performance is considered as an indication of a style that corresponds to performance.
- Get feedback from learners about their experiences in the learning process and adjust the learner model accordingly.
- Observe the learners' interactions with the system and determine a corresponding learning style [5].

2.3 Models of Learning Style

Researchers generally agree that learning styles play a vital and significant role in providing effective learning experience for learners. For example, Lang [9] suggested that "research in learning styles attempts to categorize individuals into different categories by the patterns they use to take in, perceive, and interpret situations". Landry [10] then claimed that "everyone has a unique learning style, and instruction should be designed to best accommodate different methods of learning."

There are many models for detecting learning styles. Some learning style's models that we have gathered include [11] Kolb; Felder-Silverman; Dunn and Dunn; Myers-Briggs Type Indicator (MBTI); Right- and Left-brained; VARK; Gregory's Mind; Honey and Mumford; Herrmann "Whole Brain" and Grasha-Riechmann. Due to space limitation, we shall not deliberate each of these learning styles models but only discuss those that are significant to our work.

Fleming's [12] VARK Model. VARK is considered to be one of the classical learning theories in the educational field. The VARK model categorizes individuals as

Visual, Auditory, Read/Write, and Kinesthetic. Usually Visual learners prefer to be treated with symbols and charts but Auditory learners use listening to contact with the world. Read/Write learners use text while Kinesthetic learners are active in nature. Sometimes, there can be a mixture of models (i.e. multiple models) which is called multi-model learners.

In VARK, identifying a learner's mode involves using an instrument to detect the learning preference. The instrument consists of 16 questions and there is only one answer for each question. Every answer corresponds to one of the four classes of learning styles. The responses are compiled by category and the maximum value is used to determine the respondent's learning style. If two or more categories share the same highest score, then the respondent is considered to be a multi-modal learner. A multi-modal learner learns well in more than one category of the model.

Personalized Learning Model. Researches in personalized service in e-Learning systems focus on two main directions: Adaptive Educational Systems and Intelligent Tutors (agents).

Adaptive educational systems enable the selection of learning resources according to the learner's profile which comprises information such as learning style, knowledge, background, and goals. In a quite similar fashion, intelligent tutors offer relevant educational activities and provide feedback on those activities based on the learner's profile [6].

3 Limitation of Existing Applications

Most systems that we reviewed only consider learners' responses to a specific questionnaire and detect learning styles from learner's behaviors and actions. These systems do not exploit other information such as the learners' social surrounding to detect learning styles. To overcome these problems and provide a near-perfect model of a learner's learning style, existing learning models need to be extended. A strategy that exploits a learner's social surrounding needs to be conceived and the words used by learners need to be analyzed. Our argument for such proposition is due to the fact that we should not ignore the influence of the environment that shapes a learner's personality including his/her learning style. There is a cliché that says "for each person, there are three ideas about him: first, what he thinks about himself; second, what people said about him; and the third is the real truth".

Vazire and Carlson [13] support this notion in an article, "There are aspects of personality that others know about us that we don't know ourselves, and vice-versa. To get a complete picture of a personality, you need both perspectives."

Consequently, we posit that learning styles may also be detected not only from the behavior and profile of the learners but also from resources that are close to the learners such as their parents, friends, and teachers. These people could provide the added information such as specific learners' characteristics that are relevant to their learning styles. We perceive the possible advantages of using these resources from the many years of contact that these people have with the learners. In this paper, we exploit the second statement of the cliché and propose a framework that utilizes the statement to enhance the learning styles detection.

In another perspective, none of the algorithms that we reviewed consider the words used by a learner as an important resource to identify his/her learning style. Figure 1 shows the visual map of neuro-linguistic programming (NLP) that demonstrates that speech is one of the important features that represent every person and their styles. We argue that there is a very strong relation between the personality of a person and the language he/she uses. Most human communication tells us something about the people doing the saying [14]. We consider such conversational attribute as another dimension for our framework.

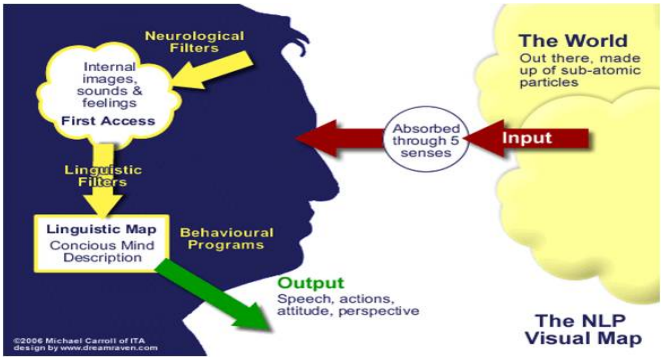


Fig. 1. Neuro-Linguistic Programming Visual Map [15]

4 The Proposed Multi-agent Framework

To satisfy a near-accurate learning style for learners, we analyze the dimensions that are commonly considered when detecting learning styles, but we take into account two additional dimensions: the social surrounding of learners and the words they used. These could be investigated by soliciting information from parents, friends, and teachers and analyzing the words used by the learners while describing some event.

Consequently, we propose a multi-agent framework which examines learners' learning styles using three intelligent software agents; each is responsible for detecting the learning styles depending on its knowledge base. These agents communicate their decisions to each other before a final decision is reached. The framework proposes that a learner attempts the VARK test [3], in which there are 16 questions and each question has four options that correspond to four different styles, i.e. Visual, Auditory, Reading/Writing, and Kinesthetic. The framework also proposes that the learner uses the e-learning system to discover his/her actions and behaviors that relates to his/her learning style. All these information are saved in the learner's profile.

To implement the framework, we introduce two algorithms. Fuzzy Logic is used by Student Agent (STA) and Social Surrounding Agent (SSA). STA saves information about a learner's behaviors, actions, used words and all information in his/her profile. The Social Surrounding Agent (SSA) solicits all the information from the learner's close contacts (parents, friends and teachers) and saves all information in its knowledge base. The Evaluation Agent (EVA) communicates with the SSA and STA and receives the results from the VARK test to provide the necessary information. We

introduce another algorithm, the Ant Colony Optimization (ACO) algorithm for EVA, which is responsible for deciding on the learning styles. Figure 2 shows the proposed framework.

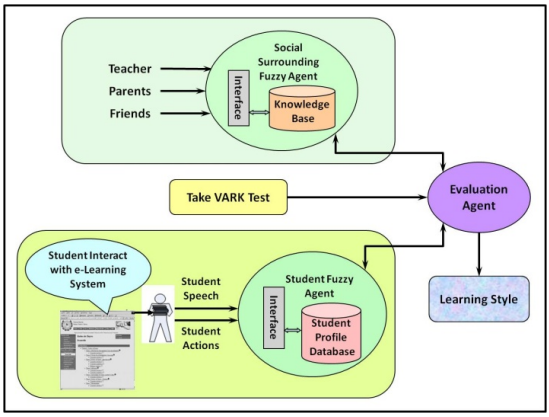


Fig. 2. The Proposed Framework

4.1 Using Fuzzy Logic

Each of the Social Surrounding Agent and Student Agent uses fuzzy logic to detect the learner's learning style. In the framework, the agents are designed as fuzzy agents. A fuzzy agent is "a software agent that implements fuzzy logic but interacts with the environment through an adaptive rule-base" [16].

We use K to denote Kinesthetic, V for Visual, A for Auditory, R for Read/Write and LS for Learning Style**.** For example, a visual learner can understand information in a chart or graph. Using this type of information is important to create rules. There are four membership functions: visual, auditory, read/write and kinesthetic. Figure 3 shows an example of fuzzy rules used by the SSA.

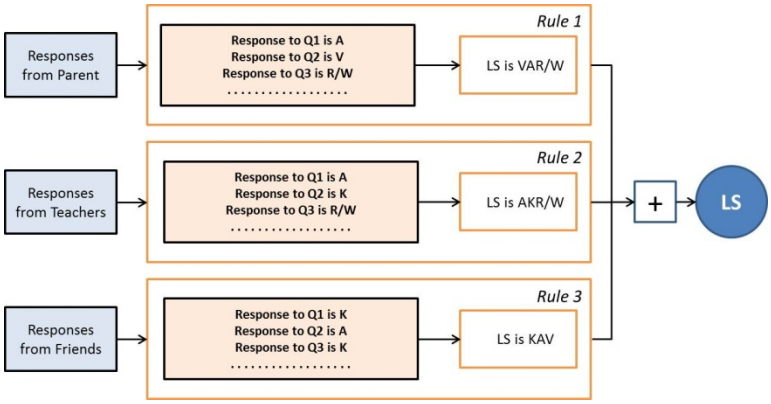


Fig. 3. The reasoning model of fuzzy agents

4.2 Using ACO Algorithm

Figure 4 depicts the proposed Ant Colony Optimization (ACO) structure which is used in conjunction with the fuzzy agents. Initially, an ant starts from the selected node (first detected by the VARK test). It then travels across the structured graph, and at each stage determines its direction.

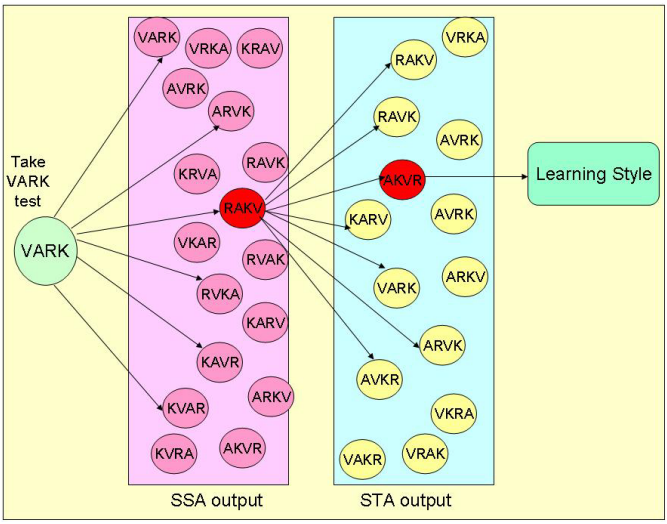


Fig. 4. The Proposed Structure of EVA's ACO

With the first four styles taken from the VARK test's results attempted by a learner, the algorithm starts from these four main styles, which we represent as a node (V, A, R, K), and uses the output of the SSA and STA to determine the final learning style. There are 24 final nodes, each node represents the VARK style in different sequences, i.e. VARK, VKRA, KVRA, . . . etc. Each one of these nodes represents a VARK style in different priority depending on the position of each character. Node VARK means that the style of the person is Visual (high percentage) and then Auditory (lower percentage), followed by Read/Write and Kinesthetic.

Figure 4 shows that each node represents a multi-modal learning style for the learner and each edge represents the evaluation of each style. The proposed algorithm consists of two stages. The initial value for the first stage is taken from the VARK test. The weight for each edge depends on the output of SSA. In the second stage, the output of the first stage is used with the output from STA to determine the ultimate learning style.

Assume that the output of the VARK test is as shown in Table 1 and the output of SSA is shown in Table 2.

Table 1. Output of VARK Test

Style	Percentage
V	40
A	30
R	20
K	10

Table 2. Output of SSA

Style	Percentage
V	30
A	20
R	40
K	10

The ACO algorithm uses each of these numbers as weight on the edge which connects the nodes. Depending on the final weight, the best node is reached as shown in Figure 5. In the figure, the V style is the dominant style (70%), followed by R (60%), A (50%) and K (20%).

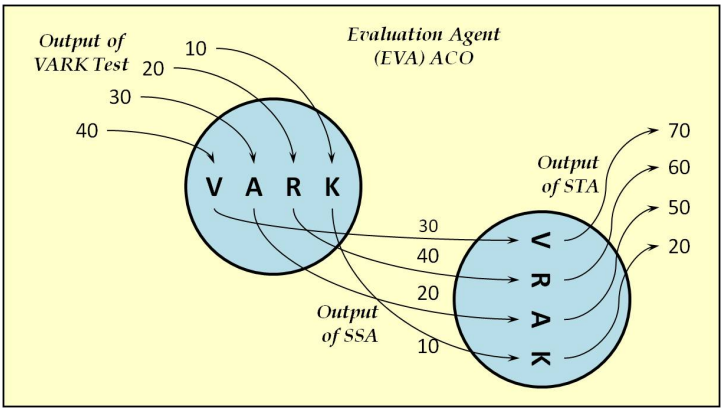


Fig. 5. An Example of the ACO Calculation

5 Conclusion and Further Work

This paper compared some state-of-the-art approaches for detecting learning styles that used multi-agent systems. We focused our attention on the internal structure of automatic detection of learning style systems which used multiple agents. We then proposed a new framework based on new sources of information from learners' social surrounding such as friends, parents, and teachers and words used by learners.

The proposed multi-agent system consists of three agents: Social Surrounding Agent (SSA), Student Agent (STA) and Evaluation Agent (EVA). Each agent is responsible for a specific set of operations. SSA and STA use fuzzy logic to decide on a learner's learning style. Fuzzy logic allows two agents (SSA and STA) to be adapted for managing ambiguous data, which is very common in many real-world problems. EVA uses the ant colony optimization algorithm for proposing a set of optimal paths to detect all possible properties of learning styles. Our focus is to develop accurate

models to detect learners' learning styles based on the knowledge from the models that use different intelligent techniques.

In our future work, we shall implement the proposed framework and develop the algorithms to detect the learning styles based on the framework. Additionally, detailed mechanisms for all the agents and the sources of information will be developed.

References

1. Alonso, E., d'Inverno, M., Kudenko, D., Luck, M., Noble, J.: Learning in Multi-agent Systems. Result of a Panel Discussion. In: Third Workshop of the UK's Special Interest Group on Multi-agent Systems (2001)
2. Boff, E., Vicari, R.M., Fagundes, M.S.: Using a Probabilistic Agent to Support Learning in Small Groups. In: Louca, L.S., Oplatková, Y.C.Z., Al-Begin, K. (eds.) The 22nd European Conference on Modeling and Simulation, ECMS (2008) ISBN: 978-0-9553018-5-8 / ISBN: 978-0-9553018-6-5 (CD)
3. Joy, M.: An Innovative Use of Learning Objects and Learning Style in Pedagogic Agent Systems. Department of Computer Science, University of Warwick (2005)
4. Pham, Q.D., Florea, A.M.: An approach for Detecting Learning Styles in Learning Management Systems based on Learners' Behaviors. In: International Conference on Education and Management Innovation IPEDR, vol. 30. IACSIT Press, Singapore (2012)
5. Popescu, E.: Diagnosing Students' Learning Style in an Educational Hypermedia System. Software Engineering Department, University of Romania (2008)
6. Schiaffino, S., Garcia, P., Amandi, A.: eTeacher: Providing Personalized Assistance to e-Learning Students. Elsevier Ltd. (2008), doi:10.1016/
7. Rabbat, R.R.: Bayesian Expert Systems and Multi-Agent Modeling for Learning-Centric eb- wbased Education. PhD Thesis, American University of Beirut (2005)
8. Peña, C.-I., Marzo, J.-L., de la Rosa, J.-L.: Intelligent Agents in a Teaching and Learning Environment on the Web. University of Girona, Spain (2002)
9. Lang, T.K.: The Effect of Learning Styles. Computer Attitude and Classroom Technology on Student Performance and Motivation, Doctoral Dissertation, Auburn University, AL (2004)
10. Landry, J.M.: Learning Styles of Law Enforcement Officers. Does Police Work Affect How Officers Learn? Capella University (2011)
11. Graf, S.: Adaptively in Learning Management Systems Focusing on Learning Styles. PhD Thesis, Vienna University of Technology, Faculty of Informatics (2007)
12. Fleming, N.: VARK a guide to learning style, Copyright 2001 - 2012 Neil Fleming, http://www.vark-learn.com/english/index.asp
13. Nauert, R.: Friends May Know You Better Than You Know Yourself, http://psychcentral.com/news/2011/05/09/friends-may-know-you-better-than-you-know-yourself/26009.html
14. Holtzman, N.S., Yarkoni, T.: More Personality Residues in Language Use: A Primer. The Online Newsletter for Personality Science Issue 5 (2010)
15. Harding, D.: The Science of Communication and the Art of Change (2012), http://www.deehardinglifecoach.com/page4.htm
16. Epstein, J.G., Möhring, M., Troitzsch, K.G.: Fuzzy-Logical Rules in a Multi-Agent System. In: SimSoc VI Workshop, Groningen, September 19–21 (2003)

Author Index