Jeng-Shyang Pan
Shyi-Ming Chen
Ngoc Thanh Nguyen (Eds.)

LNAI 6421

Computational Collective Intelligence

Technologies and Applications

Second International Conference, ICCCI 2010
Kaohsiung, Taiwan, November 2010
Proceedings, Part I

1 Part I

Springer

Lecture Notes in Artificial Intelligence 6421

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

Subseries of Lecture Notes in Computer Science

Jeng-Shyang Pan Shyi-Ming Chen
Ngoc Thanh Nguyen (Eds.)

Computational Collective Intelligence

Technologies and Applications

Second International Conference, ICCCI 2010
Kaohsiung, Taiwan, November 10-12, 2010
Proceedings, Part I

Springer

Volume Editors

Jeng-Shyang Pan
National Kaohsiung University of Applied Sciences
Department of Electronic Engineering
415 Chien-Kung Road, Kaohsiung 807, Taiwan
E-mail: jspan@cc.kuas.edu.tw

Shyi-Ming Chen
National Taiwan University of Science and Technology
Department of Computer Science and Information Engineering #43, Sec.4
Keelung Rd., Taipei, 106,Taiwan
E-mail: smchen@mail.ntust.edu.tw

Ngoc Thanh Nguyen
Wroclaw University of Technology, Institute of Informatics
Str. Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
E-mail: ngoc-thanh.nguyen@pwr.wroc.pl

Library of Congress Control Number: 2010937276

CR Subject Classification (1998): I.2, I.2.11, H.3-4, C.2, D, H.5

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN	0302-9743
ISBN-10	3-642-16692-X Springer Berlin Heidelberg New York
ISBN-13	978-3-642-16692-1 Springer Berlin Heidelberg New York

springer.com

Printed in Germany

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

Preface

This volume composes the proceedings of the Second International Conference on Computational Collective Intelligence—Technologies and Applications (ICCCI 2010), which was hosted by National Kaohsiung University of Applied Sciences and Wroclaw University of Technology, and was held in Kaohsiung City on November 10-12, 2010. ICCCI 2010 was technically co-sponsored by Shenzhen Graduate School of Harbin Institute of Technology, the Tainan Chapter of the IEEE Signal Processing Society, the Taiwan Association for Web Intelligence Consortium and the Taiwanese Association for Consumer Electronics. It aimed to bring together researchers, engineers and policymakers to discuss the related techniques, to exchange research ideas, and to make friends. ICCCI 2010 focused on the following themes:

- Agent Theory and Application
- Cognitive Modeling of Agent Systems
- Computational Collective Intelligence
- Computer Vision
- Computational Intelligence
- Hybrid Systems
- Intelligent Image Processing
- Information Hiding
- Machine Learning
- Social Networks
- Web Intelligence and Interaction

Around 500 papers were submitted to ICCCI 2010 and each paper was reviewed by at least two referees. The referees were from universities and industrial organizations. 155 papers were accepted for the final technical program. Four plenary talks were kindly offered by: Gary G. Yen (Oklahoma State University, USA), on “Population Control in Evolutionary Multi-objective Optimization Algorithm,” Chin-Chen Chang (Feng Chia University, Taiwan), on “Applying De-clustering Concept to Information Hiding,” Qinyu Zhang (Harbin Institute of Technology, China), on “Cognitive Radio Networks and Its Applications,” and Lakhmi C. Jain (University of South Australia, Australia), on “Intelligent System Design in Security.”

We would like to thank the authors for their tremendous contributions. We would also express our sincere appreciation to the reviewers, Program Committee members and the Local Committee members for making this conference successful. Finally, we would like to express special thanks for the financial support from the National Kaohsiung University of Applied Sciences, Kaohsiung City Government, National Science Council and Education Ministry, Taiwan, in making ICCCI 2010 possible.

Novermber 2010

Ngoc Thanh Nguyen
Jeng-Shyang Pan
Shyi-Ming Chen
Ryszard Kowalczyk

ICCCI 2010 Conference Organization

Honorary Chair

Chun-Hsiung Fang	National Kaohsiung University of Applied Sciences, Taiwan
Jui-Chang Kung	Cheng Shiu University, Taiwan

General Chair

Ngoc Thanh Nguyen	Wroclaw University of Technology, Poland

Program Committee Chair

Jeng-Shyang Pan	National Kaohsiung University of Applied Sciences, Taiwan
Shyi-Ming Chen	National Taiwan University of Science and Technology, Taiwan
Ryszard Kowalczyk	Swinburne University of Technology, Australia

Special Session Chairs

Bao-Rong Chang	National University of Kaohsiung, Taiwan
Chang-Shing Lee	National University of Tainan, Taiwan
Radoslaw Katarzyniak	Wroclaw University of Technology, Poland

International Advisory Chair

Bin-Yih Liao	National Kaohsiung University of Applied Sciences, Taiwan

International Publication Chair

Chin-Shin Shieh	National Kaohsiung University of Applied Sciences, Taiwan
Bing-Hong Liu	National Kaohsiung University of Applied Sciences, Taiwan

Local Organizing Committee Chair

Mong-Fong Horng	National Kaohsiung University of Applied Sciences, Taiwan

ICCCI 2010 Steering Committee

Chair

Ngoc Thanh Nguyen	Wroclaw University of Technology, Poland

Co-chair

Ryszard Kowalczyk	Swinburne University of Technology, Australia
Shyi-Ming Chen	National Taiwan University of Science and Technology, Taiwan
Adam Grzech	Wroclaw University of Technology, Poland
Lakhmi C. Jain	University of South Australia, Australia
Geun-Sik Jo	Inha University, South Korea
Janusz Kacprzyk	Polish Academy of Sciences, Poland
Ryszard Tadeusiewicz	AGH-UST, Poland
Toyoaki Nishida	Kyoto University, Japan

ICCCI 2010 Technical Program Committee

Jeng Albert B.	Jinwen University of Science and Technology, Taiwan
Gomez-Skarmeta Antonio F.	Murcia University, Spain
Shih An-Zen	Jinwen University of Science and Technology, Taiwan
Andres Cesar	Universidad Complutense de Madrid, Spain
Hsieh Cheng-Hsiung	Chaoyang University of Technology, Taiwan
Lee Chin-Feng	Chaoyang University of Technology, Taiwan
Badica Costin	University of Craiova, Romania
Godoy Daniela	Unicen University, Argentina
Barbucha Dariusz	Gdynia Maritime University, Poland
Greenwood Dominic	Whitestein Technologies, Switzerland
CAPKOVIC Frantisek	Slovak Academy of Sciences, Slovakia
Yang Fuw-Yi	Chaoyang University of Technology, Taiwan
Huang Hsiang-Cheh	National University of Kaohsiung, Taiwan
Chang Hsuan-Ting	National Yunlin University of Science and Technology, Taiwan
Lee Huey-Ming	Chinese Culture University, Taiwan
Deng Hui-Fang	South China University of Technology, China
Czarnowski Ireneusz	Gdynia Maritime University, Poland
Lu James J.	Emory University, USA
Kacprzyk Janusz	Polish Academy of Sciences, Poland

Marecki Janusz	IBM T.J. Watson Research, USA
Sobecki Janusz	Wroclaw University of Technology, Poland
Jung Jason J.	Yeungnam University, South Korea
Nebel Jean-Christophe	Kingston University, USA
Dang Jiangbo	Siemens Corporate Research, USA
Huang Jingshan	University of South Alabama, USA
Chang Jui-fang	National Kaohsiung University of Applied Sciences, Taiwan
Nunez Manuel	Universidad Complutense de Madrid, Spain
Gaspari Mauro	University of Bologna, Italy
Khurram Khan Muhammad	King Saud University, Saudi Arabia
Sheng Quan Z.	University of Adelaide, Australia
Katarzyniak Radoslaw	Wroclaw University of Technology, Poland
Unland Rainer	University of Duisburg-Essen, Germany
Ching Chen Rung	Chaoyang University of Technology, Taiwan
Shen Rung-Lin	National Taipei University, Taiwan
Yang Sheng-Yuan	St. John's University, Taiwan
Yen Shu-Chin	Wenzao Ursuline College of Languages, Taiwan
Chen Shyi-Ming	National Taiwan University of Science and Technology, Taiwan
Zadrozny Slawomir	Polish Academy of Sciences, Poland
Hammoudi Slimane	ESEO, France
Hong Tzung-Pei	National University of Kaohsiung, Taiwan
Hsu Wen-Lian	Academia Sinica, Taiwan
Pedrycz Witold	University of Alberta, Canada
Baghdadi Youcef	Sultan Qaboos University, Oman
Lo Yu-lung	Chaoyang University of Technology, Taiwan
Cheng Yuh Ming	Shu-Te University, Taiwan
Huang Yung-Fa	Chaoyang University of Technology, Taiwan
Ye Yunming	Harbin Institute of Technology, China

Keynote Speakers

Gary G. Yen	Oklahoma State University, USA
Lakhmi C. Jain	University of South Australia, Australia
Chin-Chen Chang	Feng Chia University, Taiwan
Qinyu Zhang	Harbin Institute of Technology Shenzhen Graduate School, China

Program Committee of Special Sessions

Dariusz Barbucha	Gdynia Maritime University, Poland
Bao-Rong Chang	National University of Kaohsiung, Taiwan
Hsuan-Ting Chang	National Yunlin University of Science and Technology, Taiwan

Chuan-Yu Chang	National Yunlin University of Science and Technology, Taiwan
Rung-Ching Chen	Chaoyang University of Technology, Taiwan
Shyi-Ming Chen	National Taiwan University of Science and Technology, Taiwan
Kazimierz Choroś	Wrocław University of Technology, Poland
Mohamed Hassoun	ENSSIB Villeurbanne, France
Mong-Fong Horng	National Kaohsiung University of Applied Sciences, Taiwan
Chien-Chang Hsu	Fu-Jen Catholic University, Taiwan
Wu-Chih Hu	National Penghu University of Science and Technology, Taiwan
Chien-Feng Huang	National University of Kaohsiung, Taiwan
Tien-Tsai Huang	Lunghwa University of Science and Technology, Taiwan
Huey-Ming Lee	Chinese Culture University, Taiwan
Che-Hung Lin	Cheng Shiu University, Taiwan
Lily Lin	China University of Technology, Taiwan
Piotr Jędrzejowicz	Gdynia Maritime University, Poland
Jeng-Shyang Pan	National Kaohsiung University of Applied Sciences, Taiwan
Chia-Nan Wang	National Kaohsiung University of Applied Sciences, Taiwan

Table of Contents – Part I

Collective Intelligence in Web Systems – Web Systems Analysis

Intelligent Computing for Data Security

Smart Clustering Techniques

Web Service

Quantitative Management of Technology (I)

Agent System

CCI-Based Optimization Models

Quantitative Management of Technology (II)

Fuzzy Logic and Its Applications

Table of Contents – Part II

Social Networks

Innovations in Computation and Application

Intellignet Signal Processing for Human-Machine Interaction (I)

Novel Approaches to Intelligent Applications

Intelligent Technologies for Medical Related Applications

Intellignet Signal Processing for Human-Machine Interaction (II)

Novel Approaches to Collective Computations and Systems

Intelligent Systems

Advanced Knowledgement Management (I)

Table of Contents – Part III

Intelligent Computing for Image Analysis (I)

Intelligent Digital Watermarking and Pattern Recognition

Advanced Knowledgement Management (II)

Intelligent Computing for Image Analysis (II)

Innovative Information System and Application

Intelligent Computing for Networks

Soft Computing to Industrial Management and Applications

Innovations in Pervasive Computing

Biological Computing

Neuro-Fuzzy Models in Global HTTP Request Distribution

Krzysztof Zatwarnicki

Faculty of Electrical Engineering, Automatic Control and Computer Science,
The Opole University of Technology, Opole, Poland
k.zatwarnicki@po.opole.pl

Abstract. In the paper, the new method GARD of HTTP request distribution, in globally distributed clusters of Web Servers, is presented. The method described uses neuro-fuzzy models, enabling the estimation of the HTTP request response time. Both, the description of the method and the research environment as well as the obtained results are described in the paper.

Keywords: HTTP request distribution, Neuro-Fuzzy model, Quality of web service, Request response time estimation, Distributed system.

1 Introduction

The Internet is a medium, thanks to which varied activities, such as: fun, entertainment, acquiring knowledge can be undertaken by a man. The aforementioned activities, redirected onto the virtual field of the Internet, cause an increase in the users' number of the informative, entertaining and social services. It enforces the need of improving the quality of the Web services, which is connected with a higher standard of efficiency in processing the requests at the level of Web-based centers, as well as the level of choosing the centre to service the requests on the Internet. Therefore, the Web systems processing the HTTP requests can be divided into systems of local processing and global Web-based systems, whose centres are located on the vast geographical area.

In the paper, the method of HTTP request distribution, in the globally distributed clusters of the Web servers, is presented. The method was described initially in [3]. However, the presented paper describes in detail the concept of neuro-fuzzy models, enabling estimation of the response time to the HTTP request. Also, the results of comparatory research of the two variations of presented method are shown herewith. The conducted research, concerning request distribution, is presented in section 2 and the GARD method is described in point 3. Point 4 ilustrates the research environment and the results of conducted simulation experiments. Point 5 contains a short summary.

2 Related Works

The issue of a request distribution in the Web systems has been raised in the number of papers. Most of them, however, concern the request distribution in the local clusters of

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 1–10, 2010.

the Web servers e.g. [1, 14, 5]. The request distribution in the globally distributed systems is realized mostly with the use of an adequate DNS system e.g. [4, 8, 9]. There are also globally distributed systems employing brokers to redirect the request, located in a close proximity to a client. The presented method uses neuro-fuzzy models enabling estimation of the client's request response time. The models were previously applied in the methods of the HTTP request distribution in the local clusters of the Web Server e.g. [1] and in the aforementioned global distribution with the brokers [2]. There are also other papers describing a fuzzy approach to the distribution [10] and the HTTP request scheduling [6].

3 Description of GARD Method

The GARD method is designed to distribute the HTTP requests in the Web services working in a Wide Area Network, especially in the Internet. GARD is an acronym and stands for Global Adaptive Request Distribution.

The scheme of the request distribution system applying the GARD method is presented in Fig. 1.

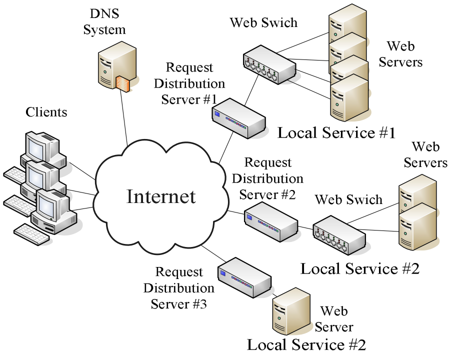


Fig. 1. The scheme of the request distribution system GARD

The components of the GARD system are: clients, Local services, request distribution servers, DNS system.

Local services (LS) are the clusters of the Web servers, consisting of: a web switch and a certain number of the Web servers or a single Web server. Each of the Local service in the system is able to process every request of the request set admissible by the Web service.

Request distribution servers are located nearby the Local servers and connected with them through the local network. Exactly one request distribution server falls on every single Local service and has its own IP address. The servers are responsible for the clients' HTTP request distribution.

Every client, who wants to download an HTTP object from the Web service for the first time, is required to obtain the IP address on the basis of a mnemonic address. The DNS system translates the mnemonic address of the service into the IP address of the request distribution server, with the use of an appropriate DNS server. The DNS system is permitted to select any address of the request distribution server. The further part of the paper presents the research on the GARD system, in which the DNS system selects the request distribution server located closest to the client or the server allocated to the local service with the lowest workload rate.

Next, the client sends the HTTP request to the request distribution server, which measures the connection time. As follows, the request is dispatched by the request distribution server to the local service, to which it's been allocated. The answer is obtained in a form of the HTTP object. If the object is an HTML file, it is modified by the request distribution server in a way, so that the addresses of the embedded objects would point the same objects located on other local services. At least one object should be downloaded by the client from each of the local services, whereas the rest of the objects should be downloaded from the original local service. The request distribution server forwards to the client the HTTP response containing requested object and a demand to set a specific cookie values: local service identifier and the connection time, using a cookie mechanism. When the client sends the requests to the remaining local services, in order to download specific objects embedded in the Web site, each of the respective servers places its measured connection time in the cookie field.

When the client connects with the server again, in order to download a consecutive HTML file, the file itself is modified, so that the addresses of the objects embedded in the Web site and the links would indicate local services with the shortest response time. The request distribution server estimates the response time for the requested HTTP objects, on the basis of the connection sent in the cookie and the knowledge of the workload rate of the other local services.

The individual request distribution servers exchange both, the information on the workload of the local services and the knowledge required to estimate the time of the request processing.

In order to estimate the response time, the model of the local service was constructed, whose scheme is presented in Fig 2.

The response time is defined as the time measured from the moment, when the first byte of the HTTP request is obtained by the request distribution server, up to receiving a confirmation about the last byte being obtained by the client. The confirmation is sent by the client in the TCP protocol, with the ACK flag set in the TCP segment.

The module inputs in the local service model are: the address u_{ig} of the HTTP object and the workload of the local service $O_i = [a_i, \bar{t}_i]^T$ described by the number of requests processed simultanously by the local service a_i and the connection time $\bar{t}_i$. The lower index i points that data in the request distribution server are most recent, at the time of i-th request income. The workload of the local service is described as the number of the HTTP requests, being processed actively by this service, and its value is recorded at specified time intervals. The model of the Web service comprises of four functional modules: a classification module, an estimation mechanism, an adaptation mechanism and a load state module.

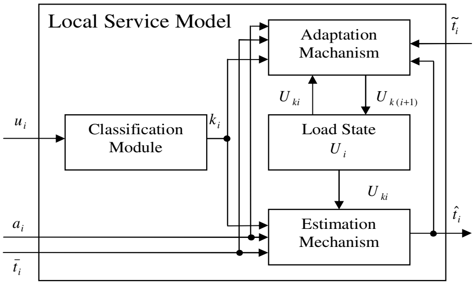


Fig. 2. Local service model

The classification mechanism, classifies all objects requested by the client. Its input is the address u_i of the requested object, where i is the number of the consequtive requests incoming to the Web service $i = 1,2,3,4,....$. The classification mechanism contains information on the sizes and the types of the objects offered by the service. The objects are classified in a way that the response times for objects belonging to the same class are similar. Static objects, that are files offered by the service, are classified according to their size. Dynamic objects, which are the objects created by the Web service at the moment of the client's request income, are classified so that each of the objects was allocated to a different class. On the output of the classification mechanism, class k_i of the requested object is obtained, where $k_i \in \{1,2,3,...,K\}$, and K is the number of classes of the objects processed by the Web service.

The load state module stores data being the information U_i, which enables estimation of the request response time. The load state was denoted as $U_i = [U_{1i},...,U_{ki},...,U_{Ki}]^T$, where U_{ki} is data, concerning the requests of the k-th number class and $U_{ki} = [\Lambda_{ki}, \Gamma_{ki}, \Phi_{ki}]$, whereas $\Lambda_{ki} = [a_{0ki},...,a_{lki},...,a_{(L\text{-}1)ki}]$, $\Gamma_{ki} = [t_{0ki},...,t_{mki},...,t_{(M-1)ki}]$, $\Phi_{ki} = [y_{1ki},..., y_{jki},..., y_{Jki}]$. The parameters a_{lki}, t_{mki} and y_{jki} are the parameters of the membership function of an input and an output of fuzzy-sets in the neuro-fuzzy model described below.

The estimation mechanism is applied to estimate the response time $\hat{t}_i$ to i-th request. One of its input is the workload of the Web service a_i, the second input is the connection time $\bar{t}_i$. The estimation mechanism, according to the class of the requested object, receives the information U_{ki} from the load state module, required to estimate the response time. When the calculations are complete, the estimated response time $\hat{t}_i$ is returned. The adaptation mechanism modifies the information U_{ki} included in the load state module U_i, concerning k-th class of the objects. The modification is processed on the basis of the information on the following: the load state of the Web service a_i, the connection time $\bar{t}_i$, class of requested object k_i and measured response time $\tilde{t}_i$.

The estimation and adaptation mechanisms form a neuro-fuzzy model presented in Fig. 3a. We can distinguish 3 blocks in the model: Fuzzification Layer, Rule Layer and Defuzzification Layer.

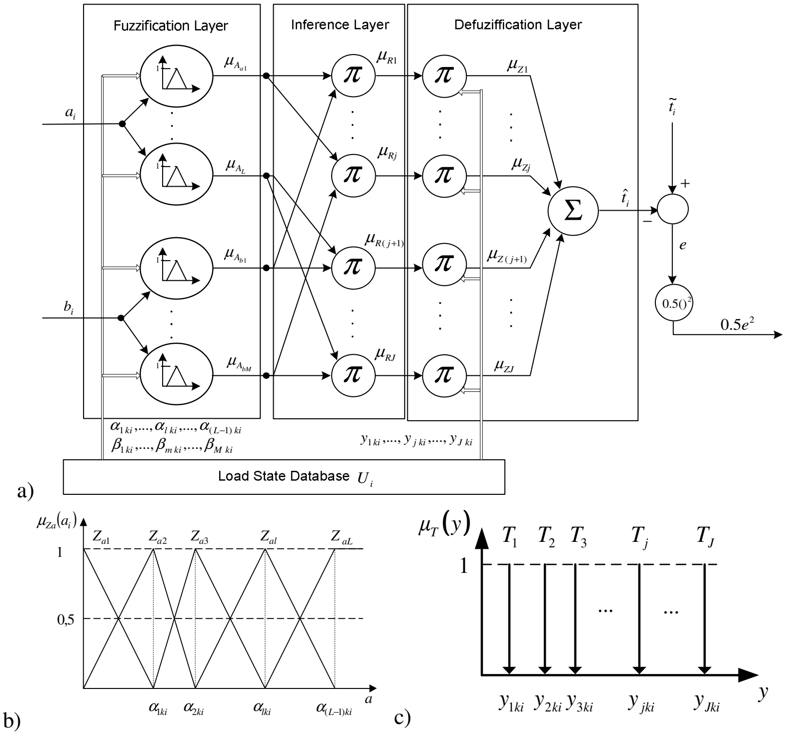


Fig. 3. a) The neuro-fuzzy model of the service model; b) Fuzzy sets for p_{ki}; c) Fuzzy sets for the output

For the input a_i we define L fuzzy sets Za1,…,ZaL and M fuzzy sets Zb1,…,ZbM for $\bar{t}_i$. Membership functions for both inputs are triangular and feature a partition of unity. Their shaping is shown in Fig. 3b (more specifically, for a_i), where parameters α_{lki} define the shaping of the membership functions for a_i. The similar parameters β_{mki} define the shaping of the membership functions for $\bar{t}_i$. The fuzzy sets of the outputs T1,…,TJ are assumed to be the singletons, pointing at

$y_{1ki},\ldots,y_{jki},\ldots,y_{Jki}$ values (Fig. 3c). The parameters are used as the weights of neurons and are tuned during the adaptation process.

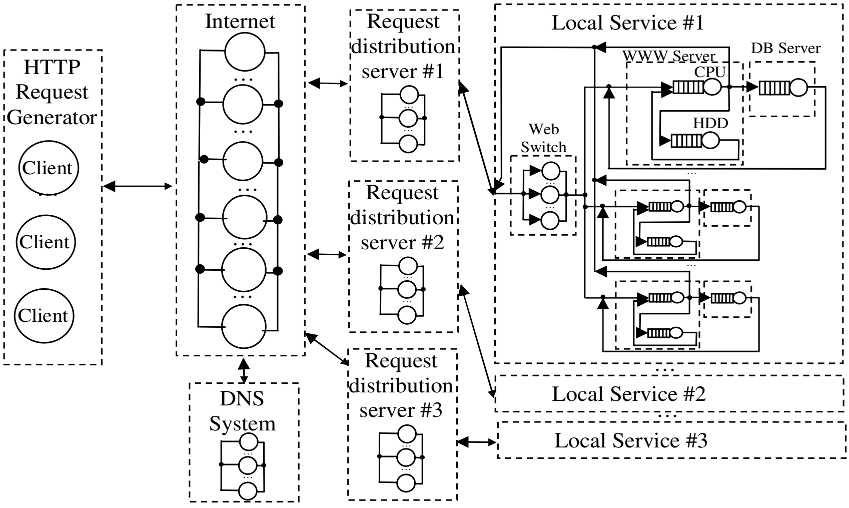
The individual rules are formed as follows R j: IF (a = Zal) AND ($\bar{t}$ =Zbm) THEN (y=Tj), where j is the number of the fuzzy rule and equals to $j = 1,2,3,\ldots,M$, l is the number of the fuzzy set of the a_i input, m is the number of the fuzzy set of the $\bar{t}_i$ input, a is a linguistic variable of a_i input, $\bar{t}$ is the linguistic variable of $\bar{t}_i$ input.

PROD fuzzy operator is used as AND connective, thus the degree μ_{Rj} of the membership of antecedent of rule Rj is equal to $\mu_{R_j}\left(a_i ,\bar{t}_i\right) = \mu_{Z_{al}}\left(a_i\right)\cdot \mu_{Z_{bm}}\left(\bar{t}_i\right)$. PROD operator is used for the implication function. The result of the fuzzy inference is a fuzzy value that has to be retransformed (i.e. defuzzified) into a crisp value. The "weight" method, which favors the rule with the highest output value, is used in this operation. $\tilde{t}_i$ is used to tune the weights of the neurons $\alpha_{\phi kg}$, $\beta_{\theta ki}$ and y_{jki} which are modified using the back propagation method each time after the completion of the request. ϕ and θ are the indexes of the membership functions for inputs a_i and $\bar{t}_i$, respectively, $\phi = 1,2,\ldots,L-1$ and $\theta = 1,2,\ldots,M-1$. The mean square error $E_i = 0.5(e_i)^2$ is calculated for the neuron in the output layer. Then the new weight values are calculated as $\alpha_{\phi k\ (i+1)} = \alpha_{\phi \mathrm{k}\ (i)} + \eta_a\left(\tilde{t}_i - \hat{t}_i\right)\cdot y_{jk\mathrm{i}} \cdot \sum_{j=1}^{J}\left(\mu_{Z_{bm}i}\right)\cdot \sum_{l=1}^{L}\left(\frac{\partial \mu_{Z_{al}i}}{\partial \alpha_{\phi k\ (i)}}\right)$, $y_{jk(i+1)} = y_{jk(i)} + \eta_y\left(\tilde{t}_i - \hat{t}_i\right)\cdot \mu_{R_j i}$, $\beta_{\ k\ (i+1)} = \beta_{\ k\ (i)} + \eta_b\left(\tilde{t}_i - \hat{t}_i\right)\cdot y_{jk\mathrm{g}} \cdot \sum_{j=1}^{J}\left(\mu_{Z_{al}i}\right)\cdot \sum_{m=1}^{M}\left(\frac{\partial \mu_{Z_{bm}i}}{\partial \beta_{\ k\ (i+1)}}\right)$. The values of η , L, M should be found established in preliminary research.

4 Simulation Model and Experiment Results

Conducted research let us evaluate the quality of the Web system, working under GARD supervision. The research was run as experimental tests. The simulators were built with the use of the CSIM19 package [13], which enabled the construction of process orientated simulation models. The simulation programme included the following modules: a request generator, the Internet, DNS system, a request distribution server and a local service. The scheme of the simulation model is presented in Fig. 4a.

The clients' requests were generated in the module in a way, to make the nature of generated HTTP request flow correspond with the one observed in the Internet, and to be characterized by burst and self-similarity. Fig. 4b and 4c present the probability distribution and its parameters, applied when constructing the HTTP request generator [4].



a)

Category	Distribution	Parameters
Requests per session	Inverse Gaussian	μ=3.86, λ=9.46
User think time	Pareto	α=1.4, k=1
Objects per request	Pareto	α=1.33, k=2

HTML object size	Lognormal
	Pareto
		μ=7.630, σ=1.001
		α=1, k=10240

Embedded object size	Lognormal	μ=8.215, σ=1.46

b)

Type	Mean service time	Frequency
High intensive	20 ms	0.85
Medium intensive	10 ms	0.14
Low intensive	5 ms	0.01

c)

Fig. 4. a) A simulation model; b) Workload model parameters; c) Workload model parameters of dynamic objects

The Internet module of the simulator was used to model the latency observed on the Internet. The data transmission time for sending the request and receiving the response of the request is calculated as follows [12]:

$$data\_transmission\_time = RTT + \frac{object\_size + HTTP\_response\_header\_size}{throughput}$$

where RTT (Round Trip Time) is the transmission time of a single datagram IP sent to the local service and back again. The average size of the HTTP header $HTTP\_response\_header\_size$ was assumed to amount 290 B, and a $throughput$ was calculated as a number of data bytes, transmitted in the net per time unit. Respective experiments were run on the Internet, which contributed to obtaining an ordered set of the RTT values and throughput. The tests were based on sending requests to fifty different Web services, at specified time intervals and downloading a document RFC1832 of the size equal to 47 KB from them. Fifty different scenarios of the client's behaviour on the Internet, were generated in this process.

Each simulated client possessed behaviours chosen randomly from the fifty behaviours. The number of behaviours for each client was the same as the number of clusters in the system modeled. In the simulator, also the local cluster was modeled and its module comprised of: a Web dispatcher, WWW servers and database servers. The role of the Web dispatcher was to distribute the requests to an appropriate WWW server in accordance with the assumed algorithm of the request distribution. It worked under the FNRD adaptation algorithm [1], one of the best algorithms of the request distribution for the local cluster systems.

The model of the WWW server contained a processor, a hard drive and a cache working according to the LRU algorithm (Last Recently Used). When the processor is considered, the following were taken into account: the connection time, the time required by the WWW daemon to prepare the response and data transmission time. In case of the hard drive, the time of positioning the head over the appropriate cylinder was considered and data transmission time.

It was possible in simulator to process static and dynamic requests. The dynamic requests were processed by the WWW and the database servers, whereas the static requests only by the WWW server with the use of the cache. The dynamic requests were categorised into 3 classes, similarly to [4]: highly intensive, medium intensive and low intensive.

The time of processing the dynamic requests by the database server was modeled according to the hyper-exponential distribution. The distributions parameters of particular types of the dynamic requests are presented in Fig. 4c. In the research it was assumed that 80% of the sites is static and 20% dynamic.

In the discussed solution, the DNS system was responsible for taking decisions on the requests distribution. Therefore, four algorithms of the request distribution were implemented in the DNS module: Round-Robin (RR); Weighted Round-Robin Load (WRR_L), the weights in the WRR_L algorithm were selected on the basis of the local service workload rate, measured as the number of requests processed actively; Round Trip Time (RTT), used by the DNS system to point the address of the local service, for which the RTT time (from the client to the local service) was the shortest.

The simulation module of the request distribution server could work in two modes. In the first mode, it was passing all incoming requests directly to a local cluster. In the second mode, GADR method was involved in the request distribution process. When the request distribution server worked in the first mode the DNS system was responsible for distributing requests according to mentioned three algorithms: RR, WRR and RTT. When the distribution server worked according to the GARD method, the DNS system could work according to the chosen policy of the request distribution. The two policies were considered: RTT and WRR_L and therefore two types of the GARD system were created: GARD_RTT and GARD_L respectively.

The research was run in two configurations of the local service. In the first one, the number of the servers for every cluster was equal and amounted 3. In the second configuration, the number of the servers for the first cluster was 1, for the second cluster was 3 and for the third was equal to 5. In order to determine the properties of the new methods, individual clusters were loaded in a different way. Thus, during the first experiment the number of clients in the closest proximity to individual clusters was equal for all of the clusters (the experiment was marked as follows 3/33%,3/33%,3/33%). In the second experiment, the number of the clients in the closes proximity to different clusters was different and amounted: for the first cluster 11%, second 33%, third 55% (marked as: 3/11%,3/33%,3/55%). The cluster workload in the second configuration (for a different number of the servers in clusters) during the first experiment amounted: 11%, 33%, 55% for the first, second and third cluster respectively (marked as: 1/11%,3/33%,5/55%). In the last experiment, the workload was as follows: 55%, 33%, 11% (marked as: 1/55%,3/33%,5/11%). Figure 5 presents the experiment results for both configurations and different scenarios of the service workload. The graphs show the relationship between the mean response time to a HTTP request and the number of new clients incoming within one second.

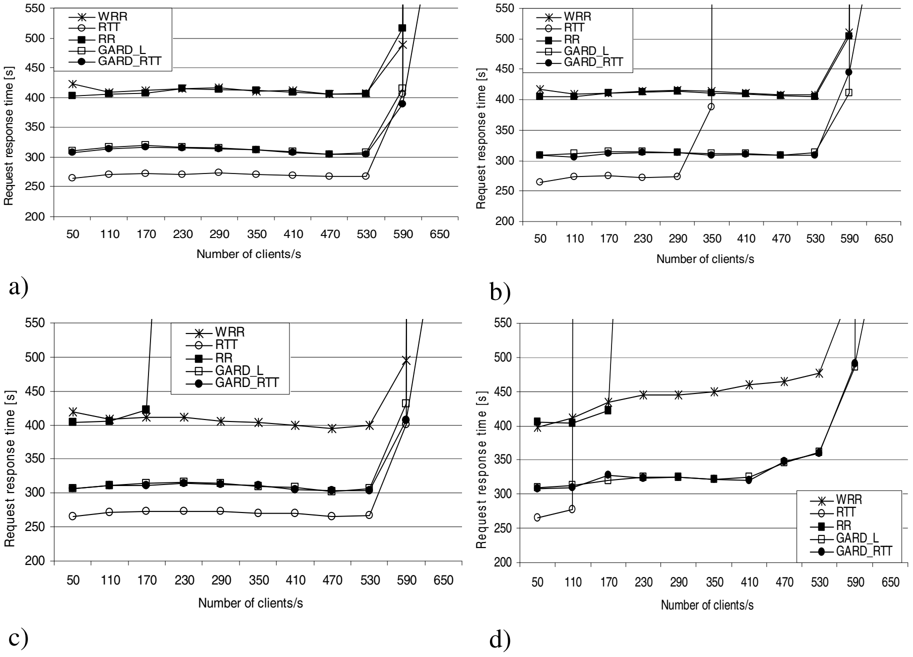


Fig. 5. HTTP request response time vs. number of clients per second: a) 3/33%, 3/33%, 3/33% run, b) 3/11%, 3/33%, 3/55% run, c) 1/11%, 3/33%, 5/55% run, d) 1/55%, 3/33%, 5/11% run

The results indicate, that the systems working according to the methods GARD_RTT and GARD_L, dispatch the requests efficiently in the environment of globally distributed systems, obtaining the shortest response times, almost in all of the tests. The results obtained for the RTT method were moderately good, but only in a perfectly balanced environment (Fig. 5a), which is impossible in real conditions. As presented in the graphs, the response time for both methods: GARD_L and GARD_RTT were similar, therefore it is recommended to use the GARD_L method, as cheaper and much easier in application, as it does not require the complicated DNS system to be employed with.

5 Conclusion

In the paper, a new GARD method of the HTTP request distribution in the globally distributed clusters of the Web servers, is presented. Both, the description of the method using neuro-fuzzy models and the research environment are described, alongside the results for the two types of presented distribution method: GARD_RTT and GARD_L. Presented results point, that it is useful to continue further studies concerning the GARD method, which could be introduced in productive Web systems.

References

1. Borzemski, L., Zatwarnicki, K.: Fuzzy-Neural Web Switch Supporting Differentiated Service. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 195–203. Springer, Heidelberg (2006)
2. Borzemski, L., Zatwarnicki, K., Zatwarnicka, A.: Adaptive and Intelligent Request Distribution for Content Delivery Networks. Cybernetics and Systems 38(8), 837–857 (2007)
3. Borzemski, L., Zatwarnicki, K.: CDNs with Global Adaptive Request Distribution. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part II. LNCS (LNAI), vol. 5178, pp. 117–124. Springer, Heidelberg (2008)
4. Cardellini, V., Casalicchio, E., Colajanni, M., Mambelli, M.: Web Switch Support for Differentiated Services. ACM Perf. Eval. Rev. 29(2), 14–19 (2001)
5. Casalicchio, E., Colajanni, M.: A Client-Aware Dispatching Algorithm for Web Clusters Providing Multiple Services. In: Proc. WWW 2010, pp. 535–544 (2001)
6. Cherkasova, L., Phaal, P.: Session-based admission control: A mechanism for peak load management of commercial Web sites. IEEE Transactions on Computers 51(6) (2002)
7. Cisco, Boomerang Control Protocol, TX (2009), `http://www.cisco.com/warp/public/cc/pd/cxsr/cxrt/tech/ccrp_wp.htm#wp17824`
8. Hong, Y.S., No, J.H., Kim, S.Y.: DNS-Based Load Balancing in Distributed Web-server Systems. In: The Fourth IEEE Workshop on Software Technologies for Future Embedded and Ubiquitous Systems, and the Second International Workshop on Collaborative Computing, Integration, and Assurance (SEUS-WCCIA 2006) (2006)
9. Pan, J., Thomas Hou, Y., Li, B.: An overview of DNS-based server selections in content distribution networks. Computer Networks 43(6), 695–711 (2003)
10. Jianbin, W., Cheng-Zhong, X.: QoS: Provisioning of client-perceived end-to-end QoS guarantees in Web servers. IEEE Trans. on Computers 55(12) (December 2006)
11. Lee, K.-M., Kwak, D.-H., Leekwang, H.: Tuning of fuzzy models by fuzzy neural networks. Fuzzy Sets and Systems 76(1), 47–61 (1995)
12. Menasce, D., Almeida, V.: Capacity Planning for Web Performance. Metrics, Models, and Methods. Prentice Hall, New York (1998)
13. Mesquite Software Inc. CSIM19 User's Guide. Austin, TX, `http://www.mesquite.com`
14. Pai, V.S., Aron, M., Banga, G., Svendsen, M., Druschel, P., Zwaenpoel, W., Nahum, E.: Locality-Aware Request Distribution in Cluster-Based Network Servers. SIGOPS Oper. Syst. Rev. 32(5), 205–216 (1998)

Real Anomaly Detection in Telecommunication Multidimensional Data Using Data Mining Techniques

Kazimierz Choroś

Institute of Informatics, Wrocław University of Technology,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
kazimierz.choros@pwr.wroc.pl

Abstract. The tremendous amount of data are generated and collected by telecommunication companies. These data include call detail data which describe the calls traversing the telecommunication networks as well as network and customer data which mainly describe incomes of telecommunication companies. The amount of data is so huge that manual analysis of these data is impossible. The need to automatically handle such large volumes of data has led to the development of special algorithms and technologies such as data mining, intelligent computer agents, knowledge-based expert systems, etc. Telecommunication companies are strongly interested not only in identifying fraudulent phone calls and identifying network faults but also in forecasting the preferred directions of customer calls or the incomes of the companies. The paper presents a communication real anomaly detection framework, which uses data mining technologies using OLAP cube built for telecommunication data.

Keywords: anomaly detection, telecommunication data, multidimensional data, OLAP cube, data mining.

1 Introduction

The amount of telecommunication data is so huge that manual analysis of these data is very difficult or even impossible. Furthermore, the data are growing all the time. It creates a great demand for efficient data mining techniques in order to help identify telecommunication patterns, catch fraudulent activities, make better use of resources, and in consequence improve the quality of service.

Telecommunication data are of multidimensional nature. It is impossible to analyze them by using only statistics. The main features and technical parameters are: type of call, calling time, duration of call, call destination, location of caller, etc. The multidimensional analysis of such data can be used to forecast the data traffic, user group behavior, company profit, and so on. For example, an analyst in the telecommunication company may wish to regularly view charts regarding calling source, destination, volume, and time-of-day usage patterns. Therefore, it has been suggested to consolidate telecommunication data into large data warehouse and then perform multidimensional analysis using OLAP technique. One of the most frequent purpose of multidimensional analysis is the detection of anomalies, such as abnormal customer traffic, abnormal call

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 11–19, 2010.

duration, and mainly sudden, unpredictable decreases of calls to certain destinations. Anomaly detection is based on models of normal data and points out any deviation from the normal model in the observed data. The goal is to determine whether the data examined belong to a normal observed earlier or to an anomalous behavior. An action, an event or a behavior is treated as anomalous if its degree of deviation in relation to the profile of characteristic behavior of the customer, defined by the model of normality, is high enough. The anomaly detection techniques have the advantage that they can detect new types of unwanted actions as deviations from normal usage. Unfortunately, their weakness is rather high false alarm rate.

Formally, an anomaly detection in a customer behavior B can be defined as a pair

$$B = (M,S), \tag{1}$$

where: M is the model of normal behavior of a customer,
S is a similarity measure that allows obtaining the degree of coincidence or deviation that such behavior has with regard to the model M.

The anomaly detection systems are frequently subsystems of intrusion detection systems. In our telecommunication analyses an anomaly will be mainly seen as a significant increase or decrease of calls to a certain destination.

The paper is organized as follows. The next section describes the main related works in the area of data mining techniques applying for the analysis of telecommunication data. The third section presents the specificity of telecommunication data. In the sections 4 the proposed OLAP structure is described and the experimental results for the tested data are reported. Section 5 presents the analysis of results of tests leading to profit prognosis formulation as well as anomaly detections. The final conclusion and the future research work areas are discussed in the last 6th section.

2 Related Works

There are many recent investigations towards automatic analysis of telecommunication data [1, 2], many of the data mining techniques have been proposed and tested on telecommunication data [1, 3-5]. Also techniques based on metaheuristics (optimization/search/learning algorithm skeletons) and nature-inspired paradigms can provide really efficient solutions to a wide variety of problems, arising in operations research, parallel computing, telecommunications, data networks, and many other disciplines [6].

It has been also proposed [7] to automate the data mining process of telecommunication data using cognitive and cooperative capabilities of intelligent agents and Web services encapsulating specific data mining models throughout the data mining process cycle.

Optimizing telecommunication network traffic is a special case of routing with great value due to the large volume. The problem is extremely difficult because traffic lead and network topology vary with time in unpredictable ways and the lack of central coordination. World leading telecommunication companies like France Telecom, British Telecom, and the former MCI WorldCom have successfully used ant colony optimization (ACO) algorithm [8].

One of the most attractive procedure for telecommunication companies and network suppliers is an anomaly detection. The methods used to detect anomalies are classified into four categories [9, 10]: statistical anomaly detection, classifier based anomaly detection, anomaly detection using machine learning, and finite state machine anomaly detection. It has been also proposed a novel anomaly detection framework [11], which uses data mining technologies to build four independent detection models. A detection framework is dedicated to detecting HTTP request exploitations. The framework is composed of four detection models. The classifiers calculate the anomalous probabilities for HTTP requests upon detection to decide whether they are anomalous or not.

In [12] an anomaly detection method for mobile telecommunication networks is presented that describes the normal states of the system with a so-called self-organizing map identified from the data. Large deviation in the data samples from the self-organizing map nodes is detected as anomalous behavior. It has been detected using global thresholds. The method has proved to be useful in practice in finding new phenomena in network system. Furthermore, it is noted that data collected from an operational GSM network during only one day consist of several gigabytes of data. It is obvious that it is impossible for network operators to analyze all these data manually, especially in multivariate space, where it is impossible to visualize all the dimensions of the data space simultaneously. Therefore, automated multivariate methods are needed to analyze the data sets.

In anomaly detection, anomalous data records are typically detected as deviations from normal data modeled as unimodal data distributions [13].

The data mining techniques are also applied for detecting anomalies in many other huge sets of data such as for example railway data [14].

3 Multidimensional OLAP Cubes

Telecommunication data have their own specificity. The two main sets of data are: invoice date and traffic data. The analysis of the invoice data should permit to forecast a profit of the telecommunication company. Whereas, the analysis of traffic data may prevent the loss of profit in a given destination due to the detection of anomalies.

The data usually included in invoices are: invoice period, date of issue, caller number, caller personal data, i.e. name, address, tax identification number, and then cost of every type call, total cost.

Whereas, the traffic data describe connections in telecommunication networks. Every single connection generates a record with such data as: caller number, number dialed, start time of a call, call end time, call duration, call rate, type of call.

A special application has been implemented to analyze and to forecast incomes and to forecast telecommunication network load, and as well as to detect any anomaly. An OLAP (OnLine Analytical Processing) cube [15, 16] has been applied in which more than 3 hundred million records of telecommunication connections from 40 months have been gathered. These connections concerned about 25 thousand customers from 7 market segments living in 4644 localities and over 11 million invoices have been made out. Nevertheless, these data have been deprived of personal parts with names or addresses. Because the telecommunication data used in the tests are secret only some fragmentary numerical results will be presented in the paper.

An OLAP cube is defined as the structure enabling data manipulation and analyzes from multiple perspectives. The arrangement of data into cubes overcomes a limitation of relational databases encountered in case of large amounts of data. The OLAP keeps data in an optimized multi-dimensional array storage. Therefore, it requires the pre-computation and storage of information in the cube. Each cell in a multidimensional structure contains aggregated data related to elements along each of its dimensions. These numeric data are called measures which are categorized by dimensions. A dimension is a data element that categorizes each item in a data set into non-overlapping regions. The main function of dimensions is to provide filtering, grouping and labeling.

Two OLAP cubes have been created. The first cube has been defined for income data and the second one for telecommunication traffic data. The process of the OLAP tubes constructions has started by establishing connections between the data base where all telecommunications data were gathered. Then the data source views have been defined where the adequate tables could be chosen for further analysis. Next the dimensions of multidimensional structures have been proposed. In Table 1 the dimensions in these two OLAP tubes of telecommunication data are presented.

Table 1. Dimensions in the telecommunication data OLAP tubes

Cube	Dimensions	Relations/Descriptions
Invoices	connection type	type of the connection
	client address	client address with number zone
	client name	full name of the client
	client type	category of the client
	call type	type of the call
	period	month of connections invoiced
Traffic	client type	category of the client
	connection type	type of the connection
	client number	number of the client who dials
	dialed number	number of dialed client
	date	date of the connection
	month	month of the connection
	week	week of the connection

Finally the data objects have been introduced to two defined OLAP tubes and at the same time the aggregations have been calculated for different levels of the dimension hierarchy.

4 Anomaly Detection – Tests on Real Data

The tests have been conducted with a large data set collected in one of Polish telecommunication companies. Now the processing of anomaly detection will be reported. In the anomaly process detection the traditional control charts methods used in a statistical process control have been applied. Statistical process control consists in the application of statistical methods to the monitoring and control of a process. The goal is to be able to examine a process and the sources of variation in that process and in consequence to early detect and prevent a problem before it occurs. The control chart is a primary tool and a graphical representation of descriptive statistics for specific quantitative measurements of a process. These descriptive statistics are usually displayed in the control chart in comparison to their "in-control" sampling distributions. The comparison detects any unusual variation in the process, which we call an anomaly. Control charts are also used with product measurements to analyze process capability and for continuous process improvement efforts.

The computer application with classes presented in Table 2, used in the tests on real data, has looked for the adequate data in the OLAP tubes to form time series. The time series data have been analyzed to find probable anomalies.

Table 2. Description of the classes in anomaly detection application

Class	Description
BaseTimeSeries	Basic class for time series with method of time series generation as well as the adequate values and dates extraction
TimeSeries	Methods of time series analysis and anomaly detection
BaseCubeConnection	Methods of connections with OLAP tubes
CubeConnection	Methods of time series generation
Anomaly	Class defining anomaly and generating anomaly diagram
TimeSeriesPlot	Class defining diagram of a time series with an anomaly
Calendar	Class with calendar indicating working days
Statistics	Statistical methods for the calculation of mean values, standard deviation, minimal and maximal values

The standard approach based on control charts using 3-sigma limits was applied. Let *w* be an observed value, for example the amount on the invoice. Then μ_W is a mean value with the standard deviation σ_W. The control line is equal to the mean value. An upper line for the upper control limit (UCL) is define as follows:

$$UCL = \mu_w + 3\ \sigma_w \tag{2}$$

and a lower line for the lower control limit (LCL):

$$LCL = \mu_w - 3\ \sigma_w \tag{3}$$

Because the values μ_W and σ_W are unknown we estimate them by the mean value of the sample

$$\mu_w = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4}$$

and by the root of unbiased variance estimator

$$\sigma_w = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \mu_w\right)^2}{n-1}} \tag{5}$$

where: x_i is a value from the sample,
n is the sample size.

If one of the data points lies outside the UCL and LCL, i.e. outside the 3-sigma limits, we should investigate this special case. We call it an anomaly which should be carefully examined.

The control charts may also show us an upward or downward trends. A trend is usually a series of consecutive data points which lie on one side of the mean. There is a controversy as to how long a run of observations, all on the same side of the centre line, should count as a trend, the most frequently it is 6, 7, 8, or 9 points.

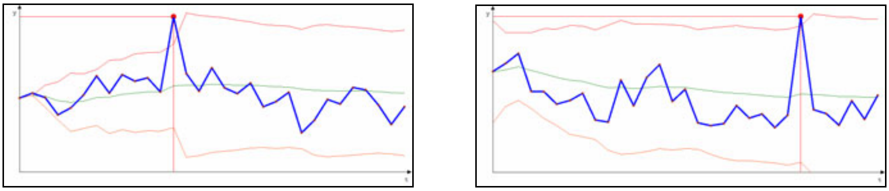


Fig. 1. Different examples of the anomalies in income real data for different connection directions – points above the mean value

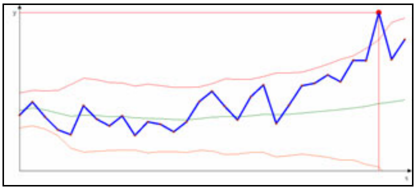


Fig. 2. Example of the trend and the anomaly in income real data for different connection directions – points above the mean value

Figures 1 and 2 present several examples of control charts of income data. They illustrate the invoice amounts in consecutive months for one of the segments of customers. The first two charts in Figure 1 show simple deviation from average, whereas in the example in Figure 2 we can also observe an upward trend. For a telecommunication

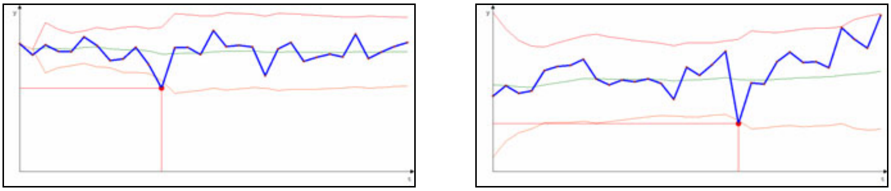


Fig. 3. Different examples of critical anomalies in income real data for different connection directions – points below the mean value

company anomalies above the mean in income data are not disquieting, but anomalies below the mean should be a warning. They are critical.

In two charts in Figure 3 we observe simple deviations. The next charts in Figure 4 present the cases which should alarm the telecommunication company. We observe not only anomalies but also downward trends. It is a warning that a customer makes fewer and fewer connections what leads to the decrease of incomes.

These examples show that the two main kinds of anomalies: above and below the mean values should be treated in different ways. It may lead to the strategy of significant reduction of faulty alarms when detecting anomalies.

The Figure 5 presents the numbers of anomalies classified according to the type of calls. The figure shows that for local calls the number of anomalies above and below the mean is similar but for all other directions contrary these numbers significantly differ. Fortunately, the numbers of anomalies above the mean - which are not so disquieting - are much greater.

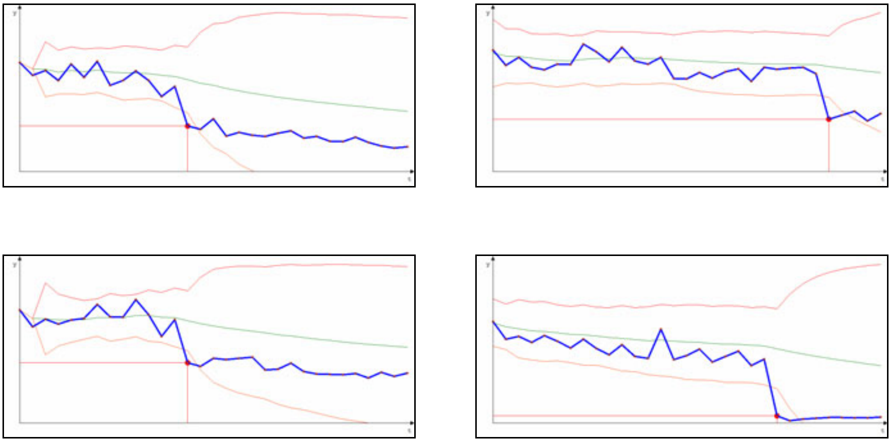


Fig. 4. Different examples of downward trends and of critical anomalies in income real data for different connection directions – points below the mean value

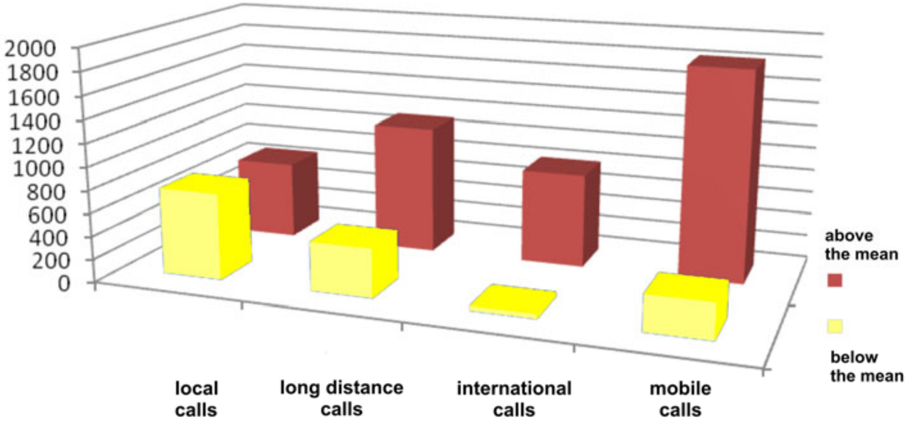


Fig. 5. Numbers of anomalies classified according to the type of calls (main directions)

5 Final Conclusion and Further Studies

The analysis of a huge amount of data is much easier when the data are introduced into the multidimensional structures of OLAP type. This is also the case with telecommunication data. The OLAP structure enables among others the detection of anomalies in traffic and invoice data. The anomalies have been categorized into two groups: critical and non-critical anomalies. The fast detection of critical anomalies may suggest to a telecommunication company to undertake some activities to prevent lost of a customer. Furthermore, the numbers of anomalies for different segments of customers significantly differ. It means that some customer segments are much more stable than others.

The tests have been shown how we can detect anomalies. The analysis of control charts may significantly reduce the number of faulty alarms. But then it is also possible for example to analyze in which months of the year the greatest number of anomalies are observed. Generally, the period of summer holidays as well as Christmas time is well known as a time of different abnormalities. The first statistical analysis show that the majority of anomalies has been detected for summer months. The problem remains how to discriminate real anomalies between all of them.

Further studies will be undertaken to automatically detect these anomalies which were classified as real and critical.

References

1. Sumathi, S., Sivanandam, S.N.: Introduction to Data Mining and its Applications. In: Part 24. Data Mining in Telecommunications and Control. Studies in Computational Intelligence (SCI), vol. 29, pp. 615–627 (2006)
2. Weiss G.M.: Data Mining in Telecommunications. In: Maimon, O., Rokach, L.: Data Mining and Knowledge Discovery Handbook, ch. 56, pp. 1189–1201. Springer, US (2005)

3. Sumathi, S.: Data mining and data warehousing. In: Sumathi, S., Esakkirajan, S. (eds.) Fundamentals of Relational Database Management Systems. Studies in Computational Intelligence (SCI), vol. 47, pp. 415–475 (2007)
4. Berry, M.J.A., Linoff, G.S.: Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, 2nd edn. John Wiley & Sons, Chichester (2004)
5. Berry, M.W., Castellanos, M. (eds.): Survey of Text Mining II. Clustering, Classification, and Retrieval. Springer, Heidelberg (2007)
6. Alba, E., Talbi, E., Zomaya, A.Y.: Nature-inspired distributed computing. Computer Communications 30, 653–655 (2007)
7. Rocha-Mier, L.E., Sheremetov, L., Batyrshin, I.: Intelligent agents for real time data mining in telecommunications networks. In: Gorodetsky, V., Zhang, C., Skormin, V.A., Cao, L. (eds.) AIS-ADM 2007. LNCS (LNAI), vol. 4476, pp. 138–152. Springer, Heidelberg (2007)
8. Kordon, A.K.: Applying Computational Intelligence. In: How to Create Value. Part 6. Swarm Intelligence: The Benefits of Swarms, pp. 145–174. Springer, Berlin (2010)
9. Patcha, A., Park, J.-M.: An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer Networks 51, 3448–3470 (2007)
10. Zhang, W., Yang, Q., Geng, Y.: A survey of anomaly detection methods in networks. In: International Symposium on Computer Network and Multimedia Technology, pp. 1–3 (2009)
11. Wang, X.-F., Zhou, J.-L., Yu, S.-S., Cai, L.-Z.: Data mining methods for anomaly detection of HTTP request exploitations. In: Wang, L., Jin, Y. (eds.) FSKD 2005. LNCS (LNAI), vol. 3614, pp. 320–323. Springer, Heidelberg (2005)
12. Kumpulainen, P., Hätönen, K.: Local anomaly detection for mobile network monitoring. Information Sciences 178, 3840–3859 (2008)
13. Lazarevic, A., Srivastava, N., Tiwari, A., Isom, J., Oza, N., Srivastava, J.: Theoretically optimal distributed anomaly detection. In: IEEE International Conference on Data Mining Workshops, pp. 515–520 (2009)
14. Rabatel, J., Bringay, S., Poncelet, P.: SO_MAD: SensOr Mining for Anomaly Detection in railway data. In: Perner, P. (ed.) Advances in Data Mining. Applications and Theoretical Aspects. Lecture Notes in Computer Science, LNAI, vol. 5633, pp. 191–205. Springer, Heidelberg (2009)
15. Wrembel, R., Koncilia, C. (eds.): Data Warehouses and OLAP: Concepts, Architectures, and Solutions. Idea Group Inc., Hershey (2007)
16. O'Brien, J.A., Marakas, G.M.: Management Information Systems, 9th edn. McGraw-Hill/Irwin, Boston (2009)

Time Series Forecasting of Web Performance Data Monitored by MWING Multiagent Distributed System

Leszek Borzemski and Maciej Drwal

Institute of Informatics
Wrocław University of Technology
Wrocław, Poland
{leszek.borzemski,maciej.drwal}@pwr.wroc.pl

Abstract. This paper presents analysis of Web performance data collected by MWING active Internet measurement system. MWING collects traces of HTTP transactions' performance and allows for estimation of important quality of Web user's experience metrics such as resource download time or goodput (application level throughput). We compare data collected by distributed measurement agents and analyze the responsiveness of common Web servers for different clients. We examine two time series prediction techniques: ARIMA models and pattern matching predictor.

Keywords: Web performance, distributed systems, time series, prediction.

1 Introduction

Measurements are important to understand network behavior. They have been used to track performance issues since the beginning of Internet and World Wide Web development. There are no commonly acceptable measurement standards. There is a stimulus paradigm: passive versus active approaches. Both approaches have their values and should be considered in conjunction with one another. But in passive approach we can gather only network traffic that goes by the measurement sensors whereas active experiments are attractive as the traffic which is sent to the network is under control and the measurements can be done in a distributed way within the Internet. However non active monitoring may cause over utilization of transmission network and computer resources used in the experiment what in turn may have non-acceptable negative influence on the Internet as the whole. Therefore, each active monitoring is a challenge requiring a careful experiment design and realization.

In this paper we deal with active measurements performed to study user perceived latency while surfing the World Wide Web. The latency is the time spent by a user while waiting for a requested Web page. The agents placed at the different geographic network localizations continuously conduct synchronous

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 20–29, 2010.

measurements to the group of Web sites to evaluate how different users perceive performance of the same Web site at the same moment. The measurements are done in an application layer measurement system called MWING to monitor Web sites performance at the HTTP level [1].

We study the Web performance for three Polish academic domains and an academic domain in the USA. Usually, the conclusions based on real-life measurements may be critical in the analysis of Web site behavior and its network connectivity. Our aim is to evaluate the network conditions (we called it the network weather) as it is perceived by the users of that domain the agent represents. All agents are located at university networks which are connected to high bandwidth national networks. Three agents representing users from the following academic domains: Wroclaw University of Technology (TUW) in Wroclaw, Institute of Theoretical and Applied Informatics of the Polish Academy of Sciences (ITAI) in Gliwice and Gdansk University of Technology (GUT), in Gdansk, respectively are localized in the same domain ,´. They are wired with the same Internet Service Providers broadband infrastructure, that is to the National Research and Education Network PIONIER network – a nationwide broadband optical network for e-science in Poland – which utilizes own fiber optics technology and features 10GbE transmission internally and via external connections to the European research network GANT2 and the Internet [9]. These agents allow us to ask a question whether the domains they represent feature similar Web access capabilities. The fourth agent in a campus network of University of Nevada in Las Vegas (UNLV) in USA can be considered as a control agent, giving the opportunity to compare the results including the answer to the question how Web access in Poland is different to the conditions that can be met among academic Web users in USA. Network accessibility at UNLV is provided with high bandwidth Internet2 national network [10]. Inside the campus networks the agents were wired to uncongested 100 Mb/s Ethernet LANs as the most of other users.

The advantage of our measurement approach is that we can continuously observe the actual behavior of the Web access as experienced by each of agent and develop short as well as long time range performance forecasts in the name of the whole community of users of the domain where the agent is installed [2], [3] and use this knowledge to redirect user requests to better auctioning Web site when the server selection problem occurs. Further, we use HTTP protocol, as the real-life browsers, therefore, we can consider the goodput rather than throughput [5]. The goodput is an application measure of transmission quality and better describes users satisfaction than throughput. The solution does not require any changes needed to be made on the Web site side, including Web server and Web pages modifications, therefore, all Web sites can be evaluated.

The domain users may be interested in performance changes over time for their own domains as well as for other domains. This may be employed, for example, to chose the proper time periods to run effectively common network projects. The answer to that question may be given based on the results of time series analysis.

Therefore, in this paper we present the time series analysis performed for measurements collected in our experiments. Especially, we are interested in the similarities in HTTP performance between different agents and forecasting.

The remainder of this paper is organized as follows. Section 2 overviews experimental results obtained in our measurements. Comparison of similarity of network performance is described in Section 3. Time series forecasting methods are discussed in Section 4. Section 5 concludes the paper.

2 Experimental Results

The purpose of the study was to examine the similarities in HTTP transfer performance between different clients. The measurements were performed by downloading a sample resource (about 140 Kb) from multiple web servers. Each client requested the resource approximately at the same time, with fluctuations of delays of up to 30 seconds. Probing each server by each client was repeated in 30 minutes intervals and lasted for many months. For the purpose of analysis we examine about one month of subsequent observations (resulting in about 1000 samples).

Figures 1 and 2 show the typical download duration time series obtained for 3 different clients. The measurements times were synchronized and cases with high delays were removed. The original values were transformed by logarithm function due to extremely high variance. This occurs due to the fact that transfer time probability distributions are heavy tailed. The server from Figure 1 from domain is located in Poland, so are the first two clients (Wroclaw and Gdansk). At average it is accessed faster by these two clients. The server from Figure 2 was located in USA and it responds faster to the requests from Las Vegas.

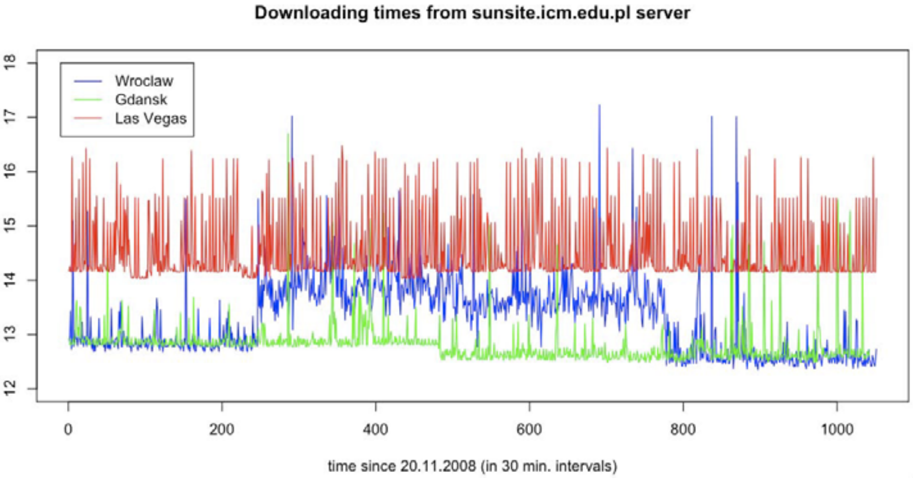


Fig. 1. Transfer times comparison for 3 clients. Logarithms of values are used.

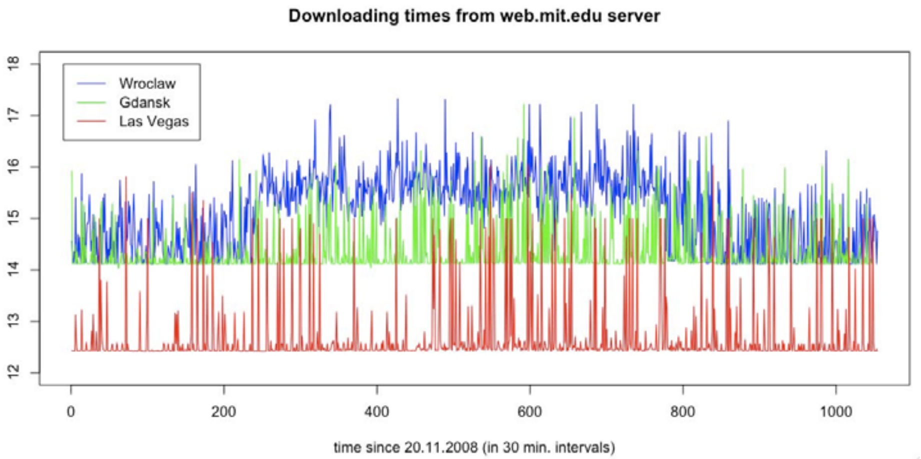


Fig. 2. Transfer times comparison for 3 clients. Logarithms of values are used.

Since the measured transaction times were short, no similarities were observed in overall traffic shaping between these 3 clients. However, there were statistically significant similarities between clients from Gdansk and Gliwice.

Long term characteristics can be observed, especially for Wroclaw client (to small extent also for Gdansk client). There is a significant change in the middle parts of the plots, where the average transfer time for Wroclaw increases, and after few days drops again. Such effects may be caused by long term change in network structure or the way the outgoing traffic is routed from Wroclaw client local network. In result, the considered time series are non-stationary, which brings additional factor of unpredictability.

3 Similarity in Network Performance Reception between Clients

The only observed statistically significant similarity was between clients located in Gdansk and Gliwice. Figure 3 presents the excerpt of measured time series. It can be noticed that many local changes are reproduced in both series. This is due to the fact that both local networks have very similar access to the backbone network and in result their requests are handled in very similar way.

Table 1 presents the detailed similarity analysis. The observed traffic was split into 3 classes: good, medium and bad. From the practical point of view this is enough to describe user's experience. Due to assumed experimental setup (small resource size, probing period of 0.5 hour) more detailed analysis may be too inaccurate. The choice of quality classes is somewhat arbitrary. For example, here we can consider download time as [illegible] if it is below the median value of all measured times. The class [illegible] would contain the cases of transfer time above the 90% of all samples (9/10-quantile of the population). All other cases

Table 1. This table summarizes the similarity in perceiving web servers by clients from Gdansk and Gliwice. Numbers of observations with the same quality class (good, medium, bad) observed in both clients are listed. 3 different class categorizations were used to set up class boundaries: (1) median and 9/10-quantile; (2) median and 14/15-quantile; (3) mean value and mean + variance. Along with the number of cases there is also "profit" column, which shows the percentage of cases which could be predicted from one client for another, after subtracting the cases which could be selected by random guess.

server	(1) cases	(1) profit	(2) cases	(2) profit	(3) cases	(3) profit
www.embed.com.cn	588	31%	609	34%	759	55%
www.networksorcery.com	615	35%	644	39%	589	31%
robsite.net	472	14%	496	18%	645	38%
www.teco.uni-karlsruhe.de	586	30%	629	36%	714	48%
uni-osnabrueck.de	468	14%	491	17%	708	47%
uni-tuebingen.de	445	11%	462	13%	574	28%
hea-www.harvard.edu	501	18%	532	23%	276	0%
www.isi.edu	482	16%	505	19%	697	46%
web.mit.edu	466	14%	490	17%	696	45%
www.teco.edu	501	19%	524	22%	603	33%
ftp.ics.uci.edu	491	17%	513	20%	463	13%
abcdrfc.online.fr	572	28%	608	34%	812	62%
wigwam.sztaki.hu	476	15%	488	17%	792	59%
web.fis.unico.it	491	17%	509	20%	487	17%
omega.di.unipi.it	807	61%	846	67%	716	49%
nagoya-u.ac.jp	460	13%	482	16%	492	17%
www.potaroo.net	609	34%	611	34%	956	82%
www.cs.vu.nl	723	50%	740	52%	618	35%
www.ii.uib.no	452	13%	479	16%	573	30%
www.freesoft.org	490	17%	515	21%	685	44%
ietfreport.isoc.org	524	22%	573	29%	717	49%
paginas.fe.up.pt	717	49%	752	54%	737	52%
csie.ncu.edu.tw	518	21%	538	24%	738	52%
csx.cam.ac.uk	570	28%	592	31%	452	12%
ftp.univie.ac.at	545	25%	566	28%	975	85%
tecfa.unige.ch	478	16%	511	20%	717	49%
www2.cs.uh.edu	471	19%	487	22%	632	43%
www.watersprings.org	491	18%	523	22%	552	26%
simplefailover.com	488	17%	509	20%	635	37%
curl.rtin.bz	466	14%	485	17%	588	31%
plan9.aichi-u.ac.jp	626	37%	630	37%	870	70%
www.ihelpers.co.kr	533	23%	558	27%	694	46%
sunsite.icm.edu.pl	527	23%	553	27%	784	58%

are considered as [illegible]. This setup corresponds to the Figure 3. Table 1 also presents two other boundary setups: median with 14/15-quantile and with two dividing points equal to mean value and mean plus variance.

To evaluate the profit from predicting quality of experience class of one client based on the measured class of other client, we estimated what would be the

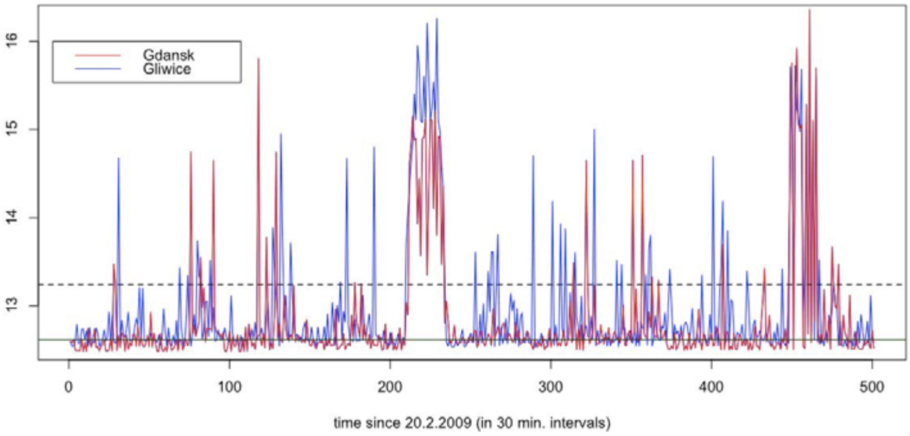


Fig. 3. Transfer times comparison for Gdansk and Gliwice clients. Solid horizontal line denotes median, dashed horizontal line denotes 90-percentile. High correlation can be seen.

number of correctly predicted classes, assuming no correct class was selected by coincidence. With 3 classes we can evaluate [illegible] probability from the total numbers of cases corresponding to each class. Then, selecting the class with highest probability would account to "blind guess". Table 1 contains the percentage of correctly classified cases after subtracting the number of cases coming from such guessing. In result it describes the real obtained knowledge after measuring the performance of one client.

4 Forecasting

Considered type of time series is predictable only to certain extent. While it is impossible to forecast these series with very high accuracy due to predominant pure stochastic effects, a limited degree of effectiveness is achievable.

We consider two forecasting strategies: standard Box-Jenkins framework [4] and novel pattern matching based algorithm. For the study of machine learning approach to forecast these time series see [2], where we used auto-correlation along with cross-correlation (TCP level round trip time measurements) and exogenous factors (time of day, day of week).

Autoregressive Integrated Moving Average (ARIMA) models are used to represent the non-stochastic information contained in time series with a set of parameters.

Let $y_1, y_2, \ldots, y_t$ be the subsequent samples of stochastic process. A non-seasonal $ARIMA(p, d, q)$ model can be written as:

$$(1 - B^d)y_t = c + \phi(B)y_t + \theta(B)\epsilon_t$$

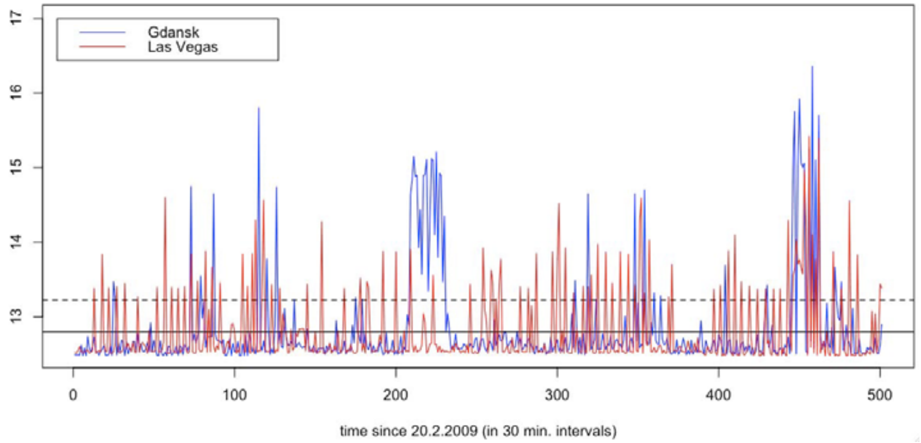


Fig. 4. Transfer times comparison for Gdansk and Las Vegas clients. Solid horizontal line denotes median, dashed horizontal line denotes 90-percentile. Series were transformed in such way that both have equal expected value (originally Gdansk client had expected value 12.8, while Las Vegas client 14.5). They are almost completely uncorrelated.

Here ϕ and θ are polynomials of degree p and q respectively, corresponding to the order of autoregressive part and moving average part. ϵ_t is gaussian random variable with zero mean and variance σ^2, since it is assumed that stochastic component of the process is white noise. B is the backshift operator for which d represents the order of observable trend (non-stationarity). This kind of process can approximate many typical real world time series data.

The selection of parameters p, d and q is performed by fitting several different ARIMA models to the data, and evaluating certain quality criterion: usually Akaike Informantion Criterion or Bayesian Information Criterion.

We used automatic ARIMA model selection algorithm similar to the one proposed in [7] (using AIC). It turns out that linear autoregressive models and exponential smoothing perform rather poorly in terms of long term prediction on these series. Since only very short term prediction with these models gives acceptable accuracy, each prediction was made 1 step ahead. Forecast for 100 points is presented in Figures 5 and 6 as seen on the actual data. These results are fairly good. Although predicted ARIMA series do not resemble the actual data, the approximation is good in terms of mean squared error. The main reason for this is that ARIMA assume gaussian noise to model randomness while this cannot be true for heavy tailed series.

Figures 6 shows the prediction in the presence of rapid concept change. Such changes are quite apparent in all considered datasets. Simple ARIMA models can fail to detect these changes without very careful parameter selection.

Because it is difficult to deal with the type of randomness contained in these data series with standard statistical models, we decided to use a different approach.

Instead of fitting parametric model to the data with assumed noise distribution, we may try to learn and memorize frequent patterns appearing repeatedly in the series directly. We propose pattern matching predictor, which is similar in idea to predictors proposed in [8] and [6].

The predictor selects decision X^*_{n+1} after estimating conditional probability of symbol occurrence after observing k previous symbols:

$$X^*_{n+1} = \arg\max_{a \in A} Pr[X_{n+1} = a | X_{n-k} = x_{n-k}, \ldots, X_n = x_n]$$

The set of symbols is finite thus the values of time series need to be appropriately discretized. Each probability is estimated as the fraction of occurrences of a pattern of length k in the previous observations. This allows us to represent the whole time series only using a small set of frequent patterns, in a similar way to compression algorithms. Here we decided to use approximate pattern matching (instead of exact matching) which corresponds to lossy compression, as we expect significant noise in the data which should be neglected.

Two patterns $A = (a_1, \ldots, a_k)$ and $B = (b_1, \ldots, b_k)$ of length k match approximately if:

$$d(A, B) = \sum_{i=1}^{k} (a_i - b_i)^2 < \epsilon$$

The only two parameters used by this predictor are k and ϵ. The predictor may need special strategies for selection of these parameters, but exhaustive search was used in this work.

We can observe in Figures 5 and 6 that pattern matcher can reconstruct the real process quite accurately. The resulting prediction may have mean squared error similar (or slightly larger) to ARIMA models, but is capable of reconstructing the shape of real process. It also handles very well non-stationarity. Probably the best property of this predictor is the capability of making very accurate long term predictions (even tens of steps ahead).

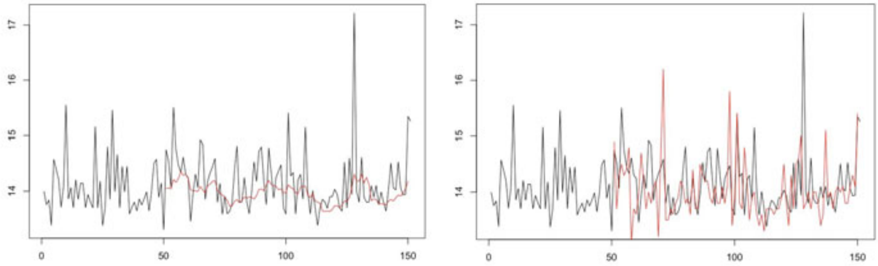


Fig. 5. Left: forecasting of ARIMA(3,0,1) model with lowest AIC for stationary part (100 points in one step ahead mode). Right: the same data predicted using pattern matcher (100 steps ahead after learning 200 previous observations).

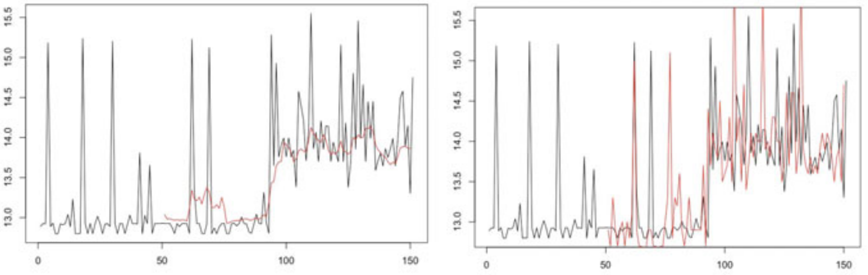


Fig. 6. Left: one step ahead ARIMA model prediction in the presence of rapid concept shift. Right: 100 steps ahead from pattern matcher on the same data.

5 Conclusions

While it is possible to predict the performance of HTTP transaction after collecting the history of transfer time measurements for given client [2], the same servers can be seen completely different by other clients at the same time. For short web transactions, like the ones performed by MWING, the server performance role is negligible. The most important is how the HTTP responses are routed in the Internet. The data we have analyzed contain only one pair (Gdansk–Gliwice) which is served in very similar way (although their geographical distance is fairly big). This is clearly seen in the measured time series. They not only react similarly for temporary deviations from stationary behavior. They also have very similar round trip times on the application level, which is not the case for two other clients – Wroclaw and Las Vegas.

All the analyzed time series share special characteristics. Taking long observations (several days, up to one month) shows their non-stationarity. Considering the slices with stationarity, they have very many short spikes. The data sources of such series are best modeled by heavy tailed distributions, e.g. Pareto distributions. In their histograms we can see that events very far from the mean value occur with significant frequency. The whole series are best modeled by combinations of heavy tailed distributions.

References

1. Borzemski, L., Cichocki, Ł., Kliber, M.: Architecture of Multiagent Internet Measurement System MWING Release 2. In: Håkansson, A., Nguyen, N.T., Hartung, R.L., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2009. Lecture Notes in Computer Science, LNAI, vol. 5559, pp. 410–419. Springer, Heidelberg (2009)
2. Borzemski, L., Drwal, M.: Prediction of Web goodput using nonlinear autoregressive models. In: García-Pedrajas, N., Herrera, F., Fyfe, C., Benítez, J.M., Ali, M. (eds.) Trends in Applied Intelligent Systems. Lecture Notes in Computer Science, LNAI, vol. 6097, pp. 347–356. Springer, Heidelberg (2010)

3. Borzemski, L., Drwal, M.: Statistical analysis of active Web performance measurements. In: Proceedings of 6th Working Conference on Performance Modelling and Evaluation of Heterogenous Networks, pp. 247–258 (2010)
4. Box, G., Jenkins, G.M., Reinsel, G.C.: Time Series Analysis: Forecasting and Control, 3rd edn. Prentice-Hall, Upper Saddle River (1994)
5. Celandroni, N., Potorti, F.: Maximizing single connection TCP goodput by trading bandwidth for BER. International Journal of Communication Systems 16(1), 63–70 (2003)
6. Fua, T., Chunga, F., Luka, R., Ng, C.: Stock time series pattern matching: Template-based vs. rule-based approaches. Engineering Applications of Artificial Intelligence 20, 347–364 (2007)
7. Hyndman, R.J., Khandakar, Y.: Automatic time series forecasting: The forecast package for R. Journal of Statistical Software 26(3) (2008)
8. Jacquet, P., Szpankowski, W., Apostol, I.: A Universal Predictor Based on Pattern Matching. IEEE Transactions on Information Theory 48(6) (2002)
9. PIONIER: The National Research and Education Network (2010), http://www.pionier.net.pl
10. UNLV Office of Information Technology (2010), http://oit.unlv.edu

Comparison and Analysis of the Eye Pointing Methods and Applications

Piotr Chynał and Janusz Sobecki

Institute of Informatics, Wroclaw University of Technology
Wyb.Wyspianskiego 27, 50-370 Wroclaw, Poland
Piotr.Chynal@student.pwr.wroc.pl, Janusz.Sobecki@pwr.wroc.pl

Abstract. In the paper we present comparison and analysis of the selected eye pointing methods and applications. Eye pointing enables communication with the computer system using human eye movements tracking. We analyze not only available eyetracking solutions but also those developed at our laboratory (i.e. ljo_gazetracker) are using specialized gazetracking devices. Those solutions were compared by means of series of experiments carried out with a group of users. These experiments showed that Dasher is the most effective gazewriting application, and that ljo_gazetracker is the most precise eye pointing system. However, it needs application of quite sophisticated and pretty expensive technology. Thus Opengazer seems to be a good freeware alternative.

Keywords: Eyetracking, Human-Computer Interaction.

1 Introduction

In modern applications, interfaces are set to enable fast choices and selections, and the most popular control devices are a computer mouse, a touch-pad and a touch-screen [9]. However the development of the technology enables us to deliver new ways of interaction with users. The choice of element using a computer mouse is limited just to moving over the target and pressing a button. With eyetracking device it is even more simple. The line of gaze goes over a desired target, and it is automatically selected. The research shows that this way is faster and more effective than using a traditional computer mouse [8]. Therefore more and more applications with eyetracking interfaces are being developed. Using eyetracking instead of computer mouse enables people with physical disabilities or busy hands to gain access to the computer [7]. For some people it might be the only way to contact the outside world. This work presents chosen eyetracking solutions in human-computer interaction, and their recommendations, based on needs and financial conditions. To reach this target we needed to analyze existing eyetracking applications, projects and also create our own solutions. Then we have carried out two experiments with users that enabled us to choose the best applications and methods for writing with gaze, controlling the computer and for gaze detection. Additional information about eyetracking, its history and other applications can be found in [1], [3], [4].

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 30–38, 2010.

2 Eyetracking Tools Used in the Experiments

Experiments were conducted in the Software Quality Laboratory, at the Wrocław University of Technology. The main equipment used during this research was ASL 6000 eyetracking module [2]. It consists of two computers (one for user, and one for controlling with EyeTrac 6000 software), Pan Tilt module with two cameras, and ASL module with monitor. One of the Pan Tilt eyetracking cameras was using infrared to detect pupil and cornea reflection, and the second one was tracking face and head movement. Other equipment used in the experiments was casual web cam Logitech Quick Cam Pro 9000. Software used in those experiments:

1. Piotr Zieliński Opengazer[1] .Net port made by Przemysław Nibyłowicz[2]. This application enables gazetracking using ordinary webcam. It is freeware open source software. After selecting feature face points on the video image, user calibrates the program by looking at the appearing squares. Next, when calibration is finished, line of gaze is tracked by the program.
2. MyEye. It is an application developed by Marcelo Laginestra[3]. It enables gazetracking and controlling of the pointing device. To enable eyetracking in this application we need to set different parameters so that we get blue circle around the pupil. After looking in every corner of the screen the calibration is complete. This program is dedicated to the infrared cameras, so we had some problem to get the blue circle around the pupil right with ordinary Logitech Quick Cam Pro 9000. Therefore we decided to use Opengazer .Net port during the experiments.
3. Dasher. This application was created by Interfence Group[4]. Version of Dasher used in experiments was 4.11. In this application user can input text using any controlling device, such as mouse, touchpad, touch screen etc. Writing with Dasher is very easy. User chooses letters appearing on the screen, with cursor. Dasher uses zooming interface, so the letters, on which the cursor is pointing are zooming in. Those letters form words. System uses probalistic methods to give the most probable letters more space [5].
4. GazeTalk5. This application is a complex system that enables to control many different elements, such as browsing internet, listening to the music and many more. It was developed by Eye Gaze Interaction Group at IT University in Copenhagen[5]. Writing in GazeTalk5 is performed by choosing the adequate field, which contains letter or option. System uses prediction mechanisms to suggest the words that he thinks we want to write. Choosing the letter can be done by one of the three ways. First of them is dwell time, which means, that when we look at desired letter, it will be accepted after some time. We can set that time in the options menu. Other options of choosing are standard clicks, and by scanning all the fields, and selecting the highlighted letter when clicked. Besides its rich functionality GazeTalk5 enables running Dasher straight from the main menu, so we can write using that application, and send the text straight to GazeTalk5.

[1] http://www.inference.phy.cam.ac.uk/opengazer/

[2] http://netgazer.sourceforge.net

[3] http://myEye.jimdo.com

[4] http://www.inference.phy.cam.ac.uk/dasher/

[5] http://www.gazegroup.org/research/15

5. FlashKeyboard created by Piotr Chynał. This application enables writing without using a keyboard. It was developed in Adobe Flash CS4 with ActionScript 3.0. Selecting the letters is possible in two ways. User can move cursor over desired letter and wait for two seconds or press space key during that time, to make faster selection. Therefore this application can be used by disabled people and by casual users.
6. Ljo_gazetracker, the application created by Dorota Molik and Wojciech Pietrowski in Software Quality Laboratory, at the Wrocław University of Technology. It was developed and improved by Piotr Chynał. It consists of two parts, first that gathers the data from ASL6000 module and the second which is a Java program that transforms received parameters from the first component and determines the position on screen user is looking at. Furthermore, this program enables control of the cursor with line of gaze.

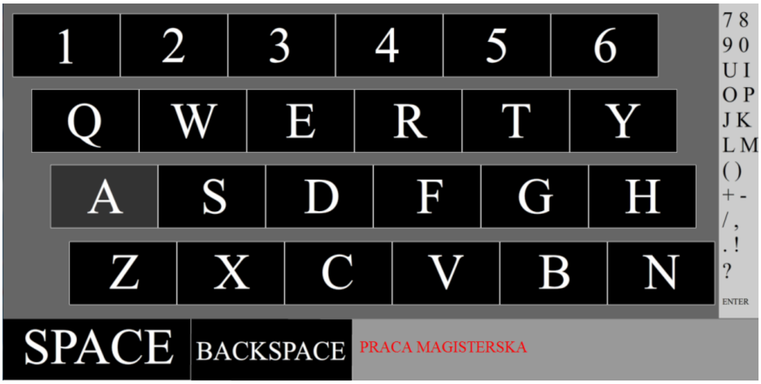


Fig. 1. Interface of the FlashKeyboard

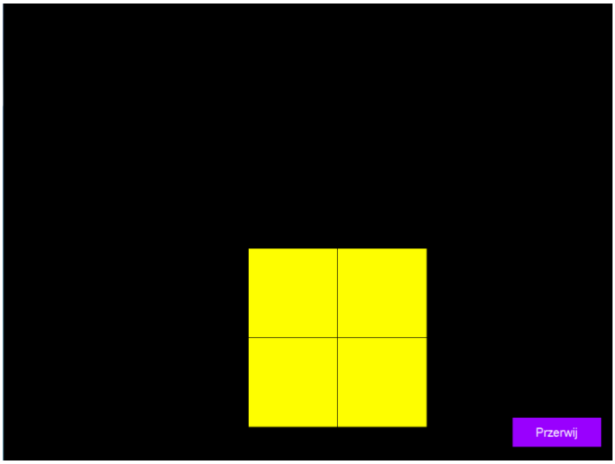


Fig. 2. Precision_test screen shot, showing the first square of the test

7. Precision_test. This program, likewise ljo_gazetracker was created by Dorota Molik and Wojciech Pietrowski. It was implemented in Adobe Flash CS3 with ActionScript 2.0. During working with this program, yellow squares are appearing on the screen. The idea of precision_test is the following, first the user look at yellow square displayed on the screen, so the cursor position is within this square, then after 1,5 seconds, the square is selected, and next the smaller one appears in different place. When square is so small that the user is not able to point gaze within it, the program stops, and prints the size of last selected square in pixels.

3 Experiments

We conducted two experiments. Goal of the first of them was to select the most effective application for writing with gazetracking. In the second experiment the objective was to analyze different eye pointing solutions.

The first experiment was carried out with ten people – students from different universities and faculties. For gazetracking we used Logitech Quick Cam 9000 Pro camera and .Net port of Opengazer. The test consisted of writing three short phrases by the respondents. The phrases were the following: "test", "hello world" and "i am writing with my eye". They were typed using the following tested programs: GazeTalk5, Dasher and FlashKeyboard that were used by each user in a random order. During the experiment the time and the number of mistakes were recorded. After the test, users were asked to describe verbally their reflections on the used programs.

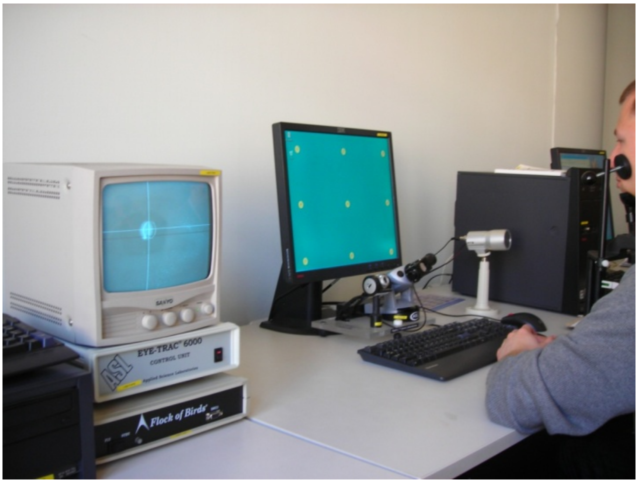


Fig. 3. ASL 6000 eyetracking module in Software Quality Laboratory, at the Wroclaw University of Technology

The second experiment was also conducted on ten users. They all were students of the Wroclaw University of Technology. We have created two stands for this experiment. In the first of them, the respondents worked on Pan Tilt camera with ljo_gazetracker program. Second stand was with Logitech Quick Cam pro 9000 camera and .Net Opengazer port.

In this experiment, users were asked to perform identical tests on both stands. The first test was performing precision test with Flash precision test application. Next test was to write phrase “eye writing test” with Dasher. Last task was to input the web address of one of the Polish Internet portals in GazeTalk5, and then go to mailbox bookmark. During those tests, the results were noted. For precision test it was the result printed by the application. For Dasher and GazeTalk5 we measured time. After the experiment users were asked to express their thoughts on the tested applications and cameras.

4 Results of the Experiments

We present the results of both tests in Tables 1 to 4. They present particular results and comparisons between programs and eyetracking solutions.

Table 1. Results of the first experiments, with sum of time from all three tasks, sum of mistakes made by users in all three tasks and average time of writing

	Dasher		GazeTalk5		FlashKeyboard	
	Time [m:s:ms]	Amount of mistakes	Time [m:s:ms]	Amount of mistakes	Time [m:s:ms]	Amount of mistakes
Person 1	01:29:91	0	05:41:93	1	04:11:49	1
Person 2	03:24:50	0	07:20:76	2	04:16:67	0
Person 3	03:03:71	0	06:42:97	3	05:07:54	1
Person 4	02:37:82	0	07:46:98	1	05:13:58	0
Person 5	02:00:72	0	07:12:04	2	05:58:25	2
Person 6	02:45:33	0	08:18:26	4	05:12:99	0
Person 7	02:25:25	0	09:00:04	3	06:17:64	1
Person 8	02:48:30	0	07:38:68	1	05:55:44	2
Person 9	02:19:78	0	07:24:06	3	04:54:97	0
Person 10	02:45:68	0	06:59:04	2	06:32:73	1
Average total time[m:s:ms]	02:25:31		07:33:66		05:32:71	
Sum of mistakes	0		22		8	

Table 2. Comparison of the applications tested during the first experiment

	Dasher	GazeTalk5	FlashKeyboard
Speed of writing	Very fast, sum of average times for all phrases was – 02:25:31minutes	Slow, sum of average times for all phrases was - 04:52:03 minutes	Fast, average time for all phrases was 03:32:54 minutes
Number of mistakes	No mistakes	Lots of mistakes, caused mainly by selecting wrong letters	Little amount of mistakes, caused by low precision of cursor
Complexity of usage	At first it was hard, but after some time very easy	Easy, but sometimes it is hard to find particular letter, when application suggests other letters, than those that we want to write	Simple and intuitive, created on base of computer keyboard
Ease of use	Very comfortable	Not too comfortable, it wears eyes and the „Midas touch” problem occurs	Comfortable
Interaction technique	Selecting further letters with cursor, combining them to words	Selecting by moving the cursor on desired letter and waiting for two seconds	Selecting letters by pressing space key or by moving the cursor over letter and waiting for two seconds for its selection
Other remarks	Best ranked application by respondents, very good algorithms suggesting most adequate words, working even with worse calibration and precision	Not the best of interfaces. This application has many modules, and the possibility to write using Dasher	All the letters were visible during the work, intuitive interface

Table 3. Results of the second experiment with results of the precision test, time of writing test, time of Internet portal test and average values of all of them

	Camera Pan Tilt + ljo_gazetracker			Camera Logitech + Opengazer		
	Result [px]	Writing time [m:s:ms]	Mailbox time [m:s:ms]	Result [px]	Writing time [m:s:ms]	Mailbox time [m:s:ms]
Person 1	33	01:05:05	02:05:07	75	02:08:15	03:07:25
Person 2	25	01:16:98	02:31:80	120	01:15:37	03:21:47
Person 3	33	00:52:14	03:12:95	150	00:47:41	02:54:34
Person 4	120	01:35:18	02:34:60	150	00:51:07	03:18:03
Person 5	40	00:44:33	01:49:38	66,67	00:49:18	04:01:13
Person 6	75	01:10:07	02:55:41	120	00:56:66	03:12:83
Person 7	37,5	01:15:85	01:29:04	75	02:40:82	02:47:71
Person 8	40	01:05:70	02:11:15	120	00:49:95	02:52:31
Person 9	40	00:50:89	02:50:28	150	01:09:13	03:10:70
Person10	33	01:02:56	01:46:17	200	00:40:21	03:21:24
Average precision [px]	47,65			122,67		
Average writing time [m:s:ms]	00:59:86			01:12:79		
Average time to get to Onet.pl mail box [m:s:ms]	02:18:59			03:12:70		

Table 4. Comparison of the eyetracking solutions tested during the second experiment

	Camera Pan Tilt + ljo_gazetracker	Camera Logitech + Opengazer
Precision	Very precise	Quite precise
Gaze-tracking	Very long and arduous setting of the camera: to get it directly on the eye, and set up pupil and cornea reflection	After selecting the feature points on the face and quick calibration it is ready
Calibration	Fast, user looks at nine points on the screen	Fast, it is possible to set up the custom number and coordinates of the calibration points.
Losing of the tracking	It does not loose image of the eye, small head movement does not make any problems.	Often loses points on the face even during little head movement and light changes.
Other remarks	After it is calibrated correctly once, other people can start working immediately	Calibration needs to be repeated few times for better results. Losing of gaze tracking is very burdensome

5 Summary

After the experiments we gathered all results and observations, and analyzed them. Comparing the time and the amount of mistakes made by users in the first experiment, we can see that the fastest and most reliable application for gaze-writing is Dasher. GazeTalk5 was the slowest and had the most mistakes, mostly due to the "Midas touch" problem [6]. In the opinion of the respondents, Dasher was also the best application. Firstly writing in it was hard to manage, but after a while, it was easy to write in it for everyone.

Using the FlashKeyboard application, we were able to compare selecting letters by using space key and by moving cursor over the target and waiting for two seconds. The result was that using a key is much faster. Still, in case of people with motor disabilities usually they can only use applications which use dwell time for selection. Using the key could be used by people who have the ability to move at least one finger. The keyboard key could be replaced by a button.

Depending on needs and financial predispositions we can present some eyetracking recommendations. In case we want to control the pointing device with large precision and speed, we should get a professional eyetracking camera worth tens of thousands of Euro. There are many professional solutions available, like for example Tobii products. Still such solutions are recommended for companies that use eyetracking for professional purposes, that need its equipment to work fast and reliably. Therefore for handicapped and casual users it is recommended to use freeware solutions and webcam. The precision of such solution is not high, but as the research showed, using such solution we can write in Dasher as fast as with expensive professional camera. We should suspect, that using more advanced camera than Logitech Quick Cam 9000 Pro with, for example infrared LEDs could produce better results. In conclusion of experiments conducted in this work, we can say, that the best solution for handicapped or casual user is usage of combination Opengazer plus GazeTalk5. Opengazer is the best freeware software for eyetracking with ordinary webcam. GazeTalk5 enables complex control of PC with eyetracking. Using it we can browse the Internet, listen to the music etc. Moreover, it allows the user to run Dasher directly from the main menu. Thanks to that we can write with GazeTalk5 using the fastest available gaze-writing applications and be able to do many other things besides writing.

Comparing the different methods of gazetracking, clearly the best one is that used in Pan Tilt camera. It obtains simultaneously the position of pupil and cornea reflection, and traces the line of gaze on the screen very precisely. In case of freeware solutions we have few different approaches. After the experiments we can say that the best of them is the solution used in Opengazer. It captures the key points on users face and extracts them from video image. This solution enables fast calibration and descent precision. There are some different solutions, that use for example cameras with night vision, or like myEye infrared light, but they have no appliance with ordinary web cameras.

References

[1] Mohamed A.O., Perreira Da Silva M., Courbolay V., A history of eye gaze tracking (2007), http://hal.archives-ouvertes.fr/docs/00/21/59/67/PDF/Rapport_interne_1.pdf (March 12, 2010)

[2] Applied Science Group, Eye Tracking System Instructions. ASL Eye-Trac 6000, Pan Tilt/Optics (2006)

[3] Duchowski, A.T.: A Breadth-First Survey of Eye Tracking Applications, http://brm.psychonomic-journals.org/content/34/4/455.full.pdf (March13,2010)

[4] Duchowski, A.T.: Eye tracking methodology: Theory and practice, pp. 205–300. Springer, London (2003)

[5] Information about Dasher project, http://www.inference.phy.cam.ac.uk/dasher/DasherSummary2.html

[6] Jacob, R.: Eye Tracking in Advanced Interface Design, Washington D.C., Human-Computer Interaction Lab Naval Research Laboratory, http://www.cs.tufts.edu/~jacob/papers/barfield.pdf (March12, 2010)

[7] Sesin, A., Adjouadi, M., Cabrerizo, M., Ayala, M., Barreto, A.: Adaptive eyegaze tracking using neural-network-based user profiles to assist people with motor disability. Journal of Rehabilitation Research & Development 45(6), 801–818 (2008)

[8] Zhai, S., Morimoto, C.H., Ihde, S.: Manual and gaze input cascaded pointing, ACM SIGHCI-Human Factors Comput. Syst. Conference, http://www.almaden.ibm.com/u/zhai/papers/magic/magic.pdf (March 19, 2010)

[9] Zhang, X., MacKenzie, S.: Evaluating Eye Tracking with ISO 9241 - Part 9, Department of Computer Science and Engineering York University, Toronto, http://www.yorku.ca/mack/45520779.pdf (March 12, 2010)

Abstraction Relations between Internal and Behavioural Agent Models for Collective Decision Making

Alexei Sharpanskykh and Jan Treur

VU University Amsterdam, Department of Artificial Intelligence
De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
{sharp,treur}@few.vu.nl
http://www.few.vu.nl/~{sharp,treur}

Abstract. For agent-based modelling of collective phenomena individual agent behaviours can be modelled either from an agent-internal perspective, in the form of relations involving internal states of the agent, or from an agent-external, behavioural perspective, in the form of input-output relations for the agent, abstracting from internal states. Illustrated by a case study on collective decision making, this paper addresses how the two types of agent models can be related to each other. First an internal agent model for collective decision making is presented, based on neurological principles. It is shown how by an automated systematic transformation a behavioural model can be obtained, abstracting from the internal states. In addition, an existing behavioural agent model for collective decision making incorporating principles on social diffusion is described. It is shown under which conditions and how by an interpretation mapping the obtained abstracted behavioural agent model can be related to this existing behavioural agent model for collective decision making.

1 Introduction

Agent models used for collective social phenomena traditionally are kept simple, and often are specified by simple reactive rules that determine a direct response (output) based on the agent's current perception (input). However, in recent years it is more and more acknowledged that in some cases agent models specified in the simple format as input-output associations are too limited. Extending specifications of agent models beyond the format of simple reactive input-output associations essentially can be done in two different manners: (a) by allowing more complex temporal relations between the agent's input and output states over time, or (b) by taking into account internal processes described by temporal (causal) relations between internal states. Considering such extended formats for specification of agent models used to model collective social phenomena, raises a number of (interrelated) questions:

(1) When agent models of type (a) are used in social simulation, do they provide the same results as when agent models of type (b) are used?
(2) How can an agent model of type (a) be related to one of type (b)?
(3) Can agent models of type (a) be transformed into agent models of type (b), by some systematic procedure, and conversely?

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 39–53, 2010.

Within the context of modelling collective social phenomena, the internal states in agent models of type (b) do not have a direct impact on the social process; agent models that show the same input-output states over time will lead to exactly the same results at the collective level, no matter what internal states occur. This suggests that for modelling social phenomena the internal states could be hidden or abstracted away by transforming the model in one way or the other into a model of type (a). An interesting challenge here is how this can be done in a precise and systematic manner.

The questions mentioned above are addressed in this paper based on notions such as ontology mappings, temporal properties expressed in hybrid (logical/numerical) formats, and logical and numerical relations between such temporal properties. Here the idea to use ontology mappings and extensions of them is adopted from [15] and refined to relate (more abstract) agent models of type (a) to those of type (b), thus addressing question (2) above. Moreover, addressing question (1), based on such a formally defined relation, it can be established that at the social level the results for the two agent models will be the same. This holds both for simulation traces and for the implied temporal properties (patterns) they have in common. It will be discussed how models of type (b) can be abstracted to models of type (a) by a systematic transformation, implemented in Java, thus also providing an answer to question (3).

The approach is illustrated by a case addressing the emergence of group decisions. It incorporates from the neurological literature the ideas of somatic marking as a basis for individual decision making (cf. [4, 6, 7]), and mirroring of emotions and intentions as a basis for mutual influences between group members (cf. [10, 11, 12]).

The paper is organized as follows. Section 2 presents an internal agent model **IAM** for decision making in a group, based on neurological principles, and modelled in a hybrid logical/numerical format; cf. [3]. In Section 3 an existing behavioural agent model **BAM** for group decision making is briefly described, and specified in hybrid format. In Section 4 first the internal agent model **IAM** introduced in Section 2 is abstracted to a behavioural model **ABAM**, and next in Section 5 it is shown how this behavioural agent model **ABAM** can be related the the behavioural agent model **BAM**, exploiting ontology mappings between the ontologies used for **ABAM** and **BAM**. Section 6 concludes the paper.

2 The Internal Agent Model IAM for Group Decision Making

This case study concerns a neurologically inspired computational modeling approach for the emergence of group decisions, incorporating somatic marking as a basis for individual decision making, see [4], [6], [7] and mirroring of emotions and intentions as a basis for mutual influences between group members, see [10], [11], [12]. The model shows how for many cases, the combination of these two neural mechanisms is sufficient to obtain on the one hand the emergence of common group decisions, and, on the other hand, to achieve that the group members feel OK with these decisions.

Cognitive states of a person, such as sensory or other representations often induce emotions felt within this person, as described by neurologist Damasio [5], [6]. Damasio's *Somatic Marker Hypothesis* (cf. [4], [6], [7]), is a theory on decision making which provides a central role to emotions felt. Within a given context, each represented decision option induces (via an emotional response) a feeling which is used to mark the option. Thus the Somatic Marker Hypothesis provides endorsements or valuations for the different options, and shapes an individual's decision process.

In a social context, the idea of somatic marking can be combined with recent neurological findings on the *mirroring function* of certain neurons (e.g., [10], [11], [12]). Such neurons are active not only when a person prepares for performing a specific action or body change, but also when the person observes somebody else intending or performing this action or body change. This includes expressing emotions in body states, such as facial expressions. The idea is that these neurons and the neural circuits in which they are embedded play an important role in social functioning and in (empathic) understanding of others; (e.g., [10], [11], [12]). They provide a biological basis for many social phenomena; cf. [10]. Indeed, when states of other persons are mirrored by some of the person's own states that at the same time are connected via neural circuits to states that are crucial for the own feelings and actions, then this provides an effective basic (biological) mechanism for how in a social context persons fundamentally affect each other's actions and feelings, and, for example, are able to achieve collective decision making.

Table 1. State ontology used

notation	*description*
SS	sensor state
SRS	sensory representation state
PS	preparation state
ES	effector state
BS	body state
c	observed context information
O	option
c(O)	tendency to choose for option O
b(O)	own bodily response for option O
g(b(O))	other group members' aggregated bodily response for option O
g(c(O))	other group members' aggregated tendency to choose for option O

Given the general principles described above, the mirroring function can be related to decision making in two different ways. In the first place *mirroring of emotions* indicates how emotions felt in different individuals about a certain considered decision option mutually affect each other, and, assuming a context of somatic marking, in this way affect how by individuals decision options are valuated. A second way in which a mirroring function relates to decision making is by applying it to the *mirroring of intentions* or *action tendencies* of individuals for the respective decision options. This may work when by verbal and/or nonverbal behaviour individuals show in how far they tend to choose for a certain option. In the internal agent model **IAM** introduced below both of these (emotion and intention) mirroring effects are incorporated.

An overview of the internal model **IAM** is given in Fig. 1. Here the notations for the state ontology describing the nodes in this network are used as shown in Table 1, and for the parameters as in Table 2. Moreover, the solid arrows denote internal causal relations whereas the dotted arrows indicate interaction with other group members. The arrow from PS(A, b(O)) to SRS(A, b(O)) indicates an as-if body loop that can be used to modulate (e.g., amplify or suppress) a bodily response (cf. [5]).

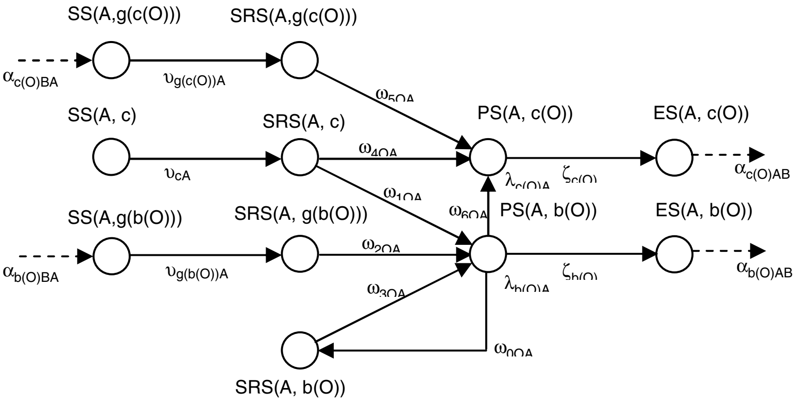


Fig. 1. Overview of the internal agent model **IAM**

Table 2. Parameters for the internal agent model **IAM**

description	*parameter*	*from*	*to*
strengths of connections within agent A	υ_{SA}	SS(A, S)	SRS(A, S)
	ω_{0OA}	PS(A, b(O))	SRS(A, b(O))

	ω_{1OA}
	ω_{2OA}
	ω_{3OA}
		SRS(A, c)
		SRS(A, g(b(O)))
		SRS(A, b(O))
			PS(A, b(O))

	ω_{4OA}
	ω_{5OA}
	ω_{6OA}
		SRS(A, c)
		PS(A, g(b(O)))
		SRS(A, c(O))
			PS(A, c(O))

	ζ_{SA}	PS(A, S)	ES(A, S)
strength for channel for Z from agent B to agent A	α_{ZBA}	sender B	receiver A
change rates for states within agent A	$\lambda_{b(O)A}$	change rate for PS(A, b(O))	
	$\lambda_{c(O)A}$	change rate for PS(A, c(O))	

This internal agent model **IAM** can be described in a detailed manner in hybrid logical/numerical format (cf. [3]) as follows.

IP1 From sensor states to sensory representations

$\text{SS}(A, S, V) \twoheadrightarrow \text{SRS}(A, S, \upsilon_{SA}V)$

where S has instances c, g(c(O)) and g(b(O)) for options O.

IP2 Preparing for an emotion expressed in a body state

$\text{SRS}(A, c, V_1)\ \&\ \text{SRS}(A, g(b(O)), V_2)\ \&\ \text{SRS}(A, b(O), V_3)\ \&\ \text{PS}(A, b(O), V)$
$\twoheadrightarrow \text{PS}(A, b(O), V + \lambda_{b(O)A}\, g(\omega_{1OA}V_1, \omega_{2OA}V_2, \omega_{3OA}V_3, V)\, \Delta t)$

IP3 Preparing for an option choice

$\text{SRS}(A, c, V_1)\ \&\ \text{SRS}(A, g(c(O)), V_2))\ \&\ \text{PS}(A, b(O), V_3)\ \&\ \text{PS}(A, c(O), V)$
$\twoheadrightarrow \text{PS}(A, c(O), V + \lambda_{c(O)A}\, h(\omega_{4OA}V_1, \omega_{5OA}V_2, \omega_{6OA}V_3, V)\, \Delta t)$

IP4 From preparation to effector state

$\text{PS}(A, S, V) \twoheadrightarrow \text{ES}(A, S, \zeta_{SA}\, V)$

where S has instances b(O) and c(O) for options O.

IP5 From preparation to sensory representation of body state

$\mathrm{PS}(S, V) \twoheadrightarrow \mathrm{SRS}(S, \omega_{0OA}V)$

where S has instances $b(O)$ for options O.

Here the functions $g(X_1, X_2, X_3, X_4)$ and $h(X_1, X_2, X_3, X_4)$ are chosen, for example, of the form $th(\sigma, \tau, X_1 + X_2 + X_3) - X_4$, where $th(\sigma, \tau, X) = 1/(1 + e^{-\sigma(X - \tau)})$.

Next the following transfer properties describe the interaction between agents for emotional responses $b(O)$ and choice tendencies $c(O)$ for options O. Thereby the sensed input from multiple agents is aggregated by adding, for example, all influences $\alpha_{b(O)BA}V_B$ on A with V_B the levels of the effector state of agents $B \neq A$, to the sum $\Sigma_{B\neq A}\ \alpha_{b(O)BA}V_B$ and normalising this by dividing it by the maximal value $\Sigma_{B\neq A}\ \alpha_{b(O)BA}\zeta_{b(O)B}$ for it (when all preparation values would be *1*). This provides a kind of average of the impact of all other agents, weighted by the normalised channel strengths.

ITP Sensing aggregated group members' bodily responses and intentions

$\wedge_{B\neq A}\ \mathrm{ES}(B, S, V_B) \twoheadrightarrow \mathrm{SS}(A, g(S), \Sigma_{B\neq A}\ \alpha_{SBA}V_B / \Sigma_{B\neq A}\ \alpha_{SBA}\zeta_{SB})$

where S has instances $b(O)$, $c(O)$ for options O.

Based on the internal agent model **IAM** a number of simulation studies have been performed, using MathLab. Some results for two simulation settings with 10 homogeneous agents with the parameters as defined in Table 3 are presented in Figure 3. The initial values for $SS(A, g(c(O)))$, $SS(A, c)$, $SS(A, g(b(O))$ are set to 0 in both settings. Note that a number of the connections strengths have been chosen rather low; for this reason also the activation levels shown in Fig. 3 are relatively low.

Table 3. The values of the parameters of model IAM used in two simulation settings

description	*parameter*	*setting 1*	*setting 2*
strengths of connections within agent A	$\upsilon_{g(c(O))A}$	0.5	0.9
	υ_{cA}	0.6	0.8
	$\upsilon_{g(b(O))A}$	0.9	0.7
	ω_{0OA}	0.8	0.6
	ω_{1OA} ω_{2OA} ω_{3OA}	0.3 0.3 0.4	0.5 0.2 0.3
	ω_{4OA} ω_{5OA} ω_{6OA}	0.6 0.2 0.2	0.4 0.3 0.3
	$\zeta_{c(O)A}$	0.6	0.8
	$\zeta_{b(O)A}$	0.9	0.4
strength for channel for Z from any agent to any other agent	α_{ZBA}	1	1
change rates for states within agent A	$\lambda_{b(O)A}$	0.7	0.9
	$\lambda_{c(O)A}$	0.4	0.9
parameters of the combination function based on threshold function $th(\sigma, \tau, X) = 1/(1 + e^{-\sigma(X - \tau)})$	σ	4	4
	τ	1.4	1.4

As one can see from Fig. 2, in both simulation settings the dynamics of the multi-agent system stabilizes after some time. Furthermore, in the stable state the agents from setting 1 demonstrate their emotional state more expressively than their intention to choose the option. In setting 2, the opposite situation can be observed in Fig. 2.

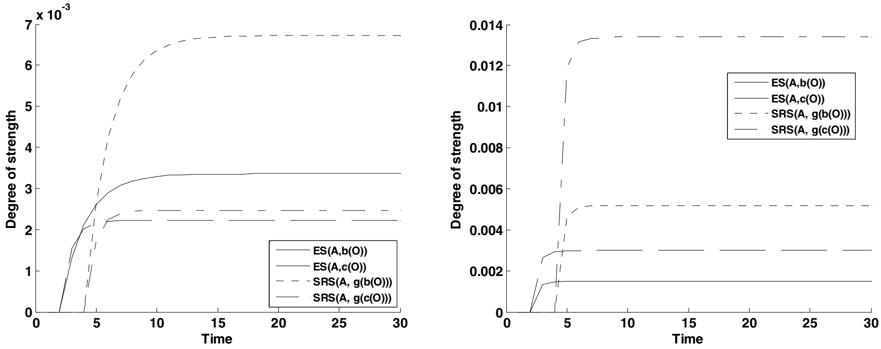


Fig. 2. The dynamics of ES(A, b(O)), ES(A, c(O)), SRS(A, g(b(O))) and SRS(A, g(c(O))) states of an agent A from a multi-agent system with 10 homogeneous agents over time for simulation setting 1 (left) and setting 2 (right) with the parameters from Table 3

3 A Behavioural Agent Model for Group Decision Making: BAM

In [9], an agent-based model for group decision making is introduced. The model was designed in a manner abstracting from the agents' internal neurological, cognitive or affective processes. It was specified in numerical format by mathematical (difference) equations and implemented in MatLab. As a first step, the contagion strength for mental state *S* from person *B* to person *A* is defined by: $\gamma_{SBA} = \varepsilon_{SB} \cdot \alpha_{SBA} \cdot \delta_{SA}$ (1). Here ε_{SB} is the personal characteristic *expressiveness* of the sender (person *B*) for *S*, δ_{SA} the personal characteristic *openness* of the receiver (person *A*) for *S*, and α_{SBA} the interaction characteristic *channel strength* for *S* from sender *B* to receiver *A*. The expressiveness describes the strength of expression of given internal states by verbal and/or nonverbal behaviour (e.g., body states). The openness describes how strong stimuli from outside are propagated internally. The overall contagion strength γ_{SA} from the group towards agent *A* is $\gamma_{SA} = \sum_{B\neq A} \gamma_{SBA} = (\sum_{B\neq A} \varepsilon_{SB} \cdot \alpha_{SBA}) \cdot \delta_{SA}$ (2). This value is for the aggregated input $s_{g(S)A}(t)$ of the other agents upon state *S* of agent *A*:

$$s_{g(S)A}(t) = \sum_{B\neq A} \gamma_{SBA} \cdot q_{SB}(t) \,/\, \gamma_{SA} = \sum_{B\neq A} \varepsilon_{SB} \cdot \alpha_{SBA} \cdot q_{SB}(t) \,/\, (\sum_{B\neq A} \varepsilon_{SB} \cdot \alpha_{SBA}) \qquad (3)$$

How much this external influence actually changes state *S* of the agent *A* may be determined by additional personal characteristics of the agent, for example, the tendency η_{SA} to absorb or to amplify the level of a state and the positive or negative bias β_{SA} for the state *S*. The dynamics of the value $q_{SA}(t)$ of *S* in *A* over time given as:

$$q_{SA}(t + \Delta t) = q_{SA}(t) + \gamma_{SA}\ \boldsymbol{c}(s_{g(S)A}(t), q_{SA}(t))\ \Delta t \qquad (4)$$
$$\text{with } \boldsymbol{c}(X, Y) = \eta_{SA}\cdot[\beta_{SA}\cdot(1 - (1\text{-}X)\cdot(1\text{-}Y)) + (1 - \beta_{SA})\cdot XY] + (1 - \eta_{SA})\cdot X - Y$$

Note that for ***c****(X, Y)* any function can be taken that combines the values of *X* and *Y* and compares the result with Y. For the example function ***c****(X, Y)*, the new value of the state is the old value, plus the change of the value based on the contagion. This change is defined as the multiplication of the contagion strength times a factor for the amplification of information plus a factor for the absorption of information. The absorption part (after $1 - \eta_{SA}$) considers the difference between the incoming contagion and the current level for *S*. The amplification part (afterη_{SA}) depends on the bias of the agent towards more positive (part of equation multiplied by β_{SA}) or negative (part of equation multiplied by $1 - \beta_{SA}$) level for *S*. Table 4 summarizes the most important parameters and state variables within the model (note that the last two parameters will be explained in Section 3.2 below).

Table 4. Parameters and state variables

$q_{SA}(t)$	level for state *S* for agent *A* at time *t*
$e_{SA}(t)$	expressed level for state *S* for agent *A* at time *t*
$S_{g(S)A}(t)$	aggregated input for state *S* for agent *A* at time *t*
ε_{SA}	extent to which agent *A* expresses state *S*
δ_{SA}	extent to which agent *A* is open to state *S*
η_{SA}	tendency of agent *A* to absorb or amplify state *S*
β_{SA}	positive or negative bias of agent *A* on state *S*
α_{SBA}	channel strenght for state *S* from sender *B* to receiver *A*
γ_{SBA}	contagion strength for *S* from sender *B* to receiver *A*
$\omega_{c(O)A}$	weigth for group intention impact on *A* 's intention for *O*
$\omega_{b(O)A}$	weigth for own emotion impact on *A* 's intention for *O*

This generalisation of the existing agent-based contagion models is not exactly a behavioural model, as the states indicated by the values $q_{SA}(t)$ are internal states and not output states. After multiplication by the expression factor ε_{SA} the behavioural output states $e_{SA}(t)$ are obtained that are observed by the other agents. The model can be reformulated in terms of these behavioural output states $e_{SA}(t)$, assuming that time taken by interaction is neglectable compared to the internal processes:

$$s_{g(S)A}(t) = \textstyle\sum_{B\neq A} \alpha_{SBA} \cdot e_{SB}(t) \;/\; (\textstyle\sum_{B\neq A} \varepsilon_{SB} \cdot \alpha_{SBA}\,) \tag{5}$$

$$e_{SA}(t + \Delta t) = e_{SA}(t) + \; \varepsilon_{SA}\, \gamma_{SA}\, \mathbf{c}(s_{g(S)A}(t),\, e_{SA}(t)/\varepsilon_{SA})\, \Delta t \tag{6}$$

To obtain an agent-based social level model for group decision making, the abstract agent-based model for contagion described above for any decision option *O* has been applied to both the emotion states *S* for *O* and intention or choice tendency states *S'* for *O*. In addition, an interplay between the two types of states has been modelled. To incorporate such an interaction (loosely inspired by Damasio's principle of somatic marking; cf. [4], [7], the basic model was extended as follows: to update $q_{SA}(t)$ for an intention state *S* relating to an option *O*, both the intention states of others for *O* and the $q_{S'A}(t)$ values for the emotion state *S'* for *O* are taken into account. Note that in this model a fixed set of options was assumed, that all are considered. The emotion and choice tendency states *S* and *S'* for option *O* are denoted by *b(O)* and *c(O)*, respectively. Then the expressed level of emotion for option *O* of person *A* is $e_{b(O)A}(t)$, and of choice tendency or intention for *O* is $e_{c(O)A}(t)$. The combination of the own (positive) emotion level and the rest of the group's aggregated choice tendency for option *O* is made by a weighted average of the two:

$$s_{g(c(O))A}{}^{*}(t) \;=\; (\omega_{c(O)A}/\omega_{OA})\; s_{g(c(O))A}(t) \;\; + (\omega_{b(O)A}/\omega_{OA})\, e_{b(O)A}(t) \,/\varepsilon_{SA}$$

$$\gamma_{c(O)A}{}^{*} = \omega_{OA}\; \gamma_{c(O)A}$$

where $\omega_{c(O)A}$ and $\omega_{b(O)A}$ are the weights for the contributions of the group choice tendency impact and the own emotion impact on the choice tendency of A for O, respectively, and $\omega_{OA} = \omega_{c(O)A} + \omega_{b(O)A}$. Then the behavioural agent-based model for interacting emotion and intention (choice tendency) contagion expressed in numerical format becomes:

$$s_{g(b(O))A}(t) = \sum_{B\neq A} \alpha_{b(O)BA} \cdot e_{b(O)B}(t) \ / \left(\sum_{B\neq A} \varepsilon_{b(O)B} \cdot \alpha_{b(O)BA} \ \right) \tag{7}$$

$$e_{b(O)A}(t + \Delta t) = e_{b(O)A}(t) + \ \varepsilon_{b(O)A}\ \gamma_{b(O)A}\ \mathbf{c}(s_{g(b((O))A}(t),\ e_{b(O)A}(t)/\varepsilon_{b(O)A})\ \Delta t \tag{8}$$

with as an example

$$\mathbf{c}(X,\ Y) = \eta_{b(O)A}\cdot[\beta_{b(O)A}\cdot(1 - (1\text{-}X)\cdot(1\text{-}Y)) + (1 - \beta_{b(O)A})\cdot XY\,] \ + \ (1 - \eta_{b(O)A})\cdot X - Y$$

$$s_{g(c(O))A}(t) = \sum_{B\neq A} \alpha_{c(O)BA} \cdot e_{c(O)B}(t) \ / \left(\sum_{B\neq A} \varepsilon_{c(O)B} \cdot \alpha_{c(O)BA} \ \right) \tag{9}$$

$$e_{c(O)A}(t + \Delta t) = e_{c(O)A}(t) +$$
$$\varepsilon_{c(O)A}\ \omega_{OA}\ \gamma_{c(O)A}\ \mathbf{d}((\omega_{c(O)A}/\omega_{OA})\ s_{g(c(O))A}(t) + (\omega_{b(O)A}/\omega_{OA})\ e_{b(O)A}(t)/\varepsilon_{b(O)A},\ e_{c(O)A}(t)/\varepsilon_{c(O)A})\Delta t \tag{10}$$

with as an example

$$\mathbf{d}(X,\ Y) = \eta_{c(O)A}\cdot[\beta_{c(O)A}\cdot(1 - (1\text{-}X)\cdot(1\text{-}Y)) + \ (1 - \beta_{c(O)A})\cdot XY\,] \ + (1 - \eta_{c(O)A})\cdot X - Y$$

To be able to relate this model expressed by difference equations to the internal agent model **IAM**, the model is expressed in a hybrid logical/numerical format in a straightforward manner in the following manner, using atoms has_value(x, V) with x a variable name and V a value, thus obtaining the behavioural agent model **BAM**. Here s(g(b((O)), A), s(g(c((O)), A), e(b(O), A) and e(c(O), A) for options O are names of the specific variables involved.

BP1 Generating a body state
has_value(s(g(b(O)), A), V_1) & has_value(e(b(O), A), V)
↠ has_value(e(b(O), A), V + $\varepsilon_{b(O)A}\ \gamma_{b(O)A}$ **c**(V_1, V/$\varepsilon_{b(O)A}$) Δt)

BP2 Generating an option choice intention
has_value(s(g(c(O)), A), V_1) & has_value(e(b(O), A), V_2) & has_value(e(c(O), A), V)
↠ has_value(e(c(O), A), V + $\varepsilon_{c(O)A}\ \omega_{OA}\ \gamma_{c(O)A}$ **d**(($\omega_{c(O)A}/\omega_{OA}$) V_1 + ($\omega_{b(O)A}/\omega_{OA}$) $V_2/\varepsilon_{b(O)A}$, V/$\varepsilon_{c(O)A}$) Δt)

BTP Sensing aggregated group members' bodily responses and intentions
$\wedge_{B\neq A}$ has_value(e(S, B), V_B) ↠ has_value(s(g(S), A), $\Sigma_{B\neq A}\ \alpha_{SBA}V_B / \Sigma_{B\neq A}\ \alpha_{SBA}\varepsilon_{SB}$)

In Section 5 the behavioural agent model **BAM** is related to the internal agent model **IAM** described in Section 2. This relation goes via the abstracted (from **IAM**) behavioural agent model **ABAM** introduced in Section 4.

4 Abstracting Internal Model IAM to Behavioural Model ABAM

First, in this section, from the model **IAM** by a systematic transformation, an abstracted behavioural agent model **ABAM** is obtained. In Section 5 the two behavioural agent models **ABAM** and **BAM** will be related. In [14] an automated abstraction transformation is described from a non-cyclic, stratified internal agent model to a behavioural agent model. As in the current situation the internal agent model is not assumed to be noncyclic, this existing transformation cannot be applied. In particular, for the internal agent model considered as a case in Section 2 the properties IP2 and IP3 are cyclic by themselves (recursive). Moreover, the as-if body loop described by properties IP2 and IP5 is another cycle. Therefore, the transformation introduced here exploits a different approach. The two main steps in this transformation are: elimination of sensory representation atoms, and elimination of preparation atoms (see also Fig. 4).

1. Elimination of sensory representation atoms

It is assumed that sensory representation atoms may be affected by sensor atoms, or by preparation atoms. These two cases are addressed as follows

a) Replacing sensory representation atoms by sensor atoms

- Based on a property $\mathrm{SS}(A, S, V) \twoheadrightarrow \mathrm{SRS}(A, S, \upsilon V)$ (such as IP1), replace atoms $\mathrm{SRS}(A, S, V)$ in an antecedent (for example, in IP2 and IP3) by $\mathrm{SS}(A, S, V/\upsilon)$.

b) Replacing sensory representation atoms by preparation atoms

- Based on a property $\mathrm{PS}(A, S, V) \twoheadrightarrow \mathrm{SRS}(A, S, \omega V)$ (such as IP5), replace atoms $\mathrm{SRS}(A, S, V)$ in an antecedent (for example, in IP2) by $\mathrm{PS}(A, b(O), V/\omega)$.

Note that this transformation step is similar to the principle exploited in [14]. It may introduce new occurrences of preparation atoms; therefore it should preceed the step to eliminate preparation atoms. In the case study this transformation step provides the following transformed properties (replacing IP1, IP2, IP3, and IP5; see also Fig. 4):

IP2* Preparing for a body state

$\mathrm{SS}(A, c, V_1/\upsilon_{cA})\ \&\ \mathrm{SS}(A, g(b(O)), V_2/\upsilon_{g(b(O))A})\ \&\ \mathrm{PS}(A, b(O), V_3/\omega_{0OA})\ \&\ \mathrm{PS}(A, b(O), V)$
$\twoheadrightarrow\ \mathrm{PS}(A, b(O), V + \lambda_{b(O)A}\, \mathbf{g}(\omega_{1OA}V_1, \omega_{2OA}V_2, \omega_{3OA}V_3, V)\, \Delta t)$

IP3* Preparing for an option choice

$\mathrm{SS}(A, c, V_1/\upsilon_{cA})\ \&\ \mathrm{SS}(A, g(c(O)), V_2/\upsilon_{g(c(O))A}))\ \&\ \mathrm{PS}(A, b(O), V_3)\ \&\ \mathrm{PS}(A, c(O), V)$
$\twoheadrightarrow\ \mathrm{PS}(A, c(O), V + \lambda_{c(O)A}\, \mathbf{h}(\omega_{4OA}V_1, \omega_{5OA}V_2, \omega_{6OA}V_3, V)\, \Delta t)$

2. Elimination of preparation atoms

Preparation atoms in principle occur both in antecedents and consequents. This makes it impossible to apply the principle exploited in [14]. However, it is exploited that preparation states often have a direct relationship to effector states:

- Based on a property $\mathrm{PS}(A, S, V) \twoheadrightarrow \mathrm{ES}(A, S, \zeta V)$ (such as in IP4), replace each atom $\mathrm{PS}(A, S, V)$ in an antecedent or consequent by $\mathrm{ES}(A, S, \zeta V)$.

In the case study this transformation step provides the following transformed properties (replacing IP2*, IP3*, and IP4; see also Fig. 4):

IP2* Preparing for a body state

$\mathrm{SS}(A, c, V_1/\upsilon_{cA})\ \&\ \mathrm{SS}(A, g(b(O)), V_2/\upsilon_{g(b(O))A})\ \&\ \mathrm{ES}(A, b(O), \zeta_{b(O)A} V_3/\omega_{0OA})\ \&\ \mathrm{ES}(A, b(O), \zeta_{b(O)A} V)$
$\twoheadrightarrow\ \mathrm{ES}(A, b(O), \zeta_{b(O)A} V + \zeta_{b(O)A} \lambda_{b(O)A}\, \mathbf{g}(\omega_{1OA}V_1, \omega_{2OA}V_2, \omega_{3OA}V_3, V)\, \Delta t)$

IP3* Preparing for an option choice

$\mathrm{SS}(A, c, V_1/\upsilon_{cA})\ \&\ \mathrm{SS}(A, g(c(O)), V_2/\upsilon_{g(c(O))A}))\ \&\ \mathrm{ES}(A, b(O), \zeta_{b(O)A} V_3)\ \&\ \mathrm{ES}(A, c(O), \zeta_{c(O)A} V)$
$\twoheadrightarrow\ \mathrm{ES}(A, c(O), \zeta_{c(O)A} V + \zeta_{c(O)A} \lambda_{c(O)A}\, \mathbf{h}(\omega_{4OA}V_1, \omega_{5OA}V_2, \omega_{6OA}V_3, V)\, \Delta t)$

By renaming V_1/υ_{cA} to V_1, $V_2/\upsilon_{g(b(O)A}$ to V_2 , $\zeta_{b(O)A} V_3/\omega_{0OA}$ to V_3, $\zeta_{b(O)A} V$ to V (in IP2*), resp. $V_2/\upsilon_{g(c(O))A}$ to V_2¸$\zeta_{b(O)A} V_3$ to V_3, and $\zeta_{c(O)A} V$ to V (in IP3*), the following is obtained:

IP2 Preparing for a body state**

$\mathrm{SS}(A, c, V_1)\ \&\ \mathrm{SS}(A, g(b(O)), V_2)\ \&\ \mathrm{ES}(A, b(O), V_3)\ \&\ \mathrm{ES}(A, b(O), V)$
$\twoheadrightarrow\ \mathrm{ES}(A, b(O), V + \zeta_{b(O)A} \lambda_{b(O)A}\, \mathbf{g}(\omega_{1OA}\upsilon_{cA} V_1, \omega_{2OA}\upsilon_{g(b(O))A} V_2, \omega_{3OA}\omega_{0OA} V_3/ \zeta_{b(O)A}, V/\zeta_{b(O)A})\, \Delta t)$

IP3 Preparing for an option choice**

$\mathrm{SS}(A, c, V_1)\ \&\ \mathrm{SS}(A, g(c(O)), V_2)\ \&\ \mathrm{ES}(A, b(O), V_3)\ \&\ \mathrm{ES}(A, c(O), V)$
$\twoheadrightarrow\ \mathrm{ES}(A, c(O), V + \zeta_{c(O)A} \lambda_{c(O)A}\, \mathbf{h}(\omega_{4OA}\upsilon_{cA} V_1, \omega_{5OA}\upsilon_{g(c(O))A} V_2, \omega_{6OA}V_3/\zeta_{b(O)A}, V/\zeta_{c(O)A})\, \Delta t)$

Based on these properties derived from the internal model **IAM** the specification of the abstracted behavioural model **ABAM** can be defined; see also Fig. 3, lower part.

Hybrid Specification of the Abstracted Behavioural Agent Model ABAM

Note that in IP2** V_2 and V have the same value, so a slight further simplification can be made by replacing V_3 by V. After renaming of the variables according to

ABP1			ABP2		
V_1	→	W_0	V_1	→	W_0
V_2	→	W_1	V_2	→	W_1
V_3	→	W	V_3	→	W_2
V	→	W	V	→	W

the following abstracted behavioural model **ABAM** for agent A is obtained:

ABP1 Generating a body state

$\mathrm{SS}(A, c, W_0)\ \&\ \mathrm{SS}(A, g(b(O)), W_1)\ \&\ \mathrm{ES}(A, b(O), W)$

$\twoheadrightarrow\ \mathrm{ES}(A, b(O), W + \zeta_{b(O)A}\, \lambda_{b(O)A}\, \mathbf{g}(\omega_{1OA}\upsilon_{cA}\, W_0,\ \omega_{2OA}\upsilon_{g(b(O))A}\, W_1,\ \omega_{3OA}\omega_{0OA}\, W / \zeta_{b(O)A},\ W/\zeta_{b(O)A})\ \Delta t)$

ABP2 Generating an option choice intention

$\mathrm{SS}(A, c, W_0)\ \&\ \mathrm{SS}(A, g(c(O)), W_1)\ \&\ \mathrm{ES}(A, b(O), W_2)\ \&\ \mathrm{ES}(A, c(O), W)$

$\twoheadrightarrow\ \mathrm{ES}(A, c(O), W + \zeta_{c(O)A}\, \lambda_{c(O)A}\, \mathbf{h}(\omega_{4OA}\upsilon_{cA}\, W_0,\ \omega_{5OA}\upsilon_{g(c(O))A}\, W_1,\ \omega_{6OA}\, W_2/\zeta_{b(O)A},\ W/\zeta_{c(O)A})\ \Delta t)$

ITP Sensing aggregated group members' bodily responses and intentions

$\wedge_{B\neq A}\ \mathrm{ES}(B, S, V_B)\ \twoheadrightarrow\ \mathrm{SS}(A, g(S), \Sigma_{B\neq A}\ \alpha_{SBA}V_B / \Sigma_{B\neq A}\ \alpha_{SBA}\zeta_{SB})$

where S has instances $b(O)$, $c(O)$ for options O.

Note that as all steps made are logical derivations, it holds **IAM** ⊢ **ABAM**. In particular the following logical implications are valid (shown hierarchically in Fig. 3):

IP1 & IP5 & IP2 ⇒ IP2*   IP4 & IP2* ⇒ ABP1
IP1 & IP3 ⇒ IP3*   IP4 & IP3* ⇒ ABP2

The transformation as described is based on the following of assumptions:

- Sensory representation states are affected (only) by sensor states and/or preparation states
- Preparation atoms have a direct relationship with effector atoms; there are no other ways to generate effector states than via preparation states
- The time delays for the interaction from the effector state of one agent to the sensor state of the same or another agent are small so that they can be neglected compared to the internal time delays
- The internal time delays from sensor state to sensory representation state and from preparation state to effector state within an agent are small so that they can be neglected compared to the internal time delays from sensory representation to preparation states

The transformation can be applied to any internal agent model satisfying these assumptions. The proposed abstraction procedure has been implemented in Java. The automated procedure requires as input a text file with a specification of an internal agent model and generates a text file with the corresponding abstracted behavioural model as output. The computational complexity of the procedure is $O(|M|*|N| + |L|*|S|)$, where M is the set of srs atoms in the **IAM** specification, N is the set of the srs state generation properties in the specification, L is the set of the preparation atoms and srs atoms in the loops in the specification, and S is the set of the effector state generation properties in the specification.

Using the automated procedure the hybrid specification of **ABAM** has been obtained. With this specification simulation has been performed with the values of parameters as described in Table 3. The obtained curves for ES(A, c(O)) and ES(A, b(O)) are the same as the curves depicted in Fig. 3 for the model **IAM**. This outcome confirms that both the models **ABAM** and **IAM** generate the same behavioural traces and that the abstraction transformation is correct.

5 Relating the Behavioural Agent Models BAM and ABAM

In this section the given behavioural agent model **BAM** described in Section 3 is related to the behavioural agent model **ABAM** obtained from the internal agent model **IAM** by the abstraction process described in Section 4. First the notion of interpretation mapping induced by an ontology mapping is briefly introduced (e.g., [8], pp. 201-263; [15]). By a basic ontology mapping π atomic state properties (e.g., a_2 and b_2) in one ontology can be related to state properties (e.g., a_1 and b_1) in another (e.g., $\pi(a_2) = a_1$ and $\pi(b_2) = b_1$). Using compositionality a basic ontology mapping used above can be extended to an interpretation mapping for temporal expressions. As an example, when $\pi(a_2) = a_1$, $\pi(b_2) = b_1$, then this induces a mapping π^* from dynamic property $a_2 \twoheadrightarrow b_2$ to $a_1 \twoheadrightarrow b_1$ as follows: $\pi^*(a_2 \twoheadrightarrow b_2) = \pi^*(a_2) \twoheadrightarrow \pi^*(b_2) = \pi(a_2) \twoheadrightarrow \pi(b_2) = a_1 \twoheadrightarrow b_1$. In a similar manner by compositionality a mapping for more complex temporal predicate logical relationships A and B can be defined, using

$$\pi^*(A \mathbin{\&} B) = \pi^*(A) \mathbin{\&} \pi^*(B) \qquad \pi^*(A \vee B) = \pi^*(A) \vee \pi^*(B)$$

$$\pi^*(A \Rightarrow B) = \pi^*(A) \Rightarrow \pi^*(B) \qquad \pi^*(\neg A) = \neg\, \pi^*(A)$$

$$\pi^*(\forall T\ A) = \forall T\ \pi^*(A) \qquad \pi^*(\exists T\ A) = \exists T\ \pi^*(A)$$

To obtain a mapping the given behavioural model **BAM** onto the abstracted **ABAM**, first, consider the basic ontology mapping π defined by :

$\pi(\text{has\_value}(e(S, A), V)) = ES(A, S, V)$ where instances for S are $b(O)$, $c(O)$ for options O

$\pi(\text{has\_value}(s(S, A), V)) = SS(A, S, V)$ where instances for S are $g(b((O))$, $g(c((O))$ for options O

Next by compositionality the interpretation mapping π^* is defined for the specification of the behavioural model **BAM** as follows:

Mapping BP1 Generating a body state

$$\pi^*(\text{BP1}) = \pi^*(\text{has\_value}(s(g(b(O)), A), V_1) \mathbin{\&} \text{has\_value}(e(b(O), A), V)$$
$$\twoheadrightarrow \text{has\_value}(e(b(O), A), V + \varepsilon_{b(O)A}\, \gamma_{b(O)A}\, \mathbf{c}(V_1, V/\varepsilon_{b(O)A})\ \Delta t))$$
$$= \pi(\text{has\_value}(s(g(b(O)), A), V_1)) \mathbin{\&} \pi(\text{has\_value}(e(b(O), A), V))$$
$$\twoheadrightarrow \pi(\text{has\_value}(e(b(O), A), V + \varepsilon_{b(O)A}\, \gamma_{b(O)A}\, \mathbf{c}(V_1, V/\varepsilon_{b(O)A})\ \Delta t))$$
$$= SS(A, g(b(O)), V_1) \mathbin{\&} ES(A, b(O), V) \twoheadrightarrow ES(A, b(O), V + \varepsilon_{b(O)A}\, \gamma_{b(O)A}\, \mathbf{c}(V_1, V/\varepsilon_{b(O)A})\ \Delta t)$$

Mapping BP2 Generating an option choice intention

$$\pi^*(\text{BP2}) = \pi^*(\text{has\_value}(s(g(c(O)), A), V_1) \mathbin{\&} \text{has\_value}(e(b(O), A), V_2) \mathbin{\&} \text{has\_value}(e(c(O), A), V)$$
$$\twoheadrightarrow \text{has\_value}(e(c(O), A), V + \varepsilon_{c(O)A}\, \omega_{OA}\, \gamma_{c(O)A}\, \mathbf{d}((\omega_{c(O)A}/\omega_{OA})\, V_1 + (\omega_{b(O)A}/\omega_{OA})\, V_2/\varepsilon_{b(O)A}, V/\varepsilon_{c(O)A})\ \Delta t))$$
$$= \pi(\text{has\_value}(s(g(c(O)), A), V_1)) \mathbin{\&} \pi(\text{has\_value}(e(b(O), A), V_2)) \mathbin{\&} \pi(\text{has\_value}(e(c(O), A), V))$$
$$\twoheadrightarrow \pi(\text{has\_value}(e(c(O), A), V + \varepsilon_{c(O)A}\, \omega_{OA}\, \gamma_{c(O)A}\, \mathbf{d}((\omega_{c(O)A}/\omega_{OA})\, V_1 + (\omega_{b(O)A}/\omega_{OA})\, V_2/\varepsilon_{b(O)A}, V/\varepsilon_{c(O)A})\ \Delta t))$$
$$= SS(A, g(c(O)), V_1) \mathbin{\&} ES(A, b(O), V_2) \mathbin{\&} ES(A, c(O), V)$$
$$\twoheadrightarrow ES(A, c(O), V + \varepsilon_{c(O)A}\, \omega_{OA}\, \gamma_{c(O)A}\, \mathbf{d}((\omega_{c(O)A}/\omega_{OA})\, V_1 + (\omega_{b(O)A}/\omega_{OA})\, V_2/\varepsilon_{b(O)A}, V/\varepsilon_{c(O)A})\ \Delta t)$$

Mapping BTP Sensing aggregated group members' bodily responses and intentions

$$\pi^*(\text{BTP}) = \pi^*(\wedge_{B\neq A}\ \text{has\_value}(e(S, B), V_B) \twoheadrightarrow \text{has\_value}(s(g(S), A), \Sigma_{B\neq A}\ \alpha_{SBA}V_B / \Sigma_{B\neq A}\ \alpha_{SBA}\varepsilon_{SB}))$$
$$= \wedge_{B\neq A}\ \pi(\text{has\_value}(e(S, B), V_B)) \twoheadrightarrow \pi(\text{has\_value}(s(g(S), A), \Sigma_{B\neq A}\ \alpha_{SBA}V_B / \Sigma_{B\neq A}\ \alpha_{SBA}\varepsilon_{SB}))$$
$$= \wedge_{B\neq A}\ ES(B, S, V_B)) \twoheadrightarrow SS(A, g(S), \Sigma_{B\neq A}\ \alpha_{SBA}V_B / \Sigma_{B\neq A}\ \alpha_{SBA}\varepsilon_{SB})$$

So to explore under which conditions the mapped behavioural model **BAM** is the abstracted model **ABAM**, it can be found out when the following identities (after unifying the variables V_i, V and W_i, W for values) hold.

$$\pi^*(\text{BP1}) = \text{ABP1} \qquad \pi^*(\text{BP2}) = \text{ABP2} \qquad \pi^*(\text{BTP}) = \text{ITP}$$

However, the modelling scope of **ABAM** is wider than the one of **BAM**. In particular, in **ABAM** an as-if body loop is incorporated that has been left out of consideration for **BAM**. Moreover, in the behavioural model **BAM** the options O are taken from a fixed set, given at forhand and automatically considered, whereas in **ABAM** they are generated on the basis of the context c. Therefore, the modelling scope of **ABAM** is first tuned to the one of **BAM**, to get a comparable modelling scope for both models **IAM** and **ABAM**. The latter condition is achieved by taking the activation level W_0 of the sensor state for the context c and the strengths of the connections between the sensor state for context c and preparations relating to option O can be set at *1* (so $\upsilon_{cA} = \omega_{1OA} = \omega_{4OA} = 1$); thus the first argument of g and h becomes 1. The former condition is achieved by leaving out of **ABAM** the dependency on the sensed body state, i.e., by making the third argument of g zero (so $\omega_{0OA} = 0$).

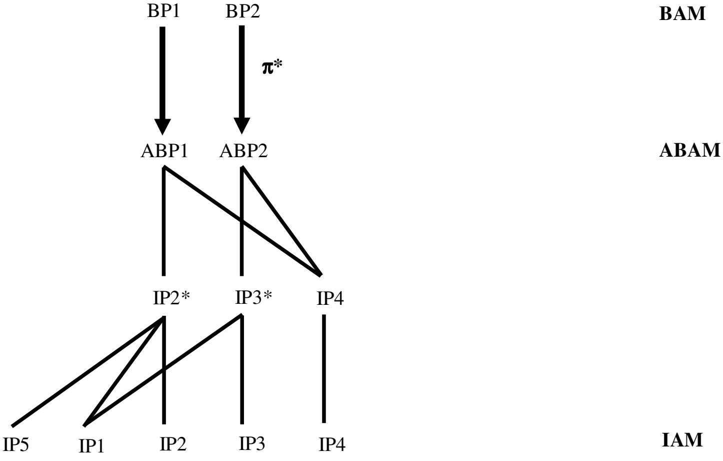


Fig. 3. Logical relations from network specification via internal agent model and abstracted behavioural model to behavioural agent model: **IAM ⊢ ABAM = π(BAM)**

Given these extra assumptions and the mapped specifications found above, when the antecedents where unified according to $V_i \leftrightarrow W_i$, $V \leftrightarrow W$ the identities are equivalent to the following identities in V, V_i

$$\varepsilon_{b(O)A}\,\gamma_{b(O)A}\,\mathbf{c}(V_1, V/\varepsilon_{b(O)A}) = \zeta_{b(O)A}\,\lambda_{b(O)A}\,\mathbf{g}(1, \omega_{2OA}\upsilon_{g(b(O))A}\,V_1, 0, V/\zeta_{b(O)A})$$

$$\varepsilon_{c(O)A}\,\omega_{OA}\,\gamma_{c(O)A}\,\mathbf{d}((\omega_{c(O)A}/\omega_{OA})\,V_1 + (\omega_{b(O)A}/\omega_{OA})\,V_2/\varepsilon_{b(O)A}, V/\varepsilon_{c(O)A}) =$$
$$\zeta_{c(O)A}\,\lambda_{c(O)A}\,\mathbf{h}(1, \omega_{5OA}\upsilon_{g(c(O))A}\,V_1, \omega_{6OA}\upsilon_{b(O)A}\,V_2/\zeta_{b(O)A}, V/\zeta_{c(O)A})$$

$$\Sigma_{B\neq A}\ \alpha_{SBA}V_B / \Sigma_{B\neq A}\ \alpha_{SBA}\varepsilon_{SB} = \Sigma_{B\neq A}\ \alpha_{SBA}V_B / \Sigma_{B\neq A}\ \alpha_{SBA}\zeta_{SB}$$

The last identity is equivalent to $\varepsilon_{SB} = \zeta_{SB}$ for all S and B with $\alpha_{SBA} > 0$ for some A. Moreover, it can be assumed that $\varepsilon_{SB} = \zeta_{SB}$ for all S and B. There may be multiple ways in which this can be satisfied for all values of V_1, U_2, U. At least one possibility is the following. Assume for all agents A

$$\begin{array}{ll} \lambda_{b(O)A} = \gamma_{b(O)A} & \upsilon_{b(O)A} = 1 \\ \lambda_{c(O)A} = \omega_{OA}\,\gamma_{c(O)A} & \upsilon_{g(S)A} = 1 \end{array}$$

for S is $b(O)$ or $c(O)$. Then the identities simplify to

$$\mathbf{c}(V_1, U) = \mathbf{g}(1, \omega_{2OA}\, V_1, 0, V)$$

$$\mathbf{d}((\omega_{c(O)A}/\omega_{OA})\, V_1 + (\omega_{b(O)A}/\omega_{OA})\, V_2, V) = \mathbf{h}(1, \omega_{5OA}V_1, \omega_{6OA}V_2, V)$$

Furthermore, taking $\omega_{2OA} = 1$, $\omega_{5OA} = \omega_{c(O)A}/\omega_{OA}$, $\omega_{6OA} = \omega_{b(O)A}/\omega_{OA}$, the following identities result (replacing $\omega_{5OA}V_1$ by V_1 and $\omega_{6OA}V_2$ by V_2)

$$\mathbf{c}(V_1, V) = \mathbf{g}(1, V_1, 0, V) \qquad \mathbf{d}(V_1 + V_2, V) = \mathbf{h}(1, V_1, V_2, V)$$

There are many possibilities to fulfill these identities. For any given functions $\mathbf{c}(X, Y)$, $\mathbf{d}(X, Y)$ in the model **BAM** the functions $\mathbf{g}$, $\mathbf{h}$ in the model **IAM** defined by

$$\mathbf{g}(W, X, Y, Z) = \mathbf{c}(W - 1 + X + Y, Z) \qquad \mathbf{h}(W, X, Y, Z) = \mathbf{d}(W - 1 + X + Y, Z)$$

fulfill the identities $\mathbf{g}(1, X, 0, Z) = \mathbf{c}(X, Z)$ and $\mathbf{h}(1, X, Y, Z) = \mathbf{d}(X+Y, Z)$. It turns out that for given functions $\mathbf{c}(X, Y)$, $\mathbf{d}(X, Y)$ in the model **BAM** functions $\mathbf{g}$, $\mathbf{h}$ in the model **IAM** exist so that the interpretation mapping $\boldsymbol{\pi}$ maps the behavioural model **BAM** onto the model **ABAM**, which is a behavioural abstraction of the internal agent model **IAM** (see also Fig. 3): $\boldsymbol{\pi}^*$(BP1) = ABP1, $\boldsymbol{\pi}^*$(BP2) = ABP2, $\boldsymbol{\pi}^*$(BTP) = ITP. As an example direction, when for $\mathbf{c}(X, Y)$ a threshold function th is used, for example, defined as $\mathbf{c}(X, Y) = \text{th}(\sigma, \tau, X+Y) - Y$ with $\text{th}(\sigma, \tau, V) = 1/(1 + e^{-\sigma(V - \tau)})$, then for $\tau' = \tau + 1$ the function $\mathbf{g}(W, X, Y, Z) = \text{th}(\sigma, \tau', W+X+Y+Z) - Z$ fulfils $\mathbf{g}(1, X, 0, Z) = \mathbf{c}(X, Z)$. Another example of a function $\mathbf{g}(V, W, X, Y)$ that fulfills the identity when $\mathbf{c}(X, Z) = 1 - (1 - X)(1 - Z) - Z$ is $\mathbf{g}(W, X, Y, Z) = W\,[1 - (1 - W)(1 - X)(1 - Z)] - Z$. As the properties specifying **ABAM** were derived from the properties specifying **IAM** (e.g., see Figs. 2 and 3), it holds **IAM** $\vdash$ **ABAM**, and as a compositional interpretation mapping $\boldsymbol{\pi}$ preserves derivation relations, the following relationships holds for any temporal pattern expressed as a hybrid logical/numerical property A in the ontology of **BAM**:

$$\mathbf{BAM} \vdash A \;\Rightarrow\; \boldsymbol{\pi}(\mathbf{BAM}) \vdash \boldsymbol{\pi}(A) \;\Rightarrow\; \mathbf{ABAM} \vdash \boldsymbol{\pi}(A) \;\Rightarrow\; \mathbf{IAM} \vdash \boldsymbol{\pi}(A)$$

Such a property A may specify certain (common) patterns in behaviour; the above relationships show that the internal agent model **IAM** shares the common behavioural patterns of the behavioural model **BAM**. An example of such a property A expresses a pattern that under certain conditions after some point in time there is one option O for which both $b(O)$ and $c(O)$ have the highest value for each of the agents (joint decision).

6 Discussion

This paper addressed how internal agent models and behavioural agent models for collective desion making can be related to each other. The relationships presented were expressed for specifications of the agent models in a hybrid logical/numerical format. Two agent models for collective decision making were first presented. First an internal agent model **IAM** derived from neurological principles modelled in a network specification **NS** was introduced with **NS** $\vdash$ **IAM**, where $\vdash$ is a symbol for derivability.

Next, an existing behavioural agent model **BAM**, incorporating principles on social contagion or diffusion, was described, adopted from (Hoogendoorn, Treur, Wal, and Wissen, 2010). Furthermore, it was shown how the internal agent model **IAM** can be systematically transformed into an abstracted behavioural model **ABAM**, where the internal states were abstracted away, and such that **IAM** $\vdash$ **ABAM**. This generic transformation has been implemented in Java. Moreover, it was shown that under certain conditions the obtained agent model **ABAM** can be related to the behavioural agent model **BAM** by an interpretation mapping π, i.e., such that π(**BAM**) = **ABAM**. In this way hybrid logical/numerical relations where obtained between the different agent models according to:

$$\mathbf{IAM} \vdash \mathbf{ABAM} \text{ and } \mathbf{ABAM} = \pi(\mathbf{BAM})$$

These relationships imply that, for example, collective behaviour patterns shown in multi-agent systems based on the behavioural agent model **BAM** are shared (in the form of patterns corresponding via π) for multi-agent systems based on the models **ABAM** and **IAM**.

References

1. Bosse, T., Duell, R., Memon, Z.A., Treur, J., van der Wal, C.N.: A Multi-Agent Model for Mutual Absorption of Emotions. In: Otamendi, J., et al. (eds.) Proc. of the 23th Eur. Conf. on Modelling and Simulation, ECMS 2009. European Council on Modeling and Simulation, pp. 212–218 (2009)
2. Bosse, T., Duell, R., Memon, Z.A., Treur, J., van der Wal, C.N.: A Multi-Agent Model for Emotion Contagion Spirals Integrated within a Supporting Ambient Agent Model. In: Yang, J.-J., Yokoo, M., Ito, T., Jin, Z., Scerri, P. (eds.) PRIMA 2009. Lecture Notes in Computer Science, LNAI, vol. 5925, pp. 48–67. Springer, Heidelberg (2009)
3. Bosse, T., Jonker, C.M., van der Meij, L., Treur, J.: A Language and Environment for Analysis of Dynamics by Simulation. Int. J. of AI Tools 16, 435–464 (2007)
4. Damasio, A.: Descartes' Error: Emotion, Reason and the Human Brain. Papermac, London (1994)
5. Damasio, A.: The Feeling of What Happens. Body and Emotion in the Making of Consciousness. Harcourt Brace, New York (1999)
6. Damasio, A.: Looking for Spinoza: Joy, Sorrow, and the Feeling Brain. Vintage books, London (2003)
7. Damasio, A.: The Somatic Marker Hypothesis and the Possible Functions of the Prefrontal Cortex. Philosophical Transactions of the Royal Society: Biological Sciences 351, 1413–1420 (1996)
8. Hodges, W.: Model theory. Cambridge University Press, Cambridge (1993)
9. Hoogendoorn, M., Treur, J., van der Wal, C.N., van Wissen, A.: Modelling the Emergence of Group Decisions Based on Mirroring and Somatic Marking. In: Zhong, N. (ed.) BI 2010. LNCS, vol. 6334, pp. 29–41. Springer, Heidelberg (2010)
10. Iacoboni, M.: Mirroring People. Farrar, Straus & Giroux, New York (2008)
11. Pineda, J.A. (ed.): Mirror Neuron Systems: The Role of Mirroring Processes in Social Cognition. Humana Press Inc. (2009)

12. Rizzolatti, G., Sinigaglia, C.: Mirrors in the Brain: How Our Minds Share Actions and Emotions. Oxford University Press, Oxford (2008)
13. Sharpanskykh, A., Treur, J.: Verifying Interlevel Relations within Multi-Agent Systems. In: Proc. of the 17th Eur. Conf. on AI, pp. 247–254. IOS Press, Amsterdam (2006)
14. Sharpanskykh, A., Treur, J.: Relating Cognitive Process Models to Behavioural Models of Agents. In: Jain, L., et al. (eds.) Proc. of the 8th Int. Conf. on Intelligent Agent Technology, IAT 2008, pp. 330–335. IEEE Computer Society Press, Los Alamitos (2008)
15. Treur, J.: On the Use of Reduction Relations to Relate Different Types of Agent Models. Web Intelligence and Agent Systems (2010) (to appear)

Quick Response System Using Collaborative Filtering on Fashion E-Business

Kyung-Yong Chung[1,*], Chang-Woo Song[2], Kee-Wook Rim[3], and Jung-Hyun Lee[2]

[1] School of Computer Information Engineering, Sangji University
dragonhci@hanmail.net
[2] Department of Information Engineering, Inha University
[3] Department of Computer and Information Science, Sunmoon University

Abstract. As fashion E-business is coming, it is becoming important to provide the analysis of preferences that is becoming increasingly more customer oriented. Consumers caused the diversification of the fashion product because they seek fashion and individuality in order to satisfy their needs. In this paper, we proposed the quick response system using the collaborative filtering on fashion E-business. The proposed method applies the developed quick response system to increase the efficiency of merchandising for the products of design styles. Collaborative filtering was adopted in order to recommend final design styles of interest for designers based on the predictive relationship discovered between the current designer and other previous designers. Ultimately, this paper suggests empirical applications to verify the adequacy and the validity of our system.

Keywords: Collaborative Filtering, QR System, Fashion E-Business, HCI.

1 Introduction

It is normal to develop fashion products through predicting purchase needs of consumers in Fashion E-business. Various products are produced at the same time and distributed efficiently in the different regions. To produce fashion products in the proper quantity and to distribute to customers who want it, precise market analysis, reasonable distribution network and fast exchange of information are essential [1],[2],[3],[4]. The quick response system allows that the company observes the consumer's needs consistently and establishes manufacture schedule rapidly so that they could prohibit the products unnecessarily stocked. Consumer's preference is collected and analyzed through the data generated by POS system, and this is provided to the related manufacturer through network in real time, so that the manufactures could merchandise rapidly, produce and deliver the products according to the consumer's need [5],[6],[7]. The merchandising of apparel product, product by prediction, retailer, purchase behavior, reduction in stock inventory and product lead-time, cooperative system, and conformation to the consumer's needs are included, through binding the new technology in fashion industry for the purpose of quick response system.

[*] Corresponding author.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 54–63, 2010.

The rest of this paper is organized as follows. Section 2 describes briefly the process of quick response system. Section 3 illustrates the proposed quick response system using the collaborative filtering in more detail. In Section 4, the experimental results are presented. The conclusions are given in Section 5.

2 Process of Quick Response System on Fashion E-Business

The life cycles of the fashion E-business are short. However, lead-time from planning to production and sales is long. And all produced products seem to be sold or unsold. Therefore the apparel industry increases the production that can actively cope with responses, which is called the market responses, and diminishes the burden of inventories [3]. Figure 1 shows the process of the quick response system.

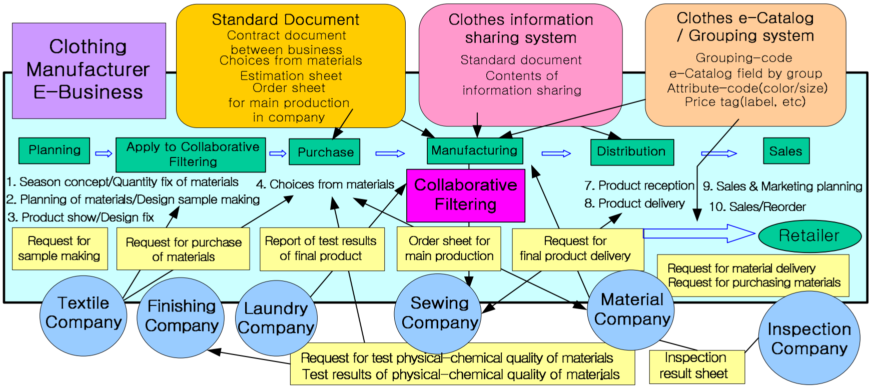
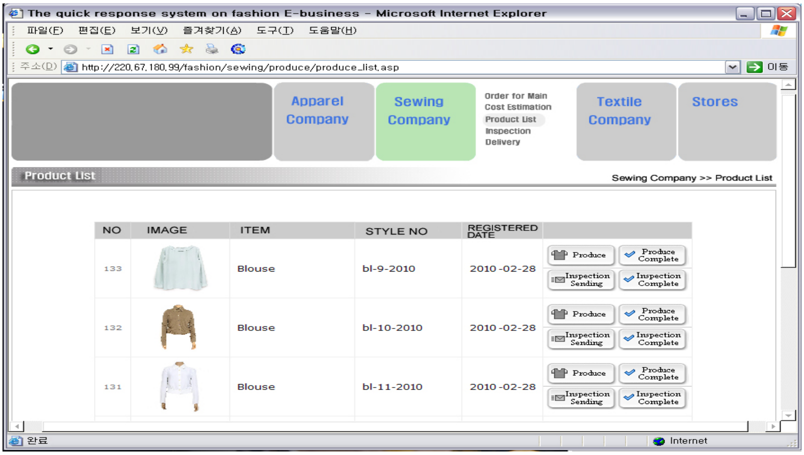

Clothing Manufacturer E–Business
Standard Document
Contract document between business
Choices from materials
Estimation sheet
Order sheet for main production in company
Clothes information sharing system
Standard document
Contents of information sharing
Clothes e–Catalog / Grouping system
Grouping–code
e–Catalog field by group
Attribute–code(color/size)
Price tag(label, etc)
Planning
Apply to Collaborative Filtering
Purchase
Manufacturing
Distribution
Sales
1. Season concept/Quantity fix of materials
2. Planning of materials/Design sample making
3. Product show/Design fix
4. Choices from materials
Collaborative Filtering
7. Product reception
8. Product delivery
9. Sales & Marketing planning
10. Sales/Reorder
Retailer
Request for sample making
Request for purchase of materials
Report of test results of final product
Order sheet for main production
Request for final product delivery
Request for material delivery
Request for purchasing materials
Textile Company
Finishing Company
Laundry Company
Sewing Company
Material Company
Inspection Company
Request for test physical–chemical quality of materials
Test results of physical–chemical quality of materials
Inspection result sheet

Fig. 1. Process of quick response system


The quick response system on fashion E-business - Microsoft Internet Explorer
파일(F) 편집(E) 보기(V) 즐겨찾기(A) 도구(T) 도움말(H)
주소(D) http://220.67.180.99/fashion/sewing/produce/produce_list.asp
이동
Apparel Company
Sewing Company
Order for Main
Cost Estimation
Product List
Inspection
Delivery
Textile Company
Stores
Product List
Sewing Company >> Product List
완료
Internet

NO	IMAGE	ITEM	STYLE NO	REGISTERED DATE	
133		Blouse	bl-9-2010	2010 -02-28	Produce Produce Complete Inspection Sending Inspection Complete
132		Blouse	bl-10-2010	2010 -02-28	Produce Produce Complete Inspection Sending Inspection Complete
131		Blouse	bl-11-2010	2010 -02-28	Produce Produce Complete Inspection Sending Inspection Complete

Fig. 2. Quick response system on fashion E-business (Sewing company-Product list)

The quick response strategic is designed to fill the gap from the production to their distribution in order to meet the needs of customers on a basis and to offer an appropriate amount of products at an appropriate cost. And the use of the system became visually convenient by reorganizing it into an apparel company, a textile company, a sewing company, stores, and outlets. As a production process, a web based system that can efficiently identify such processes as textile procurement, warehousing, stocking, and production progress is needed. This system enables suppliers, manufacturers, distributors, and retailers to share information, beyond the limitation as individual companies, and even allows small and medium sized companies to conduct international trading with their clients.

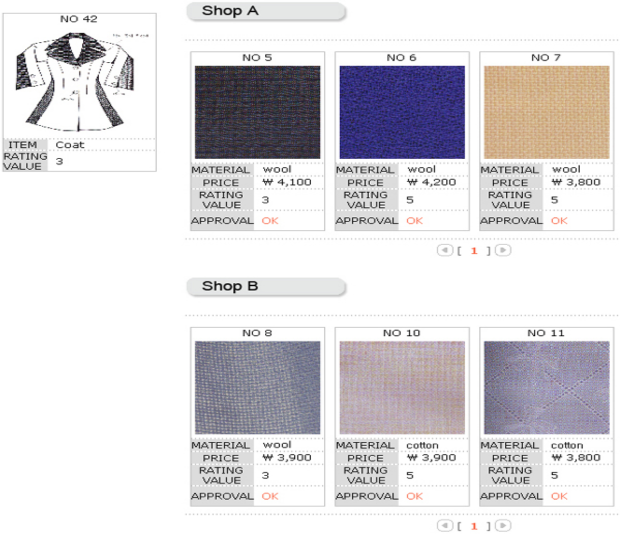
The quick response system on fashion E-business is shown as in Figure 2. Our system was conducted at http://220.67.180.99/fashion/. The specification of the server computer were IBM eServer X206, 2.8 GHz, 4GB RAM. It was developed with ASP and Microsoft SQL server 2005 to be applicable to the Web. The transaction data collected from users were stored in the database.

3 Quick Response System Using Collaborative Filtering

3.1 Apparel and Textile Module in Quick Response System

The apparel module can be classified into such sub-domains as a product show, final styles, request for sample making, cost account, process status, and outlet store management. Here, domains have their own screen pages, respectively. All these are organically connected as well. In a real product show, merchandisers and designers select the designs to be produced as samples among designs in several occasions of meetings. However, the proposed system enables to go directly to the product show page, and enter evaluation scores for each style. Also, the evaluation for each style becomes possible online from a long distance. The page which shows the processing status - the page showing the current location of the process - enables to confirm the current process status regarding the final styles. The point in time of process status indicates the current location of an apparel company, a sewing factory, and stores; concretely, it is demonstrated by time zone, and the following steps are included: (1) drawing up a job, (2) sending the job, (3) drawing up a cost account, (4) sending the cost account to head office, (5) modification of the cost account, (6) completion of an approval of the cost account, (7) under production, (8) sending the product to product inspection, (9) completion of product inspection, (10) warehousing to the head office, (11) delivering to stores, (12) warehousing to stores [3].

The textile module can be classified into such sub-domains such as textile shows, final textile styles, textile lists, and textile estimate list functions. The textile module is organically connected with the apparel module and sewing module, and judgment of job progress situation, through which jobs are confirmed in each module and an approval process is included in the textile module. The textile show page is demonstrated to include colors, textile tissue, and post-processing in terms of textile. Generally, textile designers should be professionals who can design using computers, and be fully aware of technical aspect in a textile producing factory. The textile designers can precisely direct to actually print and apply it to textile design, after experimenting the colors and textile tissues [3]. Figure 3 shows the textile that gained 3 points or higher among the evaluated textile fabrics. When an approval is made here, it is sent to the approval list of textile list, and textile estimate can be drawn up and modified for each approved textile. The drawn up textile estimate list is sent to the sewing module, and waits for the approval. When the textile estimate is approved in the sewing module, textile is produced in mass production and supplied in the textile module.



NO 42

ITEM	Coat
RATING VALUE	3

Shop A

	NO 5	NO 6	NO 7
MATERIAL	wool	wool	wool
PRICE	₩ 4,100	₩ 4,200	₩ 3,800
RATING VALUE	3	5	5
APPROVAL	OK	OK	OK

[1]

Shop B

	NO 8	NO 10	NO 11
MATERIAL	wool	cotton	cotton
PRICE	₩ 3,900	₩ 3,900	₩ 3,800
RATING VALUE	3	5	5
APPROVAL	OK	OK	OK

[1]

Fig. 3. Confirming textile samples page from the textile company

3.2 Sewing and Store Module in Quick Response System

The cost account, textile estimate, product list, product inspection, and delivery function were included in the sewing module under the close cooperation with an apparel company, and all these have each screen page [3]. The inspection page for sewing defects is shown as in Figure 4. The sewing defects have uneven collar, misaligned/crooked pocket, twisted pant leg, hiking at front or back, uneven hemline, uneven at back vent, 'scissoring' at front or back, fullness in fly, puckering at collar joint, sleeve hanging backward, puckering at crotch, misaligned buttons, and buttonholes causing gaps. Figure 4 is carried out focusing on a production request, and size specifications. The content to be inspected is different according to the kinds of apparel, but raw and subsidiary materials, sewing process, size consistency, finishing touch, and box confirmation are included in general. After a problem is caused, there is almost no problem solving method; therefore, the concept of quality focuses on company-wide quality control in improving work efficiency, while it is concerned with diminishing problems ranging from product design to production process. Accordingly, defect rate in sewing is automatically calculated through this system.

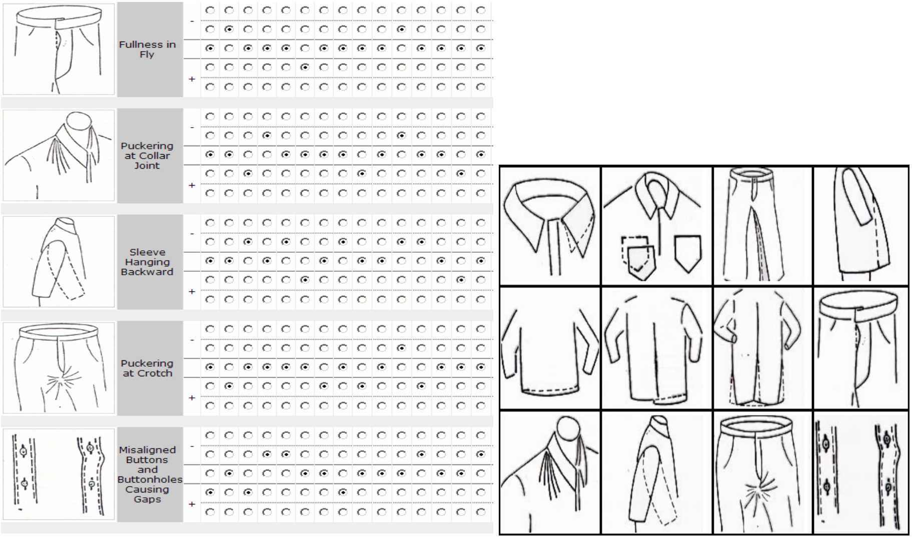
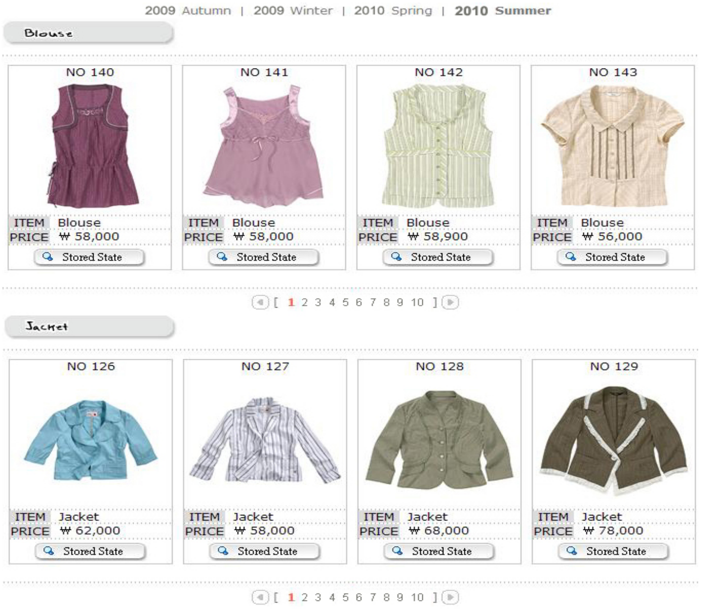

Fullness in Fly
-
+
Puckering at Collar Joint
-
+
Sleeve Hanging Backward
-
+
Puckering at Crotch
-
+
Misaligned Buttons and Buttonholes Causing Gaps
-
+

Fig. 4. Inspection page for sewing defects (Sewing company-Inspection)


2009 Autumn | 2009 Winter | 2010 Spring | 2010 Summer

Blouse

NO 140	NO 141	NO 142	NO 143
ITEM Blouse	ITEM Blouse	ITEM Blouse	ITEM Blouse
PRICE ₩ 58,000	PRICE ₩ 58,000	PRICE ₩ 58,900	PRICE ₩ 56,000
Stored State	Stored State	Stored State	Stored State

[1 2 3 4 5 6 7 8 9 10]

Jacket

NO 126	NO 127	NO 128	NO 129
ITEM Jacket	ITEM Jacket	ITEM Jacket	ITEM Jacket
PRICE ₩ 62,000	PRICE ₩ 58,000	PRICE ₩ 68,000	PRICE ₩ 78,000
Stored State	Stored State	Stored State	Stored State

[1 2 3 4 5 6 7 8 9 10]

Fig. 5. Product style list page in shopping stores

The shopping store stages are composed of internet shopping stores and outlet stores. In each store, the products are arranged by each item, and thus composed of skirts, slacks, one-piece dresses, jumpers, jackets, shirts, blouses, and so on. This is also arranged by each year again, and the data concerning these products are built up each year. The warehousing status can be confirmed by product in Figure 5. In doing so, the screen concerning the status of inventory at the head office, status, total quantity, sales volume, and sales ratio, popularity is displayed. All these data are required to identify product inventory level in the store. When warehousing confirmation is made in the shopping store module, an automatic notification regarding completion is delivered to the appropriate apparel company.

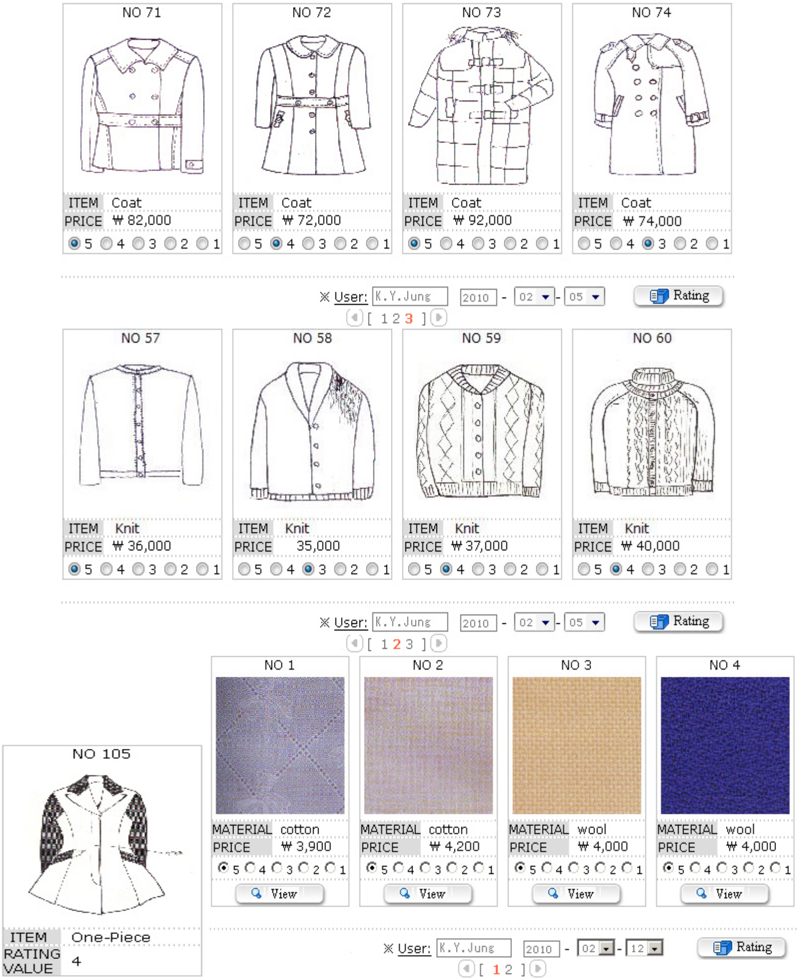
3.3 Selection of Final Design Styles Using Collaborative Filtering

In this section, collaborative filtering is used for personalized recommendation services, which provide recommendations based on ratings. Collaborative filtering recommends objects for a target user based on the opinions of other users by considering to what extent the target user and the other users have agreed on other objects in the past. This enables the technique to be used on any type of objects. A large variety of services can be created, since the collaborative filtering considers only human judgments about the value of objects. These judgments are usually expressed as numerical ratings, expressing the preference for objects. The importance of collaborative filtering is reflected in a growing number of research activities. One of the earliest was the *Grouplens* project, which focused on filtering news articles from Usenet, and more recently movie recommendation. *Ringo* was a collaborative filtering prototype for recommending music, leading to the spin-off company *Firefly* [11],[12]. Most collaborative filtering systems collect the user opinions as ratings on a numerical scale, leading to a sparse matrix rating. The Collaborative filtering technique then uses this rating matrix to predict the rating. Several algorithms have been proposed to utilize the rating matrix [2],[4]. In our fashion recommendation system, we apply a commonly used technique, also used in the *Grouplen* project and in *Ringo*, which is based on vector correlation.

In apparel and textile module, online rating is allowed in the limited score range from 1 to 5 in the product show page concerning the evaluation of each style, since there are liberal or conservative people in relation to the evaluation scores. (1 means very negative evaluation, 5 shows very positive evaluation). The design styles recommended by collaborative filtering are selected as the final styles. Figure 6 shows the selection stage of final design styles using the collaborative filtering.

All the participants select ‘Enter’ in the participant’s selection section, and enter their names. Then they give evaluation scores by clicking on ‘Rating’ button. If the concerned PC’s IP and evaluator’s name are the same, the system takes a measure not to build up evaluation with dual evaluation. When an evaluator completes evaluation and clicks on ‘Finish’ button, the evaluator is automatically logged out. The product show page is arranged by item (skirt, slacks, dress, jumper, jacket, shirt, blouse, etc.), year, season, gender, and all these are reconfigured with new windows [3].

In the following we describe the underlying formulas in more detail to make the general idea of automatically using other designer as expert recommenders understandable. Usually, the task of Collaborative filtering is to predict the rating of a particular designer *u* for style *s*. The system compares the designer *u*’s ratings with the rating of all other designers, who have rated the style *s*. Then a weighted average of the other designer rating is used as a prediction. If $Style_u$ is set of styles that a designer



NO 71	NO 72	NO 73	NO 74
ITEM Coat	ITEM Coat	ITEM Coat	ITEM Coat
PRICE ₩ 82,000	PRICE ₩ 72,000	PRICE ₩ 92,000	PRICE ₩ 74,000
● 5 ○ 4 ○ 3 ○ 2 ○ 1	○ 5 ● 4 ○ 3 ○ 2 ○ 1	● 5 ○ 4 ○ 3 ○ 2 ○ 1	○ 5 ○ 4 ● 3 ○ 2 ○ 1

※ User: K.Y.Jung 2010 - 02 - 05 Rating

[1 2 3]

NO 57	NO 58	NO 59	NO 60
ITEM Knit	ITEM Knit	ITEM Knit	ITEM Knit
PRICE ₩ 36,000	PRICE 35,000	PRICE ₩ 37,000	PRICE ₩ 40,000
● 5 ○ 4 ○ 3 ○ 2 ○ 1	○ 5 ○ 4 ● 3 ○ 2 ○ 1	○ 5 ● 4 ○ 3 ○ 2 ○ 1	○ 5 ● 4 ○ 3 ○ 2 ○ 1

※ User: K.Y.Jung 2010 - 02 - 05 Rating

[1 2 3]

NO 1	NO 2	NO 3	NO 4
MATERIAL cotton	MATERIAL cotton	MATERIAL wool	MATERIAL wool
PRICE ₩ 3,900	PRICE ₩ 4,200	PRICE ₩ 4,000	PRICE ₩ 4,000
● 5 ○ 4 ○ 3 ○ 2 ○ 1	● 5 ○ 4 ○ 3 ○ 2 ○ 1	● 5 ○ 4 ○ 3 ○ 2 ○ 1	● 5 ○ 4 ○ 3 ○ 2 ○ 1
View	View	View	View

NO 105

ITEM One-Piece

RATING VALUE 4

※ User: K.Y.Jung 2010 - 02 - 12 Rating

[1 2]

Fig. 6. Selection stage of final design styles using collaborative filtering

u has rated then we can define the mean rating of designer *u*. Collaborative filtering algorithms predict the rating based on the rating of similar designer. When the Pearson correlation coefficient is used, similarity is determined from the correlation of the rating vectors of designer *u* and the other designer *a* by Eq. (1).

$$w(u,a) = \frac{\sum_{s \in Style_u \cap Style_a} (r_{u,s} - \overline{r_u})(r_{a,s} - \overline{r_a})}{\sqrt{\sum_{s \in Style_u \cap Style_a} (r_{u,s} - \overline{r_u})^2 \bullet \sum_{s \in Style_u \cap Style_a} (r_{a,s} - \overline{r_a})^2}} \tag{1}$$

The value of *w*(*u*,*a*) measures the similarity between the designers' rating. If the value is 1, it means the positive relationship, and if -1, then it means the negative relationship, and if 0, it means that there's no correlation. The prediction formula is based on the assumption that the prediction is a weighted average of the other designers' rating. The weights refer to the similarity between the designer *u* and the other designers by Eq. (2). *Style$_d$* represents the designers who rated style *s*.

$$p(u,s) = \overline{r}_u + \frac{1}{\sum_{a\in Style_d} w(u,a)} \cdot \sum_{a\in Style_d} w(u,a)\cdot(r_{a,s} - \overline{r}_a) \qquad (2)$$

4 Experimental Results

In this section, to verify the efficiency of the proposed method, this implemented prototype system is named the quick response system using the collaborative filtering (QRS-CF). QRS-CF was developed because the fashion E-business did not have a system that could coordinate apparel designers and textile producers, resulting in inefficient fashion development. For experiment, we selected several well known quick responses system with easily available implementation. There are two methods of quick responses, based on two-level quick responses system for Chinese fashion distribution (2-LQRS) [8], Web-EDI [9],[13]. In order to evaluate our system, QRS-CF is compared with 2-LQRS, Wed-EDI. We calculate the T-test of three kinds of quick responses systems. The T-test is a method that uses T-distribution in a statistical verification process. The T-distribution shows bilateral symmetry like normal distribution and changes in the peak of the distribution according to the number of cases. Also, the T-test can be used to verify a possible difference in average values between two target groups. In addition, it classifies the groups as the case of independent sampling and dependent sampling.

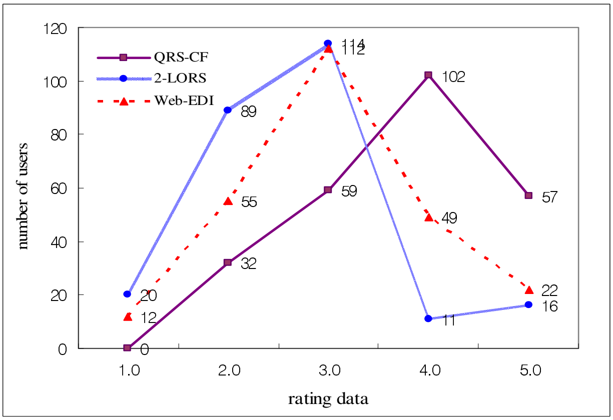


Fig. 7. Distribution of the rating data for the satisfaction

This study established evaluation data based on the results of the survey performed through off/on-line in order to verify the effectiveness and validity of QRS-CF. By attending 250 users, the evaluation data was used to evaluate the specific satisfaction for the recommended final design styles list ranged from 1(negative) to 5(positive) with the interval of 1 using QRS-CF. Whereas, in the number ranged from 1 to 5, 1 means very negative evaluation and 5 shows very positive evaluation. The survey collected 750 evaluation data for 21 days. Figure 7 illustrates the distribution of the rating data for the satisfaction. In the rating data, the distribution of the satisfaction obtained by the proposed method showed better scores more than 4 points compared to that of 2-LQRS, Web-EDI. It showed that users positively satisfied the quick response system using the collaborative filtering compared to that of different methods without using collaborative filtering.

Table 1. QRS-CF/2-LQRS/WEB-EDI paired samples statistics

Method	Mean	N	Std. Deviation	Std. Error Mean
QRS-CF	3.7360	250	0.95391	0.06033
2-LQRS	2.6560	250	0.92796	0.05869
Web-EDI	3.0640	250	0.99995	0.06324

Table 2. QRS-CF/2-LQRS/WEB-EDI paired samples test

Paired Variables	Paired Samples Test			
	Mean	T	95% Confidence Interval of the Difference	
			Lower	Upper
QRS-CF/2-LQRS	1.0800	13.568	0.92323	1.23677
QRS-CF/Web-EDI	0.67200	8.075	0.50810	0.83590

Table 1, 2 show the results of the T-test for the paired samples of QRS-CF, 2-LQRS, Web-EDI. As the significant level of α was 0.05, the critical value for making a decision was T <{0.9232, 0.5081} or T >{1.2367, 0.8359}. Because the values of T was presented as 13.568>1.2367 and 8.075>0.8359 in the T-test of the evaluation data in this paper, Thus, "there are statistical differences in the satisfaction of the QRS-CF and {2-LQRS, Web-EDI}" was accepted. Also, it was verified that the satisfaction of the QRS-CF showed a high level of 1.0800 or 0.6720, which is the difference in the average value of the evaluation data, compared to the 2-LQRS or Web-EDI.

5 Conclusion

The quick response system allows that the companies observe consumers needs consistently and design products and establish manufacture schedule rapidly so that they could prohibit the products to be unnecessarily stocked. Up and down stream in the long process of apparel industry needs to shorten the process time that include the communication and instruction. In this paper, we proposed the quick response system using collaborative filtering on fashion E-business. In our system, off/on-line voting is allowed in the limited score range show page concerning the evaluation of each design style, since there are liberal or conservative people in relation to the evaluation scores. The design styles recommended by collaborative filtering are selected as the final design styles. This process can integrate the recommendation to be generated at apparel and textile module by real time and it can analyze and support fashion E-business decision making in various environments. The results are encouraging and provide empirical evidence that the use of the proposed method can lead to improved performance in product by prediction, recommendation, product lead-time, and reduction in stock inventory. Future research includes to design the quick response system specifically and to study the algorithm to manage the changes of the preference efficiently.

Acknowledgements

This research was supported by the MKE, Korea, under the ITRC support program supervised by the NIPA. (NIPA-2010-C1090-1031-0004).

References

1. Oh, H.N.: Internet Usage for the Implementation of Quick Response as Supply Chain Management across B2B Electronic Commerce in Textile and Apparel Industry. J. Costume Culture 9(1), 100–110 (2001)
2. Jo, J.S., Lee, J.Y.: The Essential Information Items to be included in the E-catalogues for B2B Commerce of Textile Materials. J. Korean Society of Clothing and Textiles 26(9-10), 1366–1377 (2002)
3. Jung, K.Y., Kim, J.H., Lee, J.H., Na, Y.J.: Development of Quick Product Planning System for Textile and Fashion E-Business. In: Jacko, J.A. (ed.) HCI 2007. LNCS, vol. 4551, pp. 1138–1146. Springer, Heidelberg (2007)
4. Jung, K.Y., Lee, J.H.: User Preference Mining through Hybrid Collaborative Filtering and Content-based Filtering in Recommendation System. IEICE Trans. on Information and Systems E87-D(12), 2781–2790 (2004)
5. Yeum, I.S.: Quick Response System of Chinese Fashion Distribution with 2-lever Location Allocation. MA Thesis, Inha Univ., Korea (2004)
6. Park, D.S.: A Study on Korean and Foreign Applications of Quick Response System in the Textile Industry. MA Thesis, Hanyang Univ., Korea (2000)
7. Cho, K.W.: Direction of Fashion Education for Globalization of Domestic Fashion Industry in 21C. KOFOTI 14 (1996)
8. Ko, E.J.: Transactions: A Study of the Model Development of Korean Quick Response System (Part 1)-Focused on the Adoption Situation and the Factors related to the Adoption. J. Korean Society of Clothing & Textiles 23(7), 1052–1063 (1996)
9. Kim, J.Y.: 3-D Trimming System for Bias-Cut Apparels. J. Science of Emotion and Sensibility 7(2), 157–161 (2004)
10. Kwon, H.S.: The Spread of Role from Changed in the Paradigm of POP. J. Korean Society for Clothing Industry 4(1), 1–4 (2002)
11. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating Collaborative Filtering Recommender Systems. J. ACM Trans. on Information Systems 22(1), 5–53 (2004)
12. Kim, T.H., Yang, S.B.: An Effective Recommendation Algorithm for Clustering-Based Recommender Systems. J. Advances in Artificial Intelligence 3809, 1150–1153 (2005)
13. Lee, J.S., Han, G.H.: A Study to improve the Linkage between Apparel Industry and University Education on Clothing Construction-Focus on Process of Pattern Making. J. Costume Culture 7(6), 972–984 (1999)

An Ownership Verification Scheme by Using Image Pinned Field and Genetic Algorithm

Mn-Ta Lee[1], Hsuan-Ting Chang[2], and Mu-Liang Wang[3]

[1] Department of Electronic Engineering, Kao Yuan University,
821 Kaohsiung County, Taiwan
mtlee@cc.kyu.edu.tw
[2] Department of Electrical Engineering, National Yunlin University
of Science and Technology, 640 Yunlin, Taiwan
htchang@yuntech.edu.tw
[3] Department of Computer Science and Information Engineering, Shu-Te University,
824 Kaohsiung County, Taiwan
mulwang@stu.edu.tw

Abstract. As a result of explosive development of the Internet, digital contents can be convenient and fast exchanged on the Internet. Hence, how to protect important digital contents from being stolen or modified has become an important issue. In this paper, an optimal ownership verification scheme for image contents based on using both the pinned field of the protected image and the genetic algorithm is proposed. The pinned field explores the texture information of the protected image and can be used to enhance the robustness of the watermark. The genetic algorithm is then used to optimize the similarity between the texture information and the protected image. Experimental results show that the proposed scheme can survive under different signal processing and geometric transformation attacks.

Keywords: ownership verification, pinned field, linear feedback shift register, content authentication, genetic algorithm.

1 Introduction

The advance of computer technology and the population of the Internet have resulted in convenient and fast exchange of multimedia contents. With the more power computer and the high speed network, many popular and useful applications such as on-line games, blogs, e-learning, video on demand, e-map, etc., are running on the Internet. These applications use many multimedia contents such as texts, images, videos, audios, and pictures. Due to the public and insecure environment of the Internet, many intruders intend to do malicious attacks on multimedia contents stored on computer severs. These malicious attacks include illegal copying, tampering, modifying, and stealing digital multimedia contents. Hence, how to provide suitable techniques to protect digital multimedia contents from malicious attacks is an important and emergency issue to conquer.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 64–73, 2010.

Digital watermarking [1, 2] has been proposed as a technique to protect digital rights. By embedding owner's watermarks such as logo, trademark, seal, or copyright information into the digital content without changing the perception of the digital content, owner can claim his ownership, intellectual property, and content authentication. According to the domain where the owner watermarks are embedded, digital watermarking technologies can be classified into two main groups, i.e. the spatial domain and frequency domain technologies. For spatial domain technologies, the watermarks are embedded into the digital content by directly modifying the coefficients of the digital content. On the other hand, the frequency domain technologies embed the watermarks by modulating the coefficients of the digital content. After the embedding process, the embedded owner watermarks can be extracted from the digital content for identifying the copyright owner.

In the literatures [3, 4], the owner's watermarks are embedded in the spatial domain. Watermarks embedded in the spatial domain are straightforward methods and have the advantages of low complexity and easy implementation. However, they have disadvantages that image processing operations may easily destroy the watermarks. In the literatures [5, 6], the owner's watermarks are embedded in the frequency domain. Watermarks embedded in the frequency domain are more robust than watermarks embedded in the spatial domain. For watermarks embedded in the frequency domain, there are many technologies used such as the discrete Fourier transform (DFT), discrete cosine transform (DCT), and discrete wavelet transform (DWT). Watermarks embedded in the frequency domain are time-consuming because the pixel values of the protected image must be transformed into corresponding frequency domain.

Due to the reason that watermark embedding procedure usually produces a slight degradation in the digital multimedia contents, it is not suitable for valuable and sensitive digital multimedia contents. Therefore, schemes combining signature with digital watermarking-like techniques had been proposed in the literatures [7-9] to conquer this challenge. The general model of these combining schemes is inducted in our previous paper [10]. Based on this general model, a scheme for image copyright protection by using the pinned field of the protected image is also proposed, according to the observation that the robustness of watermarks could be enhanced by using the feature of the protected image. The pinned field reflects the texture information of the image and is used to enhance the robustness of watermark. In this paper, we propose an optimal ownership verification scheme for image contents based on using both the pinned field of the protected image and the genetic algorithm. After deciding the image pinned field, the genetic algorithm is then used to decide the optimal threshold values for the texture information of the protected image to optimize the similarity between the texture information and the protected image. Experiments show that the robustness of watermark can be further enhanced through this optimization process when compares with [10].

The rest of this paper is organized as follows: In Section 2, the image pinned field is illustrated, and then the genetic algorithm is described. The proposed novel scheme for ownership verification is illustrated in Section 3. Experimental results are presented in Section 4. Finally, we conclude our works in Section 5.

2 Related Works

The related works, including the image pinned field and the genetic algorithm, are described in this section.

2.1 The Image Pinned Field

Meiri and Yudilevich [11] proposed the pinned sine transform (PST) for image coders. The PST, an approximation to the pinned Karhunen-Loeve transform PKLT [12], uses the properties of the block boundaries to partition an image into two fields, namely, the boundary field and the pinned field. The boundary field depends only on the block boundary and the pinned field vanishes at the boundaries. The pinned field reflects the texture information of the image.

The PST divides the image *X* into non-overlapping blocks of size *k*x*r* pixels [13]. A typical block $X_{m,n}$, where *m* and *n* are the coordinates of this block, is shown in Fig. 1.

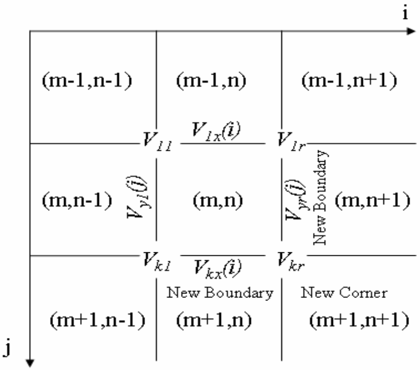


Fig. 1. The dual-field decomposition in PST for a typical block

Considering this typical block, we can find that each corner is shared by four blocks and each boundary is shared by two. The boundary field $B_{m,n}$ of block $X_{m,n}$ is obtained by the pinning function [11] and the pinned field $P_{m,n}$ is then achieved by

$$P_{m,n}(j,i)=X_{m,n}(j,i)-B_{m,n}(j,i) \tag{1}$$

An example shown the image pinned field was illustrated in our previous paper [10].

2.2 Genetic Algorithm

Genetic algorithm (GA) [14] is a strategy to solve optimization problems that simulates biological evolutions to obtain an optimal solution. GA is widely used in various fields such as pattern recognition, decision support and the nearest optimization problem. There are mainly five components, i.e., random number generator, fitness evaluation, reproduction operation, crossover operation, and mutation operation in the GA.

In general, GA starts at an initial population, called first generation which is generated with some randomly selected genes. Each individual in the population corresponding to a solution in the problem domain being addressed is called chromosome. Associated with each chromosome is a fitness value computed by the fitness function. The fitness value is used to evaluate the quality of each chromosome. The chromosomes with high quality will have greater probability to survive and form the population of the next generation. Through the operation of reproduction, crossover and mutation, a new generation is regenerated from chromosomes with high fitness values in order to find the best solution. The new generation will repeatedly apply the evaluation, reproduction, crossover and mutation operations. After a constant number of iterations are reached, or a predefined condition is satisfied, the overall process will be terminated and the approached optimal solution will be found.

In our previous work [10], a scheme for image copyright protection by using the pinned field of the protected image was proposed, according to the observation that the robustness of watermarks could be enhanced by using the feature of the protected image. The pinned field reflects the texture information of the image. By transforming the texture information of the protected image into a binary feature image, the robustness of watermark can be enhanced by using the gotten binary feature image. But how to decide the threshold values for the texture information of the protected image while transforms into binary feature image, it's an important issue. In this paper, the GA is used to solve this issue.

If the binary feature image is more similar to the protected image, the robustness of watermark is more intensive. Based on this observation, the proposed scheme in this paper uses the GA to search the optimal threshold values while transforms the texture information of the protected image into binary feature image. The two-dimensional correlation value (corr$_{BC}$), defined as in (2) and representing the similarity between the binary feature image B and the protected image C, is calculated. The sizes of images B and C are W_cxH_c pixels, respectively. The correlation value is then used to calculate the fitness value f$_{val}$ of a solution in the population of GA. The fitness function to be minimized is defined as in (3).

$$corr_{BC} = \frac{\sum_{j=1}^{H_C}\sum_{i=1}^{W_C} B(j,i)C(j,i)}{\sum_{j=1}^{H_C}\sum_{i=1}^{W_C} C(j,i)^2} \qquad (2)$$

$$f_{val} = 1/corr_{BC} \qquad (3)$$

3 The Proposed Scheme

This section illustrates our proposed scheme, based on the general model inducted in [10]. There are two similar procedures in our proposed scheme. Let the cover image be the image which needs to be protected. The signature procedure uses the genetic algorithm and the pinned field of the cover image to generate a feature-based signature; however, the authentication procedure verifies the copyright of the questioned image using the feature-based signature. These two procedures are subsequently described in the following sections.

3.1 The Signature Procedure

Assume that the cover image *C* and the watermark *T* are grayscale images of size W_cxH_c and W_txH_t pixels, respectively. The signature procedure uses the pinned field of the cover image *C* to generate a feature signature. Figure 2 shows the block diagram of the signature generation procedure. The main steps are described below:

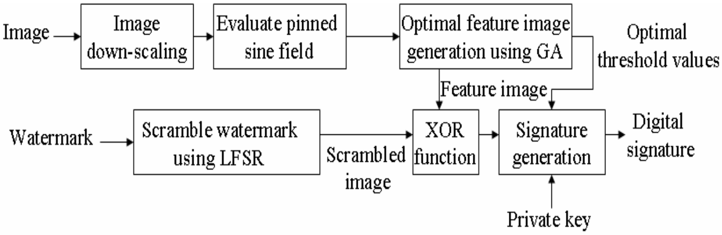


Fig. 2. Block diagram of the signature procedure

1) First, the cover image is down-scaled into the same size as the logo watermark *T*.

2) The down-scaled image is divided into non-overlapping blocks of size *k*x*r* pixels. Then, the pinned field of the down-scaled image is calculated in order to get the texture information of the image and enhance the robustness.

3) Use the genetic algorithm to decide the optimal feature threshold values *P'* of the pinned field. Then, by using optimal feature threshold values on the pinned field, an optimal feature image *F'* can be created.

4) In order to survive under geometric attacks, the watermark *T* is scrambled by using a “linear feedback shift register (LFSR),” which generates a random sequence with a random seed *R*, to form a scrambled image *T*’, i.e.,

$$T'=\{\ t'(j,i)=\mathrm{LFSR}(t(j',i')),\ 1\leqq j,\ j'\leqq H_t,\ 1\leqq i,\ i'\leqq W_t\ \}, \tag{4}$$

where LFSR denotes the linear feedback shift register scrambling function and pixel *t*(*j'*, *i'*) is scrambled to pixel *t'*(*j*, *i*) in a random order.

5) After obtaining the scrambled image *T*’, an exclusive-or (XOR) operation is applied to the scrambled image *T'* and the feature image *F'* to create the signature image *S'*, i.e.,

$$S'=T'\ \mathrm{XOR}\ F'. \tag{5}$$

6) The signature image *S'*, the random seed *R* and the optimal feature threshold values *P'* of the pinned field are then signed by using the normal signature generation system with the owner’s private key *PK* to generate a digital signature *DS'*, i.e.,

$$DS'=\mathrm{SIGN}(S', R, P', PK), \quad (6)$$

where SIGN denotes the signature generation function.

3.2 The Authentication Procedure

In order to verify the exact copyright of the questioned image, the verifier can perform the authentication procedure to prove the ownership of the questioned image. The authentication procedure does not require the presence of the cover image and is similar to the steps of the signature procedure. The block diagram of the authentication procedure is shown in Fig. 3. The main steps are described as follows:

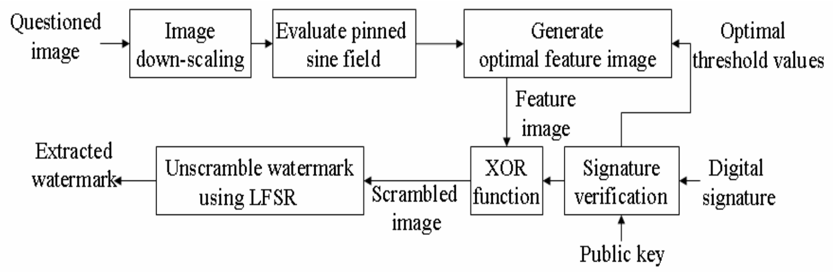


Fig. 3. Block diagram of the authentication procedure

1) Derive a down-scaled image from the questioned image using the same way described in the signature procedure.

2) Divide the down-scaled image into non-overlapping blocks of size *k*x*r* pixels. Then, the pinned field of the down-scaled image is evaluated.

3) Use the normal signature verification system with the owner's public key *UK* to verify the correctness of the digital signature *DS'*, i.e.,

$$YN^{*}=\mathrm{VERI}(DS', UK), \quad (7)$$

where VERI denotes the signature verification function and YN* is the verification result. If the verification result YN* is correct, then the signature image *S'*, the random seed *R* and the optimal feature threshold values *P'* of the pinned field are valid. Otherwise, the authentication procedure is terminated.

4) Using optimal feature threshold values *P'* on the pinned field of the down-scaled image, an optimal feature image *F** can be got.

5) Appling an XOR operation to the signature image *S'* and the feature image *F**, the result forms a scrambled logo watermark *T**, i.e.,

$$T^{*}=S' \text{ XOR } F^{*}. \quad (8)$$

6) Inversely scramble the logo watermark *T ** with seed *R*, thus generate a visual logo watermark image *WT**, i.e.,

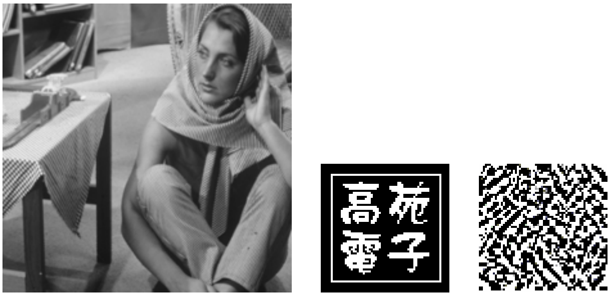
$$WT^{*}=\{wt'(j, i)=LFSR^{-1}(t^{*}(j', i')),\ 1\leqq j, j'\leqq H_t,\ 1\leqq i, i'\leqq W_t\}, \tag{9}$$

where LFSR^{-1} denotes the inverse linear feedback shift register scrambling function by using the seed *R*.

4 Experimental Results

In this section, the experiments of applying external attacks, including signal processing attacks and geometric transformation attacks on the proposed method, are demonstrated to evaluate the performance of the proposed scheme. Figures 4(a) and 4(b) show the cover image and the watermark of size 512x512 and 64x64 pixels, respectively. All of them are grayscale images. The pinned field of the reduced cover image is evaluated by dividing the reduced cover image into non-overlapping blocks of size 4x4 pixels and is shown in Fig. 4(c).

In the representation of GA individuals for a population, each individual consists of 256 variables because the reduced cover image is divided into 256 non-overlapping blocks of size 4x4 pixels. Every variable represents a possible threshold value for one block of the texture information of the reduced cover image. The 10 individuals with highest fitness value are reserved for the new population of the next generation. The number of generations for each experiment is set to 300. The values finally used for population size, mutation rate and crossover rate are 120, 0.01, 0.5.



(a) (b) (c)

Fig. 4. (a) The cover image (b) The watermark (c) The pinned field of the down-scaled image from 4(a)

The peak signal-to-noise ratio (PSNR) is used to evaluate the quality between the cover image and the attacked image. For the grayscale cover image *C* with size W_cxH_c pixels, the PSNR is defined as the follow:

$$PSNR=10\log_{10}\frac{255^2}{\frac{1}{H_cW_c}\sum_{j=1}^{H_C}\sum_{i=1}^{W_C}\left|C(j,i)-A(j,i)\right|^2}dB, \tag{10}$$

where *C*(*j*,*i*) and *A*(*j*,*i*) denote the grayscale values of the cover image *C* and the attacked image *A* at the point (*j*,*i*), respectively.

Besides, the retrieved watermark is recognizable in our method. The verifier can compare the extracted result with the original watermark. The similarity measurement between the referenced watermark *T* and extracted watermark *T'* is evaluated to estimate the robustness of our ownership verification scheme under different attacks. The similarity is evaluated by the use of the watermark retrieval rate (RR), which is the percentage of the correct pixels recovered and defined as

$$\mathrm{RR} = \frac{\sum_{j=1}^{H_T}\sum_{i=1}^{W_T}\overline{T(j,i)\ \mathrm{XOR}\ T'(j,i)}}{H_T * W_T}, \tag{11}$$

where *T*(*j*, *i*) denotes the grayscale value of the (*j*, *i*)th pixel in *T*. It is obvious that the higher *RR* is, the higher similarity between *T* and *T'* can be obtained.

Table 1 shows the experimental results of the watermarks extracted from the proposed scheme under different attacks. From these experiments, the retrieved watermark is still recognizable even though the PSNR value of the attacked image is seriously reduced. Here the *RR* values corresponding to various attacks are summarized as follows:

Attack 1) Image Blurring: A Gaussian filter with 9x9 kernel coefficients, which is used to the cover image, generates a blurring image.

Attack 2) Quarter Cropping: A quarter cropping is used to the cover image to generate a quarter cropping image.

Attack 3) Noising: The Gaussian white noise with a zero mean value and the variance value 0.01 is applied to the cover image to generate a noising image.

Attack 4) JPEG lossy compression: A JPEG lossy compression is applied to the cover image to generate a JPEG image.

Attack 5) Scaling: The cover image is first resized to 256x256 pixels and then enlarged to 512x512 pixels.

Attack 6) Sharpening: A linear mapping is applied to the cover image to generate a sharpening image.

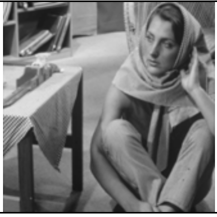
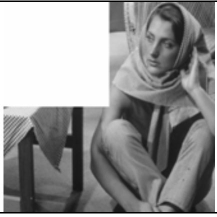
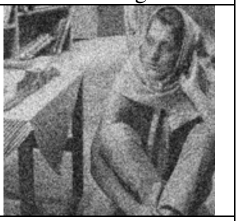
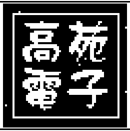
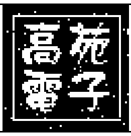
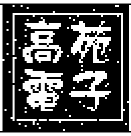
Attack 7) Median filtering: A median filter with 9x9 kernel coefficients is used to the cover image to generate a median filtering image.

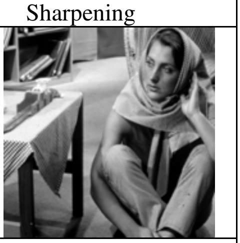
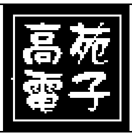
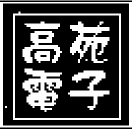
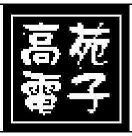
Attack 8) Gamma correction: A gamma correction (0.7) is applied to the cover image to generate a gamma correction image.

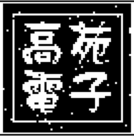
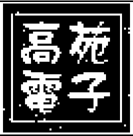
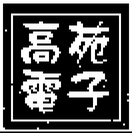
Attack 9) Histogram equalization: A histogram equalization that is uniform is applied to the cover image to generate a histogram equalization image.

As shown in Table 1, all the *RR* values are greater than 97%, which represent that the recovered watermarks are highly correlated with the original watermark. Therefore, by embedding the watermark into the pinned field of the protected image and optimizing the similarity between the texture information and the protected image through genetic algorithm is an efficient way and robust to different types of attacks.

Table 1. The Attacked Images, The Corresponding PSNR Values, The Retrieved Logos, and The Corresponding RR Values

	Image Blurring	Quarter Cropping	Noising
Attacked Image			
PSNR(dB)	33.48	12.05	11.47
Retrieved logo			
RR	99.85%	98.66%	97.44%

	JPEG	Scaling	Sharpening
Attacked Image			
PSNR(dB)	30.86	19.33	18.07
Retrieved logo			
RR	99.95%	99.93%	99.95%

	Median filtering	Gamma correction	Histogram equalization
Attacked Image			
PSNR(dB)	23.01	13.26	14.82
Retrieved logo			
RR	98.94%	99.44%	99.51%

5 Conclusions

Based on a general model, a novel scheme combining the image pinned field with genetic algorithm for image ownership verification has been proposed in this paper. The pinned field partially reflects the texture information of the images and can be used to enhance the robustness of watermark. The genetic algorithm is then used to optimize the similarity between the texture information and the protected image. The signature procedure and the authentication procedure have been described. The results of the experiments demonstrate that the proposed scheme can resist and survive under different signal processing and geometric transformation attacks, such as blurring, cropping, noising, and JPEG lossy compression, etc.

Acknowledgments

This work was partially supported by National Science Council, Taiwan under the contract number 99-2221-E-224-004.

References

1. Hsu, C.T., Wu, J.L.: Hidden Digital Watermarks in Images. IEEE Transactions on Image Processing 8(1), 58–68 (1999)
2. Podilchuk, C.I., Delp, E.J.: Digital Watermarking: Algorithms and Application. IEEE Signal Process. Mag. 18, 33–46 (2001)
3. Pitas, I., Kaskalis, T.H.: Applying Signatures on Digital Images. In: Proc. IEEE Nonlinear Signal and Image Processing, pp. 460–463 (1995)
4. Bruyndonckx, O.J., Quisquater, J., Macq, B.: Spatial Method for Copyright Labeling of Digital Images. In: Proc. IEEE Nonlinear Signal and Image Processing, pp. 456–459 (1995)
5. Koch, E., Zhao, J.: Toward Robust and Hidden Image Copyright Labeling. In: Proc. IEEE Nonlinear Signal and Image Processing, pp. 452–455 (1995)
6. Chang, H.T., Tsan, C.L.: Image Watermarking by Use of Digital Holography Embedded in DCT Domain. Applied Optics 44(29), 6211–6219 (2005)
7. Chen, T.H., Horng, G., Lee, W.B.: A Publicly Verifiable Copyright-proving Scheme Resistant to Malicious Attacks. IEEE Transactions on Industrial Electronics 52(1), 327–334 (2005)
8. Lee, W.B., Chen, T.H.: A Publicly Verifiable Copy Protection Technique for Still Images. The Journal of Systems and Software 62(3), 195–204 (2002)
9. Chang, C.C., Lin, P.Y.: Adaptive Watermark Mechanism for Rightful Ownership Protection. The Journal of Systems and Software 81(7), 1118–1129 (2008)
10. Lee, M.T., Chang, H.T., Wang, M.L.: Watermarking Mechanism for Copyright Protection by Using the Pinned Field of the Pinned Sine Transform. In: The 10th International Symposium on Pervasive Systems, Algorithms and Networks, pp. 502–507 (2009)
11. Meiri, A.Z., Yudilevich, E.: A Pinned Sine Transform Image Coder. IEEE Trans. Communications 29(12), 1728–1735 (1981)
12. Meiri, A.Z.: The Pinned Karhunen-Loeve Transform of a Two Dimensional Gauss-Markov Field. In: Proc. SPIE Conf. Image Processing, San Diego, CA (1976)
13. Ho, A.T.S., Zhu, X., Guan, Y.L.: Image Content Authentication Using Pinned Sine Transform. EURASIP Journal on Applied Signal Processing 14, 2174–2184 (2004)
14. Aslantas, V.: A Singular-value Decomposition-based Image Watermarking Using Genetic Algorithm. Int J. Electron. Commun. (AEU) 62, 386–394 (2008)

Multiple-Image Multiplexing Encryption Based on Modified Gerchberg-Saxton Algorithm and Phase Modulation in Fractional Fourier Transform Domain

Hsuan-Ting Chang[1] and Hone-Ene Hwang[2,*]

[1] Photonics and Information Laboratory, Department of ElectricalEngineering,
National Yunlin University of Science and Technology, Douliu Yunlin, 64002 Taiwan R.O.C.
[2] Department of Electronic Engineering, Chung Chou Institute of Technology,
Yuan-lin, 510 Taiwan R.O.C.
n741@ms26.hinet.net

Abstract. A technique, based on a modified Gerchberg-Saxton algorithm (MGSA) and a phase modulation scheme in the fractional Fourier-transform (FrFT) domain, is proposed to reduce crosstalks in multiple-image encryption and multiplexing. First, each plain image is encoded into a phase function by using the MGSA. Then all the created phase functions are multiplexed, with different fractional order of FrFT, and phase-modulated before being combined together into a single phase only function (POF). Simulation results show that the crosstalks between multiplexed images have been significantly reduced, compared with prior methods [1, 2], thus presenting high promise in increasing the multiplexing capacity and encrypting graylevel and color images.

Keywords: Modified Gerchberg-Saxton algorithm, fractional Fourier-transform, multiple-image multiplexing encryption.

The use of opticall multiplexing to achieve multiple-image encryption has been popular for a long time [1-4]. Differing from storing thousands of images in a single photorefractive crystal [5-8], multiple-image encryption uses two phase only function (POFs) to record several images [3, 4]. Image encryption using two statistically independent POFs is conventionally based on the Fourier-transform (FT) domain [9, 10], Fresnel-transform (FrT) domain [11, 12], or fractional Fourier-transform (FrFT) domain [13].

For multiple-image encryption purpose, the most important issue is to increase the multiplexing capacity (i.e., the number of images that can be encrypted simultaneously), or, to reduce the crosstalks on extracting the desired information encrypted therein. Situ and Zhang proposed the schemes of wavelength multiplexing [3] and position multiplexing [4] for binary images. Their methods, however, present limited applicability if the crosstalks can not be further reduced. The annoying crosstalk also prevents Situ and Zhang's schemes from applications to graylevel images.

[*] Corresponding author.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 74–80, 2010.

A novel multiple-image encryption scheme is proposed here to overcome the above crosstalk problem, aiming to increase the multiplexing capacity and enable the encryption of grayscale, or even color, images. To simplify the system complexity, we propose a modified Gerchberg-Saxton algorithm (MGSA) [15-17], operating on the FrFT domain (rather than the FT domain for conventional GSA [15, 16]), to retrieve the phase function of an image. The retrieved phase functions for all images to be encrypted are then modulated and combined (in a multiplexing manner) to form a single POF for storage.

Figure 1 shows the block diagram of the proposed MGSA. The algorithm starts with performing the inverse FrFT (abbreviated as IFrFT) on the input target image $g(x_1, y_1)$, which then gets an intermediate phase function $\psi_g(x_0, y_0)$. Next, the phase function $\psi_g(x_0, y_0)$ is constrained with a unity amplitude and then Fresnel-transformed to obtain an approximation $\hat{g}(x_1, y_1)$ with a phase function $\psi_{\hat{g}}(x_1, y_1)$ can be obtained. Again, the target image $g(x_1, y_1)$ with an updated phase function $\psi_{\hat{g}}(x_1, y_1)$ is inversely Fresnel-transformed. The above process is iterated until a required correlation (similarity) between $g(x_1, y_1)$ and $\hat{g}(x_1, y_1)$ is achieved. The converged $\psi_g(x_0, y_0)$ is then determined as the retrieved phase of $g(x_1, y_1)$, i.e., $\psi_g(x_0, y_0)$ will satisfy:

$$\begin{aligned}
&\mathrm{FrFT}\left\{\exp\left[j\psi_g(x_0, y_0)\right];\alpha\right\}\\
&= (1 - j\cot\alpha)\iint \exp\left[j\psi_g(x_0, y_0)\right]\ \exp\left[\frac{-j2\pi x_0 x_1 + j\pi(x_0^2 + x_1^2)\cos\alpha}{\sin\alpha}\right]\\
&\qquad\times\exp\left[\frac{-j2\pi y_0 y_1 + j\pi(y_0^2 + y_1^2)\cos\alpha}{\sin\alpha}\right]\mathrm{d}x_0\mathrm{d}y_0\\
&= \hat{g}(x_1, y_1)\exp\left[j\psi_{\hat{g}}(x_1, y_1)\right],
\end{aligned} \tag{1}$$

where $\alpha = \pi p/2$ and p is a fractional order of FrFT. When the input POF recorded with $\psi_g(x_0, y_0)$ which is FrFTed, the approximation image $\hat{g}(x_1, y_1)$ will be reconstructed at the (x_1, y_1) plane.

Figure 2(a) illustrates the multiple-image encryption process based on the proposed MGSA. First, each individual image $g_n(x_1, y_1)$, $n = 1 \sim N$, is encrypted into its corresponding phase function $\psi_{g_n}(x_0, y_0)$. Then, the iteration process of generating phases based on MGSA in accordance with different fractional order p_n of FrFT is performed to obtain $\psi_{p_n}(x_0, y_0)$. For different fractional order p_n multiplexing, each $\psi_{p_n}(x_0, y_0)$ satisfies:

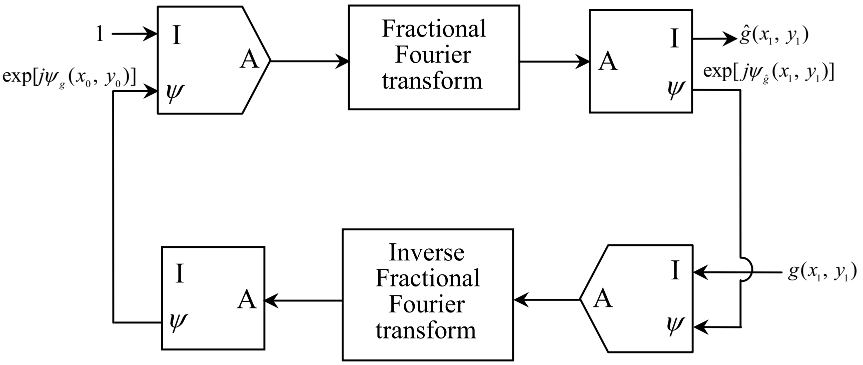

1
$\exp[j\psi_g(x_0, y_0)]$
I
A
ψ
Fractional Fourier transform
A
I
ψ
$\hat{g}(x_1, y_1)$
$\exp[j\psi_{\hat{g}}(x_1, y_1)]$
I
A
ψ
Inverse Fractional Fourier transform
A
I
ψ
$g(x_1, y_1)$

Fig. 1. Block diagram of the proposed MGSA based on FrFT domain

$$\mathrm{FrT}\left\{\exp\left[j\psi_{p_n}(x_0,y_0)\right];p_n\right\}=\hat{g}_n^p(x_1,y_1)\exp\left[j\psi^p_{\hat{g}_n}(x_1,y_1)\right], \tag{2}$$

where different fractional order p_n satisfies the relation: $\alpha_n = \pi p_n/2$ and $\psi^p_{\hat{g}_n}(x_1,y_1)$ is the accompanied phase term. These N different fractional order p_n multiplexed phase functions, $\psi_{p_n}(x_0,y_0)$, $n=1\sim N$, can be recorded together into one POF. Each encrypted image $g_n(x_1,\ y_1)$ can then be extracted or recovered from the POF as the approximation $\hat{g}_n^p(x_1,y_1)$ in Eq.(2). However, since crosstalks exist between the encrypted images which are different fractional order p_n multiplexed, the error of $\hat{g}_n(x_1,y_1)$ may be perceivable even the key for deciphering is correct. To reduce the annoying crosstalks among multiplexed images, the N encrypted images $\hat{g}_n(x_1,y_1)$ are spatially translated to different positions by using the phase modulation property of FrFT:

$$\begin{aligned}&\mathrm{FrFT}\left\{\exp\left[j\psi'_{p_n}(x_0,y_0)\right];p_n\right\}\\&\quad=\hat{g}_n^p(x_1-\mu_n\sin\alpha_n,\ y_1-\nu_n\sin\alpha_n)\exp\left[j\phi(x_1,y_1)\right],\end{aligned} \tag{3}$$

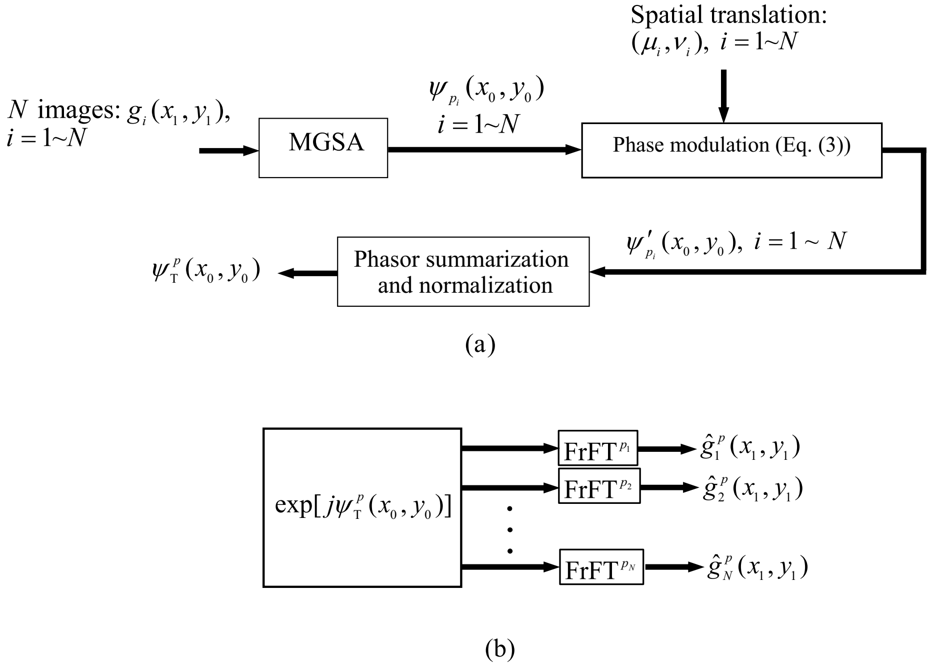
where $\psi'_{p_n}(x_0,y_0)=\psi_{p_n}(x_0,y_0)+2\pi(\mu_n x_0+\nu_n y_0)$, (4)

$\phi(x_1,y_1)$ is the accompanied phase term, and μ_n and ν_n denote the respective shift amounts of $\hat{g}_n^p(x_1,y_1)$ in the x_1 and y_1 direction, respectively, at the output plane. It is obvious from Fig. 2(b) that crosstalks can be reduced significantly with a proper arrangement of (μ_n,ν_n)'s .

To synthesize a POF for the purpose of multiple-image encryption, phasors corresponding to $\psi'_{p_n}(x_0,y_0)$, $n=1\sim N$, obtained from Eq.(4) are summed to get $\exp\left[j\psi_{\mathrm{T}}^p(x_0,y_0)\right]$:

$$\psi_{\mathrm{T}}^{p}(x_0, y_0) = arg\left\{\frac{\sum_{n=1}^{N}\exp\left[j\psi'_{p_n}(x_0, y_0)\right]}{\left|\sum_{n=1}^{N}\exp\left[j\psi'_{p_n}(x_0, y_0)\right]\right|}\right\}, \tag{5}$$

where *arg* denotes the argument operator. To the best of our knowledge, this method is new for multiplexing (encrypting) *N* images with only one POF!


Spatial translation:
(μ_i, ν_i), $i = 1 \sim N$
N images: $g_i(x_1, y_1)$, $i = 1 \sim N$
MGSA
$\psi_{p_i}(x_0, y_0)$ $i = 1 \sim N$
Phase modulation (Eq. (3))
$\psi'_{p_i}(x_0, y_0)$, $i = 1 \sim N$
Phasor summarization and normalization
$\psi_{\mathrm{T}}^{p}(x_0, y_0)$
$\exp[j\psi_{\mathrm{T}}^{p}(x_0, y_0)]$
FrFT^{p_1}
FrFT^{p_2}
FrFT^{p_N}
$\hat{g}_1^p(x_1, y_1)$
$\hat{g}_2^p(x_1, y_1)$
$\hat{g}_N^p(x_1, y_1)$

(a)

(b)

Fig. 2. (a) Block diagram of the proposed multiple-image encryption method. The optical/digital decryption system based on one POF in the fractional Fourier domain can be performed by: (b) the different FrFT order based de-multiplexing (with different order p_n).

The image decryption (extraction) process with different fractional order p_n can be expressed, respectively, as

$$\begin{aligned}&\left|\mathrm{FrFT}\left\{\exp\left[j\psi_{\mathrm{T}}^{p}(x_0, y_0)\right]; p_n\right\}\right| \\ &= \left|\hat{g}_n^p(x_1 - \mu_n, y_1 - \nu_n)\exp\left[j\psi_{\hat{g}_n}^{p}(x_1 - \mu_n, y_1 - \nu_n)\right] + n_{p_n}(x_1, y_1)\right| \\ &\approx \left|\hat{g}_n^p(x_1 - \mu_n, y_1 - \nu_n)\right| + \left|n_{p_n}(x_1, y_1)\right|,\end{aligned} \tag{6}$$

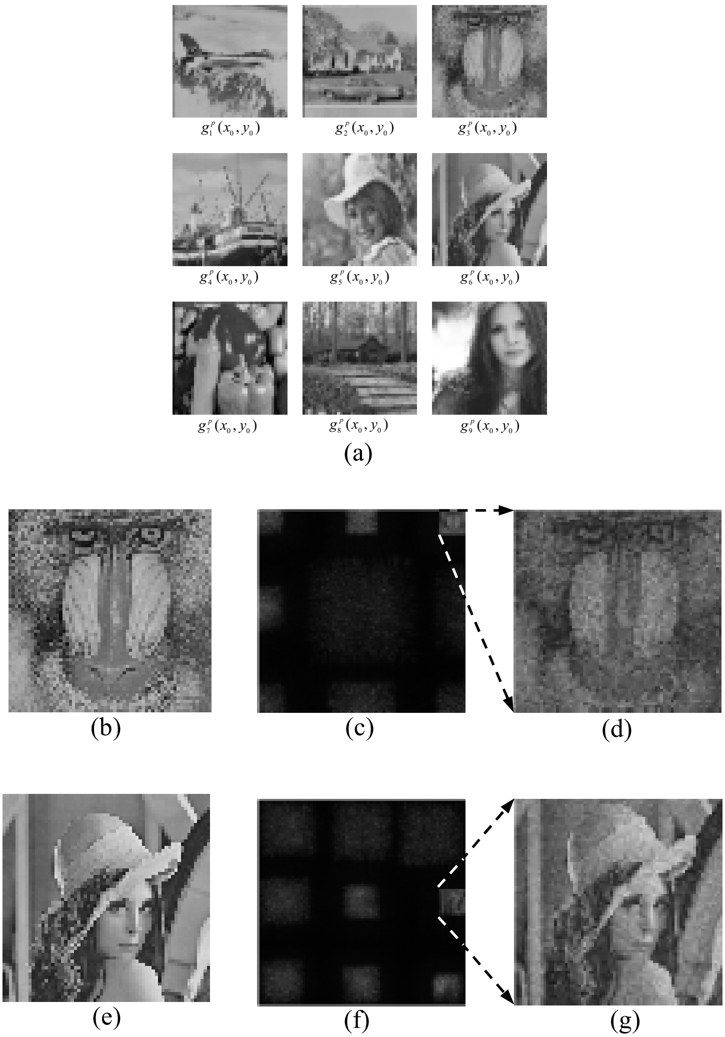


Fig. 3. (a) Nine images $g_1(x_1, y_1) \sim g_9(x_1, y_1)$ for encryption; (b) and (e) are $g_3(x_1, y_1)$ and $g_6(x_1, y_1)$ chosen for FrFT order $p_3 = 1.25$ and $p_6 = 2.0$ multiplexing, respectively; (c) and (f) are the decryption results corresponding to images in (b) and (e); (d) and (g) are the enlarged version of the selected regions in (c) and (f) , respectively.

where $n_{p_n}(x_1, y_1)$ represents the noise terms or crosstalks resulting from deciphering of the remaining images with incorrect keys. Fortunately, the proposed technique based on Eq. (6) can recover the encrypted images, $\hat{g}_n^p(x_1, y_1)$, with different spatial translations to artfully avoid the crosstalks $n_{p_n}(x_1, y_1)$.

Computer simulations are performed to verify our proposed method. Figure 3(a) shows nine original grayscale images of size 64×64 pixels. For different fractional Fourier transform order p_n based multiplexing, the order p_n are varied as $p_n = 0.5+0.25n$, where $n = 1,\ldots,9$. Figures 3(b) and 3(c) show the original $g_3(x_1, y_1)$ and the decrypted $\hat{g}_3^p(x_1, y_1)$ ($p_3 = 1.25$), respectively, and Figs. 3(e), (g) depict the original $g_6(x_1, y_1)$ and the decrypted $\hat{g}_6^p(x_1, y_1)$ ($p_6 = 2.0$), respectively. Comparing Fig. 3(b) with Fig. 3(d) (the enlarged version of one part in Fig. 3(c)), a correlation coefficient of ρ =0.95 is obtained. A similar performance can be achieved (ρ =0.94) for order $p_6 = 2.0$ multiplexing. The shifting amounts are designated to be $(\mu_n, \nu_n) = (aD, bD)$, where a and b are integers within the range [–3, 3] and $D = 64$. Figure 4 shows the comparison on the correlation coefficient between the original and the decrypted images for our proposed and the methods in Refs. [3, 4]. The proposed method evidently causes lower crosstalks (i.e., larger correlation coefficient) and hence achieves a higher storage capacity (i.e., larger N at a specified crosstalk).

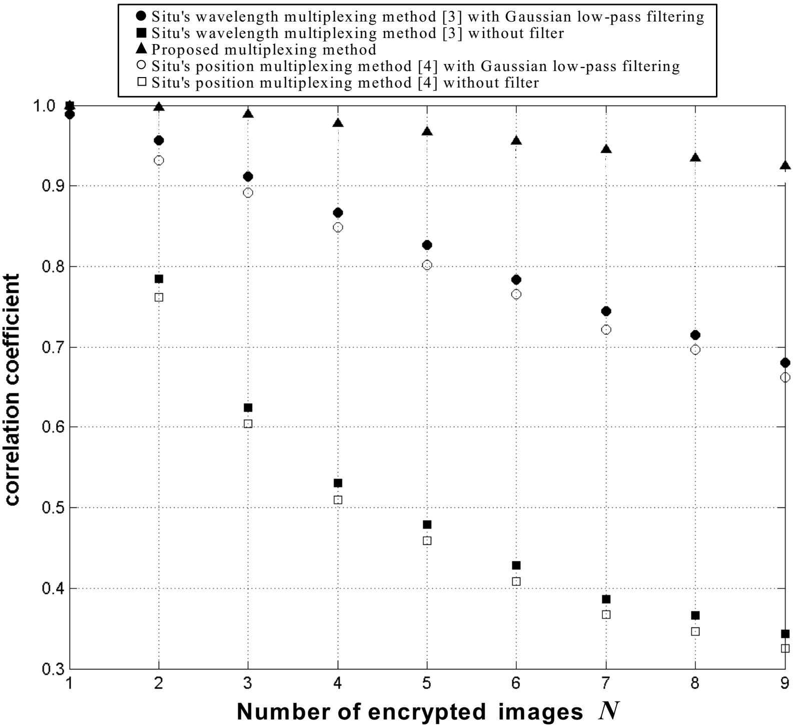


Fig. 4. Comparison of the proposed method with the Situ's methods [3, 4] in terms of correlation coefficient

In conclusion, our proposed method is new and efficient (low crosstalks) for multiplexing (encrypting) *N* images with only one POF (in contrast to traditional works which require two POFs). By the way, a lensless optical system based on FrFT is more compact, simpler and easier to implement owing to its minimization of the hardware requirement. Optical experiments will be soon conducted in our future research.

Acknowledgements

This research is supported in part by National Science Council under contract number NSC98-2221-E-235-002-MY2. Hone-Ene Hwang's e-mail address is n741@ms26.hinet.net.

References

1. Nomura, T., Mikan, S., Morimoto, Y., Javid, B.: Secure optical data storage with random phase key codes by use of a configuration of a joint transform correlator. Appl. Opt. 42, 1508–1514 (2003)
2. He, M.Z., Cai, L.Z., Liu, Q., Wang, X.C., Meng, X.F.: Multiple image encryption and watermarking by random phase matching. Opt. Commun. 247, 29–37 (2005)
3. Situ, G., Zhang, J.: Multiple-image encryption with wavelength multiplexing. Opt. Lett. 30, 1306–1308 (2005)
4. Situ, G., Zhang, J.: Position multiplexing for multiple-image encryption. J. Opt. A: Pure Appl. Opt. 8, 391–397 (2006)
5. Denz, C., Pauliat, G., Roosen, G., Tschudi, T.: Volume hologram multiplexing using a deterministic phase encoding method. Opt. Commun. 85, 171–176 (1991)
6. Heanue, J.F., Bashaw, M.C., Hesselink, L.: Encrypted holographic data storage based on orthogonal-phase-code multiplexing. Appl. Opt. 34, 6012–6015 (1995)
7. Taketomi, Y., Ford, J.E., Sasaki, H., Ma, J., Fainman, Y., Lee, S.H.: Incremental recording for photorefractive hologram multiplexing. Opt. Lett. 16, 1774–1776 (1991)
8. Zhang, X., Berger, G., Dietz, M., Denz, C.: Unitary matrices for phase-coded holographic memories. Opt. Lett. 31, 1047–1049 (2006)
9. Réfrégier, P., Javidi, B.: Optical image encryption based on input plane and Fourier plane random encoding. Opt. Lett. 20, 767–769 (1995)
10. Javidi, B., Zhang, G., Li, L.: Encrypted optical memory using double-random phase encoding. Appl. Opt. 36, 1054–1058 (1997)
11. Situ, G., Zhang, J.: A lensless optical security system based on computer-generated phase only masks. Opt. Commun. 232, 115–122 (2004)
12. Situ, G., Zhang, J.: Double random-phase encoding in the Fresnel domain. Opt. Lett. 29, 1584–1586 (2004)
13. Liu, Z., Liu, S.: Double image encryption based on iterative fractional Fourier transform: Opt. Comm. 272, 324–329 (2007)
14. Chen, L., Zhao, D.: Optical color image encryption by wavelength multiplexing and lensless Fresnel transform holograms. Opt. Express 14, 8552–8560 (2006)
15. Gerchberg, R.W., Saxton, W.O.: Phase determination for image and diffraction plane pictures in the electron microscope. Optik 34, 275–284 (1971)
16. Gerchberg, R.W., Saxton, W.O.: A practical algorithm for the determination of phase from image and diffraction plane pictures. Optik 35, 237–246 (1972)
17. Hwang, H.E., Chang, H.T., Lie, W.N.: Fast double-phase retrieval in Fresnel domain using modified Gerchberg-Saxton algorithm for lensless optical security systems. Opt. Express 17, 13700–13710 (2009)

Optical Image Encryption Based on Joint Fresnel Transform Correlator

Hsuan-Ting Chang* and Zong-Yu Wu

Photonics and Information Laboratory, Institute of Communications Engineering,
National Yunlin University of Science and Technology,
Douliu Yunlin, 64002 Taiwan R.O.C.

Abstract. Optical verification systems can make image hide in high-security phase key through the phase retrieval operatioñ Conventionally, optical verification systems based on the joint transform correlator usually use Fourier transform. In this paper we propose the joint Fresnel transform architecture and use the methods of the projection onto constraint set and nonlinear transform to determine the pure phase key. As shown in our simulation results, the advantages of the proposed method include that our system is lensless and the additional wavelength and distance parameters enhance the system security.

Keywords: Image encryption, Fresnel transform, phase retrieval, nonlinear transform.

1 Introduction

In current communication systems, cryptography [9-11] is an important issue because the information is easily attacked, stolen, or forgery. Encryption for information is indispensable. The developing of systems and technologies in optic can create more developmental and application. Compared to the previous digital encryption, optical encryption is faster, the computing speed equivalent to the light, and having characteristics of parallel processing in image signal. The other, the optical encryption can store in the form of phase and amplitude. In contrast, the optical encryption has a large space for development. The demand on the device, optical system needing more precise of instruments in encrypting and decrypt, so the invasion risk is reduced.

Previous joint transform correlators (JTCs) calculate far-field projection by using the optical Fourier transform (FT), then through CCD(Charge-Coupled Device) to receive image intensity, and implement IFT (Inverse Fourier Transform) to get target image. Reference to this framework, we propose FrT (Fresnel Transform) [14] to calculus near-field projection, thus can making lensless and increase distance parameters. In the latter segment of this system, we using POCS (Projection Onto Constraint Set) [2], [8], [12], [13] to iterate the reconstruct information to pure phase mask [3~7]. To make all of the operations based on pure phase, we using nonlinear transform and normalize to convert amplitude to phase. We reference to the iterative encrypt system

[*] Corresponding author. `htchang@yuntech.edu.tw`

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 81–89, 2010.

for one image and one random phase mask based on FrT, and two types of nonlinear conversion. According to the idea of using near-field projection and the algorithm of pure phase mask, we proposed the joint Fresnel transform system as shown in Fig. 1.

Section 2 is divided into optical iterative encryption and nonlinear transform, we will introduce the calculus process of POCS algorithm for enhance the quality of reconstruction, and the use of nonlinear transformation. The experiment of Section 3, we using image example to test this algorithm by MATLAB, and using the results of MSE (Mean square error) and CC (Correlation coefficient) to verify the quality of this algorithm. The final section, we referred to conclusions of this algorithm and the development of future applications.

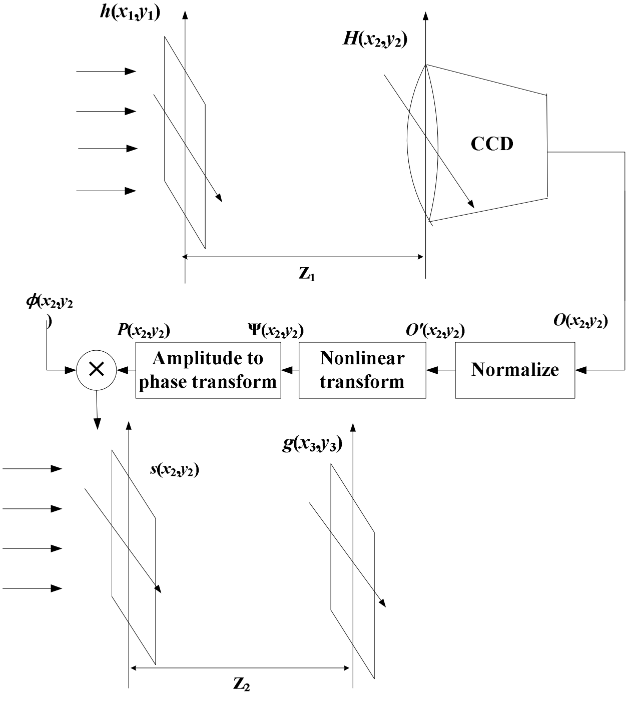


Fig. 1. The proposed joint Fresnel transform system for optical image encryption

2 Encryption algorithm

2.1 POCS

The flow chart of our system divided into the iteration part (red block) and the transform part (green block), which are shown in Fig. 2. In conversion operation of signals, we use the near-field projection for FrT.

$$\mathrm{FrT}\{H(x,y);z\} = \frac{e^{ikz}}{i\lambda z} \iint \mathrm{h}(x',y') e^{\frac{ik}{2z}\left[(x-x')^2+(y-y')^2\right]} dx'dy', \text{ where } k = \frac{2\pi}{\lambda}.$$

In the iteration part, we give an original image g and it is transformed with the FrT, the transform distance is z_2 (the block diagram is shown in Fig. 3), such as Eq. (1):

$$t(x_2,y_2)\exp[js(x_2,y_2)] = \mathrm{IFrT}\{g(x_3,y_3);z_2\} \tag{1}$$

t is the part of amplitude and s is the part of phase. We extract s to convert by FrT, such as Eq. (2):

$$g'(x_3,y_3)\exp[\,j\psi(x_3,y_3)] = \mathrm{FrT}\{\exp[\,js(x_2,y_2)];z_2\} \tag{2}$$

We make the amplitude g' to normalize, judge with g for pixel error. If some pixel error is greater than threshold that we enter, the pixel of g' can replace by the same position pixel of g. After the judge, we using the new image to multiplied by ψ, and continue to execute the next loop by back to Eq. (1), and so on.

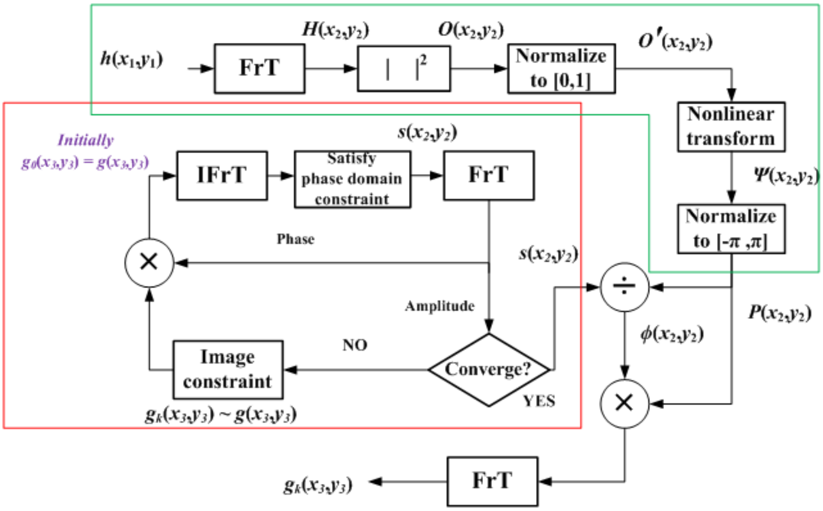


Fig. 2. The systematic block diagram of the proposed joint Fresnel transform system

2.2 Nonlinear Transform

In the nonlinear transform operation, we use to two common schemes:

(1) Power-law transform: $\psi\{x\} = y = x^b$

(2) Log-sigmoid transform: $\psi\{x\} = y = \frac{1}{1+e^{-ax+c}}$

In convert part, we using the random phase key h to convert by FrT is shown in Eq. (3):

$$H(x_2,y_2)\exp[\,j\psi_H(x_2,y_2)] = \mathrm{FrT}\{\exp[\,jh(x_1,y_1)];z_1\} \tag{3}$$

Then we use the part of amplitude H to calculate the signal intensity O, which is shown in Eq. (4):

$$O(x_2, y_2) = |\, H(x_2,y_2)\, |^2 \qquad (4)$$

Then we normalize O to [0, 1] to obtain O'. Finally, let O' convert to Ψ by using the nonlinear transform and normalize Ψ to $[-\pi, \pi]$. Then we store the result to P.

2.3 Combination Calculation

After the previous of pure phase s and the target image g', in order to combine the results, we divided s to P to obtain the phase mask φ is shown in Eq. (5):

$$\exp[\, j\phi(x_2, y_2)] = \frac{\exp[\, js(x_2, y_2)]}{\exp[\, jP(x_2, y_2)]} \qquad (5)$$

After the phase ϕ is obtained, that can make our operation s to achieve a coherent. Entering a Correct phase key get through calculus will obtain the target image g' in this system, we get keys of distance parameters z_1, z_2 and the modulation parameters used in the nonlinear transform.

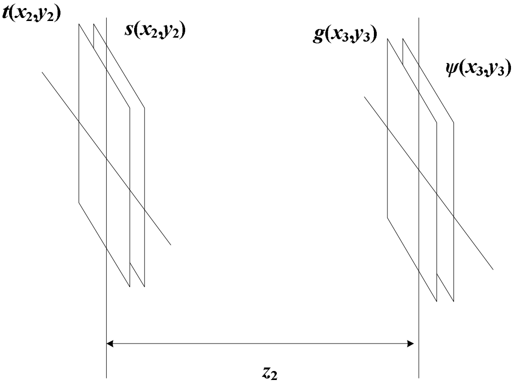


Fig. 3. Optical Fresnel transform between two planes

3 Simulation Results

The computer simulation is done by using MATLAB. The size of target image is 256 x 256, such as Fig. 4(a). The distance of FrT, z_2 = 20cm, wavelength λ = 632.8 nm. The iteration number is set to 100, the threshold value of MSE is set as five.

After the above calculation obtain phase mask s and approximated image g' is shown in Fig. 4(b) and 4(c), respectively. The CC values between g' and g is 0.9928. We use a random phase to be the input phase key, such as Fig. 5(a). Set the convert distance z_1 to 12 cm and the wavelength λ to 632.8 nm for FrT. After the signal intensity is obtained,

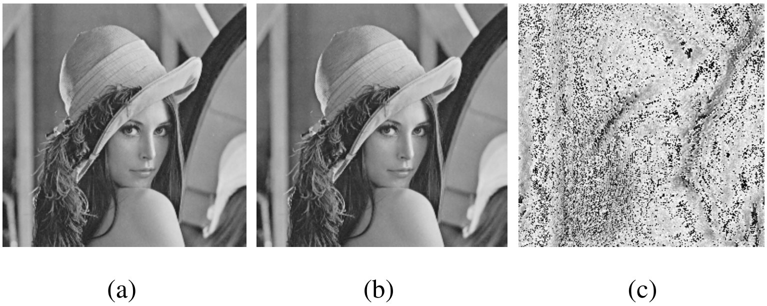


Fig. 4. (a) The target Len image; (b) the reconstructed image *g'*; (c) Phase key *s*

we use the Log-sigmoid transform as the nonlinear transformation, in which the parameters are a = -5, c = 10. After the processing of signal conversion, the pure phase P can be obtained, such as that shown in Fig. 5(b). Then following Eq. (5), we can obtain the other phase φ, which is shown in Fig. 5(c). After an overall operation, the obtained image is shown in Fig. 5(d) which is similar to that in Fig. 4(b).

Consider the sensitivity of different system parameters. Following the previous setting, if we enter the distance parameter $z_1 = 19$ cm, and 15 cm, the reconstructed images are shown in Figs. 6(a) and 6(b), respectively. Figure 7 shows the variation of

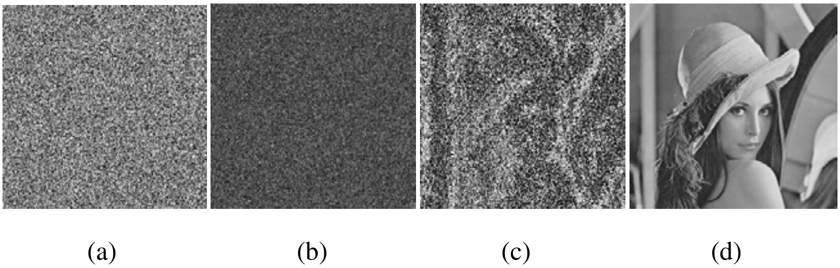


Fig. 5. (a) The random phase key; (b) Phase key P; (c) Phase key ϕ; (d) The verified image *g*'

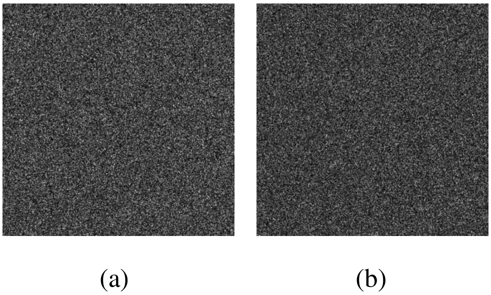


Fig. 6. The reconstructed images under incorrect distance parameters; (a) z_1 = 13cm; (b) z_1 = 17cm

CC values between the original image and reconstructed image for z_1 in the range from 1cm to 50 cm (Take a sample of each cm). If we enter the distance parameter $z_2 =$ 19 cm and 15 cm, the reconstructed images are shown in Figs. 8(a) and 8(b), respectively. We plot the variation of the CC values between the original image and reconstructed image for the z_2 parameter from 1 cm to 50 cm (Take a sample in each cm), which is shown in Fig. 9.

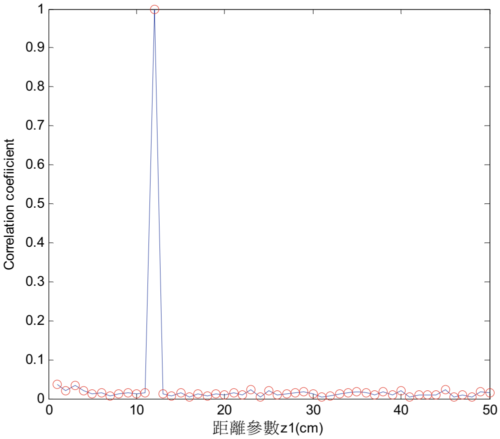
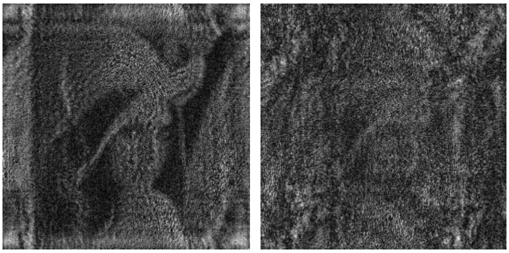


Fig. 7. Variation of the CC values between original image and reconstructed image for the z_1 parameter



(a) (b)

Fig. 8. The reconstructed images for the incorrect distance parameters: (a) z_2 = 19cm; (b) z_2 = 15cm

Considering the wavelength λ = 682.8nm, the reconstructed image is shown in Fig. 10(a). When the wavelength λ = 832.8nm, the reconstructed image is shown in Fig. 10(b). We plot the variation of the CC values between the original image and reconstructed image for the wavelength λ ranging from 332.8 nm to 832.8 nm (Take a sample in each 10nm), which is shown in Fig.11.

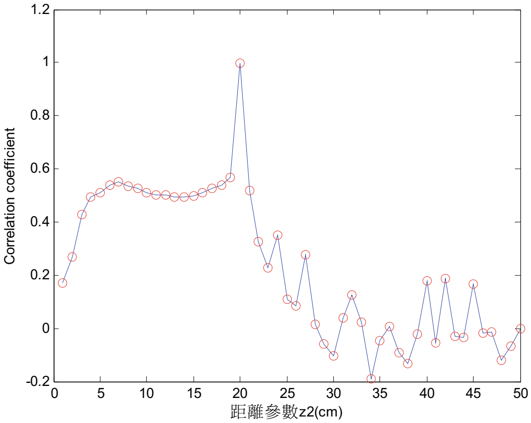


Fig. 9. Variation of the CC values between the original image and reconstructed image for the z_2 parameter

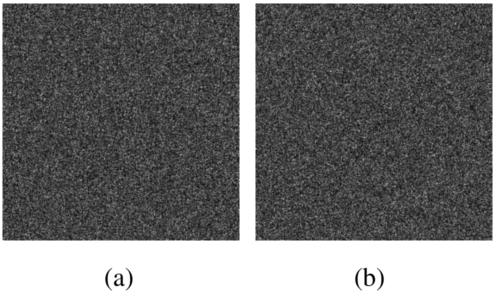


Fig. 10. The reconstructed images of two wavelengths: (a) λ = 682.8nm; (b) λ = 832.8nm

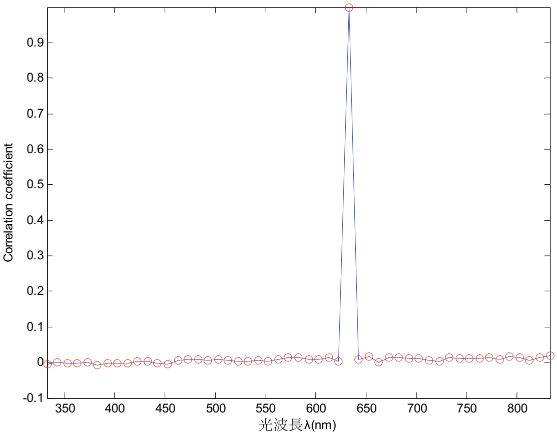


Fig. 11. Variation of the CC values between the original image and reconstructed image for the λ parameter

However, we found that the parameter can't use for all real numbers in the nonlinear transform. Inappropriate parameters will make the system wrong and operational failure, finally decrease the safety. Trough the computer simulation, we try to found the appropriate interval of parameters. We use the incorrect phase key to replace the phase key h_1 to test for Log-sigmoid transform. In processing of parameters, if a = 0, converted values will make no effects in system. We plot the varies of MSE and CC between original image and reconstructed image for a in -10 to 10 and c in -10 to 10, which are shown in Figs. 12 and 13, respectively.

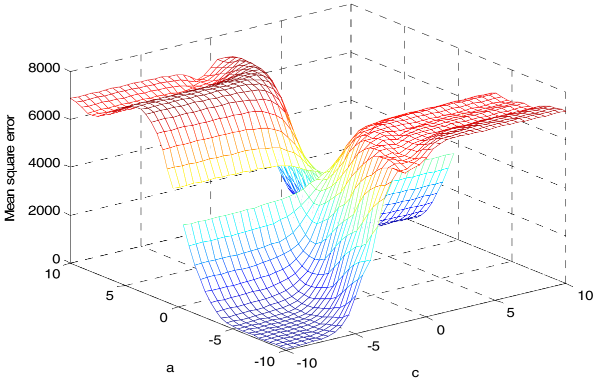


Fig. 12. Variation of the MSE values between original image and reconstructed image under the various nonlinear parameters a and c

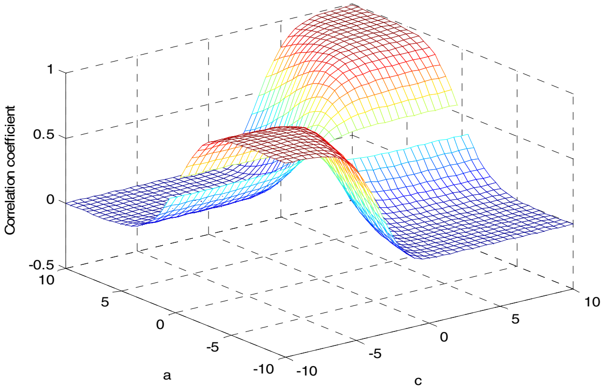


Fig. 13. Variation of the CC values between the original image and reconstructed image under the various nonlinear parameters a and c

4 Conclusion

In this paper, we proposed an optical encrypt system based on the joint Fresnel transform. The frequency part of an image is converted to a pure phase mask. Then, with the nonlinear transform methods, the target can be reconstructed by using two phase keys in the overall optical system. The proposed structure is simpler than that in conventional joint transform correlator architecture because of using no lenses in the proposed system. Moreover, the wavelength and the distances between two phases are two additional encryption parameters which can further enhance the system security level. The quality of system is estimated using the values of MSE and CC between the decrypted and original images.

Acknowledgment

This work is partially supported by National Science Council, Taiwan under the contract number NSC 99-2221-E-224-004.

References

[1] Rosen, J., Javidi, B.: Security optical systems based on a joint transform correlator with significant output images. Opt. Eng. 40(8), 1584–1589 (2001)

[2] Joseph, R.: Learning in correlators based on projections onto constraint sets. Opt. Lett 18(14), 2165–2171 (1993)

[3] Chang, H.T., Lu, W.C., Kuo, C.J.: Multiple-phase retrieval for optical security systems by use of random-phase encoding. Appl. Opt. 41(23), 4825–4834 (2002)

[4] Chang, H.T.: Image encryption using separable amplitude-based virtual image and iteratively retrieved phase information. Opt. Eng 40(10), 2165–2171 (2001)

[5] Refregier, P., Javidi, B.: Optical image encryption based on input plane and Fourier plane random encoding. Opt. Lett 20(7), 767–769 (1995)

[6] Towghi, N., Javidi, B., Luo, Z.: Fully phase encrypted image processor. J. Opt. Soc. Am. A 16(8), 1915–1927 (1999)

[7] Fienup, J.R.: Phase retrieval algorithm: a comparison. Appl. Opt. 22(15), 2758–2769 (1982)

[8] Sun, H., Kwok, W.: Concealment of damaged block transform coded images using projection onto convex sets. IEEE Transactions on Image Processing 4(4), 470–477 (1995)

[9] Chang, Y.C., Chang, H.T., Kuo, C.J.: Hybrid image cryptosystem based on dyadic phase displacement in the frequency domain. Optics Communications 236(4-6), 245–257 (2004)

[10] Yamamoto, H., Hayasaki, Y., Nichida, N.: Securing information display by use of visual cryptography. Optics Letters 28(17), 1564–1566 (2003)

[11] Chuang, C.H., Lin, G.S.: Optical image cryptosystem based on adaptive steganography. Optical Engineering 47(4), 470021–470029 (2008)

[12] Chang, H.T., Chen, C.C.: Fully phase asymmetric image verification system based on joint transform correlator. Optics Express 14(4), 1458–1467 (2006)

[13] Nomura, T., Javidi, B.: Optical encryption using a joint transform correlator architecture. Optical Engineering 39(8), 2031–2035 (2000)

[14] Chen, L., Zhao, D.: Optical color image encryption by wavelength multiplexing and lensless Fresnel transform holograms. Optics Express 14(19), 8552–8560 (2006)

Communication and Trust in the Bounded Confidence Model

M.J. Krawczyk[1], K. Malarz[1,⋆], R. Korff[2], and K. Kułakowski[1]

[1] Faculty of Physics and Applied Computer Science, AGH University of Science and Technology, al. Mickiewicza 30, PL-30059 Kraków, Poland

[2] Mainland Southeast Asian Studies, Universität Passau, Innstrasse 43, D-94032 Passau, Germany

Abstract. The communication process in a situation of emergency is discussed within the Scheff theory of shame and pride. The communication involves messages from media and from other persons. Three strategies are considered: selfish (to contact friends), collective (to join other people) and passive (to do nothing). We show that the pure selfish strategy cannot be evolutionarily stable. The main result is that the community structure is statistically meaningful only if the interpersonal communication is weak.

Keywords: mass opinion, computer simulations, social networks.

1 Introduction

Communication is a key process in social systems and it is of interest from various perspectives. For a political scientist the feedback between media and the audience in democratic systems determines the outcome of political decisions [1]. For a system-oriented sociologist, communication enables a social system to maintain its identity by reduction the complexity of its environment [2]. In between, in social psychology, people communicate to create meanings and these meanings determine human actions [3]. In psychology, communication is investigated in terms of transactional analysis [4], a Freudist scheme to decipher human attitudes. In mathematics, the physical concept of entropy was generalized within communication theory [5], which is at the foundations of computer science. In game theory, communication allows for new solutions in non-zero-sum games which yet remain unsolved [6]. It makes sense to look for interdisciplinary models of the communication processes, which could integrate at least some of mathematical formulations used in different fields.

In theory of Artificial Intelligence, Hebbian learning theory [7] and the neurological perspective [8] were used to equip simulated agents with human-like beliefs and emotions [9]. The outcome of this formulation was a set of differential equations, where the time-dependent variables were the level of belief, the

[⋆] Corresponding author, malarz@agh.edu.pl, http://home.agh.edu.pl/malarz/

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 90–99, 2010.

stimulus, the feeling and the preparation of the body state. The relations between these variables were assumed to depend on some additional parameters, as learning rate from feeling to belief etc. A detailed description of these models can be found in [9,10]. A similar model of dynamics of public opinion was formulated by Zaller [1] in terms of master equations for conditional probabilities. There, a set of independent agents was subject to a stream of messages coming from media. The variables used were the probabilities that messages are received and that they are accepted; additional parameters were introduced to describe the credibility of messages, the awareness on resistance to persuasion, the predisposition on resistance to persuasion etc. A concise description of this mathematical formulation can be found in Section II of [11]. The Zaller approach to theory of public opinion was reformulated in [11] to a geometrical scheme, similar to the bounded confidence model [12]. In this new scheme, an agent was represented by a set of messages he received in the past. The messages were expressed as points on a plane of issues, and the probability that a message was received by an agent was postulated to depend on the position of the point with respect to the messages received by him previously. Later, the communication between agents was added to the model [13].

The aim of this paper is to use the same geometrical model [11,13] to describe how trust emerges from beliefs of agents and the communication between them [9]. A new element here is the time dependence of the threshold value, which measures the trust between agents. We adopt the definition of trust formulated in [14]: "[illegible]". This formulation is close to the idea of conditional probability used by Zaller [1], that the information is accepted. Similar time-dependent couplings between the probabilities of different actions of agents were found to be efficient in modeling social systems [15]. Also, the influence of one agent on the action of another one is a direct reference of the sociological definition of power [16]; here we interpret it as an enhancement of a positive self-evaluation or the receiver by sending to him a message which he is willing to accept.

The sociological context we bear in mind is that some amount of agents is faced with an unexpected, dangerous situation. The degree of the danger can vary from a direct threat, as an evacuation of people from a building or a train after an explosion, to an anxiety of ship passengers of an unverified possibility of storm. In any case, agents can remain passive or they can act in a selfish way or they can join in a collective action. Each solution brings some risk and inconvenience, and the necessity to decide is itself an unwanted circumstance, which leads to stress and lowers the agents' self-estimation. We intend to discuss their dilemma in terms of the Scheff theory of pride and shame [17], which seems to provide a set of ideas which are particularly appropriate for our emergency scenario.

This text is organised as follows. In the next section we provide a short description of the emergency scenario in terms of the Scheff theory. Section 3 is devoted to a mathematical formulation of two models, which are used here to evaluate the efficiency of contacting with relatives or friends by phone; doing this is considered here to be the selfish strategy. In the same section we demonstrate numerical results, which state the conditions when this selfish strategy is not helpful. Section 4 defines the Zaller–Deffuant model of bounded confidence [12,13] with the time-dependent trust levels. The main result of this paper is the time dependence of the structure of communities of the evacuated people, described in Section 5. This dynamics is governed by the communication process. Last section is devoted to conclusions.

2 Scheff Theory of an Emergency Scenario

A catastrophe, a terrorist attack, or any other strong threat is obviously a strong disturbance of current plans of its participants, accompanied perhaps with a direct threat for their health or even life. As such, it must give raise to strong negative emotions, triggered by fear. Here we are interested in the communication between the participants. For each one, there are three options: i) to communicate by phone with friends or relatives, who do not participate in the situation, ii) to contact with other participants and in a search for a collective strategy, or iii) to withdraw and do nothing. These three strategies will be termed as S (selfish), C (collective) and P (passive), respectively. It might be surprising that during the 7/7 London bombings, the passive strategy was observed as a quite common [18].

According to the theory of pride and shame by Thomas J. Scheff [17], the strategy selected by the participants depends largely on conscious or unconscious messages which they send and receive. Namely, they try to evaluate how they are perceived by other participants: with respect or not. In the former case their positive self-estimation is enhanced, what enables the interpersonal attunement and the social solidarity. Consequently, they are willing to select the strategy C, what stabilizes the mutual respect. This cycle can be termed as the loop of pride. On the contrary, disrespectful signals lead to an enhancement of the negative self-evaluation, a shame and its repression and a hostility. As a consequence the lack of respect is stabilized and chances for C decrease; what is being selected is S or P. This behaviour can be termed as the loop of shame. The scheme of two loops in the Scheff theory was described in [19].

As we see, a positive feedback is present in both loops. However a switching is possible from the stage of shame in the loop of shame to the stage of an interpersonal attunement in the loop of pride [19]. This switching is possible by means of a realization of the shame. According to Scheff, it is an unconscious shame what breaks interpersonal bonds. Consequently, getting in contact with other participants of the scenario helps to realize, that the threat is common and the solidarity can be restored. Below we investigate the dynamics of the possible spread of this realization, from those who are already conscious to those who are not yet. Once they get into the loop of pride, they propagate this state further.

3 Models of a Phone Communication Jam

The selfish strategy S is natural in the case of large catastrophes, where everybody wants to get know if the relatives or friends remain alive. However, the payoff heavily depends on the amount of persons who select this strategy. It is straightforward to expect that once the amount of phone speakers excesses some percentage of the population, successive attempts to contact by phone must fail even if the phone network works properly. It is just almost impossible to have simultaneously more than one phone call. We demonstrate the effect by an evaluation of the percentage p_2 of agents who successfully talk to each other, and not only try to get connection; the probability of the latter state is denoted as p_1 from now on. These probabilities are to be calculated against the percentage p of people who try to get connection.

The calculation is performed within two different approaches. First one is a cellular automaton with parallel updating. A single variable $s_{i,j}$ is assigned to each site of the square lattice with the helical boundary conditions. The list of possible states of $s_{i,j}$ consists $0, L, R, U, D, L^*, R^*, U^*$ and D^*. Agents in the state '0' are silent. With the probability p an agent picks up his phone and starts to calling to his/her randomly selected neighbour. Agents in states L, R, D, U are trying to connect to his/her nearest neighbour situated on left, right, down or up, respectively. They stop unsuccessful attempts to connect with probability p'. Agents in states L^*, R^*, D^*, U^* talk with his/her left, right, down or upper neighbour, respectively. They finish their talk and return to the state '0' with the probability $\lambda = 1/\tau$. The probability p_1 is calculated as the amount of agents in one of the states L, R, D, U, while the probability p_2 is calculated as the amount of agents in one of the states L^*, R^*, D^*, U^*.

In the second approach, agents are distributed at nodes of Erdős–Rényi network, and the updating is sequential. The variables used $w_{i,j}$ deal with the states of bonds between agents; $w_{i,j} = 1$ means that i tries to phone to j, while $w_{i,j} = 0$ means that he does not try. The probability p_2 is found as the percentage of agents involved in mutual talks. The probability p_1 is the amount of agents i such that $w_{i,j} = 1$, and $w_{j,i} = 0$. The simulation goes as follows. Once i is selected randomly with repetitions, we ask if there is a neighbour $j(i)$ such that $w_{i,j} = 1$. If yes, we check if $w_{j,i} = 1$. If yes, i and j are talking; the talk is broken with probability λ. Coming back, if $w_{j,i} = 0$, we check if there is any neighbour $k(j)$ such that $w_{j,k} = 1$. If not, $w_{j,i}$ is set to one; this means that j starts to talk with i. If yes, $w_{i,j}$ is set to zero; then i selects randomly one of his neighbours n and tries to connect him with probability p; again $w_{i,n}$ is set to one. Next possibility about i is that $w_{i,j} = 0$ for all his neighbours $j(i)$. Then, if any $w_{j,i} = 1$, $w_{i,j}$ is set to one. If $w_{j,i} = 0$ for all i's neighbours $j(i)$, again i tries to communicate with one of his neighbours.

These approaches are very different in details. In the cellular automaton we use von Neumann neighborhoods of four cells, and the topology is just a plane. On the contrary, in the case of the Erdős–Rényi network the topology is random with the small world property. Also, in the case of network unsuccessful calls initiated in one step are stopped only after second selection of the same node.

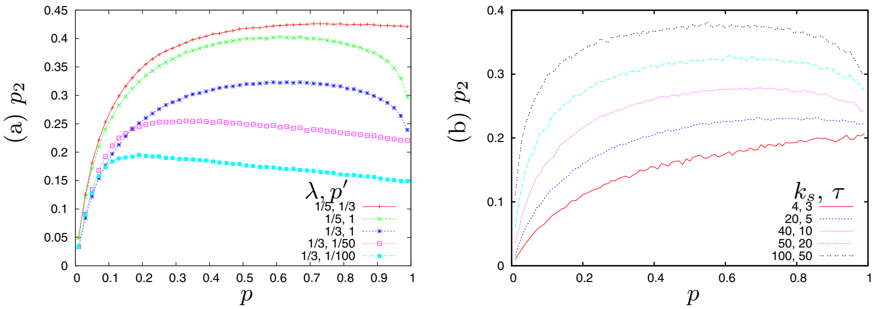


Fig. 1. The probability of a successful talk $p_2(p)$ against the probability p of intention to talk, obtained within (a) the cellular automaton and (b) network approach. (a) The size of the lattice is $L \times L$, where $L = 10^3$. The parameters p' and λ are the probabilities of stopping unsuccessful and successful connection, respectively. (b) The size of the network is $N = 100$ except the lowest curve, where $N = 1000$. The results are the same for larger N, except that the fluctuations decrease. The parameter k_s is the mean degree of the Erdős–Rényi network, and the parameter τ is the mean time of successful talks.

Still, as we see in Figs. 1(a) and 1(b), the result is qualitatively the same: in both models the curves $p_2(p)$ do not increase above some value p^* of p. As this result is obtained within two entirely different models, it can be considered as validated. Although we do not specify the payoffs, it is clear that the strategy S can be efficient only if it is chosen by a minority. This effect allows to expect that other strategies, P or C, can be active.

4 Bounded Confidence Model

The model of communication used here is a slightly modified version of the Zaller–Deffuant model [13]. There are two kinds of messages: those from media and those from the agents themselves. All messages are represented by points on a plane. A new message is accepted by an agent if the position of the message is not too far from the messages accepted by this agent in the past. This means that the system is characterized by a critical distance μ in the message space; ‘not too far’ means ‘the distance is shorter, than μ’. In the original formulation of the bounded confidence model [12], the parameter μ meant the tolerance for distant opinions. We are willing to maintain this meaning of μ in this formulation.

The system is subject to a stream of messages from media. Besides of that, each agent reproduces one of messages accepted by him previously; he does so with the probability r, which we select to be $r = 10/N$, where N is the number of agents [13]. The modifications of the model with respect to the original version [13] are, that.) we do not distinguish between the messages received and accepted – this can be done, because we do not discuss final decisions of the agents;..) agents send particular messages, and not their time averages;...) what is most important, the value of the critical distance $a_{i,j}$ depends both on the sender j and on the receiver i and it varies in time; in this way we formalize the definition of trust, given in [14]. On the contrary to this variation, the critical distance μ for the messages sent by media remains constant.

The time dependence of $a_{i,j}$ is determined by the following rules. The initial values of $a_{i,j}$ are equal to μ. Once a message from j is accepted by i, $a_{i,j}$ is transformed to $a_{i,j}/2 + \mu$. Once a message is not accepted, $a_{i,j}$ is transformed into $a_{i,j}/2$. In this way, the matrix elements of trust between agents vary between zero and 2μ. In this variation, more recent messages matter more; after three subsequently accepted messages, the respective matrix element $a_{i,j}$ increases from zero to 0.875 of the maximal possible value. In other words, the memory of the system remains finite. These rules were found to be useful in cooperation modeling [15].

The message positions are limited to a square on a plane (x, y), where both coordinates vary between -1 and $+1$. Actually, the critical distance is defined with respect to the size of this square. After a sufficiently long time, each agent is going to accept each message [11,13]; this is assured by the increase of the area around the accepted messages from media. Here we are not interested in the asymptotic regime, but rather in the transient process of filling the square of particular agents by accepted messages. The role of the interpersonal communication in this process is encrypted in the time dependence of the trust matrix $a_{i,j}$. At the asymptotic stage, i.e. for time long enough, $a_{i,j} = 2\mu$ for all i, j. In the transient time, the role of the communication between agents and the role of the messages from media can depend on the value of the threshold μ. We are interested in the structure of communities of agents where the mutual trust is established. These results are described in the next section.

5 Communities

The plot shown in Fig. 2 shows the time dependence of the mean value $\langle a_{i,j} \rangle$ of the trust matrix elements for different values of the trust parameter μ. The diagonal matrix elements are excluded. Although there is no clear difference between neighboring curves, two different regimes can be observed. For $\mu \geq 0.7$, the obtained curve increases almost monotoneously with time. On the contrary, for μ close to 0.2 and nearby, the curve first decreases, later increases to the value 2μ. The decrease mark the overall fall of interpersonal trust. At this stage, the process of an increase of individual areas around accepted messages is due mostly to the messages from media. The difference is then between the communication dominated by the interpersonal messages (large μ) and the one dominated by media (small μ).

The structure of the communities of mutual trust is investigated by means of two algorithms, both designed as to identify communities in networks [20]. Here the network nodes are represented by agents, and the weights of the bonds – by the matrix elements $a_{i,j}$. First algorithm is the agglomerative hierarchical clustering method proposed by Mark Newman [21]; it can be applied to symmetric as

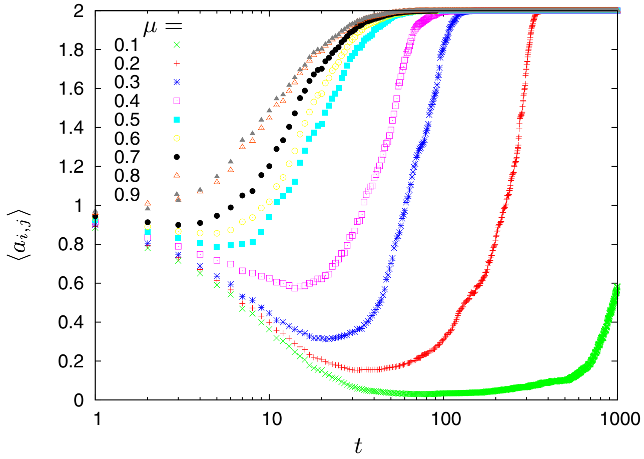


Fig. 2. The time dependence of the mean value $\langle a_{i,j} \rangle$ of the trust matrix element normalized to the range 2μ for different values of the parameter μ

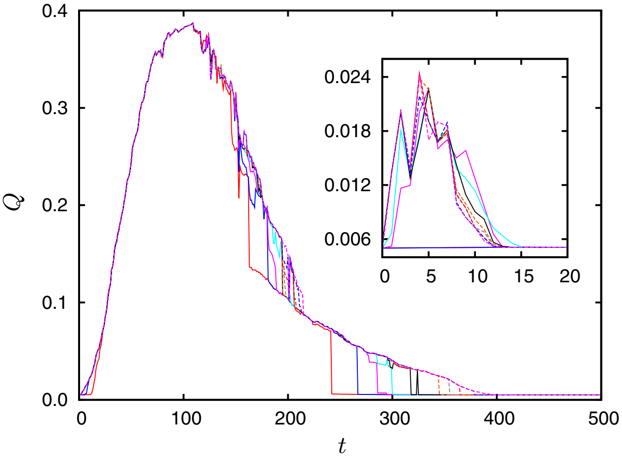


Fig. 3. The time dependence of the modularity Q for $\mu = 0.2$ and different values of the parameter β according to the method of differential equations [23]. In the inset the same plot for $\mu = 0.7$.

well as to non-symmetric matrices. Second algorithm relies on a numerical solution of a set of nonlinear differential equations [22], where the unknown variables are the same matrix elements $A_{i,j} = a_{i,j}/(2\mu)$. The equations are

$$\frac{dA_{j,k}}{dt} = \Theta(A_{j,k})\Theta(1 - A_{j,k}) \sum_i (A_{j,i}A_{i,k} - \beta) , \tag{1}$$

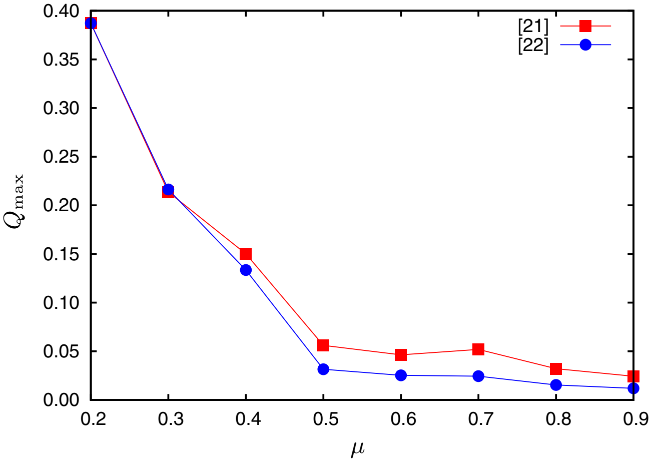


Fig. 4. The maximal value of the time dependence of the modularity $Q(t)$ against the tolerance parameter μ according to the Newman algorithm [21] (red curve) and the algorithm of differential equations [22] (blue curve)

where $\Theta(x)$ is the step function and β is a model parameter, which measures the threshold, above which the product of the matrix elements starts to increase. This method works on symmetric matrices; to apply it, we have to symmetrize the trust matrix $a_{i,j}$. Both methods make use of the modularity Q, defined as [23]

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{i,j} - \frac{k_i k_j}{2m} \right) \delta_{c_i, c_j} , \tag{2}$$

where $\delta_{i,j}$ is the symbol of the Kronecker delta, k_i is the weighted degree of node i and m is the total number of edges of the network. The search of the maximal value of the modularity allows to find the optimal structure of communities. Simultaneously, the value of Q allows to evaluate the statistical meaningfulness of the obtained structure. A large value of Q (about 0.3 or more) means that the structure differs remarkably from a random one.

When Q is small, as for example at the beginning of the simulation, the method of differential equations [22] produces one connected cluster. However, when Q is large, both applied methods give almost the same communities. A brief inspection of the obtained data allows to state that once a community appears, it persists, with some new nodes being attached during the process. In our sample of 100 nodes, new communities of 4-9 nodes are born by a separation from the whole mass, and even if modified, they can be identified. However, the most important result is not a particular structure, but its meaningfulness, measured by Q. We observe that i) the time dependence of the modularity Q displays a strong maximum, ii) its maximal value is meaningful (in the range $0.2 < Q < 0.4$) only for small value of μ. These results are shown in Figs. 3 and 4. They mean that the community structure undoubtely appears only if the interpersonal communication is weak.

6 Conclusions

Our calculations are related to the problem of an emergency scenario, when individuals select one of three strategies: a selfish communication with persons not involved into the emergency situation, a passivity, or a collective action. We used two simple models to demonstrate, that the selfish strategy cannot be rewarding, if applied by the majority. Next we concentrated on the community structure, which appears when individuals communicate to join in a collective action. The communication is modeled with using the bounded confidence theory, where the parameter of tolerance measures the individual ability to accept messages sent by other persons.

We found a qualitative difference between the communication dynamics for small and large tolerance μ. Loosely speaking, a small value of μ in our model means that people are willing to accept only these messages which are directly close to their own opinions. As shown in Fig. 2, in these conditions the mutual trust falls quickly, and it is only selected messages from media which are accepted. Once the area around a small number of accepted messages starts to widen, some neighbours can be found in a direct neighbourhood. The reconstructed trust refers only to a few neighbours, and it is strengthened often by a repetition of mutually copied messages. On the contrary, for large tolerance μ many interpersonal messages are accepted immediately, and the initial weak fall of the mutual trust is followed by its fast increase, as shown in Fig. 2. In these conditions the connections between agents include the large majority to the same cluster; as they are equally strong, communities are practically absent.

Drawing more from sociology, these conclusions can be supplemented by a note on theory of attributions [19]. People are likely to infer on the causes of events which happen to them. There are two kinds of attributions: internal, when we look for causes in our individual personalities, dispositions and attitudes, and external, when we identify causes in an external world. More than often we apply the former to our successes and the latter – to our defeats [24]. In a society divided into small communities without a communication between them, these negative external attributions are strengthened even more. In this case the collective action of groups can be mutually hostile [25].

Acknowledgements. One of the authors (K.K.) is grateful to Alexei Sharpanskykh for useful discussions. The research is partially supported within the FP7 project No. 231288 (Socionical).

References

1. Zaller, J.R.: The Nature and Origins of Mass Opinion. Cambridge UP, Cambridge (1992)
2. Luhmann, N.: A Sociological Theory of Law. Routledge, London (1985)
3. Charon, J.M.: Symbolic Interactionism: an Introduction, an Interpretation, an Integration. Pearson, Prentice Hall, New Jersey (2010)

4. Stewart, I., Joines, V.: Transactional Analysis Today, A New Introduction to Transactional Analysis. Lifespace Publ., Nottingham (2000)
5. Shannon, C.E.: A Mathematical Theory of Communication. Bell Syst. Tech. J. 27, 379–423, 623-656 (1948)
6. Straffin, P.D.: Game Theory and Strategy. Mathematical Association of America, Washington (1993)
7. Hebb, D.O.: The Organisation of Behavior. Wiley, New York (1949)
8. Damasio, A.: Descartes' Error: Emotion, Reason, and the Human Brain. Putnam, New York (1994)
9. Memon, Z.A., Treur, J.: Modeling the Reciprocal Interaction between Believing and Feeling from a Neurological Perspective. In: Zhong, N., Li, K., Lu, S., Chen, L. (eds.) BI 2009. Lecture Notes in Computer Science, LNAI, vol. 5819, pp. 13–24. Springer, Heidelberg (2009)
10. Hoogendoorn, M., Jaffry, S.W., Treur, J.: An Adaptive Agent Model Estimating Human Trust in Information Sources. In: Baeza-Yates, R., et al. (eds.) Proceedings of the 9th IEEE/WIC/ACM Int. Conf. on Intelligent Agent Technology of IAT 2009, pp. 458–465. IEEE Computer Society Press, Los Alamitos (2009)
11. Kułakowski, K.: Opinion Polarization in the Receipt-Accept-Sample Model. Physica A 388, 469–476 (2009)
12. Deffuant, G., Neau, D., Amblard, F., Weisbuch, G.: Mixing Beliefs among Interacting Agents. Adv. Compl. Sys. 3, 87–98 (2000)
13. Malarz, K., Gronek, P., Kułakowski, K.: Zaller–Deffuant Model of Mass Opinion, arXiv:0908.2519v5 [physics.soc-ph]
14. Sharpanskykh, A.: Integrated Modeling of Cognitive Agents in Socio-Technical Systems. Presented at KES AMSTA (2010)
15. Kułakowski, K., Gawroński, P.: To Cooperate or to Defect? Altruism and Reputation. Physica A 388, 3581–3584 (2009)
16. Weber, M.: Wirtschaft und Gesellschaft. Mohr, J.C.B (Paul Siebeck), Tübingen (1972)
17. Scheff, T.J.: Microsociology. Discourse, Emotion and Social Structure. The University of Chicago Press, Chicago (1990)
18. Report of the 7 July Review Committee of the London Assembly, vol. 3: Views and Information from Individuals[1]
19. Turner, J.H., Stets, J.E.: The Sociology of Emotions. Cambridge UP, Cambridge (2005)
20. Fortunato, S.: Community Detection in Graphs. Phys. Rep. 486, 75–174 (2010)
21. Newman, M.E.J.: Fast Algorithm for Detecting Community Structure in Networks. Phys. Rev. E 69, 066133 (2004)
22. Krawczyk, M.J.: Differential Equations as a Tool for Community Identification. Phys. Rev. E 77, 65701(R) (2008)
23. Leicht, E.A., Newman, M.E.J.: Community Structure in Directed Networks. Phys. Rev. Lett. 100, 118703 (2008)
24. Turner, J.H.: Face-to-Face: Towards a Sociological Theory of Interpersonal Behavior. Stanford UP, Stanford (2002)
25. Kułakowski, K., Krawczyk, M.J., Gawroński, P.: Hate—no Choice. Agent Simulations. In: Lockhardt, C.T. (ed.) Psychology of Hate, Nova Publ., New York (in print, 2010)

[1] `http://legacy.london.gov.uk/assembly/reports/7july/vol3-individuals.pdf`

Detection of Image Region-Duplication with Rotation and Scaling Tolerance

Qiumin Wu, Shuozhong Wang, and Xinpeng Zhang

School of Communication and Information Engineering, Shanghai University,
Shanghai 200072, China
{rynax_ls,shuowang,xzhang}@shu.edu.cn

Abstract. Copy-move forgery, or region-duplication, is a common type of digital image tampering. This paper presents a novel approach to detect copy-move forgery even if rotation and/or scaling took place before copying. The image under test is divided into fixed-size blocks, and log-polar Fourier transform (LPFT) is performed on the inscribed circles of these blocks. Similarity in the LPFT between different patches provides an indication of the copy-move operation. Experimental results show efficacy of the proposed method.

Keywords: digital forensics, copy-move forgery, log-polar transform, rotation and scaling invariance.

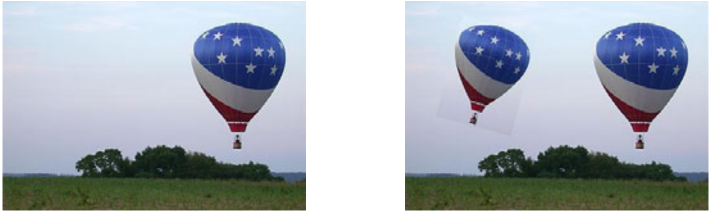
1 Introduction

With the proliferation of sophisticated image-editing software, digital images can be easily manipulated without leaving any perceptible traces. Although digital watermarking can be used in image authentication, application of watermarking is limited as a watermark must be embedded into the image before any tampering occurs. To this end, passive image forensic techniques have been developed for image authentication without the need of any precautious measures, and therefore are more flexible and practical. These techniques work on the assumption that although image forgeries may leave no visual clues of tampering, they alter the underlying statistics of an image.

Copy-move is a common type of image manipulation, which copies one part of the image and pastes into another in the same image. Detection of copy-move forgery is based on the presence of duplicated regions in the tampered image even though the image has undergone post-processing operations such as smoothing and local modifications. Exhaustive search is a straightforward way to detect copy-move, but it has very high computation complexity and is unable to deal with situations where the pasted object has been rescaled and/or rotated.

To reduce computational complexity, research has been done to use block-matching. Some related methods are based on lexicographic sorting of quantized DCT coefficients of image block [1] or PCA results of fix-sized image blocks [2] or using blur moment invariants as eigenvalues of image blocks [3]. In practice, copy-move forgery is often accompanied by scaling and rotation. For example,

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 100–108, 2010.



(a) (b)

Fig. 1. Copy-move with scaling and rotation (a) original image, and (b) forged image

the balloon in Fig. 1(a) is rotated by 15° and scaled by a factor of 0.8 before pasting into the same image to produce the result of Fig. 1(b). Features used in some previous methods such as DCT coefficients and eigenvectors of PCA bases are changed when the duplicated region is scaled/rotated, and therefore the above methods will fail in these cases.

To solve the problem, log-polar transform (LPT) may be performed on image blocks followed by wavelet decomposition [4] or lexicographic sorting [5]. Consideration of LPT comes from its scaling/rotation invariance and its successful use in other fields such as watermarking [6] and image registration [7]. Bayram et al. [8] proposed an approach based on Fourier-Mellin transform (FMT), in which each block is first Fourier transformed, and the resulting magnitude values are then mapped into log-polar coordinates. FMT values are finally projected to one dimension to give the feature vectors. Although these methods can detect copy-move forgery with scaling, they do not work well when large angle rotation of the pasted object is involved. For example, the upper limit of rotation is 10° in the method of [8].

In this paper, we propose an approach to detect copy-move image forgery with rotation/scaling tolerance. The method is based on the log-polar Fourier transform (LPFT). The image to be tested is divided into fixed-size blocks, and the inscribed circles of the blocks are taken to give circular regions of the same size. Unlike the method introduced in [8] where log-polar mapping is done in the Fourier domain of the image, we first perform a log-polar transform (LPT) on every circular region, which is in the image domain, and then take 2D Fourier transform of the LPT result. Similarity between the original and forged regions is revealed by comparing the cross-spectra of the LPFT results. Experimental results show that the proposed method is effective even if scaling and large-angle rotation occurred in the tampered regions.

2 Copy-Move Forgery Detection Using LPFT

As stated in the above, LPFT consists of two steps: LPT and FT. The purpose of LPT is to convert image scaling and rotation into translation of the log-polar representation. We identify tampered regions by comparing cross-spectrum coefficients of the LPFT magnitude spectra.

2.1 Property of LPFT

The log-polar transform is rotation and scaling invariant as rotation and scaling in the image domain corresponds to translational shift in the LPT domain. Denoting the origin in the image domain (x_0, y_0) and the coordinates of any pixel (x, y), we have the log-polar coordinates:

$$\rho = \log \sqrt{(x - x_0)^2 + (y - y_0)^2} \tag{1}$$

$$\theta = \arctan\left(\frac{y - y_0}{x - x_0}\right), \text{when } x \neq x_0 \tag{2}$$

where ρ is logarithm of the radial distance from the origin and θ the polar angle. Essentially, by applying a log-polar transform to an image, concentric circles in the image are mapped to parallel lines in the LPT domain. We will use this property to detect copy-move forgery.

Consider a source image **S** and its rotated-scaled replica **R** with a rotation angle θ_0 and scaling factor λ_0. Let f and g represent luminance values in **S** and **R** respectively. Thus,

$$f\left(x_{\mathrm{R}}, y_{\mathrm{R}}\right) = g\left(x_{\mathrm{R}}, y_{\mathrm{R}}\right) \tag{3}$$

where pixel locations in the two domains have the following relation:

$$\begin{pmatrix} x_{\mathrm{R}} \\ y_{\mathrm{R}} \end{pmatrix} = \lambda_0 \begin{pmatrix} \cos\theta_0 & -\sin\theta_0 \\ \sin\theta_0 & \cos\theta_0 \end{pmatrix} \begin{pmatrix} x_{\mathrm{S}} \\ y_{\mathrm{S}} \end{pmatrix} \tag{4}$$

Now we compute LPT of **S** and **R** to produce $\mathbf{S}_{\mathrm{LP}}$ and $\mathbf{R}_{\mathrm{LP}}$ with pixels denoted, for simplicity, by $S(\rho_{\mathrm{S}}, \theta_{\mathrm{S}})$ and $R(\rho_{\mathrm{R}}, \theta_{\mathrm{R}})$ respectively. From the definition of LPT, we can easily obtain:

$$\begin{cases} \rho_{\mathrm{R}} = \rho_{\mathrm{S}} + \log\lambda_0 \\ \theta_{\mathrm{R}} = (\theta_{\mathrm{S}} + \theta_0) \bmod 2\pi \end{cases} \tag{5}$$

Therefore, in the log-polar coordinates, the relation between $S(\rho_{\mathrm{S}}, \theta_{\mathrm{S}})$ and $R(\rho_{\mathrm{R}}, \theta_{\mathrm{R}})$ can be established:

$$R\left(\rho_{\mathrm{R}}, \theta_{\mathrm{R}}\right) = S\left(\rho_{\mathrm{S}} + \log\lambda_0, \theta_{\mathrm{R}} + \theta_0\right) \tag{6}$$

Eq. (6) indicates that **R** is a translated replica of **S** in the log-polar domain. Fig. 2 shows an example in which rotation and scaling in the Cartesian coordinates are converted into translation in the LPT domain. The abscissa and ordinate in the log-polar coordinates correspond to θ and ρ, respectively. The original image in Fig. 2(a) is rotated by 37° , and scaled by a factor of 1.2 and then cropped to keep the same size, see Fig. 2(b). Their LPT versions are shown in Figs. 2(d) and 2(e). The two LPT versions differ by a translational shift. Note that, as the image is enlarged, a part of the image in the log-polar domain is moved out of the display area and another part moved in. Fig. 2(c) is another version of

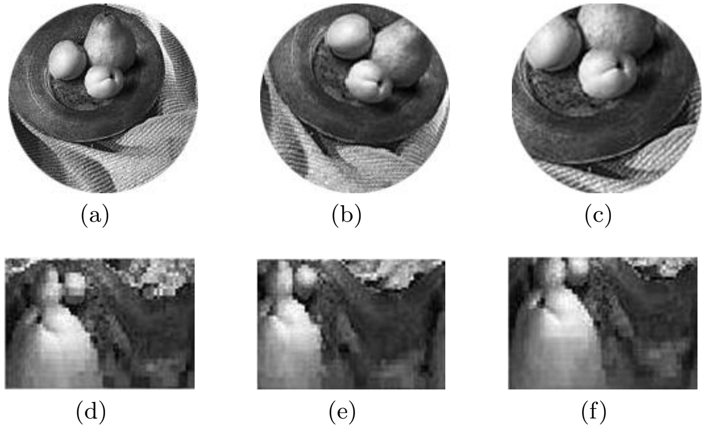


Fig. 2. Image rotation and scaling: (a) original, (b) rotated by 37° and scaled by 1.2, (c) scaled by 1.5, (d) LPT of the original, (e) LPT of the rotated-scaled image, and (f) LPT of the image scaled by 1.5

Fig. 2(a) with a scaling factor 1.5 and the same size cut, and its LPT result is shown in Fig. 2(f). Compared with Fig. 2(d), almost one-third of the content of Fig. 2(f) has changed. Scaling with a larger factor, e.g., greater than 1.5, introduces too much change of the image content in the LPT domain, therefore is not considered in the present work.

Take Fourier transform of both sides of Eq. (6):

$$F_{\mathrm{R}}(u,v) = F_{\mathrm{S}}(u,v)\exp\left[2\pi j\left(u\log\lambda_0 + v\theta_0\right)\right] \tag{7}$$

and define the normalized cross-spectrum of $\;_{\mathrm{R}}$ and $\;_{\mathrm{S}}$ as:

$$G(u,v) = \frac{\;_{\mathrm{R}}(u,v)\,F_{\mathrm{S}}^{*}(u,v)}{\left|\;_{\mathrm{R}}(u,v)\,F_{\mathrm{S}}^{*}(u,v)\right|} \tag{8}$$

The asterisk indicates complex conjugate. Since $|\;_{\mathrm{R}}| = |\;_{\mathrm{S}}|$ as can be seen from (7), $\;(u,v)$ is a two dimensional complex sinusoid:

$$G(u,v) = \exp\left[-2\pi j\left(u\log\lambda_0 + v\theta_0\right)\right] \tag{9}$$

We know that Fourier transform of a sinusoidal function is a delta function. Therefore, a peak would be present in the Fourier transform, $\;_{G}(u,v)$, if **R** is a rotated-scaled replica of **S**. For example, Fig. 3 shows $\;_{G}(u,v)$ of Figs. 2(a) and 2(b).

2.2 Copy-Move Forgery Detection

We confine our discussion to gray scale images in this work. For color images, take the luminance component of the image's YCbCr representation. The proposed forgery detection steps are:

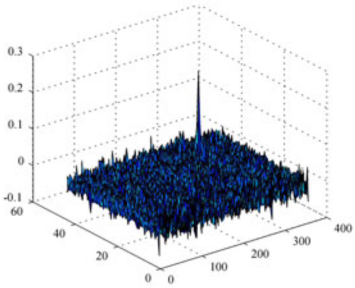


Fig. 3. $F_G(u,v)$ of the two images in Figs. 2(a) and 2(b), x and y axes correspond to u and v respectively, and z axis represents the magnitude value of $F_G(u,v)$

1. Divide the $m \times n$ image $\mathbf{I}$ such as the one shown in Fig. 4(a) into square blocks sized $2r$-by-$2r$, and take the inscribed circle of each block. Denote the circular region as $S(i,j;x,y,z)$ where (x,y) is the center, and r the radius:

$$S\left(i,j;x,y,r\right) = I(x-r+j, y-r+j),\ \sqrt{\left(x-x_0\right)^2+\left(y-y_0\right)^2} \le r \qquad (10)$$

The ranges of x and y are

$$\begin{cases} r \le x \le m-r \\ r \le y \le n-r \end{cases} \qquad (11)$$

Since x and y have $m-2r+1$ and $n-2r+1$ different values respectively, there are a total of $K=(m-2r+1)(n-2r+1)$ circles as sketched in Fig. 4(b).

2. Compute LPT of each circular region, and Fourier-transform the LPT result. The K results of LPFT are denoted as $\{\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_K\}$, see examples in Figs. 4(c)-(f).

3. Compute normalized cross-spectrum $G(u,v)$ between each pair of $\mathbf{V}_k$, and find the maximum, resulting in a total of $L = K \times (K-1)/2$ maximal values $\{p_1, p_2, \ldots, p_L\}$, each corresponding to a pair of circles in $\mathbf{I}$.

4. Compare the members in $\{p_1, p_2, \ldots, p_L\}$ with a threshold T chosen as the average absolute value of $F_G(u,v)$. If $p_l(l=1,2,\ldots,L)$ is greater than T, the two corresponding image blocks in $\mathbf{I}$ are regarded as belonging to copy-move regions.

The best choice of radius r of the circular region is related to the size of tampered region. Without any [illegible] knowledge of the image forgery, we can choose the r value according to the image size. In this work, test images are 512×512 or smaller. We let radius r be 10 pixels that is a compromise between false alarm and correct detection. Detection of any two image regions with overlapped parts is excluded to avoid false alarm caused by repeated detection of the same part in different image regions.

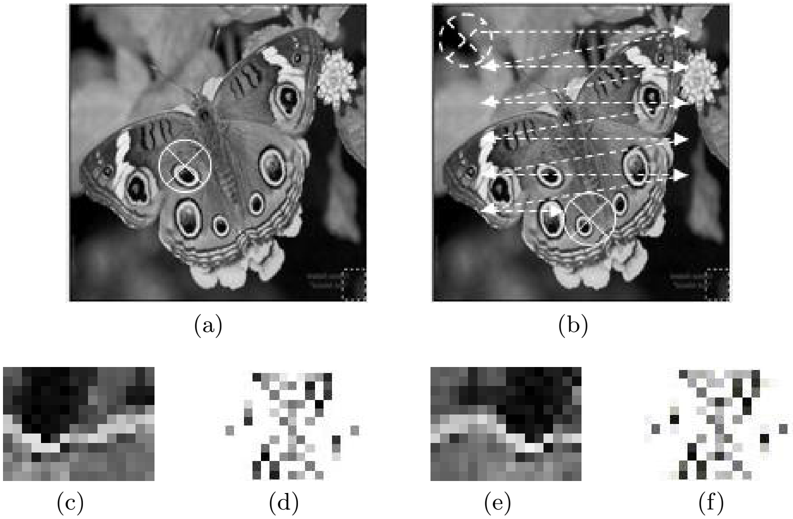
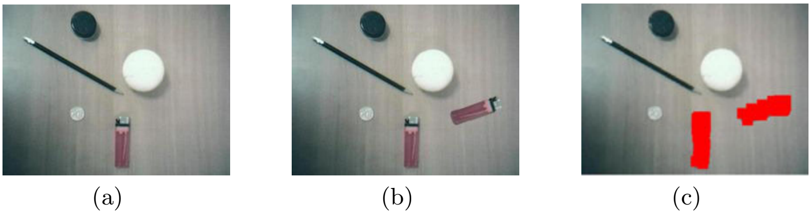


Fig. 4. Circular regions in the test image and the LPFT: (a) a circle in the image, (b) scanning over the circular regions, (c) LPT of one circle (source), (d) magnitude of LPFT of the source circle, (e) LPT of the rotated-scaled replica, and (f) magnitude of LPFT of the replica

3 Experimental Results

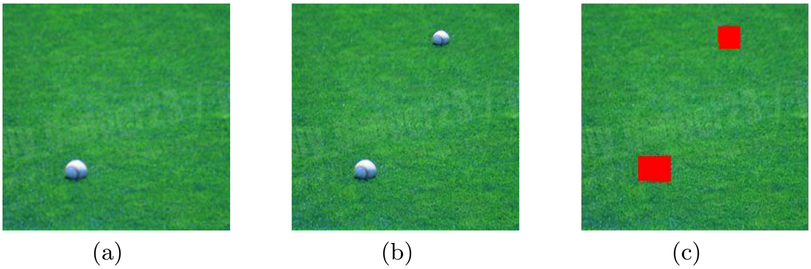
From the above discussion, there is a peak in $F_G(u, v)$ of two image regions when one is a replica of the other, as shown in Fig. 3. To set the threshold T, we perform a series of experiments for different T values from 0.1 to 0.5 with an increment 0.05. The optimal value turns out to be between 0.3 and 0.35. We normalize $F_G(u, v)$ and let $T = 0.3$ in the following experiments. If the average value of normalized $F_G(u, v)$ is less than T, which means that a peak value exists in the normalized $F_G(u, v)$, the two corresponding image regions are considered in copy-move areas, and marked with red squares.

Figs. 5(a) and 5(b) present a case of copy-move forgery with rotation. The lighter in Fig. 5(a) is rotated by 70° and pasted to another part of the image, as shown in Fig. 5(b), with the detection result in Fig. 5(c). In Fig. 6, the ball is duplicated with a scaling factor 0.67. Fig. 7 shows a rock being scaled by 0.8, copied and pasted to the bottom-left corner. Fig. 8 presents an example in which the pasted skater is taken from the image, rotated by 4° and enlarged by 1.2, and then pasted in the bottom-right corner. All the copied-moved objects and their source regions are correctly identified. In Fig. 9, the detection result of Fig. 1(b) is given, together with horizontal sections of normalized $F_G(u, v)$ with peaks indicating the copy-move operation.



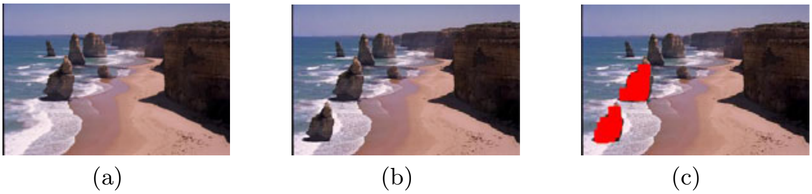
(a) (b) (c)

Fig. 5. Detection of a rotated-copied lighter: (a) original, (b) tampered with the lighter rotated by 70° and pasted to another region, and (c) detection result



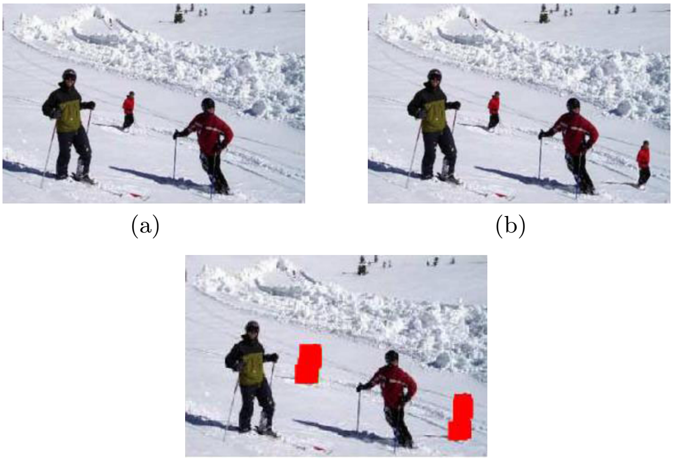
(a) (b) (c)

Fig. 6. Detection of a scaled-copied ball: (a) original, (b) tampered with the ball shrunk by 0.67 and pasted, and (c) detection result



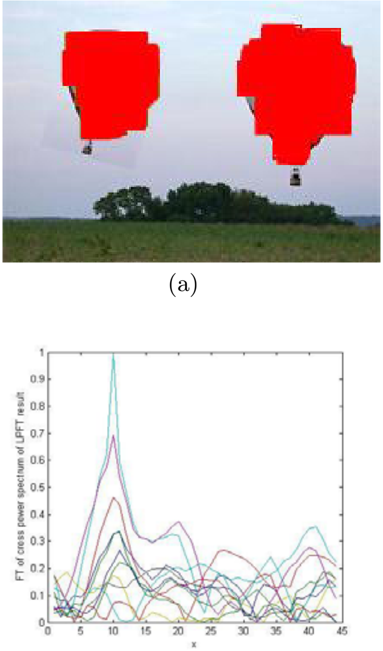
(a) (b) (c)

Fig. 7. Detection of a scaled-copied rock: (a) original, (b) tampered with the rock shrunk by 0.8 and pasted, and (c) detection result


(a) (b)

(c)

Fig. 8. Detection of a rotated-scaled-copied skater: (a) original, (b) tampered with the skier rotated by 4 degrees, enlarged by 1.2 and pasted, and (c) detection result



(a)

(b)

Fig. 9. Detection of the pasted balloon as in Fig. 1: (a) detection result, and (b) horizontal sections of normalized FT of cross spectrum of LPFT corresponding to tampered areas

4 Conclusions

We have proposed an effective method to detect copy-move image forgery with rotation and scaling tolerance. LPFT is employed to convert scaling and rotation in the image domain into translation in the log-polar domain, and peaks in the cross-spectra of LPFT pairs reveal similarity in the corresponding pairs of image blocks. This way, tampered areas in an image produced by copy-move operations, even accompanied with rotation and scaling, can be identified. Unlike FMT features in [8], the proposed LPFT is robust to large rotation, and experimental results show the efficacy of our method.

The next concern is to cope with more challenging cases containing scaling, rotation, re-sampling, cropping and other complicated post-manipulations.

Acknowledgments

This work was supported by the Natural Science Foundation of China (60773079, 60872116, and 60832010).

References

1. Fridrich, J., Soukal, D., Lukas, J.: Detection of copy-move forgery. In: Proc. of Digital Forensic Research Workshop, pp. 178–183 (2000)
2. Popescu, A.C., Farid, H.: Exposing digital forgeries by detecting duplicated image regions. Dept. Comput. Sci., Dartmouth College, Tech. Rep. TR2004-515 (2004)
3. Mahdian, B., Saic, S.: Detection of copy-move forgery using a method based on blur moment invariants. Forensic Sci. Int. 171, 180–189 (2007)
4. Myrna, A., Venkateshmurthy, M., Patil, C.: Detection of region duplication forgery in digital images using wavelets and log-polar mapping. In: Proc. of International Conference on Computational Intelligence and Multimedia Applications, pp. 371–377 (2007)
5. Bravo, S., Nunez, A.: Passive forensic method for detecting duplicated regions affected by reflection, rotation and scaling. In: Proc. of European Signal Processing Conference, pp. 824–828 (2009)
6. Zokai, S., Wolberg, G.: Image registration using log-polar mappings for recovery of large-scale similarity and projective transformations. IEEE Trans. Image Processing 14(10), 1422–1434 (2005)
7. Lin, C.-Y., Wu, M., Bloom, J., Cox, I., Lui, Y.: Rotation, scale, and translation resilient watermarking for images. IEEE Trans. Image Processing 10(5), 767–782 (2001)
8. Bayram, S., Sencar, H., Memon, N.: An efficient and robust method for detecting copy-move forgery. In: Proc. of International Conference on Acoustics, Speech, and Signal Processing, pp. 1053–1056 (2009)

An Extensible Particles Swarm Optimization for Energy-Effective Cluster Management of Underwater Sensor Networks

Mong-Fong Horng[1], Yi-Ting Chen[1], Shu-Chuan Chu[2], Jeng-Shyang Pan[1], and Bin-Yih Liao[1]

[1] Department of Electronics Engineering,
National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan
[2] Department of Computer Science and Information Engineering,
Cheng-ShiuUniversity, Kaohsiung, Taiwan
{mfhorng,ytchen,scchu,jspan,byliao}@bit.kuas.edu.tw

Abstract. Acoustic communication networks in underwater environment are the key technology to explore global ocean. There are major challenges including (1) lack of stable and sufficient power supply, (2) disable of radio frequency signal and (3) no communication protocol designed for underwater environment. Thus, acoustic so far is the only media suitable to operate for underwater communication. In this paper, we study the technology of underwater acoustic communication to support underwater sensor networks. Toward the energy-effective goal, a cluster-based sensor network is assumed. The energy-dissipation of sensor nodes is optimized by biological computing such as Particle Swarm Optimization (PSO). The objective function of sensor node clustering is formulized to constraint on the network coverage and energy dissipation. The problem of dual-objective optimization is solved by the proposed extensible PSO (ePSO). ePSOis an innovation from traditional PSO. The major innovation is to offer an extensible particle structure and to enable more flexible search for optimal solutions in space. The experimental results demonstrate that the proposed ePSO effectively and fast tackles multi-objective optimization problem. The application of ePSO on underwater acoustic communication systems shows the feasibility in real world.

Keywords: Extensible Particle Swarm Optimization (ePSO), Underwater Sensor Networks, Acoustic Communication, Energy-effective.

1 Introduction

Acoustic signal is a key to open the era of exploring ocean. The significance of acoustic signal in an underwater environment was recognized because electronic signal fails to propagate in water. There is some obstacles impedance in developing underwater communication system including 1) severe power limitations imposed by battery power; 2) severe bandwidth limitations; and 3) channel characteristics including long propagation times, multipath, and signal fading [1]. To explore the thesaurus in ocean, sensor network benefits the information acquisition from underwater environments. Of

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 109–116, 2010.

course, acoustic signal is the only media workable in this situation. In the development of underwater acoustic sensor networks, effective communication and efficient energy consumption are with the same significance [2, 8-9].

Underwater acoustic networks (UAN) are composed of fixed acoustic transceivers and autonomous underwater vehicles (AUV). As mentioned above, the varying channel characteristic impacts on the communication quality of UAN. In general, there are two categories of UAN; centralized UAN, distributed UAN and Multi-hop UAN. In a centralized UAN, there is a control node to establish the communication to the other nodes. All communication of command and data between nodes is realized by the control node. Thus the control node is very critical to the whole system. Once the control node fails, all communication functions fail simultaneously. Distributed UAN is composed of identical nodes which communicate with each other in a peer-to-peer manner. Distributed UAN has better fault tolerance against node failures. However, the lack of control node leads to the contention of communication channel among nodes. The channel utilization and energy consumption are worse than the former. Multi-hop UAN is a better solution than the others. In multi-hop UAN, all nodes are identical and organized in a two-tier hierarchy of clusters. The nodes are self-organized as non-overlapped clusters. In a cluster of nodes, the node near the cluster gravity will be selected as the cluster head to take the responsibility of a control node. After the power failure of a cluster node, the nodes in a cluster will elect a node among themselves to be a new cluster head. The lifetime of multi-hop UAN is effectively prolonged and the communication quality is improved. In this paper, the clustering of nodes in a multi-hop UAN is investigated. We first formulize the clustering of nodes as an optimization problem to be solved. To solve the formulized optimization, PSO technique is applied. PSO is one of biological computing techniques. However, the invariant particle structure of traditional PSO is not suitable for varying solution space. Thus, an extensible PSO (ePSO) is proposed to enable the variable structure of particle design. The flexibility of ePOS is helpful to solve the problem of varying solution space such as the clustering of acoustic communication nodes.

The rest of this paper is organized as follows. The previous researches related to UAN and PSO are reviewed in Section 2. In Section 3, the ePOS algorithm and its application on UAN are presented. Simulation results of the proposed approach are illustrated in Section 4. Finally we conclude this work in Section 5.

2 Related Works

There are some valuable researches on the clustering of nodes to support energy-efficiency. First, there are researches related to the issues of energy–saving in sensors networks. Chen *et. al.*[8] illustrated a tradeoff between the performance and the total energy consumption and proposed a multi-cluster structure to greatly improve performance by adopting the multi-cluster structure of the sensor network. Bao *et. al.*[9] pointed that the primary energy and balance of energy consumption of wireless sensor networks (WSNs) plays a key role in network lifetime and stability. However, the energy of each node and the total energy in wireless sensor networks are limited. Lin*et. al.* [7] presented a neural-network–based scheme of data aggregation for wireless sensor networks (WSNs) in order to reduce the number of transmissions of sensor nodes.

The coverage is also another issue of deploying sensor nodes in space. Bai*et. al* [13]. stated that the coverage of wireless sensor networks (WSNs) is one of the most important measurement criteria of QoS. Optimal coverage of sensors is propitious to the maximum possible utilization of the available sensors. It can balance node energy consumption, and elongate network lifetime. In a cluster-based sensor network, the organization of sensor nodes should be derived from the considerations of energy consumption, coverage and lifetime.

Clustering analysis is a methodology to find the cluster organization of a set of nodes according to given constraints. The legacy clustering method is to partition nodes to k clusters. In partition method, a cluster number *k* is defined before partition. Each node is placed to a specific cluster to have an optimal distribution. Typical partition methods are k-means (KM) clustering and fuzzy-c-mean (FCM. They use euclidian distance between nodes and cluster center to evaluate the belonging of each node. The major dissimilarity between them is that FCM utilizes membership functions to indicate the belonging degree of each node to all cluster centers but KM only recognizes a crisp belonging of nodes as one-to-one belonging. A major shortage of KM and FCM is the prior knowledge of cluster number *k*. In both methods, users have to know the cluster number in advance. However, in real applications, the knowledge about how to set the value of *k* is not available.

Biological computing technique (BCT) is still evolving continuously in decades. Particle Swarm Optimization (PSO) is one of the most popular BCTs. PSO was proposed by Kennedy *et. al.*[10] to emulate the behavior of bird seeking for foods. The searching behavior of each bird is affected by not only its searching experience but also its neighbors' behavior. PSO is a typical model for emulate the communication between the individuals in a group. The individuals of a group have their own search directions and speeds. They also can adapt their behavior according to their individual experience and group behavior. Thus, PSO is a good methodology for the adaptive solution of optimization problems. In this paper, we present a dynamic clustering method based on a modified PSO. The details of the proposed scheme are illustrated as follows.

3 A New Extensible Particle Swarm Optimization to Support Smart Clustering of Sensor Nodes

The particle swarm optimization (PSO) algorithm is based on the evolutionary computation technique. PSO is a population based evolutionary algorithm and has similarities to the general evolutionary algorithm. However, PSO is motivated from the simulation of social behavior which differs from the natural selection scheme of genetic algorithms. The metaphor is that of multiple collections (a swarm) of objects moving in space and thus objects are said to possess position and velocity and are influenced by the others in the swarm. PSO processes the search scheme using populations of particles which correspond to the use of individuals in genetic algorithms. Each particle is equivalent to a candidate solution of a problem. The particle will move according to the adjusted velocity that is based on the corresponding particles experience and the experience of its companions. For the D-dimensional function *f(.)*, the *i*-th particle for the *t*-th iteration can be represented as $X_i^t = (x_i^t(1), x_i^t(2),..., x_i^t(D),)$. Assume that the best previous position of the *i*-th particle for the *t*-th iteration is represented as $P_i^t = (p_i^t(1), p_i^t(2),..., p_i^t(D))$, then $f(P_i^t) \le f(P_i^{t-1}) \le ... \le f(P_i^1)$. The velocity of the *i-th* particles at the *t-th* iteration can be expressed as $V_i^t = (v_i^t(1), v_i^t(2),..., v_i^t(D))$. $G^t = (X^t(1), X^t(2),..., X^t(D))$ is defined as the best position of all particles at the *t-th* iteration. The original PSO algorithm updates the position and velocity of particles according to

$$V_i^{t+1} = V_i^t + C_1 r(P_i^t - X_i^t) + C_2 r(G_i^t - X_i^t) \tag{1}$$

and

$$X_i^{t+1} = X_i^t + V_i^{t+1}, \quad for \quad i = 0,1,...,(N-1) \tag{2}$$

Where *r*, *C*$_1$,*C*$_2$, *N*are the random number, weights and the particle size, respectively. In the original PSO algorithm, the particle size is constant. The PSO with constant structure is not suitable for some dynamic systems. Thus, we propose a variable-structure particle design to develop the ePSO algorithm for flexible applications. A particle of ePSO is expressed as $X_i^t = (x_i^t(1), x_i^t(2),..., x_i^t(k),...)$. In other words, the ePSO particle is extensible in its length. We apply this new structure to solve the problem of node clustering in sensor networks to optimize the coverage and energy-consumption.

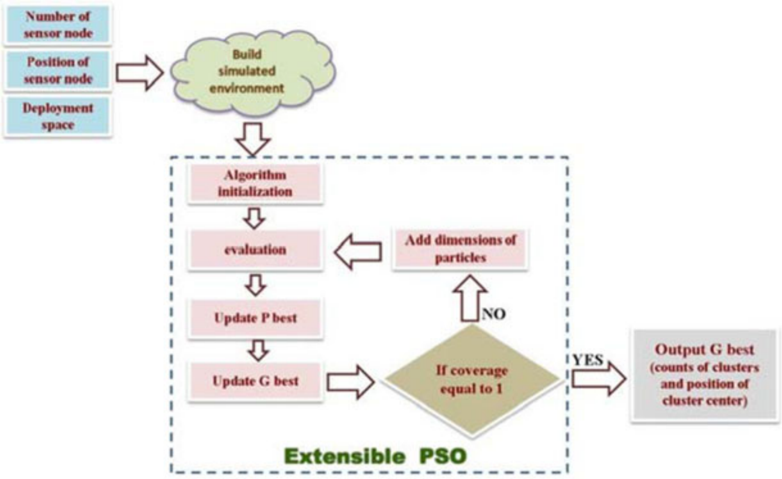


Fig. 1. Flow of Extensible Particle Swarm Optimization

The development of the proposed clustering scheme is based on two factors to search the optimal cluster organization. The first factor is the distance between nodes and their corresponding cluster center. The second is the sensing coverage of all nodes in the network. The developed ePSO will find the minimum cluster number to support maximum coverage as well as the minimum distance between nodes and centers. In

Fig.1, the flow of the proposed extensible PSO is shown. Differing from the previous clustering methods, the proposed ePSO needs no information of clustering. The number and position of sensor nodes are given at initial. According to the input data, at the first generation, the particles are randomly generated to be evaluated. After the evaluation, new P_{best}andG_{best} are obtained for updating. When the coverage is equal to 1, the resulted cluster organization is regarded as the optimal solution. The details of the proposed scheme are as follows,

Step 1. Input node number*N*, node position $\{(x_i, y_i), for\ i = 0, \ldots, (N-1)\}$

Step 2. Initialize particle population and cluster centers. The dimension of particles is set to *g* and the cluster number *n*=2. For any particle, its length is *gn*

Step 3. Calculate the fitness of each particles by

$$\mathrm{J(U,V)} = \sum_{\mathrm{j=1}}^{\mathrm{c}} \sum_{\mathrm{i=1}}^{\mathrm{n}} (\mathrm{u_{ij}})^{\mathrm{m}} \, (\mathrm{d_{ij}})^2$$

$$u_{ij} = \frac{(1/\|x_i - v_k\|^2)^{1/(m-1)}}{\sum_{k=1}^{c}(1/\|x_i - v_k\|^2)^{1/(m-1)}} j = 1,2,\ldots,c \quad i = 1,2,\ldots,n \tag{3}$$

where u_{ij} is the membership set to indicate the belonging of node $x_i$to center v_j and d_{ij} is the distance between node *i* and center v_j

Step 4. Evaluate the fitness of current particles. According the fitness evaluation as depicted in Eq. 4-5 to update the P_{best} andG_{best}.

$$\text{if } f(x_i) < f(P_i) P_{id} = x_{id} \tag{4}$$

$$\text{if } f(x_i) < f(G) \quad G = x_{id} \tag{5}$$

Step 5. Update the velocity and position of particles by

$$V_{id} = w \times V_{id} + c_1 \times rand() \times (P_{id} - x_{id}) + c_2 \times rand() \times (G_d - x_{id}) \tag{6}$$

$$x_{id} = x_{id} + V_{id} \tag{7}$$

where

V_{id}: velocity of particle x_i on dimension *d*

x_{id}: position of particle x_i on dimension *d*

w: weight

c_1, c_2 : learning factors

Rand(): a random number ranging from 0 to 1

P_{id}: current best position of particle x_i on dimension *d*

G_d: current best position of group on dimension *d*

Step 6. If the best fitness is steady and the coverage is not equal to 1, then let *n=n+1* and back to Step 1.

Step 7. Stop the evolution when the best fitness is steady and the coverage is equal to 1

4 Experiment Results

To verify the effectiveness of the proposed scheme, a 300m x 300m space is considered for node deployement. The node number in a space will determin the deployment density. There are two test cases for various node desnities, including 50-node and 150-node. The experiments of each teast case are repeated by three times. The transmission range of nodes is 50m and the power model of nodes follows the work presented in [17]. To meet the requirement of energy-effective, the energy constraint is converted to a distance constraint. Thus, the distnace contraint is specified less than 50m. In the first test case, there are 50 nodes in space. According to the proposed alogrithm, the optimal cluster number is evolving to 4 with full coverage. In the evolution of node clustering, the relationship between the coverage and cluster number is illustrated by Fig. 2. The result of node clustering is shown in Fig. 3. The red points indicate the found cluster centers. To have the minimum engery dissipation of the nodes in a cluster, the cluster gravity is the best position for cluster center. Thus the distance between the found cluster center and the cluster gravity is defined as the error of the node clustering. The corresponding errors of each node cluster are depicted in Table 1.

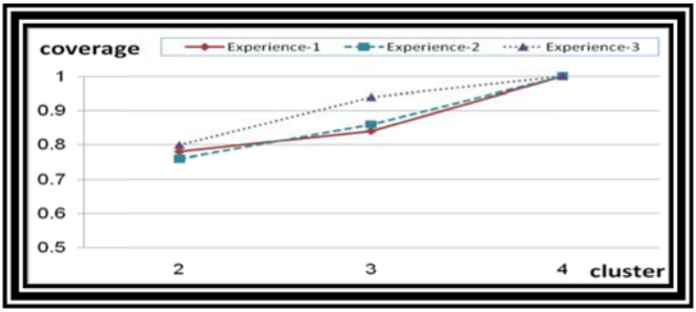


Fig. 2. Relationship between cluster number and network coverage in test case I

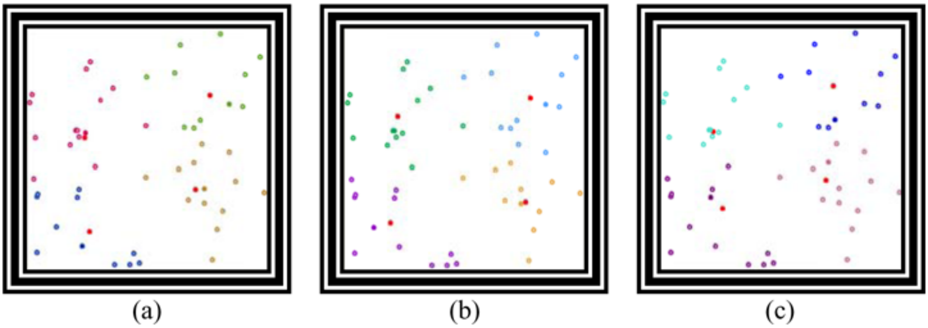


Fig. 3. The clustering of 150 nodes in three independent experiences

The other test case is to verify the clustering performance in the case of dense nodes. There are 150 nodes deployed in 300m x 300m space. As the density increaing, nodes will be benefited because they will be able to use less energy for short-distance communicaiton. Although the cluster number is slightly increased, the overall efficiency is improved.In Fig. 4, the relationship between the cluster number and network coverage is depicted. As mention above, the cluster number is increased by 1 but the power efficiency is increased. The clustering error is shown in Fig. 5

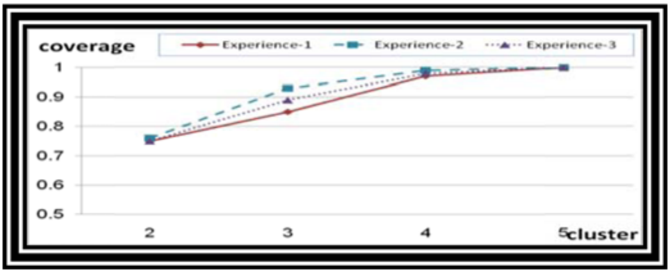


Fig. 4. Relationship between cluster number and network coverage in test case I

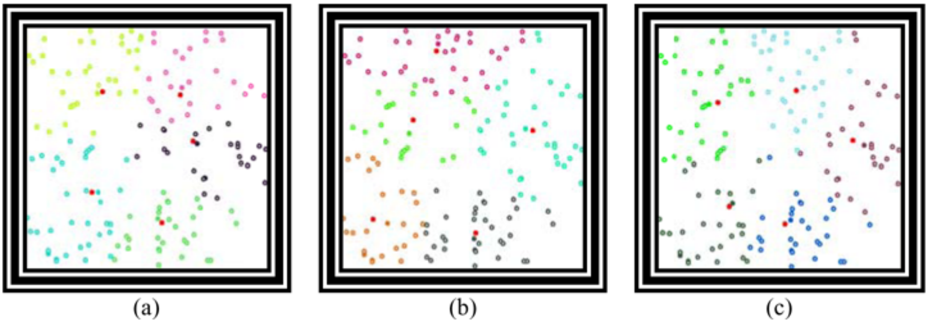


Fig. 5. The clustering of 150 nodes in three independent experiences

5 Conclusion

In this paper, an extensible particles swarm optimization for energy-effective cluster management of underwater sensor networks is proposed. The major improvement of the proposal is the extensible particle structure to enhance the performance and flexibility of PSO algorithm. Besides, the clustering of sensor nodes in underwater networks is formulized as an optimization problem with the constraints of coverage and energy-dissipation. In the clustering problem with high dynamics, tradition PSO is not suitable to solve. The proposed extensible particle is successful to solve the problem of dynamic clustering of sensor nodes in an underwater network. The clustering results confirm the benefits of high network coverage and effective energy consumption.

References

1. Berkhovskikh, L.Y.: Fundamentals of Ocean Acoustics. Springer, New York (1982)
2. Sozer, E.M., Stojanovic, M., Proakis, J.C.: Underwater Acoustic Network. J. Oceanic Eng. 25(1), 72–83 (2000)
3. Akyidiz, I.F., Pompili, D., Melodia, T.: Underwater Acoustic Sensor Networks: Research Challenges, January 2005. Ad Hoc Networks. Elsevier, Amsterdam (2005)
4. Eberhart, R., Kennedy, J.: A New Optimizer Using Particle Swarm Theory. In: Proceeding of 6th International Symposium on Micro Machine and Human Science, pp. 39–43 (1995)
5. Dorigo, M., Maniezzo, V., Colorni, A.: The Ant System: Optimization by a Colony of Cooperation Agents. IEEE Transactions of Systems, Man and Cybernetics Part-B 26(2), 29–41 (1996)
6. Aziz, N.A.B.A., Mohemmed, A.W., Alias, M.Y.: A wireless sensor network coverage optimization algorithm based on particle swarm optimization and Voronoi diagram. In: 2009 IEEE International Conference on Networking Sensing and Control, pp. 602–607 (May 2009)
7. Lin, J.W., Guo, W.Z., Chen, G.L., Gao, H.L., Fang, X.T.: A PSO-BPNN-based model for energy saving in wireless sensor networks. In: Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, pp. 948–952 (August 2009)
8. Chen, H.B., Tse, C.K., Feng, J.C.: Minimizing effective energy consumption in multi-cluster sensor networks for source extraction. In: IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, pp. 1480–1489 (March 2009)
9. Bao, X.R., Qie, Z.T., Zhang, X.F., Zhang, S.: An efficient Energy Cluster-based Routing Protocol for wireless sensor networks. In: Control and Decision Conference, pp. 4716–4721 (August 2009)
10. Kennedy, J., Eberhart, R.C.: Particle swarm optimization. Proceedings of the IEEE International Conference on Neural Networks 4, 1942–1948 (1995)
11. Tewolde, G.S., Hanna, D.M., Haskell, R.E.: Hardware PSO for sensor network applications. In: IEEE Swarm Intelligence Symposium, pp. 1–8 (November 2008)
12. Abdul Latiff, N.M., Tsimenidis, C.C., Sharif, B.S., Ladha, C.: ynamic clustering using binary multi-objective Particle Swarm Optimization for wireless sensor networks. In: IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications, pp. 1–5 (December 2008)
13. Bai, X.Z., Li, S., Jiang, C.G., Gao, Z.Z.: Coverage Optimization in Wireless Mobile Sensor Networks. In: 5th International Conference on Wireless Communications, Networking and Mobile Computing, pp. 1–4 (October 2009)
14. Low, K.S., Nguyen, H.A., Guo, H.: Optimization of Sensor Node Locations in a Wireless Sensor Network. In: Fourth International Conference on Natural Computation, pp. 286–290 (November 2008)
15. Gao, W., Kamath, G., Veeramachaneni, K., Osadciw, L.: A particle swarm optimization based multilateration algorithm for UWB sensor network. In: Canadian Conference on Electrical and Computer Engineering, pp. 950–953 (July 2009)
16. Low, K.S., Nguyen, H.A., Guo, H.: A particle swarm optimization approach for the localization of a wireless sensor network. In: IEEE International Symposium on Industrial Electronics, pp. 1820–1825 (November 2008)
17. Wang, Q., Hempstead, M., Yang, W.: A Realistic Power Consumption Model for Wireless Sensor Network Devices. In: The 7th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks SECON (2006)

The 3D Display and Analysis of the Pattern of Photolithography

Lih-Shyang Chen[1], Young-Jinn Lay[1], Lian-Yong Lin[1], Jing-Jou Tang[2], Wen-Lin Cheng[1], and Yu-Jen Lin[1]

[1] Department of Electrical Engineering, National Cheng Kung University, No.1 University Road, Tainan City 701, Taiwan
[2] Department of Electronics Engineering, Southern Taiwan University, No.1 Nan Tai St., Yong Kang City, Tainan County, Taiwan
chens@mail.ncku.edu.tw,
{bill,edwarkingsir}@mirac.ee.ncku.edu.tw,
tjj@mail.stut.edu.tw, {wenlin,zan}@mirac.ee.ncku.edu.tw

Abstract. As semiconductor technology rapidly advances, lithography with high density and fine feature size has become the challenge for nanometer-scale integrated circuit fabrication. Unfortunately the optical proximity effects (OPE) will distort the developed patterns transferred from the mask patterns. Optical proximity correction (OPC) is one of the promising resolution enhancement techniques to improve the yield of IC (Integrated Circuit) production. In this paper, we develop a system, called “3D Builder”, to enable the users to easily visualize the results of the simulation of the lithographic process and understand what the physical circuits may look like in the production line. With this visualization tool, the users can better understand the effect of OPE and how OPC should be applied to improve the yield of the IC production. As a result, the yield of the IC production can be increased.

Keywords: Visualization, 3D Display, Lithography, Optical Proximity Correction.

1 Introduction

As semiconductor technology rapidly advances, lithography with high density and fine feature size has become the challenge for nanometer-scale integrated circuit fabrication. Nowadays the critical dimension (CD) has approached the wavelength of the light source for lithography. Unfortunately the optical proximity effects (OPE) will distort the developed patterns transferred from the mask patterns [1], [2]. As a result, the yield of the production of the IC (Integrated Circuit) is decreased. Optical proximity correction (OPC) [4], [5], [6] is one of the promising resolution enhancement techniques (RETs) that can be used to reshape the mask pattern to compensate the aerial image distortion due to the light interference and diffraction[1], [2]. When the OPC technique is used, the yield of the production of the IC (Integrated Circuit) is increased. Therefore, if we can somehow predict the results of the OPE and use the OPC to correct the possible distortion, the yield of IC production can be increased dramatically [3], [4].

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 117–127, 2010.

To this end, one important aspect on increasing the yield of IC production is to somehow visualize the estimated results of the lithographic process with the OPC so that the potential errors can be found and corrected before the real IC production process. In this paper, we develop a system, called “3D Builder”, to enable the users to easily visualize the results of the simulation of the lithographic process and understand what the physical circuits may look like in the production line. With this visualization tool, the users can better understand the effect of OPE and how OPC should be applied to improve the yield of the IC production. With this system, the user can better understand how various mask patterns affect the results of the lithography and subsequently make better use of the OPC to reshape the mask patterns. As a result, the yield of the IC production can be increased.

In the second section, we will discuss some background of the simulation of the lithographic process. In the third section, we will discuss the data processing procedures of the 3D Builder system. In the fourth section, we will discuss what kinds of the defects can be detected automatically by the 3D Builder so that they can be adequately displayed and be visualized by the users. In the fifth section, we will discuss some experimental results. In the sixth section, we draw our conclusions.

2 The Background of the Simulation of the Lithographic Process

In order to predict the results of the lithographic process, normally people use some simulation software packages with some lithographic process parameters to estimate the results of the lithography. Fig. 1 shows the system of photolithography, that makes use of the light source and condenser to project the masks onto the wafer for the IC production.

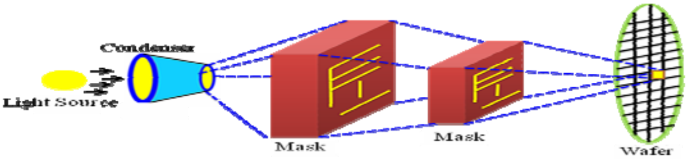


Fig. 1. The system of photolithography

In Fig. 2, we partition each “whole chip” in the wafer into several submasks each of which has some overlapping area for a complete simulation purpose. This is because a whole chip is too big for a simulation software package to accommodate huge amount of data [6]. The circuits designed in each submask are called mask patterns.

In this study, we use the MAYA[7], [8], [9] (Mask Yield Analysis System) which has been developed by the IC Design and Application Center of Southern Taiwan University and internally makes use of the SPLAT(Simulation of Projection Lens Aberrations via TCCs) [10], [11] developed by University of California, Berkeley to simulate the lithographic process. In the MAYA system, the user can plug-in some photolithographic process parameters for a particular IC production case to simulate the lithographic process and obtain the possible results.

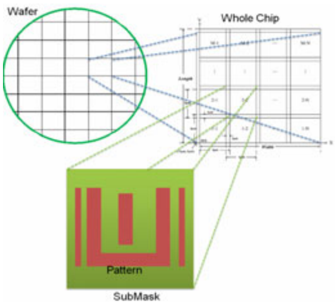


Fig. 2. Each "whole chip" in the wafer is partitioned into several submasks

When the MAYA system finishes the simulation computation for a specific set of parameters, it will output 2D images such as the one shown in Fig. 3. The blue parts are the original masks and the red parts are the results of the simulation, i.e., the results of the lithographic process. Several layers of the masks can also be shown simultaneously for the user to visualize the simulated circuits of the IC.

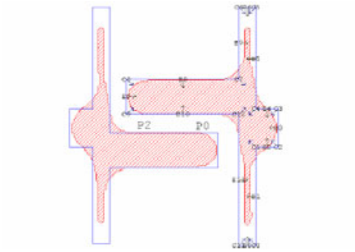
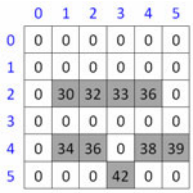


Fig. 3. The output of the MAYA system

3 The Data Processing Procedures of the 3D Builder

The 3D Builder has several modules, namely "DK" module, data pre-processing module, and 2D and 3D display modules, each of which is in charge of a certain work. The "DK" module is to compute the error types of the lithographic process and the amount of various errors which will be defined in the fourth section. Sine there is a huge amount of data produced by the MAYA system, the data pre-processing module convert and compress the data to a manageable size for further processing. 2D and 3D display modules are in charge of 2D image display and 3D object display respectively.

The simulation results generated by the MAYA system is a spo file which is a 6001*6001 two dimensional array representing a 6um*6um (1 um = 10^{-6} meter) area in a wafer, i.e., a submask area. In other words, each element in this array is a 1nm*1nm (1 nm = 10^{-9} meter) area. Each element of the array stores a value ranging from 0 to 1 and representing the light intensity of the lithographic process at the 1nm*1nm area within each submask. Each file size is about 300Mega Bytes. All the data values less than 0.2 are considered to be insignificant (since the light intensity is too weak) and can be set to 0 to simply further processing. As a result, there are a lot



	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	30	32	33	36	0
3	0	0	0	0	0	0
4	0	34	36	0	38	39
5	0	0	0	42	0	0

Fig. 4. An example of a small portion of a spo file

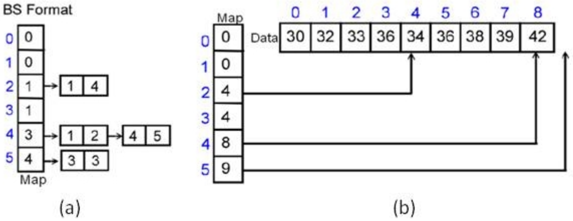


Fig. 5. The data structures of the BS format and the GS format

of 0 values in the array. We can make use of a run-length encoding scheme to compress the data without any loss and to more efficiently process the data later on.

Fig. 4 shows a simple example of the contents of the array, i.e., a small portion of a spo file and Fig. 5 shows how the data are compressed into the BS and the GS format. The BS format records the areas where the data values are not 0. For instance, the values from column 1 to column 4 in line 2 are not 0. The (1 and 4) in the BS format represent a segment of data in a scan-line. Likewise, the two segments (1 and 2), and (4 and 5) are two non-zero segments in line 4. The value in the map, called the map value, of the BS format represents the number of non-zero segments so far in that format. Therefore, the number of non-zero segments can be computed by subtracting the map value of the previous line from that of the current line. In other words, with these map values, we can compute how many segments in each line. For instance, the map value of line 4 is 3 and its previous map value is 1. For instance, we can compute that the number of segments in line 4 is 2 by subtracting 1 from 3. The GS format stores the real grey level values in the array, i.e., the light intensities of the lithographic process. All the non-zero values in the submask are stored in the GS format. For instance, the 4 non-zero values in line 2 are stored as the first 4 values. The 4 non-zero values in line 4 are stored as the values from locations 5 to 8 in the GS format. The data size can be compressed to less than 10 percents of its original size of a spo file.

There are many different 3D computer graphics primitives one can use to model objects of interest in a 3D scene. In the case we are dealing with, we want to be able to display the original mask patterns and the simulation results of the lithographic process simultaneously and compare the differences between them. The original mask patterns in general consist of many rectangular solids, each of which can be modelled by some rectangles, each of which in turn can be modelled by two triangles. The shape of the mask patterns are considered to be the ideal case of the light intensities of the lithographic process. The rectangular solids are used as graphics primitives to model the mask patterns.

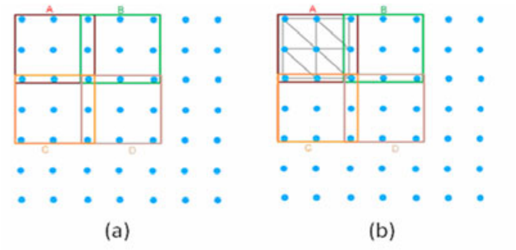


Fig. 6. (a) shows that each grid consists of a 3X3 data points and the grids overlap with each other by one data point. (b) shows that every three data points form a triangle.

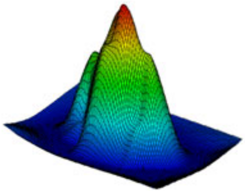


Fig. 7. The image of a typical terrain

The simulation results are the data in the spo file that has been converted to the BS and the GS compressed formats for more efficient process. The compressed formats will be un-compressed and converted back to the original 6001*6001 2D array when they are loaded into the memory. Since the 6001*6001 array is too big for a general-purpose computer screen to display and also too difficult for the user to visualize all the details simultaneously, we partition the array into several grids each of which is a 256*256 grid （a more manageable size for the system to process） in size with one pixel overlap with each other to prevent possible cracks in between when these grids are displayed. Fig. 6 shows how the partition of the grids is done. Fig. 6 (a) shows that each grid consists of a 3X3 data points and the grids overlap with each other by one data point. Every element in each gird has a light intensity value for that location. Therefore, these values in the grid can be modelled as a "terrain" shown in Fig. 7. Each data point in a grid has a light intensity value that will be displayed as the height of the terrain. Therefore, the terrain consists of many different triangles in shape. Fig. 6 (b) shows that every three data points form a triangle as a small portion of the terrain. The terrains are used as the graphics primitives to model the simulation results – the light intensities of the lithographic process.

4 Defect Analysis

With the original mask pattern data and the simulation results, we can use a software program to automatically analyze the defects that may decrease the yield in the IC production. In the lithographic process, if everything goes perfectly, the simulation results should exactly match with the original mask patterns in shape. In other words, at every location in the grid, if it is covered by a mask pattern, its light intensity should be at least higher than a threshold value. On the other hand, if it is not covered by a mask pattern, its light intensity should be zero. In this case, the IC can be produced without any defect. However, in really, this may never be the case. There are five different types of defects in the lithographic process. We define and briefly discuss these defect types here (1) Energy (2) Area deviation, (3) Bridge, (4) Hot-spot, and (5) Open.

In order to compute the defects, one should set up a threshold value to determine whether the light intensity at a location meet our expectation. If the light intensity within the mask pattern areas is higher than the threshold value, the mask patterns can be "printed" on the photoresistor. In this case, it is considered to be a successful lithographic process. Otherwise, it fails to "print" the mask pattern on the photoresistor.

For instance, in Fig. 8, the red object represents the light intensities of the simulation results of the lithographic process while the black box represents a mask pattern and is considered to be the ideal light intensities.

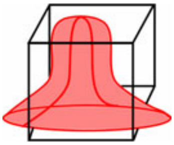


Fig. 8. The light intensity and the mask pattern

In the following, we define some basic terminologies used for our discussion.

– Energy:

In Fig. 9, we can see one of the cross-sectional images of Fig. 8 which is shown in a 2D plane. The black lines are the shape of the mask pattern - the ideal case for the light intensities and the black dotted line is the simulation result for the given mask pattern. The red region is defined as the energy underflow and the blue regions are defined as the energy overflow. The value of the distance shown in Fig. 9 can be defined by the user.

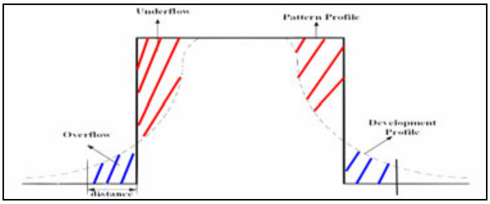


Fig. 9. The definition of the energy underflow and energy overflow

– Area Deviation:

In Fig. 10(a), the dotted lines shows a cross- sectional plane of the 3D simulation data. The resultant image of the cross-sectional plane is shown in Fig. 10 (b). The area of the red part is Area A and the area of the white rectangle is Area B. The area deviation is defined as Area A / Area B * 100%. The location of the dotted cross-sectional plane can be defined by the user. The value of the area deviation is the indication about how much the real area deviates from the ideal case. The user can move the cross-sectional plane up and down to visualize how the deviation varies from top to bottom.

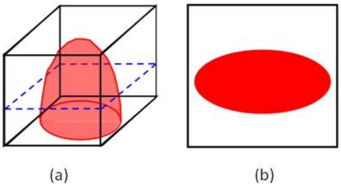


Fig. 10. The definition of the deviation

– Bridge

In Fig. 11, the blue regions with white numbers are the simulation results after the threshold operation and are called intensity areas. The red boxes with red numbers are the make patterns and are called pattern areas. If an intensity area is across two different pattern areas, it is called a bridge. In other words, the two pattern areas are supposed to have their own associated intensity areas. However, due to some OPE, the two areas are connected and become one intensity area. This bridge may eventually cause a short circuit in the IC.

– Hot-Spot:

If an intensity area does not have its own associated pattern area, it is considered to be a hot-spot and should be somehow eliminated. As shown in Fig. 11, the number 4 intensity area is a hot-spot while the number 1 intensity area is a bridge.

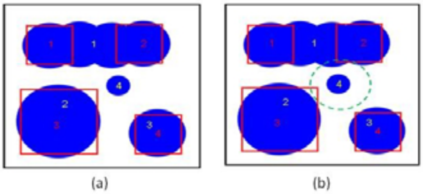


Fig. 11. The definitions of the bridge and the hot-spot

– Open:

In Fig. 12, the grey area is a pattern area and the red areas are the simulation results after the threshold operation. In this case, more than one intensity areas are associated with one pattern area. This is called an open defect. In the real circuit, this will cause an open circuit when the IC is eventually produced.

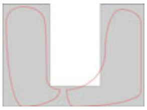


Fig. 12. The definition of an open

The system will automatically compute all the defects and their types so that the user can somehow eliminate them in some follow-up processing. The hot-spot, open, and bridge defect types are called fatal defects since they will cause failures of an IC product eventually and should be somehow eliminated by some OPC process.

When a whole chip case is selected and loaded into the system, all the submasks with data will be shown as squares on the screen as shown in Fig. 13(a) with different colors. The color of a submask indicates what type of the defects the submask has. The definitions of the colors are shown in Table 1. For instance, the green color implies that the light intensities in a submask are OK for their mask patterns. Other colors imply that the submasks have certain kinds of defects. The user can select any submask by clicking on it. The detailed mask patterns and their associated light intensities of the submask will be displayed as shown in Fig. 13(b). All the related data of the submask including the analysis of the defects, the parameters used for the defect computation described before, and so on, will be shown in the right panel. The user can change the parameters in the panel. The system will re-compute all the potential defects based on the new parameters and subsequently show the new results.

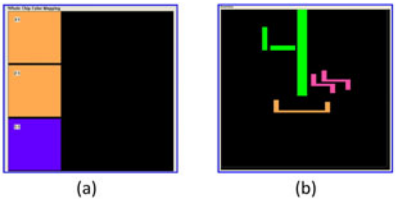


Fig. 13. (a) The display of three submasks. (b) The display of the mask patterns in a submask selected by the user.

Table 1. The definitions of the colours of the submasks

🟧 Hot-spot	🟪 Bridge	🟧 Open
🟦 Hot-spo and Bridge	🟦 Hot-spot and Open	🟪 Bridge and Open
🟥 Hot-spot, Bridge, and Open	🟩 Non-defect	🟥 User define

The user can further analyze and visualize the objects of interest in the submask in both 2D and 3D by interactively moving the objects around on the screen. The 2D and 3D images are complementary to each other in general. The 2D images allow the user to clearly visualize the original layout of the circuits since the circuits are originally designed in a 2D plane. However, in order to give the user a stronger feeling about the simulated light intensities compared with the ideal case, the 3D image may be more appropriate for this purpose. For instance, in Fig. 14(a), the original circuit layout is shown in a 2D image with different colors defined in Table1. Fig. 14(a) shows some mask patterns at a certain area of a selected submask. The user wants to see the detailed light intensities of the pink mask patterns. Therefore, he uses the mouse to draw a square on the patterns. The detailed light intensities within the square are shown in Fig. 14(b) in a 2D image. Obviously, a bridge defect exists in the circuit layout. Fig. 14 (c) and (d) show OK cases. However, although there is no defect, there are still some area deviation defect, i.e., the differences between the ideal case and the real simulation data. The red parts are the mask patterns and the green parts are light intensities. The users can also set a threshold value for the area deviation. Although the area deviation is not a fatal defect, the users still need to pay special attention to it if this value is greater than a threshold value to ensure the quality of the layout design.

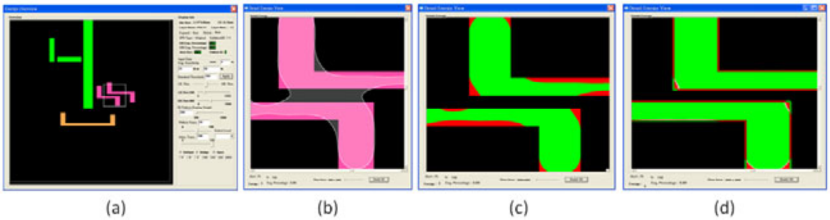


Fig. 14. The detailed view of some mask patterns

In Fig. 15, the system shows the 3D image of the simulated light intensities in white and the mask patterns in red. If the light intensities are much greater than the threshold value, a certain "terrain" may have a great height. As a result, the 3D display of both the simulated light intensities and the mask patterns may look like a huge jungle and many objects may block many other objects behind them. Therefore, in order to show the user a "clean" 3D image of all the data, for the display of a mask pattern of a certain threshold value, we only display the mask with the height of the threshold value. For the display of the simulated light intensities, we only display it with the height of （the threshold value ＋ 0.1） . In other words, the heights of simulated light intensities are at most slightly greater than those of the associated mask patterns. In this case, the user can clearly visualize all the data simultaneously.

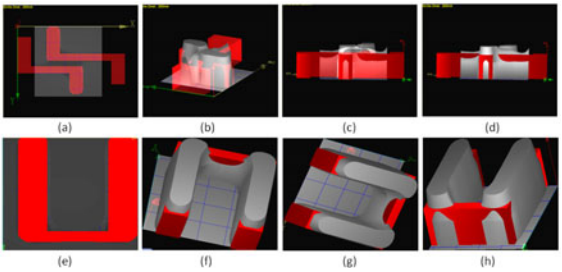


Fig. 15. The display of mask patterns and their associated terrains

The user can interactively translate and rotate the objects of interest in order to visualize them from different perspectives and better understand the data. In some cases, since the mask patterns and their associated simulated light intensities are located at almost the same locations and displayed simultaneously, it is still difficult to visualize the spatial relationships between the mask patterns and their associated light intensities. Therefore, the system enables the user to set the transparency parameters of some objects in order to better visualize their spatial relationships. Fig. 15(a) shows the original 2D layout. The user marks the white area as his region of interest so that the system will show the light intensities of that region. Fig. 15(b) and (c) show the results where some mask patterns (the red objects) are transparent while the terrain – the light intensities (the white objects) are opaque so that the users can visualize the spatial relationships between various objects. In Fig. 15(d), all the objects are opaque. In Fig. 15(e), a 2D mask pattern is shown. In Fig. 15(f), the mask pattern and its associated terrain are shown. The grid on the floor is the scale of the distance. This allows the user to know the sizes of the objects. Fig. 15(g) and (h) show the same objects from different perspectives. The users can understand how the real light intensities deviate from the original design.

5 Conclusions

The 3D Builder system can take a huge amount of data from the simulation system of lithographic process – MAYA, compressed them into a manageable data size, and display the data in both 2D and 3D so that the user can easily visualize and understand the data. Consequently, the user can somehow make correction of their design of the OPC. The experimental results turn out to be very helpful in analyzing the design of OPC. In the future, we will further explore the possibility of making use of the system to display the 3D images of the multi-layer circuit layout and detect more complicated defects in an attempt to increase the yield of the IC production.

References

1. Mack, C.A.: Field Guide to Optical Lithography. SPIE-The International Society for Optical Engineering FG 2006 (2006)
2. Tang, J.-J., Tien, T.-K., Wang, L.-Y.: Study of Pattern Transfer Quality for Nanoscale Lithography. In: 7th International Semiconductor Technology Conference (ISTC 2008), Shanghai, pp. 271–276 (2008)
3. Tang, J.-J., Liao, C.-L., Jheng, P.-C., Chen, S.-H., Lai, K.-M., Lin, L.-J.: Yield Analysis for the 65nm SRAM Cells Design with Resolution Enhancement Techniques (RET). In: 18th VLSI Design/CAD Symposium, Hualien (2007)
4. Chen, S.-H.: The Transformation Quality Analysis of Layout Patterns in Nanoscale Lithography. Master's thesis, Southern Taiwan University of Technology (2009)
5. Yong, L.-L.: Realization of Automatic Optical Proximity Correction. Master's thesis, National Chi Nan University (2009)
6. Chen, S.-H., Lin, L.-Y., Tang, J.-J.: Analysis of Mask Partition to Facilitate SoC Photolithography Simulation. In: 2008 Conference on Innovative Applications of System Prototyping and Circuit Design, pp. 86–91 (October 2008)
7. Tang, J.-J., Jheng, P.-C.: A GDSII Data Translator for Photolithography Simulator – SPLAT. In: Conference on Electronic Communication and Applications (CECA), EP036, Taiwan Kaoshiung (2007)

8. Tang, J.-J.: Development of the Graphic User Interface to Backend Design and Photolithography of Integrated Circuits. In: 2005 International Conference on Open Source (ICOS 2005), Taipei (2005)
9. Tang, J.-J., Sheu, M.-L., Lin, L.-Y.: Mask Yield Analysis System (MaYas). U-Tools Forum., `http://larc.ee.nthu.edu.tw/utool/info.php`
10. User's Guide for SPLAT Version 6.0. Electronics Research Laboratory, University of California Berkeley
11. Lin, L.-Y., Sheu, M.-L., Tang, J.-J.: A Study of Optical Lithography Simulation Using PC-Cluster. In: 5th Workshop on Grid Technologies and Applications (WoGTA 2008), pp. 73–77 (2008)

Geometrically Invariant Image Watermarking Using Scale-Invariant Feature Transform and K-Means Clustering

Huawei Tian[1], Yao Zhao[1], Rongrong Ni[1], and Jeng-Shyang Pan[2]

[1] Institute of Information Science, Beijing Jiaotong University, Beijing, China, 100044
hwtian@live.cn, {yzhao,rrni}@bjtu.edu.cn
[2] Department of Electronic Engineering, Kaohsiung University of Applied Sciences, Taiwan
jspan@cc.kuas.edu.tw

Abstract. In the traditional feature-base robust image watermarking, all bits of watermark message are bound with the feature point. If a few of points are attacked badly or lost, the performance of the watermarking scheme will decline or fail. In this paper, we present a robust image watermarking scheme by the use of k-means clustering, scale-invariant feature transform (SIFT) which is invariant to rotation, scaling, translation, partial affine distortion and addition of noise. SIFT features are clustered into clusters by k-means clustering. Watermark message is embedded bit by bit in each cluster. Because one cluster contains only one watermark bit but one cluster contains many feature points, the robustness of watermarking is not lean upon individual feature point. We use twice voting strategy to keep the robustness of watermarking in watermark detecting process. Experimental results show that the scheme is robust against various geometric transformation and common image processing operations, including scaling, rotation, affine transforms, cropping, JPEG compression, image filtering, and so on.

Keywords: robust watermark, geometric distortion, watermark synchronization, scale-invariant feature transform, k-means clustering.

1 Introduction

Digital watermarking techniques have been proposed to embed signatures in the multimedia data to identify the owner, the intended recipients, and to check the authenticity of the multimedia data. Among various problems to be solved in image watermarking, robustness against geometric transformations is a most challenging one, and many existing image watermarking algorithms are vulnerable to them.

Recently, the feature-based watermarking scheme [1-6] which is also called the second generation scheme has drawn much attention. It is an effective approach to addressing the watermark robustness against geometric distortions, because feature points provide stable references for both watermark embedding and detection. Bas et al. adopt Harris detector to extract feature points and use feature points to construct a triangular tessellation that they use to embed the watermark [1]. Tang and Hang adopt

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 128–135, 2010.

the Mexican Hat wavelet to extract feature points. Local regions are generated based on the feature points, and the watermark is embedded into the sub-blocks in DFT domain [2]. Lee et al. extract image feature points using SIFT and use them to generate a number of circular regions for watermark imbedding [3]. Wang et al. use Harris-Laplace detector to extract feature points. Local characteristic regions are constructed for embedding watermark [4]. The basic strategy of these watermarking schemes is to bind a watermark with the local region. The local region is the linchpin, upon which a watermarking scheme's success or failure depends. So, some drawbacks indwelled in current feature-based schemes restrict the performance of the robust watermarking system. First, because all bits of the watermark sequence are embedded in a local region, the robustness of the watermark scheme is close relation with the robustness of local regions. Second, the number of local regions extracted for watermark embedding is small. If a few of local regions are attacked badly or lost, the performance of the watermarking scheme will decline or fail. Third, the capacity is very low because the area of local region is usually very small. The increase of the capacity will induce a dramatic decline of the robustness.

In this paper, we propose a watermarking method, using the SIFT and k-means clustering. The SIFT is invariant to rotation, scaling, translation, partial affine distortion and partial illumination changes. Another important aspect of the SIFT is that it generates large numbers of features that densely cover the image over the full range of scales and locations [7]. In this watermarking scheme, all features are clustered to several groups according the length of the watermark sequence using k-means clustering. The robustness of the approach does not lie on a few of features, because the watermark bits are embedded into groups respectively. So, the approach shows much more robust performance although many feature points are lost. The centroids of clusters must be sent to the extractor, so the watermarking method is semi-blind.

The remainder of this paper is organized as follows. Section 2 reviews the SIFT and the k-means clustering. Section 3 describes the details of the geometrically invariant watermarking scheme. Simulation results are shown in Section 4. Section 5 concludes the paper.

2 The SIFT and the K-Means Clustering

2.1 The SIFT

The SIFT is a method for extracting distinctive invariant features from images that can be used to perform reliable matching between different views of an object or scene. The features are invariant to image scale and rotation, and are shown to provide robust matching across a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination [7].

SIFT algorithm consists of four major stages as described in [7]: (1) scale-space peak selection; (2) keypoint localization; (3) orientation assignment; (4) SIFT descriptor creation. The SIFT extracts feature points with such properties as their location (t_1, t_2), scale s, orientation θ, and a 128 element vector $\boldsymbol{ds}$ (called "SIFT descriptor"). The scale s varies proportionally with the image zoom-scale. The orientation θ varies with

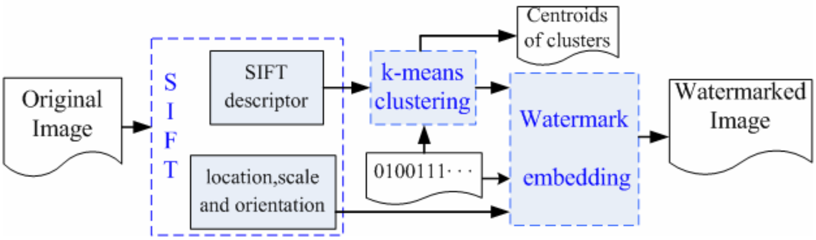


Fig. 1. Watermark embedding scheme

the image rotation-degree. The SIFT descriptor is invariant to rotation, scaling, translation, partial affine distortion and partial illumination changes.

2.2 The K-Means Clustering

K-means clustering [8] is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest centroid. Concretely, given a set of observations $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, where each observation is a d-dimensional real vector, then k-means clustering aims to partition the n observation into k ($k < n$) sets $\boldsymbol{C} = \{C_1, C_2, ..., C_k\}$ so as to minimize the within-cluster sum of squares:

$$J = \underset{\boldsymbol{C}}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x_j \in C_i} \left\| x_j - u_i \right\|^2 \tag{1}$$

where u_i is the centroid of C_j

K-means clustering algorithm is composed of the following four steps [9]: (1) Place k points into the space represented by the observations that are being clustered. These points represent initial group centroids. (2) Assign each observation to the group that has the closest centroid. (3) When all observations have been assigned, recalculate the positions of the k centroids. (4) Repeat Step (2) and (3) until the centroids no longer move. This produces a separation of the observation into groups from which the metric to be minimized can be calculated. The algorithm is significantly sensitive to the initial randomly selected cluster centers, so the algorithm can be run multiple times to reduce this effect.

3 Watermarking Scheme

3.1 Embedding Scheme

Fig.1 is an overview of our proposed watermark embedding scheme. There are three steps in the scheme:

Step 1: SIFT features including SIFT descriptor $\boldsymbol{ds}$, location (t_1, t_2) , scale s , orientation θ are extracted from the original image.

Step 2: SIFT feature points are partitioned into k (the length of watermark sequence) clusters using k-means clustering algorithm according to SIFT descriptor $\boldsymbol{ds}$. The centroids $\boldsymbol{ctd}$ of clusters will be preserved for watermark extracting.

Step 3: The watermark message is embedded bit by bit in each cluster. One cluster contains only one watermark bit.

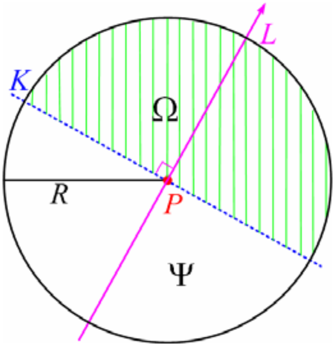


Fig. 2. SIFT disk partition

Step 3 is an elaborate process. In order to improve the robustness of the transmitted watermark sequence $W = (w_0 w_1 \cdots w_i \cdots w_k)$, where $w_i \in \{0,1\}$, only one bit watermark w_i is embedded into a cluster C_i. As shown in Fig. 2, the centre of the disk P located at (t_{i1}, t_{i2}) is a SIFT feature point x_{ij}, and the radius R of the disk equal to the SIFT scale s_i of SIFT point x_{ij}, where $x_{ij} \in C_i$. The orientation of line L represent the SIFT orientation θ_i of SIFT point x_{ij}. Line K which is the perpendicular of line L divides the disk into two zones Ω and Ψ. Because the scale s_i varies with the image zoom-scale and the orientation θ_i varies with the image rotation-degree, Ω and Ψ cover the same contents even if the image is attacked by scaling and rotation transforms.

According to watermark bit w_i, v_{i1} and v_{i2} of vector $v_i = [v_{i1} \quad v_{i2}]$ will be embedded into every pixel in zone Ω and zone Ψ respectively. If $w_i = 0$, $v_{i1} = 0$ and $v_{i2} = 1$ are embedded into every pixel of zone Ω and zone Ψ using Quantization Index Modulation (QIM) [10]. If $w_i = 1$, $v_{i1} = 1$ and $v_{i2} = 0$ are embedded into every pixel of zone Ω and zone Ψ using QIM. We construct two quantizers $Q(.;w)$, where $w \in \{0,1\}$. In this paper, we consider the case where $Q(.;w)$ is a uniform, scalar quantizer with stepsize Δ and the quantizer set consists of two quantizers shifted by $\Delta/2$ with respect to each other. Δ is pre-defined and known to both embedder and extractor, meanwhile it affects the robustness to common signal processing and the quality of the watermarked image. For zones Ω or Ψ, according to the corresponding bit v_{ij}, where $v_{ij} \in \{0,1\}$ and $j \in \{1,2\}$, each pixel $p(m,n)$ is quantized with quantizer $Q(.;v_{ij})$.

$$p_w(m,n) = Q(p(m,n); v_{ij}) \tag{2}$$

After every SIFT feature points x_{ij} is marked by v_i, the watermark embedding process is finished.

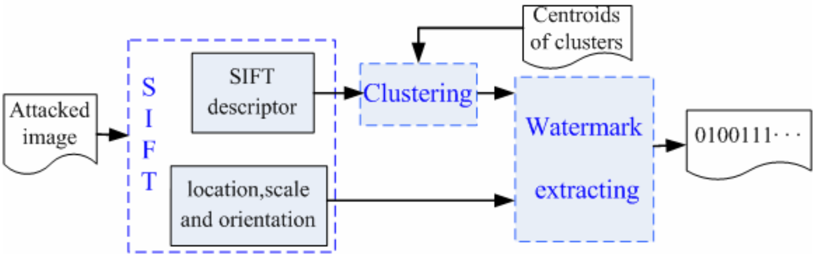


Fig. 3. Watermark detecting scheme

3.2 Detecting Scheme

As shown in Fig.3, we describe how to detect the embedded watermark message from the image that was attacked. The detecting scheme consists of three steps:

Step 1: SIFT features including SIFT descriptor $\boldsymbol{ds}$, location (t_1, t_2), scale s, orientation θ are extracted from the attacked image.

Step 2: SIFT feature points $\boldsymbol{x}$ are partitioned into k (the length of watermark sequence) clusters according to the Euclidean distances from SIFT descriptors $\boldsymbol{ds}$ to the centroids $\boldsymbol{ctd}$ which are from k-means clustering in watermark embedding scheme.

Step 3: The watermark message is extracted bit by bit in each cluster using twice voting strategy.

In Step 2, let d_{ji} represent the Euclidean distance from SIFT descriptor d_{ji} of SIFT point x_j to the centroid ctd_i. If d_{ji} is the shortest distance from x_j to all centroids of $\boldsymbol{ctd}$, the point x_j is grouped into cluster C_i. All feature points are grouped by the same method. Step 3 is elaborate. To improve the accuracy of the watermark extraction, we adopt twice voting strategy.

The $1st$ *voting:* First, we use the same method to find zone Ω and zone Ψ around a SIFT point x_j as watermark embedding scheme. Then, we extract the 2-dimension vector $\hat{v} = [\hat{v}_1 \quad \hat{v}_2]$ from Ω and Ψ. For each pixel $p_w(m,n)$ in Ω, determine the embedded bit with QIM. If $| p_w(m,n) - Q(p_w(m,n);1) | \leq | p_w(m,n) - Q(p_w(m,n);0) |$, the bit embedded in this pixel is "1". Else the watermark bit is ascertained to be "0". When geometrical distortions or/and common image processing attacks occur, even in a same sector disc, some pixels are detected to embed bit "1", and some pixels are detected to embed bit "0". Let $Num_\Omega(1)$ denote the number of pixels hiding bit "1" in Ω and $Num_\Omega(0)$ denote the number of pixels hiding bit "0" in Ω. The bit is extracted as:

$$\hat{v}_1 = \begin{cases} 1, & if\ Num_{\Omega}(1) \geq Num_{\Omega}(0) \\ 0, & if\ Num_{\Omega}(1) < Num_{\Omega}(0) \end{cases} \tag{3}$$

We can also extract $\hat{v}_2$ from zone Ψ using the same method. Let $\hat{v} = [\hat{v}_1 \quad \hat{v}_2]$. If $\hat{v} = [1 \quad 0]$, the watermark bit w_{x_j} embedded around SIFT point x_j is "1". If $\hat{v} = [1 \quad 0]$, the watermark bit w_{x_j} embedded around SIFT point x_j is "0". If $\hat{v} \neq [1 \quad 0]$ and $\hat{v} \neq [0 \quad 1]$, there is not watermark bit embedded around SIFT point x_j. So the method can not only extract the hiding bit around the SIFT point, but also abandon the points which are not presence in the original image.

The $2nd$ voting: For each SIFT point x_{ij} of cluster C_i, we use "the $1st$ voting" to get the bit embedded around the point. Let $Num_i(1)$ denote the number of SIFT point hiding bit "1" in cluster C_i and $Num_i(0)$ denote the number of SIFT point hiding bit "0" in cluster C_i. The ith bit of watermark sequence is extracted from cluster C_i as:

$$\hat{w}_i = \begin{cases} 1, & if\ Num_i(1) \geq Num_i(0) \\ 0, & if\ Num_i(1) < Num_i(0) \end{cases} \tag{4}$$

After every bit of watermark is extracted from every cluster using twice voting strategy, the watermark detecting process is finished.

The alteration of the pixels value under geometrical distortions or/and common image processing attacks is limited, because the attacked image should keep an acceptable level of visual quality. The watermark embedding and extraction are robust to such limited pixel value alteration, which attributes to the above QIM strategy and twice voting strategy.

4 Experimental Results

The watermark imperceptibility and robustness are evaluated by using 20 different 512×512 images as example images. In the experiments, 25-bits or 50-bit of watermark sequence was embedded into 25 or 50 clusters of SIFT features.

Fig. 4 shows that the PSNR values of the 20 watermarked images are between 34 dB to 45 dB. We can see from Fig.5 and Fig.6 that the watermark is very resistant to geometric deformations due to rotation and scaling. Fig. 7 shows the robustness of watermarking scheme to StirMark3.1 attacks [11] including geometric transformation and common image processing operations. From index 1 to 15, the attack is Median filter 2x2 and 3x3, Gaussian filtering 3x3, JPEG 90, 60 and 40, Removed 5 rows and 17 columns, Centered cropping 10%, Shearing (1%,1%), (0%,5%) and (5%,5%), Linear geometric transform(1.007, 0.01, 0.01, 1.012), (1.010, 0.013, 0.009, 1.011) and (1.013, 0.008, 0.011, 1.008), and Rand bend attack.

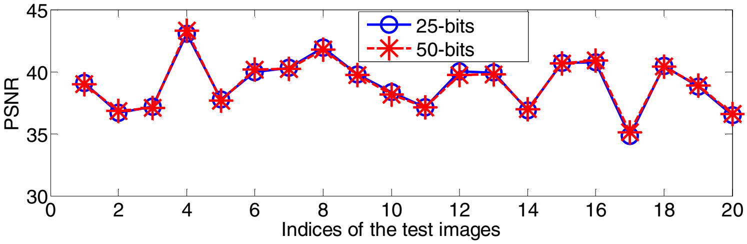


Fig. 4. The watermark distortion (In PSNR)

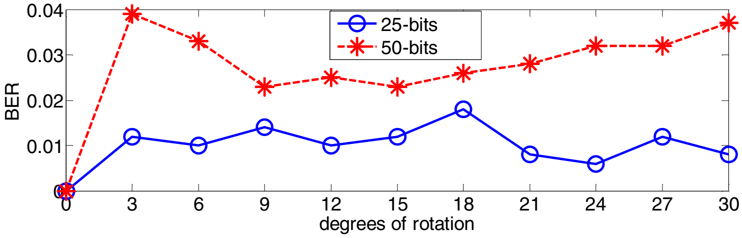


Fig. 5. Robustness to rotation

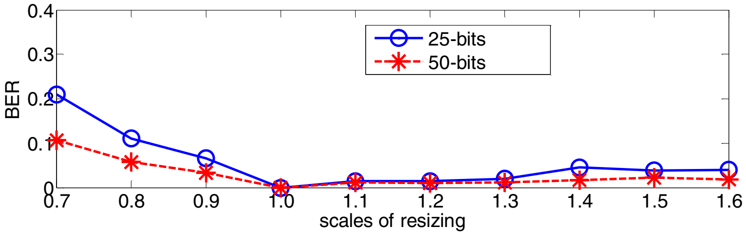


Fig. 6. Robustness to scaling

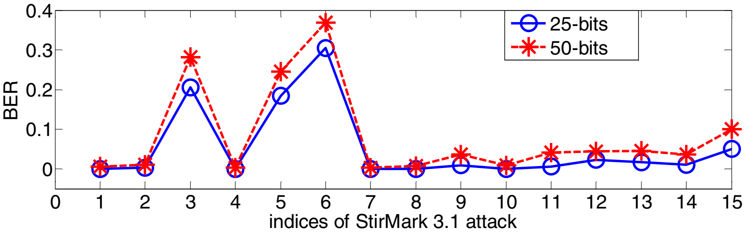


Fig. 7. Robustness to StirMark 3.1 attacks

5 Conclusion

In this paper, we have proposed a semi-blind watermarking scheme which is robust against geometrical distortions and common image processing attacks by the use of three techniques (SIFT, k-means clustering and twice voting strategy). First, The geometric invariant and muchness of SIFT is robust groundwork of watermarking scheme. Second, k-means clustering clusters SIFT features into *k* clusters, and the watermark message is embedded bit by bit in each cluster. So the robustness of watermarking will not only depend on individual feature point. Third, the exactness of the twice voting strategy ensures that the watermark message can be extracted accurately. The proposed watermarking scheme is robust against a wide variety of attacks as indicated in the experimental results. Our approach can be further improved by developing more robust clustering method than k-means clustering, so that the list of centroids would not be sent to the extracter.

Acknowledgments. Kind helps of Dr. Zhenfeng Zhu. This work was supported in part by NSF (No.60776794, No.90604032, and No.60702013), 973 program (No.2006CB303104), PCSIRT (No.IRT0707) and Beijing NSF (No.4073038)/

References

1. Bas, P., Chassery, J., Macq, B.: Geometrically invariant watermarking using feature points. IEEE Trans. Image Processing 11(9), 1014–1028 (2002)
2. Tang, C., Hang, H.: A feature-based robust digital image watermarking scheme. IEEE Trans. Image Processing 51(4), 950–959 (2002)
3. Lee, H.Y., Kim, H., Lee, H.K.: Robust image watermarking using local invariant features. Optical Engineering 45(3) 037002(1)-037002(11) (2006)
4. Wang, X., Wu, J., Niu, P.: A new digital image watermarking algorithm resilient to desynchronization attacks. IEEE Trans. Information Forensics and Security 2(4), 655–663 (2007)
5. Tian, H., Zhao, Y., Ni, R., Cao, G.: Geometrically robust image watermarking by sector-shaped partitioning of geometric-invariant regions. Optics Express 17(24), 21819–21836 (2009)
6. Li, L., Guo, B., Pan, J.-S., Yang, L.: Scale-Space Feature Based Image Watermarking in Contourlet Domain. In: Digital Watermarking. LNCS, pp. 88–102 (2009)
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
8. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. pp. 281–297. University of California Press, Berkeley (1967)
9. A tutorial on clustering algorithms, `http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/index.html`
10. Chen, B., Wornell, G.W.: Preprocessed and postprocessed quantization index modulation methods for digital watermarking. SPIE 3971, 48–59 (2000)
11. Petitcolas, F.A.P., Anderson, R.J., Kuhn, M.G.: Attacks on copyright marking systems. In: Aucsmith, D. (ed.) IH 1998. LNCS, vol. 1525, pp. 219–239. Springer, Heidelberg (1998)

DenVOICE: A New Density-Partitioning Clustering Technique Based on Congregation of Dense Voronoi Cells for Non-spherical Patterns

Jui-Fang Chang

Department of International Business,
National Kaohsiung University of Applied Sciences,
80778 Kaohsiung, Taiwan
rose@cc.kuas.edu.tw

Abstract. As data mining having become increasingly important, clustering algorithms with lots of applications have attracted a significant amount of research attention in recent decades. There are many different clustering techniques having been proposed. Some conventional partitioning-based clustering methods, such as K-means, may fail if a set of incorrect parameters is chosen, or breakdown when the objects consist of non-spherical patterns. Although density-based approaches, e.g. DBSCAN and IDBSCAN, could deliver better results, they may increase time cost when using large data bases. In this investigation, a new clustering algorithm termed DenVOICE is provided to circumvent the problems stated above. As a hybrid technique that combines density-partitioning clustering concept, the proposed algorithm is capable of resulting in precise pattern recognition while decreasing time cost. Experiments illustrate that the new algorithm can recognize arbitrary patterns, and efficiently eliminate the problem of long computational time when employing large data bases. It also indicates that the proposed approach produces much smaller errors than K-means, DBSCAN and IDBSCAN techniques in most the cases examined herein.

Keywords: data clustering, data mining, partitioning-based clustering, density-based clustering.

1 Introduction

Clustering in data mining is a critical technique of knowledge discovery. Its task is to locate the similar objects to the same group and dissimilar objects to different groups in order that identifies interesting correlations among data attributes. There are many data clustering approaches having been proposed in recent decades, and subsequently becoming a highly active discussion and attracting strong attention in the research field of data mining [1]-[8]. In general, clustering techniques can be grouped as partitioning, hierarchical, density-based, grid-based and mixed.

Most of existing partitioning-based clustering techniques have pattern recognition problems when using non-spherical patterns of cluster. Although density-based approaches could deliver better results, they may increase time cost when using large data bases. To eliminate the drawbacks of previous clustering techniques mention

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 136–144, 2010.

before, a new clustering algorithm termed **DenVOICE** based on congregation of **Den**se **VO**rono**I** **CE**lls for non-spherical patterns is proposed in this investigation by integrating with density-partitioning approaches. Experimental studies indicate that the proposed algorithm performs better than some conventional clustering algorithms.

2 Related Works

The limitations and merits of related clustering techniques is provided in this section.

K-means was proposed by McQueen in 1967 [1]. It involves three steps. (1) Choose randomly *K* objects as partition centers. (2) Locate each object to its closest centriod. (3) Re-calculate *K* partition centroids and repeat step (2) until the centroids convergence. K-means is the one of popular partitioning-based algorithm. However, it always converges to a local optimum. Moreover, K-means have annoying obstacles in precise pattern recognition when employing non-spherical or not the same size patterns.

To fulfill requirements for recognizing clusters with arbitrary shapes, density-based clustering algorithms have been proposed. Typically, those approaches identify the density of all objects and groups them based on density connectivity. DBSCAN is the one of well-know density-based approaches [2]. The algorithm first finds a core object as a set of neighbor objects consisting of more than a specified number of objects. All the objects reachable within a chain of overlapping core objects define a cluster. Although DBSCAN can accurately recognize different size clusters or non-spherical patterns and filter noise, the time complexity of the algorithm increases significantly as the number of data base increases.

Unlike DBSCAN, IDBSCAN, stands for Improved DBSCAN, samples a few representatives instead of considering all the objects in the neighborhood surrounding core object [8]. By adopting this sampling, the algorithm can significantly decrease memory usage and I/O cost while increasing the performance of DBSCAN. However, the time complexity of the algorithm is also high when using large data base.

3 The Proposed DenVOICE Algorithm

In this section, the proposed algorithm is presented in detail, which consists of four concepts, namely Determination of Bounding Zone, Forming of Candidate Dense Voronoi Cells, Noise Filtering for Dense Voronoi Cells and Hierarchic Congregation. Logically, the algorithm constructs several bounding zones through Determination of Bounding Zone and, then, divides each bounding zone into several candidate dense voronoi cells. Consequently, noise filtering process is employed to eliminate outliner within candidate dense voronoi cells. Finally, the proposed algorithm utilizes breadth-first search to congregate the dense voronoi cells to acquire whole clusters. The entire details about concepts are provided with four parts as follows:

(1) **Determination of Bounding Zone:** The main idea of this concept is to reduce computational time. Ideally, the proposed algorithm splits feature space into several bounding zones by assigning each object in data base to its closet center of bounding zone. Mathematically speaking, in the feature space there is a given object set $O = \{o_1, o_2, \ldots, o_m\}$, and a center set of bounding zones $C = \{c_1, c_2, \ldots, c_n\}$. For an object o_j picked from O, the center c_i determining process is defined as:

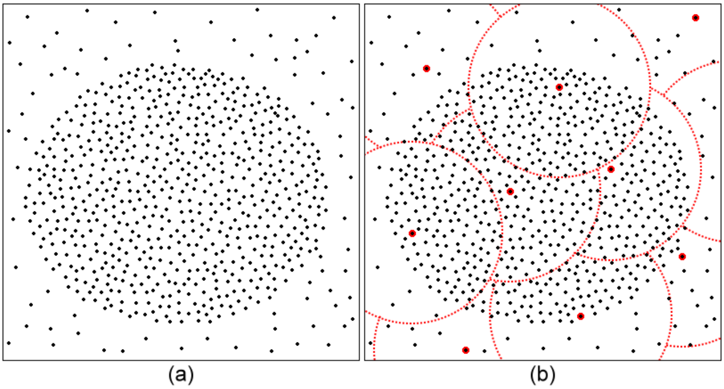


Fig. 1. (a) The original datasets (b) The situation after determining bounding zones

$$c_i = \{o_j, if\ C = \phi\} \tag{1}$$

or

$$c_i = \{o_j, if\ d(o_j, c_q) > R, c_q \in C, q = 1, \cdots, i-1\} \tag{2}$$

where R is the radius of the search zone and the Euclidean distance function $d(o_j,c_q)$ is defined as:

$$d(o_j, c_q) = \sqrt{\sum_{s=1}^{k} (o_{js} - c_{qs})^2} \tag{3}$$

where k indicates the feature dimension. If the center set C is empty or the Euclidean distance between the object o_j and each center c_q in C is greater than R, the object o_j is determined as the center of new bounding zone. Otherwise, the object o_j is assigned to the center c_q in C. As displayed in the diagram (b) of Fig. 1, each zone surrounded by dotted red circle is named "bounding zone" in which a set of larger red points indicates center of bounding zone.

(2) **Forming of Candidate Dense Voronoi Cells:** In this step, each bounding zone is divided into many candidate dense voronoi cells for later noise filtering process. At beginning of process, the proposed algorithm identifies a set of centers within each bounding zone, then, assigns each object to its closest center to form a set of candidate dense voronoi cells. The candidate dense voronoi cell can be defined as a voronoi cell that may either contain dense objects or noise objects. The determination for center set of candidate dense voronoi cells is presented as follows. There is a given object set within the bounding zone $P = \{p_1, p_2, \ldots, p_r\}$ and a center set of candidate dense voronoi cell $VC = \{vc_1, vc_2, \ldots, vc_h\}$. For each object p_j selected from P, if the distance between p_j and each center vc_i is greater than $R/2.5$, it is identified as center of candidate dense voronoi cell. Otherwise do nothing. Using previous example, displayed in the diagram (a) of Fig.2, the red solid circle represents a bounding zone that is full of areas surrounded by red dotted circle. This depicts the situation after center selecting. The goal of center determination is to make all the candidate dense voronoi

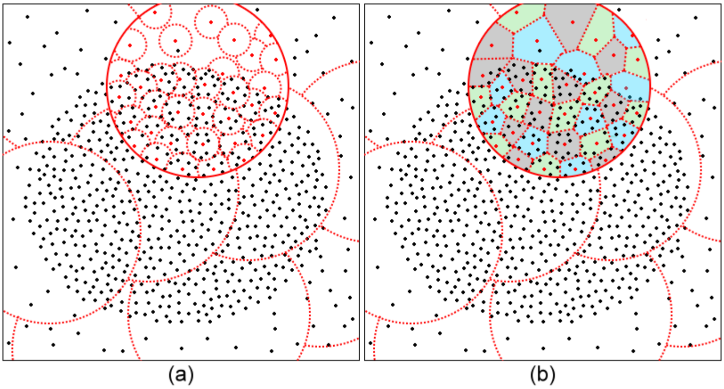


Fig. 2. The progress of constructing candidate dense voronoi cells

cells in the bounding zone much the same size. Consequently, all of objects in the bounding zone are assigned to its closest center thus a set of candidate dense voronoi cells is acquired, which is shown in the diagram (b) of Fig.2.

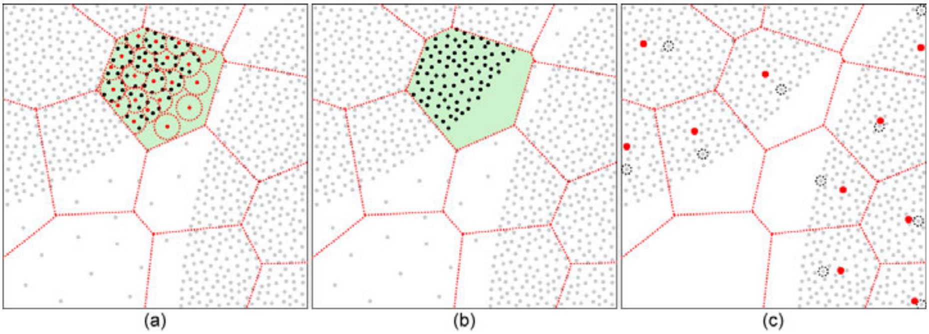


Fig. 3. Illustration of noise filtering (a) Forming of neighbor regions (b) Acquisition of dense voronoi cell (c) Recalculation of center set

(3) **Noise Filtering for Dense Voronoi Cells:** After forming processing of candidate dense voronoi cells, the noise filtering is performed on candidate dense voronoi cells to extract dense voronoi cells, which only consist of all the objects belonging to normal clusters. In order to filter the noise objects while recognizing normal objects, this investigation utilizes density function to filter the noise objects. In [3], influence function is defined as a mathematical description that the influence of an object has within its neighborhood, while the density function is defined as the sum of influence function of all objects in the region, and can be any arbitrary function. For simplicity, the Euclidean density function and Gaussian representation is applied [3]:

$$f_{Gauss}^{D}(o_i) = \sum_{j=1}^{N} e^{-\frac{d(o_i,o_j)^2}{2\sigma^2}}, \tag{4}$$

where N denotes the object number within the region, $d(o_i, o_j)$ represents the Euclidean distance between o_i and o_j, and σ is the standard deviation. It is clear that if the object locates on dense region, it will have relative high sum of influence. By this property, the objects whose sum of influence is less than a given threshold will be treated as noise and discarded. Although Gaussian density function can help the proposed algorithm to filter noise objects precisely, it may cause the time complexity of $O(n^2)$ due to performed on each object. In this work, the function is only performed on partial objects within the candidate dense voronoi cells instead of all the objects. The alternative is based on an assumption: if an object has high sum of influence derived from the formula, the neighbor objects surrounding the object also have high probability to gain the same fashion. Therefore, the time cost for influence computing will be significantly decreased. Consider a candidate dense voronoi cell displayed in the diagram (a) of Fig. 3 in which there are several red dotted circles every of which represents neighbor region. The density function is only performed on those red center objects, thus reducing time cost. Consequently, the neighbor regions containing noise objects will be pruned due to lower sum of influence. This is illustrated in diagram (b) of Fig. 3, which indicates a dense voronoi cell is gained by noise filtering. Similarly, if a whole candidate dense voronoi cell is full of noise objects, it will be pruned directly while running filtering process. At the end of this step, the proposed algorithm re-calculates the centroid for each dense voronoi cell. As represented in diagram (c) of Fig. 3, there are ten objects surrounded by black dotted circle indicating original location of centers and ten red objects denoting re-computed location of centroids. The purpose of this computation is to make sure that the re-calculated centroids locate on the central area of dense voronoi cells that is full of normal objects so that to find internally tangent radius which will be discussed later.

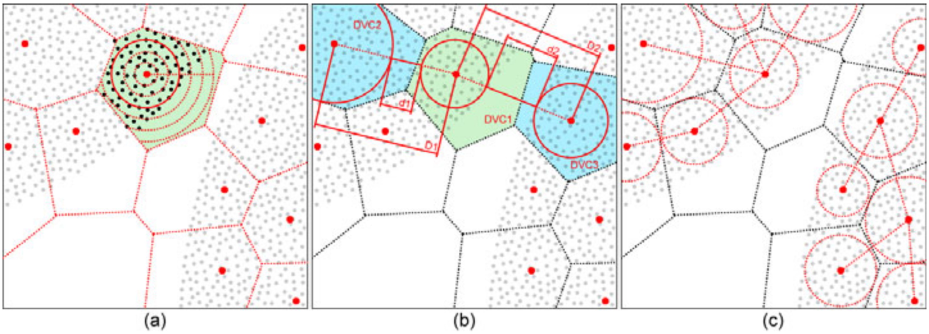


Fig. 4. Concept of hierarchic congregation

(4) **Hierarchic Congregation:** To acquire the complete clusters, each dense voronoi cell searches its neighbors and links them through the virtual edges to construct merging chain. The problem now is how to identify virtual edges short enough so that the neighbor dense voronoi cells can be merged with each other to avoid connecting the parts belonging to other clusters. To fulfill this requirement, the *ITR*, stands for the internally tangent radius, is introduced. In this investigation, the *ITR* is identified as a maximal distance centroid the center to the ring containing the highest number of objects among all others. To implement this definition, the proposed algorithm divides the

search radius *R* into several segments in order that forming many rings. Subsequently, the number of objects located on each ring is summed up. If a ring has the highest number of objects among others, the distance between the centroid and the ring is defined as *ITR*. This concept is displayed in the diagram (a) of Fig. 4, which indicates that the *ITR* is the distance between the centroid and 3$^{\text{rd}}$ ring from inside to outside, namely the red solid circle. Using previous example to explain utilization of *ITR*, consider three dense voronoi cells illustrated in the diagram (b) of Fig. 4, given as *DVC*1, *DVC*2 and *DVC*3. It is clear that both *DVC*1 and *DVC*2 belong to the same cluster, while *DVC*3 is partial of other cluster. However, the distance *D*1 between centroids of *DVC*1 and *DVC*2 is the same as the distance *D*2 between centroids of *DVC*1 and *DVC*3. It probably cause a wrong connection that *DVC*1 merges with *DVC*3 since *DVC*3 belonging other cluster. In this situation, the proposed algorithm redefines the merging distance as the original distance between the centroids of the dense voronoi cells subtracting *ITR* belonging to them separately. If the merging distance between those dense voronoi cells is less than a threshold, they will be merged into a cluster. In the diagram (b) of Fig. 4, the merging distance *d*1 between *DVC*1 and *DVC*2 is less than the merging distance *d*2 between *DVC*1 and *DVC*3. This indicates that *DVC*1 will merge with *DVC*2 instead of congregating *DVC*3. Filially, the whole cluster will be constructed by merging connectable dense voronoi cells with breadth-first search. On the other hand, a broken connection between the dense voronoi cells makes them into different clusters, which is illustrated in the diagram (c) of Fig. 4.

The complete algorithm is described as follows.

```
DenVOICE(DataSet,R,MinInf,MinMerDis)
  FOR i FROM 1 TO DataSet.Size DO
    BoundingZoneSet.determineBoundingZone(DataSet.get(i),R);
  END FOR

  DenVorCellSet = null;
  FOR i FROM 1 TO BoundingZoneSet.Size DO
    CandiDenVorCellSet = null;
    BoundingZone = BoundingZoneSet.get(i);
    FOR j FROM 1 TO BoundingZone.Size DO
      CandiDenVorCellSet.formCDVCell(BoundingZone.get(j),r);
    END FOR

    FOR j FROM 1 TO CandiDenVorCellSet.Size DO
      DenVorCell = filterNoise(CandiDenVorCellSet.get(j),MinInf);
      IF DenVorCell.Size > 0
        DenVorCell.recalcuCentroid();
        DenVorCell.calcuITR();
        DenVorCellSet.add(DenVorCell);
      END IF
    END FOR
  END FOR

  WHILE(TRUE) DO
    DenVorCell = DenVorCellSet.pop();
    IF DenVorCell == NULL
      END DenVOICE
    END IF
    changeClusterId(DenVorCell,ClusterId);
    BreadthFirstSearch(DenVorCell,ClusterId,MinMerDis);
    ClusterId++;
  END WHILE
END DenVOICE
```

DataSet represents the original dataset. R is search radius. MinInf denotes the minimal influence threshold, and MinMerDis indicates the minimal merging distance. The method is presented step by step below:

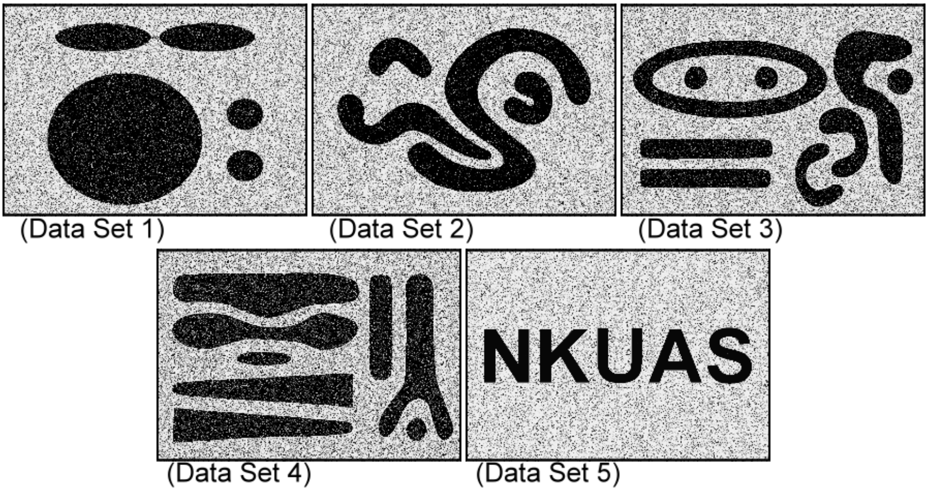
Step 1. For each object in the DataSet, examine whether the distance between the object and every center of bounding zone in the BoundingZoneSet is greater than the threshold R. In that case, set the object as a center of new bounding zone. Otherwise, assign the object to the bounding zone.

Step 2. Split each bounding zone in accordance with the threshold r, namely *R*/2.5, to form a set of candidate dense voronoi cells CandiDenVorCellSet.

Step 3. Filter noise by performing Gaussian density function on all of candidate dense voronoi cells to acquire a set of dense voronoi cells. Notably, a candidate dense voronoi cell will be pruned directly if it is full of noise objects. Consequently, the proposed algorithm re-calculates centroid and ITR for each dense voronoi cell.

Step 4. Repeat the following process until all of dense voronoi cells have been connected.

Step 5. For a dense voronoi cell, search and link its neighbors by breadth-first search in accordance with the minimal merging distance threshold MinMerDis to construct completed cluster.



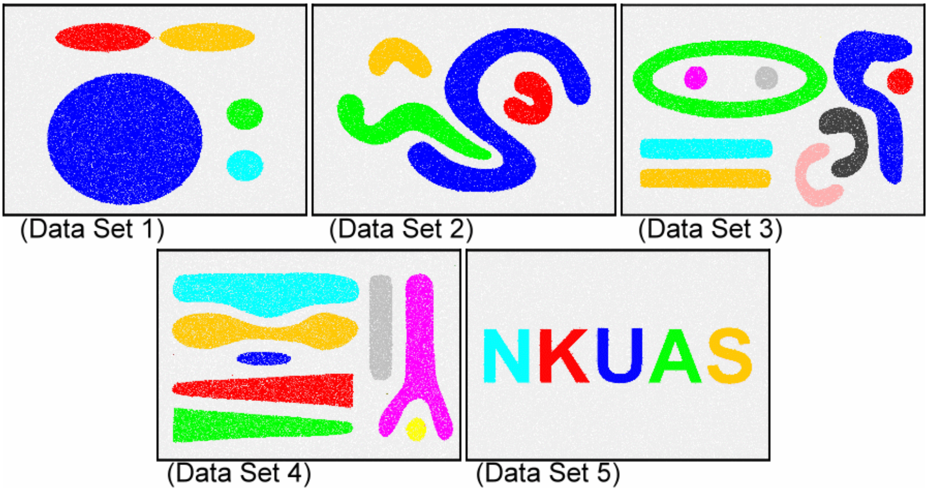
(Data Set 1) (Data Set 2) (Data Set 3)

(Data Set 4) (Data Set 5)

Fig. 5. The original datasets for experiments

4 Experiment Evaluation

In this investigation, The proposed algorithm was implemented in a Java-based program, and run on a personal computer with 4GB RAM, an Intel 2.4GHz CPU on Microsoft Windows XP professional Operational System. For simple visualization, five kinds of synthetic 2-D datasets with 11,500, 115,000, 230,000 and 575,000 objects are employed to evaluate the performance of the proposed algorithm. Fig. 5 represents the original datasets. Among these datasets, the patterns of dataset 1, 2, 3 and 4 were sampled from [2] and [5]. The results of the proposed algorithm were compared with that of K-means, DBSCAN and IDBSCAN. In the experiment, the clustering correctness rate (CCR) and noise filtering rate (NFR) are utilized. The former is the percentage of cluster objects correctly recognized by algorithm, while latter represents the percentage of noise objects correctly filtered by algorithm. Owing to the limitation of length, not all experimental results are provided. In the experiment



(Data Set 1) (Data Set 2) (Data Set 3)

(Data Set 4) (Data Set 5)

Fig. 6. The experimental results of the proposed algorithm

Table 1. Comparisons with DenVOICE, K-means, DBSCAN and IDBSCAN using 575,000 objects data sets with 15% noise; item 1 represents time cost (in seconds); item 2 denotes the CCR (%), while item 3 is NFR (%)

Algorithm	Item	DataSet-1	DataSet-2	DataSet-3	DataSet-4	DataSet-5
K-means	1	10.66	8.688	21.86	14.4	11.3
	2	57.421%	62.307%	48.175%	51.508%	67.73%
	3	0%	0%	0%	0%	0%
DBSCAN	1	N/A	N/A	N/A	N/A	N/A
	2	N/A	N/A	N/A	N/A	N/A
	3	N/A	N/A	N/A	N/A	N/A
IDBSCAN	1	192.375	158.563	181.047	224.062	82.437
	2	99.532%	99.361%	99.664%	99.387%	98.923%
	3	99.088%	98.48%	98.738%	98.985%	99.196%
DenVOICE	1	6.953	7.609	6.812	6.75	10.4
	2	99.715%	99.683%	99.49%	99.139%	99.842%
	3	98.921%	98.54%	99.246%	99.596%	99.236%

with lower size of data set, the results of DBSCAN indicate that it can accurately recognize any arbitrary pattern. However, the time cost of DBSCAN increases significantly when employing data set with 575,000 objects, which causes that Table 1 does not list the experiment results for DBSCAN (N/A means that the experiment were not performed). Although IDBSCAN increases the performance of DBSCAN when using large data sets, Table 1 demonstrates that the proposed algorithm outperforms IDBSCAN in time cost while having batter capability of handling non-spherical patterns than K-means. It is also indicates that the proposed algorithm usually yields more accurate results and has higher operational speed than K-means, DBSCAN and IDBSCAN.

5 Conclusion

This investigation presents a new density-partitioning clustering algorithm for data mining applications. By adopting the concept of congregating dense voronoi cells, the proposed clustering algorithm can handle clusters of both non-spherical and not of the same size. By integrating operation of partitioning-based clustering approach, the new algorithm can keep up performance of the density-based clustering techniques while shortening computational time. In addition, simulation results present that the new clustering approach performs better than some existing well-known methods such as the K-means, DBSCAN and IDBSCAN algorithms.

References

1. McQueen, J.B.: Some Methods of Classification and Analysis of Multivariate Observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297 (1967)
2. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, pp. 226–231 (1996)
3. Hinneburg, A., Keim, D.A.: An Efficient Approach to Clustering in Large Multimedia Databases with Noise. In: Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, pp. 58–65 (1998)
4. Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P.: Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. pp. 94–105. ACM Press, Seattle (1998)
5. Karypis, G., Han, E.-H., Kumar, V.: CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling. IEEE Computer, 68–75 (1999)
6. Krishna, K., Murty, M.N.: Genetic k-means algorithm. IEEE Transactions on Systems, Man, and Cybernetics-Part B: CYBERNETICS 29, 433–439 (1999)
7. Bandyopadhyay, S., Maulik, U.: An Evolutionary Technique Based on K-means Algorithm for Optimal Clustering in RN. Information Sciences 146, 221–237 (2002)
8. Borah, B., Bhattacharyya, D.K.: An Improved Sampling-Based DBSCAN for Large Spatial Databases. In: Proceedings of International Conference on Intelligent Sensing and Information, pp. 92–96 (2004)

Semantic Search Results Clustering

Krzysztof Strzalka and Aleksander Zgrzywa

Department of Information Systems, Wroclaw University of Technology, Poland
{Krzysztof.Strzalka,Aleksander.Zgrzywa}@pwr.wroc.pl

Abstract. Standard cluster algorithms learn without supervision (unsupervised learning), this approach allows users to impose the method of clustering concerning the preferred topic. The described method consists of the following stages: introduction of preferred topic, tagging the data based on topic and semantic relationships, using optimizing hierarchical clustering algorithm which uses the criterion function to determine the best partition method.

Keywords: document clustering, semantic tagging, ontology.

1 Introduction

Search results returned by search engines are not perfect. They contain many documents that are only loosely connected with our search topic. Often data that is interesting from our point of view is placed on further result pages, and as latest research shows – most people tend to look only at first page of results (so in case of Google with default settings it would be 10 results). Moreover, there is a problem with entities ambiguity, which is very good visible in case of people names or names of products/firms. Let us take "sun" as an example. There is Sun the firm, Sun the software, Sun the star, or Sun as the newspaper. Documents connected with all those meanings would be returned by search engine in mixed form, sorted by its internal ranking system (Page Rank in case of Google). Proposed approach functions as proxy between user and Google – it collects user's query and data delivered by Google, and next runs the clustering process to group distinct results from more than top 10 ULRs (more than the first page of results). This gives us a quick way of browsing through the results only within the interesting group (for example containing information only about the Sun as the star), and this raises the possibility of checking also those results that originally were not on the first page and probably would have been omitted. Clustering algorithm must work although at the beginning we don't know the cluster count (so the k-means method would be inappropriate), we don't know also the size of single clusters.

2 Related Works

Clusterization of a document means the use of cluster analysis in regard to results given by web search engines. It is about effective creation of sensible groups of documents with connected topics and their short description in a way understood to users. New approaches are used to cluster documents i.e. hierarchical clusterization, which is the most popular way of implementation of clusterization in search engines. The algorithm starts when there is one indivisible document collection. The documents are

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 145–151, 2010.

joined in pairs, pairs in fours etc., until we get one collection containing two classes of documents. Such collection is a tree structure which can be presented in a natural way with the use of dendrogram. However, this structure is not convenient to do comparisons, therefore its modification is more often used - so-called slices. To generate a faster and high quality clustering some researchers use the particle algorithm, but to enhance the performance of document clustering we use ontology. The COSA algorithm [5] draws the documents feature vector firstly, and then reflects words in vector for concept in concept tree of ontology. Concepts with little support will be gathered to parent concept, however concepts with big support will be separated into sub concepts. This algorithm reduces dimensionality of vectors.

3 Proposed Solution

Our wanted result is division of the search results into non overlapping clusters in such way, that for example different persons with the same name got into separate clusters. With mentioned earlier limitations, using hierarchical clustering algorithm seems to be very good solution, because it can function without a priori knowledge about the documents set. From the user's point of view, the whole process should be transparent and should not give noticeable waiting time overhead (compared to "pure" Google waiting time). Google delivers AJAX API interface, which allows us to send queries and receive result sets with additional information such as page's Page Rank (unfortunately, if we want to receive this value, the maximum returned result URL count is reduced to 65). Proposed solution taking query from user and ordering them in clusters contains the following steps:

1. Get query from user
2. Send query to search engine and get answers
3. Cluster the answers:
   a) Represent user query in form of multi attribute topic
   b) Mark documents based on topic and semantic connections
   c) Find inter document similarities
   d) Construct matrix of dissimilarity
   e) Clusterize document results with optimizing hierarchical algorithm.
4. Return grouped results

By limiting the clusterization with topic, and searching only through the words that are semantically connected to the topic we limited number of elements that need processed, and because of this dimension count of property vector is also reduced (normally it counts thousands of elements), and reduced dimensions means reduced processing time for each document.

3.1 Clusterization

Main element of described approach is the semiautomatic clustering based on user's query. To rise it's efficiency and shorten the running time we used semantic method, which allows us to skip analysis of data irrelevant to selected topic.

3.2 Ontology

To build ontology (the description of relations between words and their concepts) we used Hownet database. Hownet contains relations of hyponym, antonym, synonym type for English words. Moreover many words seems to have connections with specific semantic category. This connection creates “background” for the word. For example expression “boxing match” has “sport” as its background. This background can describe the document as well as words, and because of this we added to Hownet database expressions descriptions of backgrounds. Also the concept tree was added to main concept tree in ontology.

3.3 Document Tagging

Topic is the collection of words from user’s input. Attributes of topic are sets of concepts which describes them. If C is the collection of concepts in ontology, than attributes are subset of C - $\{p_1, p_2, \ldots, p_n\} \subset C$. Selection of attributes must allow the appropriate differentiation of found data. If we get back to the “sun” example, we can see that depending on what we have in mind by writing “sun”, we have different background. If in document we’ll find words like “star” or “space” the background would be “astronomy”. Next difference will be the place of existence/birth/living, additional element that strongly differentiates documents will be named entity – a word that exists only in a single place and have no representation in dictionary (usually proper names).

Doc, para, sen will be sets of words from document, paragraph and sentence. Distance (dis) between two words (w_1 i w_2) we can describe as:

$$dis(w_1, w_2) = \begin{cases} 1 & for\ w_1, w_2 \in sen, \\ 2 & for\ w_1, w_2 \in para, \\ 3 & for\ w_1, w_2 \in doc, \\ \infty & for\ other. \end{cases}$$

The distance equals 1 if words are in the same sentence, 2 if they are in the same paragraph, 3 if in the same document. Otherwise infinity is the distance.
Mutual information:

$$I(t_i, t_j) = \log\left(\frac{p(t_i, t_j)}{p(t_i) * p(t_j)}\right)$$

Can be used to calculate correlation between two words (t). Correlation is described as follows:

$$\xi(t_i, t_j) = \frac{I(t_i, t_j)}{dis(t_i, t_j)}$$

This equation takes into account not only frequency of words’ appearance, but also their distance. Correlation scale of words is directly proportional to their appearance frequency in document and reverse proportional to their distance – the more often words appear and the closer they are, the more correlated they are.

- We can say about words that they are correlated if word t_i describes semantic of t_j. Semantic can be derived to some degree from context.
- Documents contain several paragraphs and often we can assign one topic to one of paragraph.

Now let's use t_i as word, and T as topic. Distance between word and topic we can mark as d. Than word and topic are semantically correlated if $d \leq 2\,(d = dis(t_i,T))$. So words that do not appear together with topic in the same paragraph can't be used as tags.

If we mark document as s, T will be the topic, and $\{p_1,\ldots,p_n\}$ is the collection of attributes of T, than if we can map t_i for one of attributes, p_i and t_i are correlated with T, than $<t_i,f_i>$ can be added to vector of attributes p_i signed P_j. After checking all words in document we get set of vectors $\{P_1,\ldots,P_j\}$ (tags for whole s).

In general, the algorithm looks like:

```
a. Find paragraphs in which the topic appears → S
b. Create set from words in found paragraphs → L
c. For each word in L:
    a. For each attribute in A (A is the set of topic's attributes):
    b. Check in concept tree if given word from L can be mapped to
       given attribute from A
    c. If yes, add the word to vector
```

So aided by ontology, we check if given word fits semantically to the attribute of the topic, and if yes, we assume it's a tag, and add it to the vector. To reduce dimensions count of vector and better expose the background of topic, we add also backgrounds of added words. The complexity of such operations is $O(m \times n)$ where m is the Word Mount In document, n is the attribute Mount of topic.

3.4 The Difference Matrix

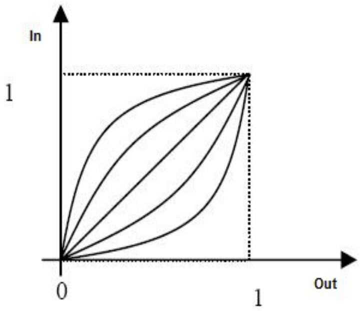
Document similarity can be written as vector similarity:

$$sim(v_1,v_2) = \frac{\sum_{k=1}^{m} W_{1k} \times W_{2k}}{\sqrt{\sum_{k=1}^{m} W_{1k}^2 \times \sum_{k=1}^{m} W_{2k}^2}}$$

The difference between them as:

$$dissim(d_1,d_2) = 1 - \frac{\sum_{i=1}^{m} f(l_i, sim(v_{1i},v_{2i}))}{n}$$

Strenghten function (f) as:



It's 1 minus the average of all similarities between two documents, where v is the vector of attributes p of document d. $sim(v_{1i}, v_{2i})$ is the measure of similarity of two vectors, and f is used to change importance (strength) of chosen similarities. The l_i parameter is used to select strength curve. Strength curve allows to rise importance of named entities, because in case of named entity even small similarity to it strongly shows where to assign the document.

3.5 Criterion Function for Optimal Dividing

Proposed criterion function used to find optimal clusterization is based on similarity metric inside one cluster and difference between different clusters and is given as:

$$dissum = \frac{\sum_{i=1}^{n} \frac{\sum_{p,q \in C_i} dis(p,q)}{|C_i|} + \sum_{C_i, C_j \in C} dis(C_i, C_j)}{|C|}$$

Where $C = \{C_1, \dots, C_n\}$ is the solution of problem, $\{C_1, \dots, C_n\}$ are groups of documents (clusters), $|C_i|$ is the count of C_i elements and $|C|$ is the overall count of clusters. Sum of distances inside i-th cluster is expressed by $\sum_{p,q \in C_i} dis(p, q)$ and $\sum_{C_i, C_j \in C} dis(C_i, C_j)$ i the sum of distances between i-th and j-th luster. Distance between two clusters is the biggest distance between their elements. Dissum function gets maximal values in two cases – when all documents are assigned to one group, or when every document gets into separate group. Both cases aren't optimal. We can see, that when we merge two groups, distance inside new, derived group will rise, but distance between groups will fall. If merging operation is optimal, outcome of dissum function will be lowered. That's why optimizing the clustering process is done by minimizing dissum.

Clustering optimization begins from assigning separate cluster for each document, next it merges the most similar clusters and checks if dissum has fallen. If yes, it saves the result and continues merging process until only one cluster (with all documents) is left. Complexity of such operation is O(m) where m is the count of documents.

4 Tests

To rate the accuracy of clusterization F-measure is used:

$$F = \frac{2 * prec * rec}{prec + rec}$$

Where *prec* is the precision given as:

$$prec = \frac{\sum_{i=1}^{k} \max\left(p\left(C_i, C_j'\right) | j = 1, \dots, n\right)}{k}$$

And *rec* is the recall:

$$rec - \frac{\sum_{i=1}^{k} \max\left(r\left(C_i, C_j'\right) | j = 1, \dots, n\right)}{k}$$

To show how close the algorithm was to the ideal resolution relative terror is used:

$$|\delta| = \frac{|\Delta|}{k} * 100\%$$

Where $|\Delta|$ is the difference between amount of clusters created by algorithm (k) and the proper value (given by expert).

First the algorithm was compared to standard TF-IDF on set of user queries such as: “sun”, “virus”, “trojan horse”, “worm”. The average F-measure of both algorithm is shown in fig. 1

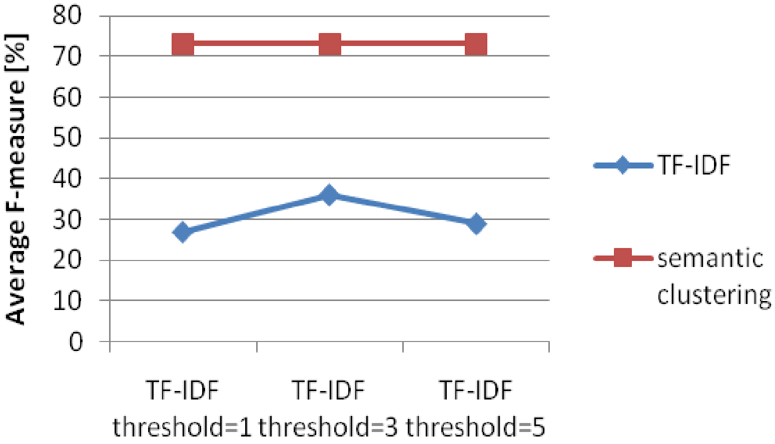


Fig. 1. Average F-measure for given problems

Than the average error of both algorithms was compared (Fig. 2.).

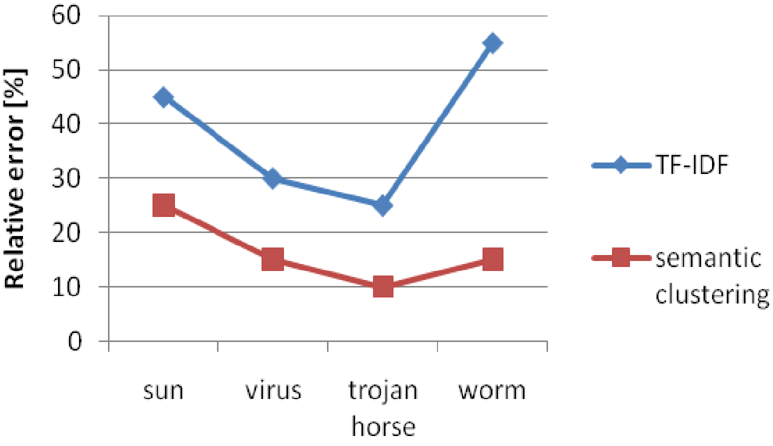


Fig. 2. Relative errors

5 Conclusion

The final results shows, that semantic approach to clusterization can give good results, and let the user see more than just the first ten usually watched links. By grouping search results into coherent clusters, user has the ability to easily discard unwanted information, and search only within interesting subgroup of (at the beginning) ambiguous names. Proposed approach is much more accurate than standard clusterization algorithm. But there's still several problems left – how to order results inside of cluster and present them to the user. Currently we order intracluster results by their Page Rank given by Google, but if there will be more than one-two pages of such results, the user mat get bored and omit parts of useful links. The other problem is with algorithm overhead, if we want to group more than first 65 results, the time needed to parse additional data will rise. It could be interesting to recluster grouped documents, to reveal many similar, or even identical documents in one group, so the user would not have to browse through several identical descriptions (for example datasheets of electronic equipment).

References

1. Zhao, Y., Krypis, G.: Topic-Driven Clustering for Document Datasets. In: SIAM International Conference on Data Mining, pp. S358–S369 (2005)
2. Wagstaff, K., Cardie, C., Rogers, S., Schroedl, S.: Constrained K-means Clustering with Background Knowledge. In: Eighteenth International Conference on Machine Learning, pp. S577–S584 (2001)
3. Basu, S., Bilenko, M., Mooney, R.J.: A Probabilistic Framework for Semi-Supervised Clustering. In: ACM SIGKDD international conference on Knowledge discovery and data mining, pp. S59–S68 (2004)
4. Qiaozhu, M., Dong, X., Hong, C., Jiawei, H., Cheng Xiang, Z.: Generating Semantic Annotation for Frequent Pattern with Context Analysis. In: 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. S337–S346 (2006)
5. Hotho, A., Staab, S., Maedche, A.: Ontology-based Text Clustering. In: IJCAI 2001 Workshop Text Learning: Beyond Supervision (2001)
6. Hung, C., Wermter, S., Smith, P.: Hybrid Neural Document Clustering Using Guided Self-Organization and WordNet. IEEE Intelligent Systems 19(2), S68–S77 (2004)
7. Fung, B.C.M., Wang, K., Ester, M.: Hierarchical Document Clustering Using Frequent Itemsets. In: Proceedings of the SIAM International Conference on Data Mining (2003)
8. Huang, C., Wermter, S., Smith, P.: Hybrid Neural Document Clustering Using Guided Self-Organization and WordNet. IEEE Intelligent Systems (March 2004)
9. Shi-Qi, Z., Ting, L., Sheng, L.: A Topical Document Clustering Method. Journal of Chinese Information Processing 21(2) (March 2007)
10. Zhao, Y., Karypis, G.: Topic-driven Clustering for Document Dataset. In: Proc. SIAM Data Mining Conference (2005)
11. Basu, S., Bilenko, M., Monney, R.J.: A probabilistic framework for semi-supervised clustering. In: Proc. of the 10 th Int'l Conference on Knowledge Discovery and Data Mining (2004)
12. Zhao, Y., Karypis, G.: Empirical and theoretical comparisons of selected criterion functions for document clustering. Machine Learning (2004)

Trend Detection on Thin-Film Solar Cell Technology Using Cluster Analysis and Modified Data Crystallization

Tzu-Fu Chiu[1], Chao-Fu Hong[2], and Yu-Ting Chiu[3]

[1] Department of Industrial Management and Enterprise Information, Aletheia University, Taiwan, R.O.C.
chiu@mail.au.edu.tw

[2] Department of Information Management, Aletheia University, Taiwan, R.O.C.
cfhong@mail.au.edu.tw

[3] Department of Information Management, National Central University, Taiwan, R.O.C.
gloria@mgt.ncu.edu.tw

Abstract. Thin-film solar cell, one of green energies, is growing at a fast pace with its long-lasting and non-polluting natures. To detect the potential trends of this technology is essential for companies and relevant industries so that the competitive advantages of companies can be retained and the developing directions of industries can be perceived. Therefore, a research framework for trend detection has been formed where cluster analysis is employed to perform the similarity measurement, and data crystallization is adopted to conduct the association analysis. Consequently, the relation patterns were identified from the relations among companies, issue years, and techniques. Finally, according to the relation patterns, the potential trends of thin-film solar cell were detected for companies and industries.

Keywords: trend detection, cluster analysis, data crystallization, thin-film solar cell, patent data.

1 Introduction

It is essential for companies and stakeholders to realize the situation of a certain technology so that companies can review their development directions of products and stakeholders can examine the suitability of their relevant investments. In technological information, up to 80% of the disclosures in patents are never published in any other form [1]. Therefore, patent analysis has been recognized as an important task for companies and industries. Through appropriate analysis, technological details and relations, business trends, novel industrial solutions, or making investment policy can be achieved [2]. Apart from those existing methods, a research framework of cluster analysis and data crystallization will be built for patent analysis in order to detect the potential trends of thin-film solar cell in the U.S.

2 Related Work

As this study is aimed to explore the technological trends of thin-film solar cell, a research framework needs to be constructed via a consideration of cluster analysis and

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 152–161, 2010.

data crystallization. In order to handle the textual nature of patent data (especially the abstract, description, and claim fields), cluster analysis is adopted to measure the similarity character between patent documents, while data crystallization is employed to measure the association character between the consisting terms of patent documents. Subsequently, the research framework will be applied to the domain of trend detection. Therefore, the related areas of this study would be trend detection, patent data, thin-film solar cell, cluster analysis, and data crystallization.

2.1 Trend Detection

Emerging trend detection (ETD) is to detect a topic area that is growing in interest and utility over time [3]. The ETD process can be viewed mainly in three phases: topic representation, identification of features, and verification of interest and utility [4]. It takes as input a collection of textual data and identifies topic areas that are either novel or are growing in importance within the corpus [3]. In this study, a research framework, formed by cluster analysis and data crystallization, will be used for conducting the trend detection upon thin-film solar cell via patent data.

2.2 Patent Data and Thin-Film Solar Cell

A patent document is similar to a general document, but includes rich and varied technical information as well as important research results [1]. Patent documents can be gathered from a variety of sources, such as the Intellectual Property Office in Taiwan (TIPO), the United States Patent and Trademark Office (USPTO), the European Patent Office (EPO), and so on. A patent document contains numerous fields, such as: patent number, title, abstract, issue date, application date, application type, assignee name, international classification (IPC), US references, claims, description, etc.

Solar cell, a sort of green energy, is clean, renewable, and good for protecting our environment. It can be mainly divided into two categories (according to the light absorbing material): crystalline silicon (in a wafer form) and thin films (of other materials) [5]. A thin-film solar cell (TFSC), also called a thin-film photovoltaic cell (TFPV), is made by depositing one or more thin layers (i.e., thin film) of photovoltaic material on a substrate [6]. The most common materials of TFSC are amorphous silicon or polycrystalline materials (such as: CdTe, CIS, and CIGS) [5]. In recent years (2003-2007), total PV production grew in average by almost 50% worldwide, whereas the thin film segment grew in average by over 80% and reached 400 MW or 10% of total PV production in 2007 [7]. Therefore, thin film is the most potential segment with the highest production growth rate in the solar cell industry, and it would be appropriate for academic and practical researchers to contribute efforts to explore and understand this rapid growing technology.

2.3 Cluster Analysis

Cluster analysis divides data into groups (clusters) that are meaningful, useful, or both [8]. Classes, or conceptually meaningful groups of objects that share common characteristics, play an important role in how people analyze and describe the world. Clusters are potential classes and cluster analysis is the study of techniques for automatically finding classes [8]. Cluster analysis will be employed in this study for measuring the similarity nature of documents, so as to divide patent data into groups and to form the technical topics of thin-film solar cell.

2.4 Modified Data Crystallization

Data crystallization, a technique of chance discovery, is used to detect the unobservable (but significant) events via inserting these unobservable events as dummy items into the given data set [9]. The unobservable events and their relations with other events are visualized by applying the KeyGraph which is a two-dimension undirected graph, consisting of event clusters, visible events, and chances [10]. A generic data crystallization algorithm can be summarized as follows [11]: (a) event identification: all events appearing in all basket data $B = \{b_i\}$ ($i \in [0, |B| - 1]$) are picked up; (b) clustering: partitions a data set into subsets, so that the data in each subset share some similarity; (c) dummy event insertion: a dummy event DE_i is inserted into a basket b_i, which results in $b_i \rightarrow \{\{e_i\}, DE_i\}$; (d) co-occurrence calculation: the co-occurrence between a dummy event and clusters is measured by equation (1), where C is the specific number of clusters; and (e) topology analysis: the dummy events having co-occurrence with multiple clusters are picked up and visualized with KeyGraph.

$$Co(DE_i, C) = \sum_{j=0}^{|C|-1} \max_{e_k \in c_j} Ja(DE_i, e_k) \tag{1}$$

Data crystallization was originally proposed to deal with unobservable events (i.e., dummy events) so as to emerge the hidden clues from existing circumstances via judging the unknown relations [12]. This method has been modified by the authors to insert extra data elements (e.g., assignee, country-code, or issued-date fields) as dummy events into the initial data records (i.e., the abstract field), so that the relations between the extra data elements and existing clusters can come out and be observed [13]. In this study, the modified data crystallization will be adapted to handle two dummy events (i.e., two extra data elements) at the same time, in order to display the combination of two different kinds of relations in one diagram simultaneously.

3 A Research Framework for Trend Detection

As this study is attempted to explore the potential trends in the thin-film solar cell technology, a research framework for trend detection, based on the cluster analysis and modified data crystallization, has been developed and shown in Fig. 1. It consists of four phases: data preprocessing, cluster analysis, modified data crystallization, and new findings; and will be described in the following subsections.

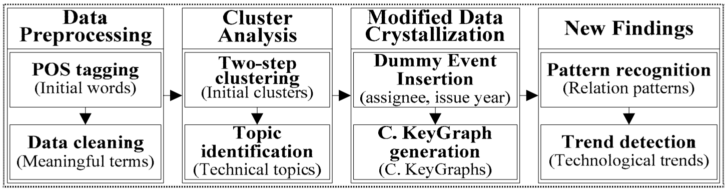


Fig. 1. A research framework for trend detection

3.1 Data Preprocessing

In first phase, the patent data of thin-film solar cell (during a certain period of time) will be downloaded from the USPTO [14]. For considering an essential part to represent a complex patent data, the abstract, assignee, and issue date fields are selected as the objects for this study. Afterward, two processes, POS tagging and data cleaning, will be executed to clean up the source textual data.

(1) POS Tagging: An English POS tagger (i.e., a Part-Of-Speech tagger for English) from the Stanford Natural Language Processing Group [15] will be employed to perform word segmenting and labeling on the patent documents (i.e., the abstract field). Then, a list of proper morphological features of words needs to be decided for sifting out the initial words.

(2) Data Cleaning: Upon these initial words, files of n-grams, stop words, and synonyms will be built so as to combine relevant words into compound terms, to eliminate less meaningful words, and to aggregate synonymous words. Consequently, the meaningful terms will be obtained from this process.

3.2 Cluster Analysis

Second phase is designed to conduct the cluster analysis via two-step clustering and topic identification so as to gain the technical topics.

(1) Two-step Clustering: In order to perform the cluster analysis, a two-step clustering is adopted from SPSS Clementine for grouping patent documents into clusters [16]. Two-step clustering is a scalable cluster analysis algorithm designed to handle very large data sets. It can handle both continuous and categorical variables (or attributes). It requires only one data pass. It has two steps: (a) to pre-cluster the cases (or records) into many small sub-clusters; (b) to cluster the sub-clusters resulting from pre-cluster step into the desired number of clusters [16]. These clusters are regarded as the initial clusters.

(2) **Topic Identification:** An initial cluster will be named via summarizing the title field of its composed patent documents and checking on the degree of its relevance to a certain technological subfield in the domain knowledge. Each named cluster is then identified as a technical topic and will be utilized in the following phases.

3.3 Modified Data Crystallization

Third phase, including dummy event insertion and crystallized KeyGraph generation, is used to find out the relations among subtopics, companies, and issue years for pattern recognition.

(1) Dummy Event Insertion: In order to conduct the data crystallization, two dummy events, namely assignee and issue date fields, need to be inserted into a patent document of the data subset which is contained in every topic (i.e., derived from the above cluster analysis).

(2) Crystallized KeyGraph Generation: After the dummy event insertion, data crystallization will be triggered to generate a crystallized KeyGraph for each topic. Firstly, using the updated data subset, a KeyGraph will be drawn to show the individual nodes, grouping nodes, and dummy nodes. Secondly, the inner clusters will be obtained by setting a threshold to the number of grouping nodes (e.g., the threshold is set to no less than 3 in this study). Thirdly, the inner cluster will be named via combining the meaning of its composed nodes and via referring to the domain knowledge, and will be regarded as a technical subtopic. Finally, this crystallized KeyGraph will be utilized to observe the relations between (or among) dummy nodes and inner clusters.

3.4 New Findings

In last phase, pattern recognition will be used to figure out the relation patterns according to the relations among subtopics, companies, and issue years; while trend detection will be utilized to explore the potential trends based on the relation patterns.

(1) Pattern Recognition: Referring to the crystallized KeyGraphs, the relations among subtopics, companies, and issue years will be applied to recognize the relation patterns by observing the linkages from a year to subtopics, a year to companies, a company to subtopics, and a subtopic to companies. Firstly, the linkages from a year to subtopics will be used to construct the "a year relating to multiple techniques" pattern (Type A). Secondly, the linkages from a year to companies will be applied to construct the "a year relating to multiple companies" pattern (Type B). Thirdly, the linkages from a company to subtopics will be utilized to construct the "a company relating to multiple techniques" pattern (Type C). Lastly, the linkages from a subtopic to companies will be utilized to construct the "a technique relating to multiple companies" pattern (Type D).

(2) Trend Detection: According to the relation patterns from the above pattern recognition and based on the domain knowledge, the potential trends of thin-film solar cell will be detected so as to facilitate the decision-making for managers stakeholders.

4 Experimental Results and Explanation

The experiment has been implemented according to the research framework. The experimental results will be explained in the following four subsections: result of data preprocessing, result of topic generation, result of crystallized KeyGraph generation, and result of pattern recognition and new findings.

4.1 Result of Data Preprocessing

As the aim of this study is to visualize the relations among issue years, technical subtopics, and companies, as well as to explore the potential trends, the patent documents of thin-film solar cell were the target data for this experiment. Mainly, the abstract, assignee, and issue year fields of patent documents were used in this study. Therefore, 160 issued patent items during year 2000 to 2009 were collected from USPTO, using key words: "'thin film' and ('solar cell' or 'solar cells' or 'photovoltaic cell' or 'photovoltaic cells' or 'PV cell' or 'PV cells')" on "title field or abstract field". The POS tagger was then triggered to do the data preprocessing upon the collected 160 patent items. Consequently, the patent documents during year 2000 to 2009 were cleaned up and the meaningful terms were obtained.

4.2 Result of Topic Generation

Using the meaningful terms from data preprocessing, the technical topics (in Table 1) were generated via the TwoStep clustering function of SPSS Clementine (version 10.1). In the table, six initial clusters were identified with the number of composed records for Cluster-1 to Cluster-6: 12, 33, 24, 21, 51, 19 respectively. According to the title fields and domain knowledge, these six clusters were named as: 'single-crystal & organic-semiconductor', 'light-absorbing-layer & CVD-method', 'roll-to-roll-process & heat-treatment', 'amorphous-film & plasma-deposition', 'microcrystalline-silicon & annealing-process', and 'sputtering & compound-thin-film', and regarded as the technical topics.

Table 1. Technical topics of thin-film solar cell

id	*name of technical topic*	*number of composed records*
topic-1	'single-crystal & organic-semiconductor'	12
topic-2	'light-absorbing-layer & CVD-method'	33
topic-3	'roll-to-roll-process & heat-treatment'	24
topic-4	'amorphous-film & plasma-deposition'	21
topic-5	'microcrystalline-silicon & annealing-process'	51
topic-6	'sputtering & compound-thin-film'	19

4.3 Result of Crystallized KeyGraph Generation

After adding dummy events (i.e., assignee and issue year fields) to the data subset of a topic, a crystallized KeyGraph of that topic (e.g., 'single-crystal & organic-semiconductor') was drawn so as to demonstrate the inserted dummy events (6 years and 8 companies) and containing inner clusters (6 subtopics) within a topic (in Fig. 2). The other crystallized KeyGraphs for 'light-absorbing-layer & CVD-method', 'roll-to-roll-process & heat-treatment', 'amorphous-film & plasma-deposition', 'microcrystalline-silicon & annealing-process', and 'sputtering & compound-thin-film' were produced successively. The crystallized KeyGraph of 'light-absorbing-layer & CVD-method' was another example shown in Fig. 3. The other four crystallized KeyGraphs were omitted because of the page limit.

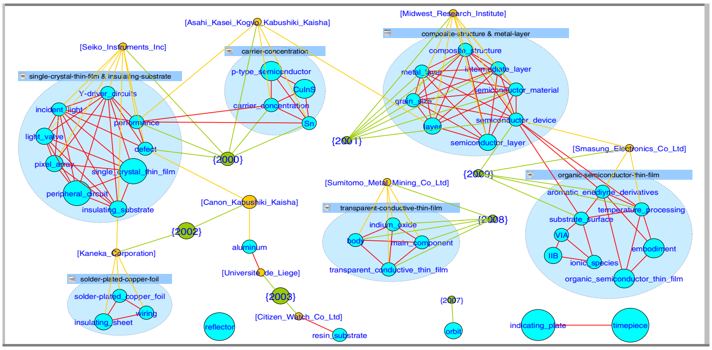


Fig. 2. A crystallized KeyGraph of 'single-crystal & organic-semiconductor'

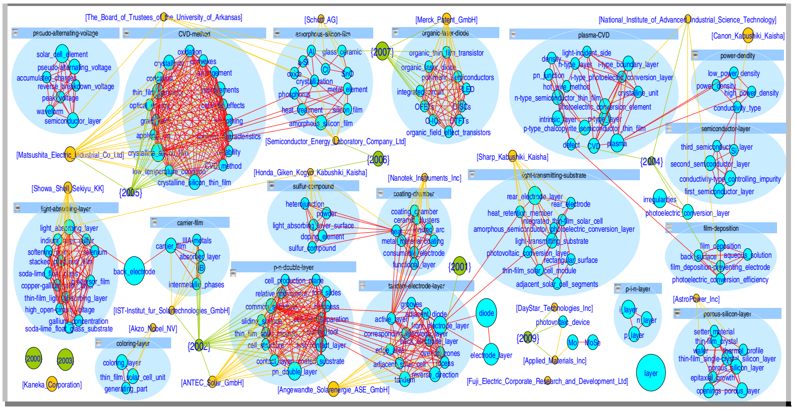


Fig. 3. A crystallized KeyGraph of 'light-absorbing-layer & CVD-method'

4.4 Result of Pattern Recognition and Trend Detection

According to the above crystallized KeyGraphs, the linkages from dummy events (i.e., years and companies) to inner clusters (i.e., subtopics) were used to recognize the relation patterns. Firstly, the relation patterns, Type A and Type B, linking from year to subtopics and companies, were summarized yearly in Table 2. Secondly, the relation patterns, Type C and Type D, were applied to define the focused techniques and significant companies if a technique linking to two or more companies and a company linking to two or more techniques, also expressed in Table 2 (in boldface and italics).

In accordance with the above Table 2 and based on domain knowledge, the potential trends of each year (2000-2009) and some specific topics were explored and depicted below.

(1) Trends from 2000 to 2009: Referring to Table 2, trends of year 2000 to 2009 were explained as follows. In 2000, the focused techniques were single-crystal-thin-film & insulating-substrate, flexible-material, and compound-thin-film & photoelectron-spectroscopy, while the significant company was Asahi-Kasei-Kogyo-Kabushiki-Kaisha. In 2001, the focused techniques were composite-structure & metal-layer, coating-chamber, and compound-thin-film & photoelectron-spectroscopy, while the significant company was Canon-Kabushiki-Kaisha. In 2002, the focused techniques were p-n-double-layer and display-device, while the significant companies were Kaneka-Corporation, Angewandte-Solarenergie-ASE-GmbH, and Canon-Kabushiki-Kaisha. In 2003, the focused technique was flexible-material, while the significant company was Sharp-Kabushiki-Kaisha. In 2004, the focused technique was laser-pulse, while the significant companies were National-Institute-of-Advanced-Industrial-Science-Technology and Kaneka-Corporation. In 2005, the focused techniques were flexible-material and RF-sputtering & CIGS, while the significant company was The-Board-of-Trustees-of-the-University-of-Arkansas. In 2006, the focused techniques were light-absorbing-layer and microcrystalline-silicon & deposition-gas, while the significant companies were Honda-Giken-Kogyo-Kabushiki-Kaisha and Midwest-Research-Institute. In 2007, the focused

Table 2. Trends of thin-film solar cell (2000-2009)

		Topic-1	Topic-2	Topic-3	Topic-4	Topic-5	Topic-6
2000	Sub-topic	*single-crystal-thin-film & insulating-substrate* carrier-concentration		SOI(silicon on insulator)-substrate			*flexible-material* *compound-thin-film & photoelectron-spectroscopy*
	Co.	Seiko_Instruments_Inc *Asahi_Kasei_Kogyo_Kabushiki_Kaisha*		Canon_Kabushiki_Kaisha	Showa_Shell_Sekiyu_KK	Matsushita_Electric_Industrial_Co_Ltd	The_Regents_of_the_University_of_California
2001	Sub-topic	*composite-structure & metal-layer*	*coating-chamber* tandem-electrode-layer		reflective-film	LED (light-emitting-diode) & multi-layer-porous-structure	*compound-thin-film & photoelectron-spectroscopy*
	Co.	Midwest_Research_Institute				*Canon_Kabushiki_Kaisha*	The_Regents_of_the_University_of_California
2002	Sub-topic		*p-n-double-layer* carrier-film		*display-device* thin-film-epitaxial-layer		glass-substrate
	Co.	*Kaneka_Corporation* Canon_Kabushiki_Kaisha	*Angewandte_Solarenergie_ASE_GmbH*		*Canon_Kabushiki_Kaisha*		Siemens_and_Shell_Solar_GmbH *Kaneka_Corporation*
2003	Sub-topic				polysilicon-thin-film etching-process		*flexible-material*
	Co.	Universite_de_Liege Citizen_Watch_Co_Ltd			ANTEC_Solar_GmbH *Sharp_Kabushiki_Kaisha*		Dutch_Space_BV
2004	Sub-topic		film-deposition plasma-CVD	In-Ga-original-substrate	*laser-pulse*		
	Co.		*National_Institute_of_Advanced_Industrial_Science_Technology*		California_Institute_of_Technology *Kaneka_Corporation*		Sharp_Kabushiki_Kaisha
2005	Sub-topic		CVD-method			silicon & conductive-film	*flexible-material* *RF-sputtering & CIGS*
	Co.		Matsushita_Electric_Industrial_Co_Ltd *The_Board_of_Trustees_of_the_University_of_Arkansas*	H_C_Starck_Ceramics_GmbH_Co_KG		Kaneka_Corporation	Cymbet_Corporation Miasole
2006	Sub-topic		*light-absorbing-layer*	In-Ga-original-substrate		*microcrystalline-silicon & deposition-gas*	
	Co.		Semiconductor_Energy_Laboratory_Company_Ltd *Honda_Giken_Kogyo_Kabushiki_Kaisha*	Interuniversitair_Microelektronica_Centrum		*Midwest_Research_Institute*	Hahn-Meitner-Institut_Berlin_GmbH
2007	Sub-topic		amorphous-silicon-film organic-laser-diode	*heating-jig-electrodeposition*	*light-emitting-diode*	TCO-(transparent-conductive-oxide)-film *polycrystalline-organic-thin-film*	
	Co.		Merck_Patent_GmbH	*Davis_Joseph_Negley*		Asahi_Glass_Company_Limited Ascent_Solar_Technologies_Inc *SolarFlex_Technologies_Inc*	The_Aerospace_Corporation
2008	Sub-topic	transparent-conductive-thin-film			*display-device*	TCO-(transparent-conductive-oxide)-film	organic-material & reentrant-substrate
	Co.	Sumitomo_Metal_Mining_Co_Ltd			Hitachi_Ltd	Asahi_Glass_Company_Limited DayStar_Technologies_Inc	*The_Trustees_of_Princeton_University* Sixtron_Advanced_Materials_Inc
2009	Sub-topic	*composite-structure & metal-layer* organic-semiconductor-thin-film		*In-Ga-original-substrate*	defective-region		*RF-sputtering & CIGS*
	Co.	*Smasung_Electronics_Co_Ltd*	DayStar_Technologies_Inc Applied_Materials_Inc	Nanosolar_Inc *Solopower_Inc*	Oerlikon_Trading_AG_Trubbach		Miasole

techniques were heating-jig-electrodeposition, light-emitting-diode, and polycrystalline-organic-thin-film, while the significant companies were Davis-Joseph-Negley and Solar-Flex-Technologies-Inc. In 2008, the focused technique was display-device, while the significant company was The-Trustees-of-Princeton-University. In 2009, the focused techniques were composite-structure & metal-layer, In-Ga-original-substrate, and RF-sputtering & CIGS, while the significant companies were Smasung-Electronics-Co-Ltd and Solopower-Inc. Additionally, during 2000 to 2009, the more focused techniques were composite-structure & metal-layer, display-device, flexible-material, compound-thin-film & photoelectron-spectroscopy, and RF-sputtering & CIGS, and the more significant companies were Kaneka-Corporation and Canon-Kabushiki-Kaisha.

(2) Trends in Topics: Learning from Table 2, trends of several specific topics could be explained as follows. In topic-1, techniques 'composite-structure & metal-layer' and 'organic-semiconductor-thin-film' were linked closely via company 'Smasung-Electronics-Co-Ltd' in 2001 and 2009. In topic-2, techniques 'plasma-CVD' and 'film-deposition' were linked together via company 'National-Institute-of-Advanced-Industrial-Science-Technology' in 2004. In topic-5, techniques 'polycrystalline-organic-thin-film' and 'TCO-(transparent-conductive-oxide)-film' were linked together via company 'SolarFlex-Technologies-Inc' in 2007 and 2008.

5 Conclusions

The research framework of cluster analysis and data crystallization for trend detection has been formed and applied to thin-film solar cell using patent data. The experiment was performed and the experimental results were obtained. The topics of thin-film solar cell during 2000 to 2009 were: 'single-crystal & organic-semiconductor', 'light-absorbing-layer & CVD-method', 'roll-to-roll-process & heat-treatment', 'amorphous-film & plasma-deposition', 'microcrystalline-silicon & annealing-process', and 'sputtering & compound-thin-film'. The potential trends in each year (from 2000 to 2009) and in several specific topics have been detected, which would be helpful for managers and stakeholders to understand the developing directions of thin-film solar cell and to facilitate the strategic planning and decision-making.

In the future work, the research framework may be joined by some other methods such as grounded theory or social network analysis so as to enhance the validity of experimental results. In addition, the data source can be expanded from USPTO to WIPO or TIPO in order to explore the potential trends of thin-film solar cell technology widely.

References

1. Blackman, M.: Provision of Patent Information: A National Patent Office Perspective. World Patent Information 17(2), 115–123 (1995)
2. Tseng, Y., Lin, C., Lin, Y.: Text Mining Techniques for Patent Analysis. Information Processing and Management 43, 1216–1247 (2007)
3. Kontostathis, A., Galitsky, L., Pottenger, W.M., Roy, S., Phelps, D.J.: A Survey of Emerging Trend Detection in Textual Data Mining. In: Berry, M. (ed.) A Comprehensive Survey of Text Mining. ch. 9, Springer, Heidelberg (2003)
4. Le, M.H., Ho, T.B., Nakamori, Y.: Detecting Emerging Trends from Scientific Corpora. International Journal of Knowledge and Systems Sciences 2(2), 53–59 (2005)

5. Solarbuzz, Solar Cell Technologies (2010), http://www.solarbuzz.com/technologies.htm
6. Wikipedia, Thin film solar cell (2010), http://en.wikipedia.org/wiki/Thin_film_solar_cell
7. Jager-Waldau, A.: PV Status Report 2008: Research, Solar Cell Production and Market Implementation of Photovoltaics, JRC Technical Notes (2008)
8. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Pearson Addison Wesley, Boston (2006)
9. Ohsawa, Y.: Data Crystallization: Chance Discovery Extended for Dealing with Unobservable Events. New Mathematics and Natural Computation 1(3), 373–392 (2005)
10. Ohsawa, Y., Benson, N.E., Yachida, M.: KeyGraph: Automatic Indexing by Co-Occurrence Graph Based on Building Construction Metaphor. In: Proceedings of the Advanced Digital Library Conference (IEEE ADL 1998), pp. 12–18 (1998)
11. Maeno, Y., Ohsawa, Y.: Stable Deterministic Crystallization for Discovering Hidden Hubs. In: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, vol. 2, pp. 1393–1398 (2006)
12. Maeno, Y., Ohsawa, Y.: Human-Computer Interactive Annealing for Discovering Invisible Dark Events. IEEE Transactions on Industrial Electronics 54(2), 1184–1192 (2007)
13. Chiu, T.F.: Applying KeyGraph and Data Crystallization to Technology Monitoring on Solar Cell. Journal of Intelligent & Fuzzy Systems 21(3), 209–219 (2010)
14. USPTO: The United States Patent and Trademark Office(2010), http://www.uspto.gov/
15. Stanford Natural Language Processing Group, Stanford Log-linear Part-Of-Speech Tagger, http://nlp.stanford.edu/software/tagger.shtml (2009)
16. SPSS, Clementine 10.1, Algorithms Guide, USA: Integral Solutions Limited (2006)

Quality of Semantic Compression in Classification

Dariusz Ceglarek[1], Konstanty Haniewicz[2], and Wojciech Rutkowski[3]

[1] Poznan School of Banking, Poland
dariusz.ceglarek@wsb.poznan.pl
[2] Poznan University of Economics, Poland,
konstanty.haniewicz@ue.poznan.pl
[3] Business Consulting Center, Poland,
wojciech.rutkowski@bcc.com.pl

Abstract. Article presents results of implementation of semantic compression for English. An idea of semantic compression is reintroduced with examples and steps taken to perform experiment are given. A task of re-engineering available structures in order to apply them to already existing project infrastructure for experiments is described. Experiment demonstrates validity of research along with real examples of semantically compressed documents.

Keywords: semantic compression, semantic network, WiSENet, clustering, natural language processing.

1 Introduction

The aim of this work is to present an implementation of semantic compression, idea presented in [1], for English. Its main contribution is that experiment was performed for new language and an introduction of new semantic net. This has been achieved by application of re-engineered WordNet as a data structure for disambiguation resolution and a set of English domain frequency dictionaries. This research was motivated by good results already achieved for Polish [2]. In order to reach broader spectrum of peers and demonstrate usability of semantic compression, authors have decided to introduce semantic compression for English language.

For completion's sake authors decided to reintroduce notion of semantic compression. Detailed setting of this technique in Information Retrieval systems is discussed in here [1]. Discussion of necessary adjustments of already crafted solutions frequently used by researchers around the world precedes description of experiment and presentation of its results.

The work has been divided into following sections: description of semantic compression, description and discussion of semantic net in SenecaNet format, process of transferring WordNet into SenecaNet format (denoted as WiSENet), evaluation experiment using semantic compression for English, conclusions and future work.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 162–171, 2010.

2 Semantic Compression and Its Applications

In consistence with what has been stated in introductory section of this article, authors decided to reintroduce semantic compression. From now on any referrals to it are to be understood in spirit of the following definition:

Definition 1. Semantic compression is a technique that allows to transform a text fragment so that it has similar meaning but it is using less detailed terms where information loss minimization is an imperative.

The most important idea behind semantic compression is that it reduces the number of words used to describe an idea. As a consequence, semantic compression allows one to identify a common thought in seemingly different communication vessels, and it uses reduced number of word-vector dimensions in involved methods, making them more efficient.

Semantic compression for English has been made possible through adoption of WordNet [11] and adjusting it to already existing tools crafted for SEIPro2S. The details of adoption and motivation of transferring WordNet to SenecaNet format is discussed in separate section.

As stated in previous works, semantic compression can be perceived as an advanced technique to replace unnecessary terms with more general ones in processed text. Unnecessary terms are understood as words too specialised for given text [10]. This terms can be replaced by more general ones that fit into text's domain. This cannot be achieved without prior domain research and definition of domain frequency dictionaries. Thus, the need of semantic net for storing relations among words (with special focus on hypernymy and synonymy) and domain frequency dictionaries to measure which words can be freely generalised without visible information loss. These two crucial elements will be explored in greater detail after picturing semantic compression's mechanism. To further visualize semantic compression, please consider following artificial demonstration-purpose reference example.

First, one shall consider sentence A^0 and then compare it with sentence A^1 . Then, proceed to sentences B^0 and B^1 . Finally, compare sentences A^1 and B^1 .

Sentence A^0. Our cherished color television and fridge went under the hammer as the family tried to make ends meet.

Sentence A^1. Our loved colour television and fridge was sold as the family tried to survive.

Sentence B^0. Our treasured color TV set and refrigerator was traded so that our household could carry on through.

Sentence B^1. Our loved colour television and fridge was sold so that our family could survive.

Comparison demonstrates sentences that can be easily matched when one is able to transform them using semantic compression (sentences A^1 and B^1 are compressed). Matching must be based on a algorithm resembling classic bag of words [12] to yield best results. Authors strongly believe that given example is clear enough to convey basic mechanism of semantic compression. Real life examples will be given in section devoted to evaluation experiment. If one is to further delve into semantic compression and its place in Information Retrieval systems, please refer to more detailed description given in [2]. As one is ready to observe, semantic compression has to impose some information loss. It is a result of generalisation process that allows to replace less frequent terms with their more frequent hypernyms.

Domain corpora are resource that is foundation of domain frequency dictionaries that are of great value in generalisation process. There are no predefined settings for the level of generalisation that yields greatest effects. It has to be computed throughout an experiment for given set of documents. One can refer to exemplary results of semantic compression for Polish which can be found in [2]. Results for English are summarised and discussed later in this work.

A set of application for semantic compression is possible.

For the authors of this work, an interesting way to apply it to real world usage scenario is to check whether an artifact overuses unquoted references to someone's work. This kind of application enables one to weed out instances of plagiarism.

Another interesting application can be search for similar work in some vast corpora. This shall be extremely useful to anyone poised against such a task, as one does not have to match actual word phrasing but can focus more on the notion to be found. This application can be treated as method for automatic creation of related search terms not based on other queries but basing on rephrasing of current one. Others are also interested in the field, refer to [3] , [4] and [5].

One should also consider application of semantic compression to verify whether community based classification are free from random misclassification. Previously referenced work on semantic compression along with verification in this article hints that overall quality of automatic categorization can be significantly better than one performed with traditional methods. Performing clustering over corpus of semantically compressed documents results in fewer errors [8].

Semantic compression can find its application in intellectual property protection systems such as SEIPro2S implemented for Polish [1]. This kind of systems focuses on the meaning of documents trespassing corporate boundaries. In knowledge oriented societies where majority of revenue is an outcome of knowledge application this kind of system is invaluable asset.

3 SenecaNet Features and Structure

As earlier emphasized, any reasonable text transformation that promises informed choices when substituting one term for another one of more general nature fitting into text's domain, must be based on a structure capable of storing a variety of semantic relations.

A number of structures ranging from simple dictionaries, through thesauri to ontologies were applied into the matter [7]. Out of them semantic net has proven to be the best solutions due to its outstanding features coupled with lack of overcomplexity.

Experiment given in [2] was based on SenecaNet. SenecaNet is semantic net that stores relations among concepts for Polish. It stores over 137000 concepts, its other features are listed and described elsewhere.

It follows a notion of semantic net in every aspect. It has chosen to represent concepts of semantic net as a list. There is a specific format that allows for fast traversal and a number of checkup optimizations, that SenecaNet implements. As mentioned before, concepts are represented as a list of entries. Entry is stored in a way that allows for referencing connected concepts in an efficient manner.

Every entry of this list conveys information on actual descriptor to be found in text, hypernyms, synonyms, antonyms and descriptors that are in unnamed relation to given descriptor.

There is additional rule that every descriptor can occur exactly one time on the leftmost part of entry when whole semantic network is considered. This restriction introduces extremely important feature. There can be no cycles in structure devised in this manner.

Each descriptor can have one or more hypernyms (a heterarchy as in [7]). Each descriptor can have one or more synonyms. Synonyms are listed only once on the right side of entry, they do not occur on the leftmost part of entry, This is additional anticycle guard. An excerpt from WiSENet format is given to illustrate described content.

Listing 1.1. SenecaNet file format

```
Barack Obama|politician,#president(USA),
car|vahicle,&engine,
gigabyte|computer memory unit,&byte,
Real Madrid|football team,@Madrid,
volume unit|unit of measurement,@volume,
Jerusalem|city,:Palestine Authority,
Anoushka Shankar|musician,#sitarist,#daughter(Ravi Shankar),
Hillary Clinton|politician,#secretary of state(USA),
```

For many solutions, structure of semantic net is transparent, it does not affect the tasks net is applied to. Nevertheless, semantic compression is much easier when descriptors are represented by actual terms and their variants are stored as synonyms.

4 WordNet to SenecaNet Conversion

When faced with implementation of semantic compression for English one has to use a solution that has similar capabilities as those on SenecaNet. Building up a new semantic net for English is a great effort surpassing authors' abilities, thus we have turned to existing solutions. WordNet has proven to be excellent resource. It was applied by numerous research teams to a great number of tasks yielding good results. Thus, it was a natural choice in authors' research.

WordNet itself is a sense-oriented semantic net structure that contains over 130000 terms grouped in synsets. Every synset is collection of words (lemmas) that are in synonymy [11]. The design choice are not to be discussed here, yet one has to emphasize that synsets are elegant solution, at the same time they are cumbersome in text processing applications.

Authors had to confront a challenge of converting synset oriented structure into cycleless semantic net operating on descriptors to be recognized as actual terms in processed text fragment. An algorithm to accomplish this has been devised. It operates on sets, taking into account data on every lemma stored in given synset and synsets (therefore their lemmas) that are hypernyms to the one processed.

Synset is understood as a group of terms that have similar meaning. Under close scrutinization a lot of terms gathered in one synset fails to be perfect synonyms to each other. They share a common sense, yet the degree to which they do that, varies. Lemma is any member of synset, it can be a single term or a group of terms representing some phrase [11].

Before algorithm is given, an example of naive approach to a problem is demonstrated. This shall enable reader to follow the process of semantic network transformation in greater detail and with less effort.

One need to drop additional data on word sense as ideally one would like to come up with a list of words. Lets consider word “abstraction” as a target term. WordNet stores “abstraction” in six different synsets. As they are numbered basing on their frequency, naive approach would suggest to start with the first sense. A generalization path leading from our chosen farthest leaf to the root can easily be obtained.

When one is to apply this kind of transformation, he shall quickly face the consequences of introducing great many circulatory graphs in his mapped structure.

In order to avoid graph cycles in target structure, authors needed to modify the way one chooses words to describe synset. The best situation is when a lemma contained in synset descriptor belongs only to this synset, ie. lemma itself is a unique synset descriptor. In other situations, authors try to find other lemma from the same synset, which satisfies the condition. Experiments have prooved that this produces desired networks, but cannot satisfy criterion of lack of losses during transformation. Obtained semantic net consisted of only 25000 terms serving as concepts, where a total of 86000 noun synsets were processed. Eventually, “synthetic” synset descriptor is developed. Introduction of synthetic descriptors is not contrary to authors’ ambitions to convert WordNet into WiSENet in a lossless manner along with usage of actual terms as concept descriptors. Synthetic descriptors are always result of untangling of some cycle thus they always can be outputted as actual terms to be found in processed text.

Please refer to figures 1 and 2 to view visualisation of this process. Notice that term approximation is contained in several synsets. Thus it fails as a concept descriptor (see 1). One can easily observe that term “bringing close together” occurs exactly once, thus can replace synthetic descriptor “approximation.n.04”.

All this is gathered in tables 1 and 2.

Table 1. Companion table for figure 1

Synset	Terms	Parent synset
change of integrity	change of integrity	change.n.03
joining.n.01	joining, connection, connexion	change of integrity
approximation.n.04	approximation, bringing close together	joining.n.01
approximation.n.03	approximation	version.n.01
approximation.n.02	approximation	similarity.n.01
estimate.n.01	estimate, estimation, approximation, idea	calculation.n.02

Table 2. Companion table for figure 2

Term	Parents
change of integrity	change.n.03
approximation	bringing close together, approximation.n.02, estimate.n.01, approximation.n.03
approximation.n.02	similarity.n.01
approximation.n.03	version.n.01
bringing close together	joining
joining	change of integrity
estimate.n.01	calculation.n.02
estimate	estimate.n.02,estimate.n.01,estimate.n.05, appraisal.n.02,estimate.n.04,compute,count on

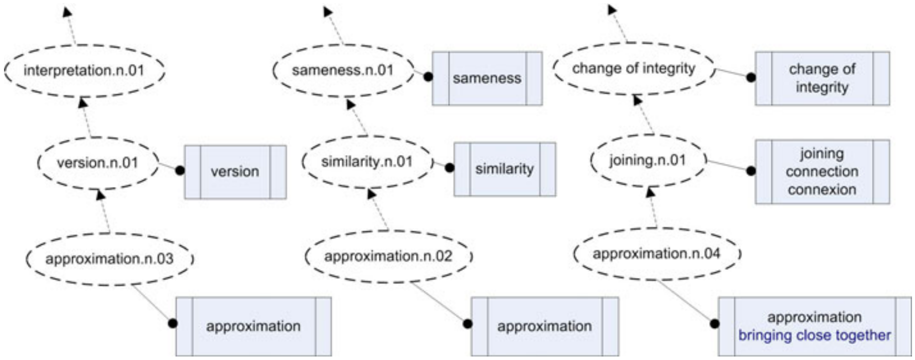


Fig. 1. WordNet synset description

The whole procedure is realized as described below.

The first step is to build a frequency dictionary (F) for lemmas, counting synsets containing a given lemma. Algorithm loops through all synsets in WordNet (WN), and all lemmas in the synsets (S), and count every lemma occurrence. In the second step, it picks a descriptor (possibly a lemma) for every synset.

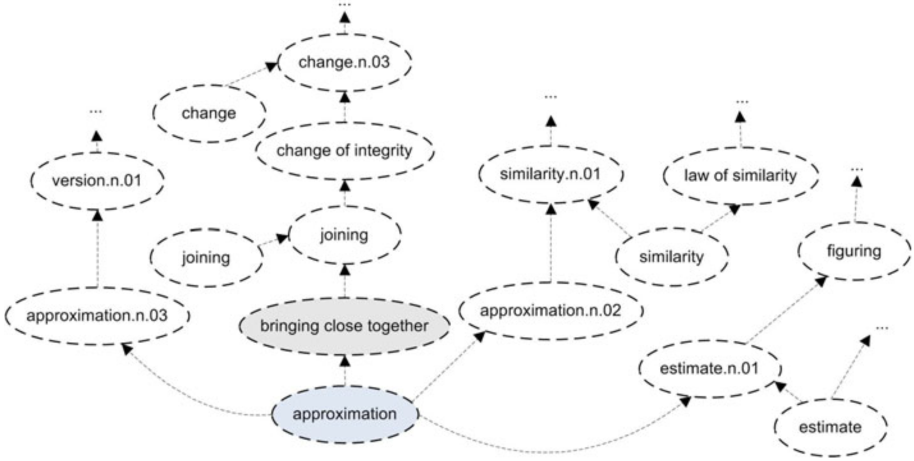


Fig. 2. Concepts description in WiSENet format

Next, it begins checking, whether synset descriptor (d) contains a satisfactory lemma. After splitting the descriptor (partition point is the first dot in synset description) and taking the first element of resulting list, algorithm examines, whether such lemma occurs exactly once throughout all synsets - if answer is positive, it can be used as a new synset descriptor. If contrary, it loops through lemmas from examined synset and checks if there is any unique lemma which can be utilised as a descriptor. In case no unique lemma can be found, a genuine WordNet descriptor is used.

5 Evaluation

As in previously conducted research for Polish we have devised an experiment that enables to verify whether semantic compression does yield better results when applied to specific text processing tasks.The evaluation experiment is performed by a comparison of clustering results for texts that were not semantically compressed with those that were [6]. Authors gathered texts coming from following domains: [illegible]

To verify the results, all documents have been initially labeled manually with a category. All documents were in English.

Clustering procedure was performed 8 times. First run was without semantic compression methods: all identified concepts (about 25000 - this is only about a fifth of all concepts in the research material) were included. Then, semantic compression algorithm has been used to gradually reduce the number of concepts. It started with 12000 and it proceeded with 10000, 8000, 6000, 4000, 2000 and 1000 concepts.

Classification results have been evaluated by comparing them with labels specified by document editors: a ratio of correct classifications was calculated. The outcome is presented in Tables 3 and 4. The loss of classification quality is virtually insignificant for semantic compression strength which reduces the number of concepts to 4000.

As briefly remarked in earlier section the conducted experiment indicates, that semantic compression algorithm can be employed in classification tasks to significantly reduce the number of concepts and corresponding vector dimensions. As a consequence, tasks with extensive computational complexity are performed faster.

A set of examples of semantically compressed text fragments (for 4000 chosen concepts) is now given. Each compressed fragment is proceeded by its original.

Table 3. Classification quality without semantic compression

Clustering features	1000	900	800	700	600	Average
All concepts	94,78%	92,50%	93,22%	91,78%	91,44%	92,11%
12000 concepts	93,39%	93,00%	92,22%	92,44%	91,28%	91,81%
10000 concepts	93,78%	93,50%	93,17%	92,56%	91,28%	92,23%
8000 concepts	94,06%	94,61%	94,11%	93,50%	92,72%	93,26%
6000 concepts	95,39%	94,67%	94,17%	94,28%	93,67%	93,95%
4000 concepts	95,28%	94,72%	95,11%	94,56%	94,06%	94,29%
2000 concepts	95,56%	95,11%	94,61%	93,89%	93,06%	93,96%
1000 concepts	95,44%	94,67%	93,67%	94,28%	92,89%	93,68%

1a The information from AgCam will provide useful data to agricultural producers in North Dakota and neighboring states, benefiting farmers and ranchers and providing ways for them to protect the environment.

1b information will provide adjective data adjective producer american_state adjective state benefit creator creator provide structure protect environment

2a Researchers trying to restore vision damaged by disease have found promise in a tiny implant that sows seeds of new cells in the eye.The diseases macular degeneration and retinitis pigmentosa lay waste to photoreceptors, the cells in the retina that turn light into electrical signals carried to the brain.

2b researcher adjective restore vision damaged by-bid disease have found predict tiny implant even-toed_ungulate seed new cell eye disease macular_degeneration retinitis_pigmentosa destroy photoreceptor cell retina change_state light electrical signal carry brain

3a Together the two groups make up nearly 70 percent of all flowering plants and are part of a larger clade known as Pentapetalae, which means five petals. Understanding how these plants are related is a large undertaking that could help ecologists better understand which species are more vulnerable to environmental factors such as climate change.

3b together two group constitute percent group flowering_plant part flowering_plant known means five leafage understanding plant related large undertaking can help biologist better understand species more adjective environmental factor such climate_change

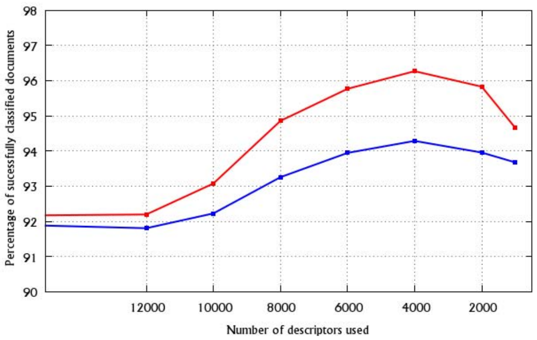


Fig. 3. Classification quality for two runs, upper line denotes results with semantic compression enabled

Table 4. Classification quality using semantic compression with proper names dictionary enabled

Clustering features	1000	900	800	700	600	Average
All concepts	94,78%	92,50%	93,22%	91,78%	91,44%	92,11%
12000 concepts	93,56%	93,39%	93,89%	91,50%	91,78%	92,20%
10000 concepts	95,72%	94,78%	93,89%	91,61%	92,17%	93,08%
8000 concepts	95,89%	95,83%	94,61%	95,28%	94,72%	94,86%
6000 concepts	96,94%	96,11%	96,28%	96,17%	95,06%	95,77%
4000 concepts	96,83%	96,33%	96,89%	96,06%	96,72%	96,27%
2000 concepts	97,06%	96,28%	95,83%	96,11%	95,56%	95,83%
1000 concepts	96,22%	95,56%	94,78%	94,89%	94,00%	94,66%

6 Conclusions and Future Work

This work has demonstrated that semantic compression is viable for English. A set of steps that were needed to make it possible has been described. Finally, an experiment has been presented along with its results.

Authors defined a number of important areas that need to be further developed. First issue that shall be tackled is state of vocabulary that is not represented in WordNet, yet is overwhelming in current culture. To exemplify, words such as these are non existent: superpipe, windsurfer, airball, blazar, biofuel, spacelab, exoplanet, wildcard, superhero, smartphone.

WiSENet is in great need of incorporating a vast corpus of geographic names and locations. This shall easily improve results for further experiments and applications. In addition, inclusion of greater number of information on actual people shall further boost generalisation results. This inclusion must focus on introduction of unnamed relation to WiSENet as it currently supports only those from original WordNet.

A great number of adjectives and adverbs can only be generalised to their type i.e. WiSENet can only tell whether it is dealing with adjective or an adverb. Authors envision addition of information whether adjective or adverb in consideration can be related to a verb or a noun. This is another feature that will improve performance of semantic compression.

Last but not least improvement is creation of vaster corpora of texts, so that WiSENet can store more new concepts.

Research has brought a number of other interesting observations. They shall be brought to reader's attention in further publications, as close to the topic of this work they shift its focus from semantic compression to semantic net's features.

References

1. Ceglarek, D., Haniewicz, K., Rutkowski, W.: Semantically Enchanced Intellectual Property Protection System - SEIPro2S. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) ICCCI 2009. LNCS, vol. 5796, pp. 449–459. Springer, Heidelberg (2009)
2. Ceglarek, D., Haniewicz, K., Rutkowski, W.: Semantic compression for specialised Information Retrieval systems. In: 2nd Asian Conference on Intelligent Information and Database Systems Studies in Computational Intelligence 283, Springer, Heidelberg (2010)
3. Baziz, M.: Towards a Semantic Representation of Documents by Ontology-Document Mapping (2004)
4. Baziz, M., Boughanen, M., Aussenac-Gilles, N.: Semantic Networks for a Conceptual Indexing of Documents. In: IR (2005)
5. Gonzalo, J., et al.: Indexing with WordNet Synsets can improve Text Retrieval (1998)
6. Hotho, A., Staab, S., Stumme, S.: Explaining Text Clustering Results using Semantic Structures. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.) PKDD 2003. LNCS (LNAI), vol. 2838, pp. 217–228. Springer, Heidelberg (2003)
7. Hotho, A., Maedche, A., Staab, S.: Ontology-based Text Document Clustering. In: Proceedings of the Conference on Intelligent Information Systems. Springer, Zakopane (2003)
8. Khan, L., McLeod, D., Hovy, E.: Retrieval effectiveness of an ontology-based model for information selection (2004)
9. Krovetz, R., Croft, W.B.: Lexical Ambiguity and Information Retrieval (1992)
10. Frakes, W.B., Baeza-Yates, R.: Information Retrieval: Data Structures and Algorithms. Prentice-Hall, Englewood Cliffs (1992)
11. Fellbaum, C.: WordNet - An Electronic Lexical Database. The MIT Press, Cambridge (1998) ISBN:978-0-262-06197-1
12. Zellig, H.: Distributional Structure. Word 10(2/3), 146–162 (1954)

Realizes Three-Tier Structure Mail System Based on the Web

Jun-Hua He and Jing-Yuan Lv

School of Computer Science and Technology,
HuangShi Institute of Technology, HuangShi City,
Pri.HuBei, China
{Jun-Hua He,hjh-6264}@163.com

Abstract. E-mail system in the B / S three-tier structure in the features: the client part of the message processing logic assigned to a feature server, no longer responsible for dealing with complex calculations and data access to key services, enabling e-mail client running efficiency is improved. Three-tiered structure also makes it clear layers of function modules, independently of each other, thus greatly simplifying maintenance and modification work.

Keywords: ASP.NET, Three-Tier Structure, Mail System.

1 Introduction

E-mail is an important part of the office automation System, an indispensable part of the existence of its business units and personnel is an essential service and therefore should be able to e-mail system for enterprise information delivery to provide a good platform.

As a computer application's part, mail receiving and dispatching has the very vital role in enterprise's office automation. The mail transmission and the receive design use the SQL sentence generally. But because is huge in the enterprise information flow's data. The mail information read is frequent. There are many shortcomings of these methods, such as: low efficiency, poor security, high costs, which the companies have brought a lot of losses. How to solve these problems, enables the mail subsystem to have the efficiency to be high, the price is small, the security is high, the stability is good, secret. Becomes has the necessity very much.

2 Realizes the Technical Background

2.1 Visual Studio .NET

Visual Studio .NET was founds .NET Framework the application procedure to provide the reliable development environment. Visual Studio.NET includes a set of default device configuration files. The equipment configuration files contain the foundation the information which needs in view of the specific equipment's application procedure. Had Visual Studio .NET, also had founded the configuration files which Pocket PC,

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 172–177, 2010.

Pocket PC 2002 and Windows CE .NET 4.1 and the higher edition's application procedure needed. These configuration files enable the user to found contain the Windows window and the ADO.NET application procedure, but also provided for the user has used ability which Web served. ASP.NET is one kind of establishment dynamic Web application procedure technology. It was a .NET frame part, ASP.NET has provided one kind of programming model and the structure, contrasted the original Web technology, it could be faster, establish, security and stability nimbly easily the application procedure. [1][2]

2.2 SQL Server

Microsoft SQL Server is produced by Microsoft, is a large database of relational database System, it has the hardware platform independent, symmetric multi-processor architecture, preemptive multi-task management, comprehensive security systems and fault tolerance, and have easy maintenance characteristics. This is also the background here subsystems.

2.3 SQL Storage Procedure

SQL storage procedure is a SQL database running on the core business, through the form of resident memory for read \ processing \ write the data to deal with the most frequent. SQL storage procedure is a set of specific functions in order to complete the SQL statements set, as compiled stored in the database. Users specify the name of the storage procedure and given parameters (If the storage procedure with parameters) to perform it. Users run the storage procedure by specifying the name of the storage procedure and given the parameters (if the storage procedure with parameters), any well-designed database applications should use storage procedures. In general, the storage procedure has the following advantages:

Storage procedure to allow the standard component-based programming;

Storage procedure to achieve a faster speed;

Storage procedure can reduce network traffic;

Storage procedure may take one kind of safety mechanism to come the full use. [3]

3 Based on B / S Architecture of E-Mail System

3.1 B / S Three-Tier Architecture

In the B / S architecture system, the user through a browser on the network to the distribution of many of the server request, the server on the browser to process the request, the user information needed to return to the browser. Three-tier architecture is between the client and the database, added a middle layer, also called component layer. The Application programs of Three-tier system put the business rules, data accessing, verify the legitimacy of such work into the middle layer of processing. Typically, the client does not directly interact with the database, but through COM / DCOM communication connection with the middle layer, and then through the middle layer and the database exchange.

Structure contains three layers: layer (USL), business logic layer (BLL), Data Access Layer (DAL).

E-mail system in the B / S three-tier structure works: the user layer to the network WEB server (business logic) to issue e-mail request, the server requests the browser to process the mail, access the database through data access layer will be e-mail users need the data returned to the browser.

E-mail system in the B / S three-tier structure in the features: the client part of the message processing logic points to the Web server, no longer responsible for dealing with complex calculations and data access and other key services, enabling e-mail client running efficiency is improved . Three-tiered structure also makes it clear layers of function modules, independently of each other, thus greatly simplifying maintenance and modification work.

3.2 Based on B / S Architecture, E-Mail System

In the mail system to send and receive e-mail Design Functions storage procedure used to store the database and extract the message content, as shown in Fig. 1. It has the following advantages:

The first is the implementation of efficient. Need to be repeatedly called for the code block, storage procedure than a batch of SQL statements to implement a lot faster. Because the storage procedure is precompiled and not to explain the implementation of the SQL statement as requested in the proposed operation be conducted only after the syntax analysis and optimization.

Second reduces the traffic between the client and database server. The client calls the storage procedure, just to the server storage procedure name and parameters can be, if it is a SQL statement, then you need to send more than, and thus greatly reduces the network traffic, reducing network load.

The third improved database security and integrity. First of all, the use of a large number of storage procedures to prevent SQL injection attacks are very effective, SQL injection attacks on the general procedure is carried out by analyzing the possible SQL statements, especially in the dynamic SQL statement, and adopted on the basis of their Add additional sentence to obtain the information you want to know. The storage procedure can be very well put an end to this loophole. [4]

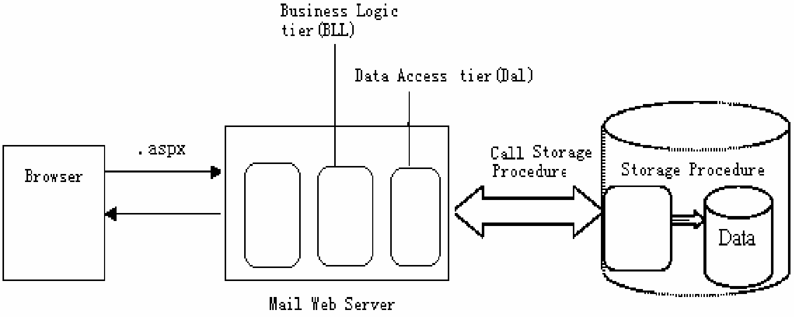


Fig. 1. The three-tier structure mail system

4 The Main Process and Code

The page code to send and receive mail at the local area network, used to send mail Functions btnSendMail_Click () and receive e-mail Functions ShowBodyDetail (). The two Functions are designed using the appropriate storage procedure to store and retrieve database SQL statement instead of the message content, as shown in Table 1.

Table 1. The storage procedures that the functions call and processing of data

SUB Function()	Call storage procedure	Data ID
btnSendMail_Click()	OA_MailSend	@MailFolderType, @MailReceiverStr , @MailSendDate, @MailSendLevel, @MailSender, @MailReceiver, @MailSubject,@MailBody, @MailCcToAddr, @MailBccToAddr, @MailReadFlag, @MailTypeFlag, @MailClassID, @MailImportance,@MailID
ShowBodyDetail()	OA_MailGetCompleteInfo	@MailID, @MailBody, @MailSender

4.1 Function to Receive E-Mail ShowBodyDetail() for Example

Intermediate OA_MailGetCompleteInfo call database storage procedures to extract data in the MAIL ID. Code is as follows:

```
Protected void ShowBodyDetail() //See mail conten
  {Kn_Mail Kn_Mail = new Kn_Mail();
   SqlDataReader dataReader = null;
   try
      {dataReader =
Kn_Mail.GetMailCompleteInfoDbreader(MailID);
// Kn_Mail.GetMailCompleteInfoDbreader call storage
procedure to extract the database MAIL ID
      }
   catch
      {Server.Transfer("../../Error.aspx");}
   if(dataReader.Read())  // See the main e-mail
database data extraction
      {this.lblSenderName.Text =
dataReader["MailSender"].ToString();
       this.lblCcToAddr.Text   =
Kn_OAUser.GetRealNameStrByUsernameStr(dataReader["MailC
cToAddr"].ToString(),0);
       string[] RecvAr =
System.Text.RegularExpressions.Regex.Split(dataReader["
MailBccToAddr"].ToString() ,",");
```

```
       for(int i=0;i<RecvAr.Length-1;i++)
          {if( RecvAr[i].ToString()==UserCookie.Value.To
String())
             {this.lblBccToAddr.Text
=Kn_OAUser.GetRealNameByUsername(UserCookie.Value.ToStr
ing());}
          }
       this.lblSubject.Text    =
dataReader["MailSubject"].ToString();
       this.lblBody.Text       =
dataReader["MailBody"].ToString();
       this.lblSendDate.Text   =
dataReader["MailSendDate"].ToString();
       this.lblReceiverStr .Text =
Kn_OAUser.GetRealNameStrByUsernameStr(dataReader["MailR
eceiverStr"].ToString(),0);
       this.lblProjectName .Text =
dataReader["classname"].ToString();
       }
   dataReader.Close();
   try
      {dataReader =
Kn_Mail.GetMailAttInfoDbreader(MailID);}
   catch
      {Server.Transfer("../../Error.aspx");}
   while(dataReader.Read())
      {lblAttachFile.Text += "&nbsp;<a
href='Download.aspx?destFileName="+Server.UrlEncode(dat
aReader[2].ToString())
+"'>"+dataReader[0].ToString()+"("+dataReader[1].ToStri
ng()+" Byte)</a><br>";
      }
   dataReader.Close();
   Kn_Mail = null;
}
```

4.2 The Storage Procedure for Receiving E-Mail Calls

```
CREATE  PROCEDURE  OA_MailGetCompleteInfo
 (@MailID varchar(100))
AS
   IF(len(@MailID)<8)
      BEGIN
         UPDATE L_TabMailList SET MailReadFlag=1 WHERE
MailId=@MailID
         SELECT MailID,(SELECT RealName FROM L_User a
where a.UserName=MailSender) as MailSender, (SELECT
RealName FROM L_User a where a.UserName=MailReceiver)
as MailRe-
ceiver,MailSendDate,MailSubject,MailBccToAddr,MailCcToA
ddr,MailBody,MailReadFlag,MailReceiverStr,MailSender,
(select ClassName from L_class b where
b.classid=L_TabMailList.classid)  as classname
         FROM L_TabMailList
         WHERE MailID = @MailID
      END
   ELSE
      declare @sql nvarchar(4000)
      select @sql = ' SELECT Subject as
MailSubject,TextContent as MailBody ,FromName as
MailSender,Ccto as MailCcToAddr,Bccto as
MailBccToAddr,SendDate as MailSendDate,''''as
MailReceiverStr,'' External e-mail'' as classname
      FROM L_TabExtMailList
      WHERE MailID ='''+ @MailID + ''''
       -- print @sql
       exec (@sql)
GO
SET QUOTED_IDENTIFIER OFF
GO
SET ANSI_NULLS ON
GO
```

5 Conclusion

In this paper, the internal mail system, through the mail system to send and receive e-mail the design of Functions to optimize the use of storage procedures, using ASP.NET and SQL SERVER realized, it embodies the advantages of .NET, and with the database the perfect combination. It has very good performance both In terms of operating efficiency and in the functional.

References

1. Pariha, M.: ASP.NET book. Electronic Industry Press, Beijing (2002)
2. Deitel, H.M.: C# programmer to develop guidelines for high-level. Tsinghua University Press, Beijing (2003)
3. Wang, S., Sa, S.-X.: An Introduction to Database Systems, 4th edn. Higher Education Press, Beijing (2006)
4. Otey, M., Coute, P.: SQL SERVER guide the development. Tsinghua University Press, Beijing (2002)

Secure Collaborative Cloud Design for Global USN Services

Tien-Dung Nguyen, Md. Motaharul Islam, Aymen Al-Saffar, Jun-Young Park, and Eui-Nam Huh

Department of Computer Engineering,
Kyung Hee University, Suwon, Korea
{ntiendung,motahar,aymen,jypark,huh}@icns.khu.ac.kr

Abstract. Nowadays, there are more and more service agents which provide variety of sensor services (Global USN). Assuming many agents implement the different applications in the different locations (e.g. e-healthcare, temperature system, camera system, etc.). In practical, service agents have to prepare too costly infrastructure to enhance the applications. When cloud service is growing rapidly, we propose a cloud service provider which gathers all sensors, applications from each agent to become the sufficient services. In this paper, we propose a system which combines all global USN Services to produce many enhanced services. Moreover, we propose a security mechanism for our proposed system as well.

Keywords: Cloud Services, Cloud Security, Secure Multicast.

1 Introduction

Cloud computing is a class of the next generation highly scalable distributed computing platform in which computing resources are offered 'as a service' leveraging virtualization and Internet technologies. Cloud-based services include software-as-a-service (SaaS) and platform as a service (PaaS). Amazon's Elastic Compute Cloud (EC2) [10] and IBM’s Blue Cloud [11] are examples of cloud computing services. These cloud service providers allow users to instantiate cloud services on demand and thus purchase precisely the capacity they require when they require based on pay-per-use or subscription-based model.

Cloud computing is receiving traction with businesses and has become increasingly popular for hosting data and deploying software and services. The attractive part of the cloud computing is that it enables customers away to increase capacity or add capabilities on the fly without upfront investment in new infrastructure, personnel training, or software licensing drastically boost their infrastructure resources, all at negligible cost.

Ubiquitous Sensor Networks (USN) [12] are an emerging business area and now used in many civilian application area, including environmental and habitat monitoring, healthcare applications, home automation, and traffic control, leading to an advanced e-Life society.

Therefore, combination of Cloud Service and USN will bring to customer more and more advantages. In Fig.1, customers can use USN (e-healthcare, temperature

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 178–187, 2010.

system, and camera system service) everywhere, anytime by Cloud Service Provider easily. Although cloud computing provides a number of advantages that include economies of scale, dynamic provisioning, increased flexibility and low capital expenditures, it also introduces a range of new security risks [14]. As cloud computing brings with it new deployment and associated adversarial models and vulnerabilities, it is imperative that security takes center stage. This is especially true as cloud computing services are being used for e-commerce applications, medical record services, and back-office business applications, all of which require strong confidentiality guarantees. Thus, to take full advantage of the power of cloud computing, end users need comprehensive security solutions to attain assurance of the cloud's treatment of security issues. This paper introduces a cloud service system model and a security mechanism based on key management with secure multicast.

The rest of the paper is organized as follows. In Section 2, we identify security concerns arising in cloud computing environments and present related work. A proposed system model is discussed in Section 3. Then we will analysis security solution in Section 4.

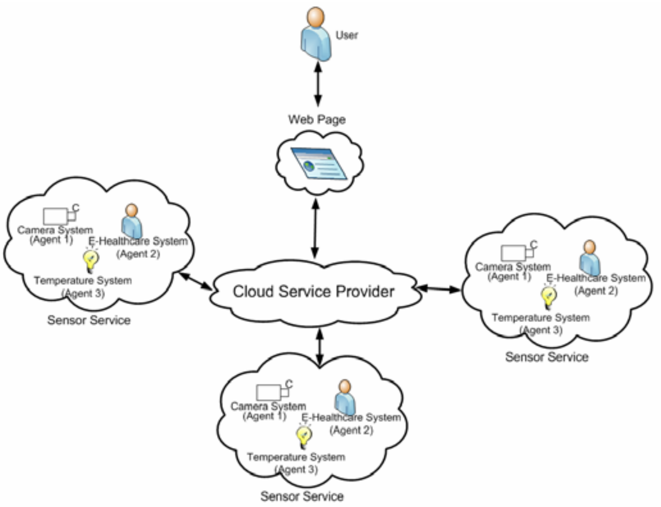

User
Web Page
Camera System (Agent 1)
E-Healthcare System (Agent 2)
Temperature System (Agent 3)
Sensor Service
Cloud Service Provider
Camera System (Agent 1)
E-Healthcare System (Agent 2)
Temperature System (Agent 3)
Sensor Service
Camera System (Agent 1)
E-Healthcare System (Agent 2)
Temperature System (Agent 3)
Sensor Service

Fig. 1. Cloud Service Provider and USN

2 Related Work

Many of the previous work in the field of cloud computing have been in the areas of its technological architecture and features, differences from other similar technologies and security issues. Regardless of what technology is being used, people generally look for the most important criterion which is security to adopt it while many other smart environments like utility computing, smart data centers, pervasive computing, automation, virtualization and intelligent networks already penetrate into our daily life [15]. Cloud computing builds different services in business, education and government sectors and becomes a new term together with the latest networking, web and software services [16]. Cloud computing inherits the advancements and limitations in other computing research areas aforementioned above. In cloud computing environment, overall security issues can be evaluated from the points of service providers and the clients. While the providers focus on the continuity of their services against configuration updates for performance and QoS, spam and virus threats and proper customer accountability, clients mainly look for the security of their data and the reliability of the provider. The basic features of cloud computing are presented and compared with the computational resources used in Grid Computing [17] together with the required security architecture [18] incorporated firewalls, intrusion detection/prevention systems, antivirus, authentication, authorization, access control, encryption and other services. A Role-Based Access (RBAC) model for authorization using secure web services [19], a context based dynamic role based access control model (CDACM) for web services [20] and proxy-based security architecture that provides authentication and authorization [21] are proposed for the management and the enhancement of security goals in web services. Reference [22] also proposes a formal model on policy-based access control framework for autonomic systems. Cloud also provides data storage in its web space for the customers. Even though the storage of user data on remote servers is not new, current emphasis on the expansion of cloud computing is whether it has drawbacks for ensuring data privacy, confidentiality and reliability. The General File System (GFS) introduced in [23] integrates different storage spaces and promises the data security.

However, there are many other security issues in [13] which did not concern before. Therefore, in this paper, we focus on encryption and key management to promote strong authentication and access control for administrative access and operations, encrypt and protect integrity of data in transit, implement strong key generation, storage and management, destruction practices.

3 Our Proposed System Model

Our system includes the architecture and security mechanism. Therefore, in order to make sense the system model, we analysis architecture which classify by layers and security model to secure the system.

3.1 Layer Model

In the Layer model, we considered a system model with Control Gateway (CG) and 3 layers as Fig.2 shown. CG is used to receive request from users, then propagate to corresponding Services in Layer 3. Layer 3 (Service Layer) is a set of images (virtual machines) which are presented for Services. One service consists of one or many applications. This layer receives requests from CG. Layer 2 (Application Layer) is a set of images (virtual machines) which are presented for Applications. Application undertakes one basic function (e.g. temperature application, electrical consumption application…). This layer receives request from Service. Layer 1 (Virtual Gateway Layer) is a set of images (virtual machines) which are presented for Virtual Gateway. Each Virtual Gateway communicates with Gateway to aggregate data from USN.

In this model, Layer 2 and 3 are classified by function of each Service or Application. Layer 1 is classified by geographical areas. This classification is advantage for the applications which collect data from many geographical areas. In practical, one service can combine one or many these applications.

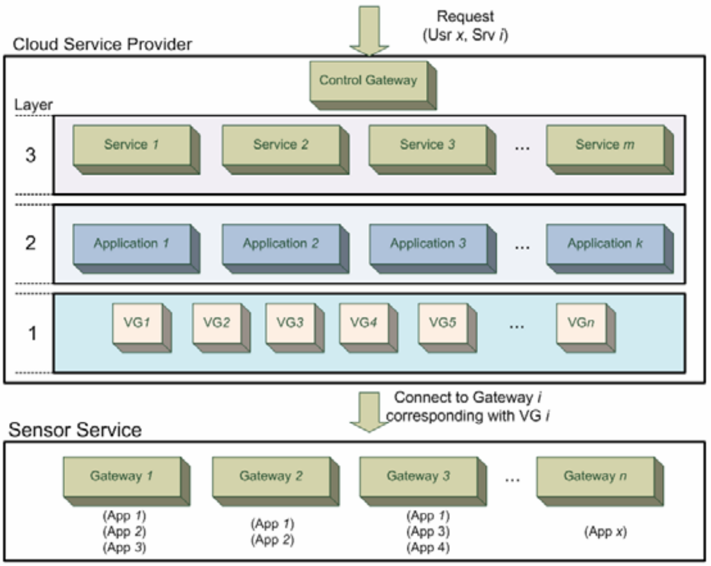


Fig. 2. Layer Model

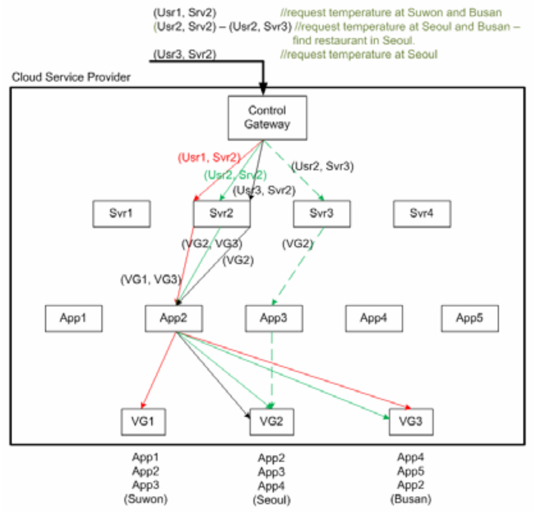
3.1.1 One Service – One Application

In this case, one service (*Svr*) includes one application (*App*). Control Gateway (*CG*) receives all requests from users will disseminate those requests to the corresponding *Svr*s. From here, *Svr*s analysis requests and collect needed data through the corresponding *App*s. At Layer 2, *App*s receive information from Layer 3 and aggregate data from suitable geographical area (*VG*). Example, in the Fig. 3, assuming Suwon includes *{App 1, App 2, App 3}*, Seoul includes *{App 2, App 3, App 4}*, Busan includes *{App 4, App 5, App 2}*. There are 4 requests from 3 users. User 1 wants to know temperature (*Svr* 2) at Suwon and Busan. User 2 finds temperature (*Svr* 2) at Seoul and Busan, and then he wants to find some restaurants (*Svr* 3) in Seoul. User 3 only requests temperature information (*Svr* 2) at Seoul. After receiving these requests, *CG* sends 3 requests to *Svr* 2, and 1 request to *Svr* 3. Request flow of each user for Service 2 can be described as Fig. 3.

Therefore, total request flows of the system for Service 2 are 4, 6, 3 for user 1, user 2, and user 3 respectively. If our system requires security, so each step needs encrypted, i.e. total encryption cost of user 1 is 4, user 2 is 6 and user 3 is 3 for Service 2.

3.1.2 One Service – Many Applications

For this case, one Service includes many Applications. Fig. 4 shows one example for one User. User 1 wants to request Service 1 which includes: *App 1* at Suwon, *App 2* at Seoul and Busan, *App 3* at Suwon and Seoul. The mechanism of Cloud Service Provider for this request follows Fig. 4. In this case, the total request flow of the system is 9. Therefore, the total encryption cost is also 9.



For User 1:

1. $CG \rightarrow Svr\,2 \quad : \{Usr\,1, Svr\,2\}$
2. $Svr\,2 \rightarrow App\,2 : \{VG\,1, VG\,3\}$
3. $App\,2 \rightarrow VG\,1 : \{App\,2\}$
4. $App\,2 \rightarrow VG\,3 : \{App\,2\}$

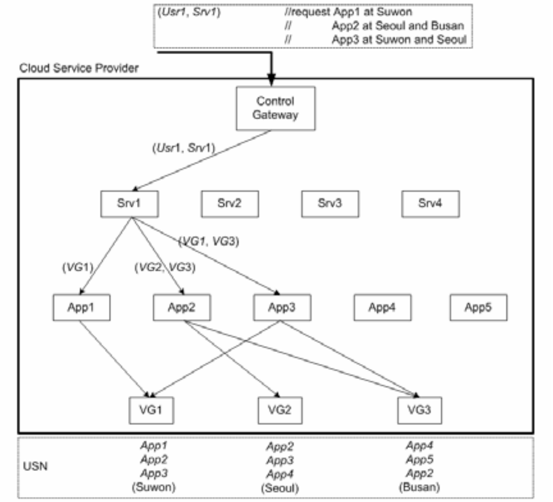
For User 2:

1. $CG \rightarrow Svr\,2 \quad : \{Usr\,2, Svr\,2\}$
2. $CG \rightarrow Svr\,3 \quad : \{Usr\,2, Svr\,3\}$
3. $Svr\,2 \rightarrow App\,2 : \{VG\,2, VG\,3\}$
4. $Svr\,3 \rightarrow App\,3 : \{VG\,2\}$
5. $App\,2 \rightarrow VG\,2 : \{App\,2\}$
6. $App\,2 \rightarrow VG\,3 : \{App\,2\}$
7. $App\,3 \rightarrow VG\,2 : \{App\,3\}$

For User 3:

1. $CG \rightarrow Svr\,2 \quad : \{Usr\,1, Svr\,2\}$
2. $Svr\,2 \rightarrow App\,2 : \{VG\,2\}$
3. $App\,2 \rightarrow VG\,2 : \{App\,2\}$

Fig. 3. One Service – One Application



1. $CG \rightarrow Svr\,1 \quad : \{Usr\,1, Svr\,1\}$
2. $Svr\,1 \rightarrow App\,1 \quad : \{VG\,1\}$
3. $Svr\,1 \rightarrow App\,2 \quad : \{VG\,2, VG\,3\}$
4. $Svr\,1 \rightarrow App\,3 \quad : \{VG\,1, VG\,3\}$
5. $App\,1 \rightarrow VG\,1 \quad : \{App\,1\}$
6. $App\,2 \rightarrow VG\,2 \quad : \{App\,2\}$
7. $App\,2 \rightarrow VG\,3 \quad : \{App\,2\}$
8. $App\,3 \rightarrow VG\,1 \quad : \{App\,3\}$
9. $App\,3 \rightarrow VG\,3 \quad : \{App\,3\}$

Fig. 4. One Service – Many Applications

3.2 Security Model

As we mention about total encryption cost, it takes a huge computation if the system receive many requests from user and especially in case a service includes many applications. The naïve security method is encrypting all requests and sends to destination. For a more concrete illustration of this point, we outline a typical procedure for securing unicast communications between a client and a server. Initially, the client and server mutually authenticate each other using an authentication protocol or service;

subsequently, a symmetric key is created and shared by them to be used for pairwise confidential communications [6], [7], [8], [9]. By this way, each service in layer 3 has an individual pair-wise key with each application in layer 2. Similarly, each application in layer 2 has a pair-wise key with each virtual gateway in layer 1. Therefore, each service may send request to each application and from each application to each virtual gateway with different key.

However, the pair-wise key mechanism is not efficient when number of user and service increase. For this reason, we propose a secure multicast mechanism combining with pair-wise key mechanism to reduce total encryption cost for whole system.

Multicast is an efficient IP communication technique when one message is to be delivered to a group of intended recipients, reducing sender overheads as well as network bandwidth consumption. In unicast, two users can provide confidentiality by encrypting data with a shared key. In multicast, group key encryption [1] is used in which the multicast traffic is encrypted with a symmetric key and every authorized member of the group is given the decryption key. By using group key encryption, it need only one encryption cost to send request to group instead of taking number of member in group encryption cost. Besides, in [2], [3], a key tree scheme known as logical key hierarchy is presented as a solution to the key management scalability problem.

Let there be a trusted key server which is given membership information to exercise group access control. When a client wants to join the group, the client and key server mutually authenticate using an authentication protocol. Having been authenticated and accepted into the group, each member shares with the server a key to be called the member's *individual key*. For group communications, the server distributes to each member a *group key* to be shared by all members of the group [4]. According to Chung et al. [4], secure multicast was just only focus on binary key graph. In [5], Wen et al. extended the binary key tree by key tree of degree *n*, and propose the optimal degree key tree when rekeying.

However, in order to apply secure multicast mechanism in our model, we propose a modified mechanism as Fig. 5.

Fig. 5 show how key server (KS) establishes and distributes the keys for layer 2 and layer 3. We implement tree matrix instead of key tree. Assume system has *n* applications and *m* services, we have *m*x*n* key matrix. Initially, KS authenticates and assigns an individual key for the client (*App* or *SVR*), e.g. individual key of *App 1* is K_{A1}. After that, KS calculates pair-wise key for each element of matrix by Hashing Message Authentication Code (HMAC) function (e.g. $\{K_{A1},K_{S1}\}=HMAC(K_{A1})$ with key K_{S1}).

For layer 2 and layer 1 security, the secure multicast is used as the optimal mechanism. One application wants to request one or many geographical area (VG), it only uses key of that group and multicast message to the group. By this way, *App* just only encrypts one times. Fig. 6 describes the example of key management for layer 2 and layer 1 with 3 VGs.

In this case, K_1, K_2, K_3 are sequent individual key of *VG1*, *VG2* and *VG3*. $K_{12}=HMAC\{K_1, K_2\}$, i.e. K_1 is encrypted by key K_2. In case Fig.6, *VG1* keeps set of key $\{K_1, K_{12}, K_{13}, K_{123}\}$, *VG2* keeps set of key $\{K_2, K_{12}, K_{23}, K_{123}\}$ and *VG3* keeps the set of key $\{K_3, K_{13}, K_{23}, K_{123}\}$. And all *App*s keep all these keys. Assume *App* want to send message to *VG1*, *VG2*, and *VG3*, it just encrypts one message by K_{123}.

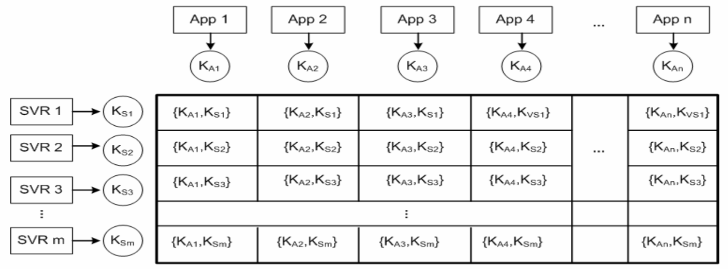

App 1
App 2
App 3
App 4
...
App n
K_{A1}
K_{A2}
K_{A3}
K_{A4}
K_{An}
SVR 1
SVR 2
SVR 3
SVR m
K_{S1}
K_{S2}
K_{S3}
K_{Sm}
$\{K_{A1},K_{S1}\}$ $\{K_{A2},K_{S1}\}$ $\{K_{A3},K_{S1}\}$ $\{K_{A4},K_{VS1}\}$ $\{K_{An},K_{VS1}\}$
$\{K_{A1},K_{S2}\}$ $\{K_{A2},K_{S2}\}$ $\{K_{A3},K_{S2}\}$ $\{K_{A4},K_{S2}\}$ ... $\{K_{An},K_{S2}\}$
$\{K_{A1},K_{S3}\}$ $\{K_{A2},K_{S3}\}$ $\{K_{A3},K_{S3}\}$ $\{K_{A4},K_{S3}\}$ $\{K_{An},K_{S3}\}$
$\{K_{A1},K_{Sm}\}$ $\{K_{A2},K_{Sm}\}$ $\{K_{A3},K_{Sm}\}$ $\{K_{A4},K_{Sm}\}$ $\{K_{An},K_{Sm}\}$

Fig. 5. Key management for layer 2 and layer 1

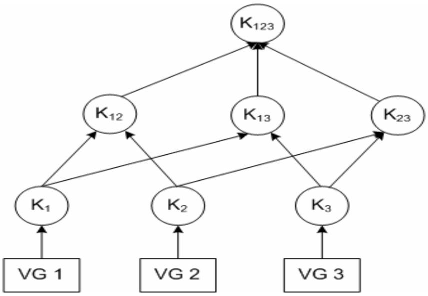


Fig. 6. Key management

```
array parent[];
program CheckParent(node[i],node[j])
  for(i = 1; i < length(parent); i++)
     if(parent[i]⊃node[i] & parent[i]⊃node[j])
        return 1;
     else return 0;
     end;
  end;
end;

program CreateKeyTree(node[n])
  for(i = 1; i < n; i++)
     for(j = 1; j < n; j++)
        if (!CheckParent(node[i],node[j]))
           parent=parent ∪HMAC(node[i],node[j]);
        end;
     end;
  end;
  Distribute parent[] to corresponding clients
  node=parent;
  parent=[];
  CreateKeyTree(node); //recursion to create new parent
end;
```

The above algorithm shows how KS establishes full tree keys mechanism for clients with *n VG*s.

4 Security Analysis

Assume system assures security between user or customer and Control Gateway, and security between Virtual Gateway and Gateway (USN). In this paper, we focus on security inside system of Cloud Service Provider.

To analysis security in the system, we propose the formula which calculates the total encryption cost. Total encryption cost is number of messages which are encrypted by *CG*, *Svr*, *App* and *VG*. In this part, we focus on case one service – many applications model. Assume that one service (*Svr*) includes λ applications (*Apps*) ($\lambda > 0$). Application i aggregates data from k_i areas (*VGs*) ($k_i > 0$).Therefore, the total encryption cost (E) for without secure multicast mechanism is:

$$E = \lambda + \sum_{i=1}^{\lambda} k_i + 1$$

From this formula, the total encryption cost in one service – many applications model is: $E = 9$ ($\lambda = 3$, $k_1 = 1$, $k_2 = 2$, $k_3 = 2$). In case one service – one application, $\lambda = 1$.

In Fig. 6, KS will create keys follows this hierarchical mechanism and distribute to each *VG*. In this case, *VG1* keeps set of key $\{K_1, K_{12}, K_{13}, K_{123}\}$, *VG2* keep set of key $\{K_2, K_{12}, K_{23}, K_{123}\}$, *VG3* keeps the set of key $\{K_3, K_{13}, K_{23}, K_{123}\}$, and all *Apps* keep set of key $\{K_1, K_2, K_3, K_{12}, K_{13}, K_{23}, K_{123}\}$. Then, the secure process of the system for Fig. 4 follows:

1. $CG \rightarrow Svr\,1$: $\{Usr\,1, Svr\,1\}K_{S1}$
2. $Svr\,1 \rightarrow App\,1$: $\{VG\,1\}K_{S1A1}$
3. $Svr\,1 \rightarrow App\,2$: $\{VG\,2,\, VG\,3\}K_{S1A2}$
4. $Svr\,1 \rightarrow App\,3$: $\{VG\,1,\, VG\,3\}K_{S1A3}$
5. $App\,1 \rightarrow VG\,1$: $\{App\,1\}K_1$
6. $App\,2 \rightarrow VG\,2,\, VG\,3$: $\{App\,2\}K_{23}$
7. $App\,3 \rightarrow VG\,1,\, VG\,3$: $\{App\,3\}K_{13}$

Therefore, the total encryption cost is 7 instead of 9. Generally, the total encryption cost when applying secure multicast is $E = 2 * \lambda + 1$.

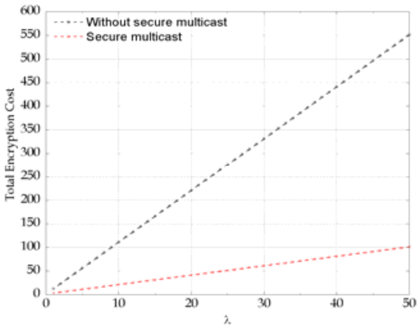


Fig. 7. Total encryption cost with λ=[1,50], k = 10

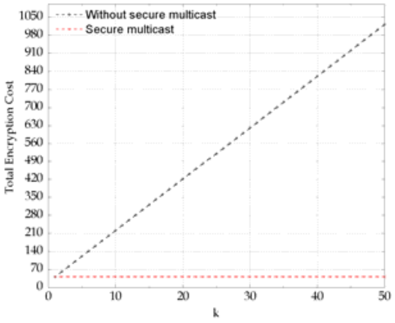


Fig. 8. Total encryption cost with λ=20, k= [1,50]

Fig. 7 shows that system with secure multicast mechanism is better than without secure multicast. With k is constant (k = 10), the total encryption cost of proposed scheme is always less than when increasing λ. In particularly, when λ= 20, total encryption cost of secure multicast and without multicast are nearly 40 and 220 respectively. In the other hand, with the fixed λ = 20, when k is scalable from 1 to 50 (Fig. 8), it is clearly that total encryption cost of secure multicast case is not effected by value of k.

5 Conclusion

In this paper, we present the system model for combining Cloud Service and Ubiquitous Sensor Network (USN) to provide more enhanced services to customers. We have proposed 3 layers system model with secure multicast mechanism. In our security scheme, total encryption cost between layer 2 and layer 1 are reduced much. However, to make our system more efficient, in the next work, we focus on reducing the total encryption cost between the layer 3 and layer 2.

Acknowledgement

This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MEST) (No. 2010-0016959).

References

1. Judge, P., Ammar, M.: Security issues and solutions in multicast content distribution: a survey. IEEE Network 17, 30–36 (2003)
2. Wong, C.K., Gouda, M., Lam, S.S.: Secure group communications using key graphs. IEEE/ACM Trans. Networking 8, 16–31 (2000)
3. Wallner, D.M., Harder, E.J., Agee, R.C.: Key management for multicast: Issues and architectures. In: RFC, vol. 2627 (June 1999)
4. Zhu, W.T.: Optimizing the Tree Structure in Secure Multicast Key Management. IEEE Commun. Lett 9(5) (May 2005)
5. Wong, C.K., Gouda, M., Lam, S.S.: Lam Secure Group Communications Using Key Graphs. IEEE/ACM Transactions on Networking 8(1) (February 2000)
6. Bird, R., Gopal, I., Herzberg, A., Janson, P., Kutten, S., Molva, R., Yung, M.: The kryptoknight family of light-weight protocols for authentication and key distribution. IEEE/ACM Trans. Networking 3, 31–41 (1995)
7. Steiner, J.G., Neuman, C., Schiller, J.I.: Kerberos: An authentication service for open network systems. In: Proc. USENIX Winter Conf., pp. 191–202 (Febuary1988)
8. Tardo, J.J., Alagappan, K.: SPX: Global authentication using public key certificates. In: Proc. 12th IEEE Symp. Research in Security and Privacy, May 1991, pp. 232–244 (1991)
9. Woo, T.Y.C., Bindignavle, R., Su, S., Lam, S.S.: SNP: An interface for secure network programming. In: Proc. USENIX 1994 Summer Technical Conf., Boston, MA (June 1994)
10. Amazon Elastic Compute Cloud, `http://aws.amazon.com/ec2` (access on October 2009)

11. Blue Cloud, IBM project,
    http://www-03.ibm.com/press/us/en/pressrelease/22613.wss/
    (access on October 2009)
12. Inoue, M.: A model and system architecture for ubiquitous sensor network businesses. In: Innovations for Digital Inclusions, 2009. K-IDI 2009. ITU-T Kaleidoscope, August 31-September 1, pp. 1–8 (2009)
13. Cloud Security Alliance, Top Threat to Cloud Computing,
    http://www.cloudsecurityalliance.org/
14. Abawajy, J.: Determining Service Trustworthiness in InterCloud Computing Environments. In: 10th International Symposium on Pervasive Systems, Algorithms, and Networks (I-SPAN 2009), Kaoshiung, Taiwan (2009)
15. Klein, C., Kaefer, G.: From smart homes to smart cities: Opportunities and challenges from an industrial perspective. In: Balandin, S., Moltchanov, D., Koucheryavy, Y. (eds.) NEW2AN 2008. LNCS, vol. 5174, p. 260. Springer, Heidelberg (2008)
16. Vouk, M.A.: Cloud computing issues, research and implementation. In: 30th International Conference on Information Technology Interface, June 23-26 pp. 31–40
17. Aymerich, F.M., Fenu, G., Surcis, S.: An approach to cloud computing network. In: 1st International Conference on the Applications of Digital Information and Web Technologies, Ostrava, Czech Republic, August 4-6, pp. 120–125 (2008)
18. Sloan, K.: Security in a virtualized world. Network Security 2009(8), 15–18 (2009)
19. Li, L., Chou, W.: Rich presence authorization using secure web services. In: 5th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE 2008), pp. 199–204 (2008)
20. Shang, C.W., Yang, Z.K., Liu, Q.T., Zhao, C.L.: A context based dynamic access control model for web service. In: EUC 2008: Proceedings of the 5th International Conference on Embedded and Ubiquitous Computing, vol. 2, pp. 339–343 (2008)
21. Wu, J., Huang, Z.M.: Proxy-based web service security". In: Proceedings of 3rd IEEE Asia-Pacific Services Computing Conference (APSCC 2008), Yilan, Taiwan, vol. 1-3, pp. 1282–1288 (December 09-12, 2008)
22. Koshutanski, H., Massacci, F.: Interactive access control for autonomic systems: from theory to implementation. ACM Transactions on Autonomous and Adaptive Systems 3(9) (August 2008)
23. Chao, H.C., Liu, T.J., Chen, K.H., Dow, C.R.: A seamless and reliable distributed network file system utilizing webspace. In: Proceedings 10th IEEE International Symposium on Web Site Evolution (WSE 2008), Beijing, Peoples R. China, pp. 65–68 (October 3-4, 2008)

Mining on Terms Extraction from Web News

Li-Fu Hsu

Hwh Hsia Institute of Technology, Department of Information Management,
235 Taipei, Taiwan
Lhhsu@cc.hwh.edu.tw

Abstract. Thousand of news stories are reported each day. How to extract the useful information from the large web news is the important technology today. However, information technology advances have partially automated to processing documents, reducing the amount of text which must be read. In this paper we present a Web News Search System, called WNSS. WNSS can discover automatically phrase extraction from large corpora of web news stories. In addition, we give concrete examples of how to preprocess texts based on the intended use of the discovered results. We also evaluate the extracted phrases can be used for important tasks.

Keywords: web news, information technology, phrase extraction, pre-process texts.

1 Introduction

In today's widespread internets, data mining in recent years is a very important issue, especially in this decade of information explosion. How to use information retrieval technology to manage and extract data to useful information is currently one of the important researches.

To effectively dig for useful information within the huge amount of data, many studies have devoted to various kinds of data mining techniques. The users may need an overall view of the web news collection: what topics are concerned, what kind of web news exist, are these news related, and so on. On the other hand, the user may want to find specific information content.

A common feature for all the tasks mentioned is that the user does not know exactly what he or she is looking for. Hence, a data mining approach should be appropriate, since by definition it is discovering interesting regularities from the data sets.

Surprisingly enough, only a few examples of data mining in text, or text mining, are available. The most notable example is the Document Explorer [1] used in mining Returns news articles.

Our study has two main aims: (1) to adopt data mining to build a model for general data mining methods are applicable to we news analysis tasks and (2) to help the web users to meet their requirements of their web news' target . In this paper, we proposed an algorithm to establish the Web News Search System (called WNSS) which can categorize the information and through the association rules establish that can help to find the useful information.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 188–194, 2010.

The article is organized as follows. In Section 2 we briefly describe the related works. In Section 3 we present the Web News Search algorithm and analyze its performance. In Section 4 we describe several experiments we did with in implementation of our algorithm on large data sets. Finally, Section 5 we present our conclusions and future work.

2 Related Work

Data mining, text mining, web news mining and other knowledge discovering techniques have become an attractive research area in the past years. The enormous amount of information often offers potential solutions to some problems. Web news mining extended the functions of traditional web search engines. Web content miners not only do simple news search but also try to extract implicit information by categorizing, filtering and interpreting web news. To achieve these functions, people either develop intelligent web agents [2] for various demands or set up multilevel databases based on the web information and web query systems [3].

Web news mining is one application of the text mining. Text mining or knowledge discovery from text (KDT) is the first time mentioned in Fedlman & Dagan [4], which deals with the machine supported analysis of text. It uses techniques from information retrieval, information extraction as well as natural language processing (NLP) and connects them with the algorithms and methods of KDD, data mining, machine learning and statistics. Thus, one selects a similar procedure as with the KDD process, it not data in general, but text documents are in focus of the analysis.

Text mining can be also defined as the application of algorithms and methods from the machine learning and statistic to texts with the goal of finding useful patterns. For this purpose it is necessary to pre-process the texts accordingly. Many authors use information extraction methods, natural language processing or some simple preprocessing steps in order to extract data from texts. To the extracted data then data mining algorithms can be applied [5],[6].

Current researches in the area of text mining classify problems of text, classification, clustering, information extraction or the search for and modelling of hidden patterns.

Following the knowledge discovery process model [7], we frequently find in literature text mining as a process with a series of partial steps, among other thins also information extraction as well as the use of data mining or statistical procedures. Hearst summarizes this in Hearst [8] in a general manner as the extraction of not yet discovered information in large collections of texts. Also Kodratoff [9] and Gomez in Hidalgo [10] consider text mining as process orientated approach on texts.

In this article, we consider web news mining mainly as text data mining. Thus, our focus is on methods that extract useful patterns from news in order to, e.g., categorize or structure news collections or to extract useful information.

3 Web News Search System

The method for web news search system will use the text data mining concept and we defined a general architecture for web news mining process in Fig. 1 that enables the implementation for the process of terms extraction from the web news. In order to meet the needs of web news terms pattern, which is different from the association-rule pattern. The mining method will first retrieve a data bank once, extract the news' terms, and use the clustering and association rules discovery the frequent items to find the useful news' term. This section describes the two components of Web News Search System module: term generation and term filtering.

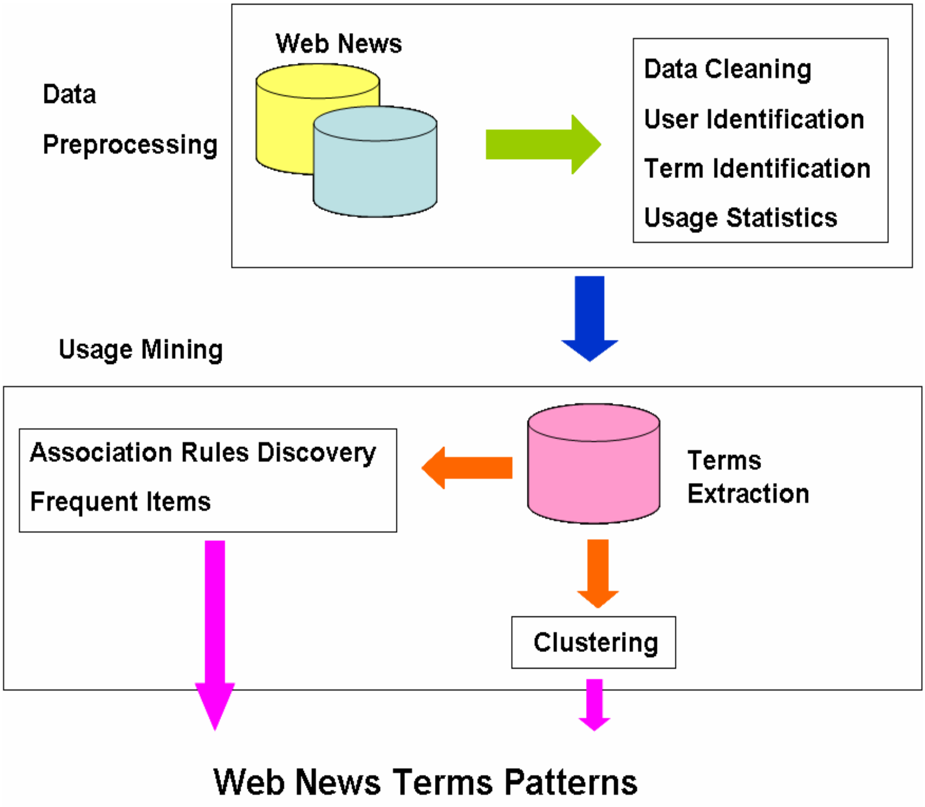
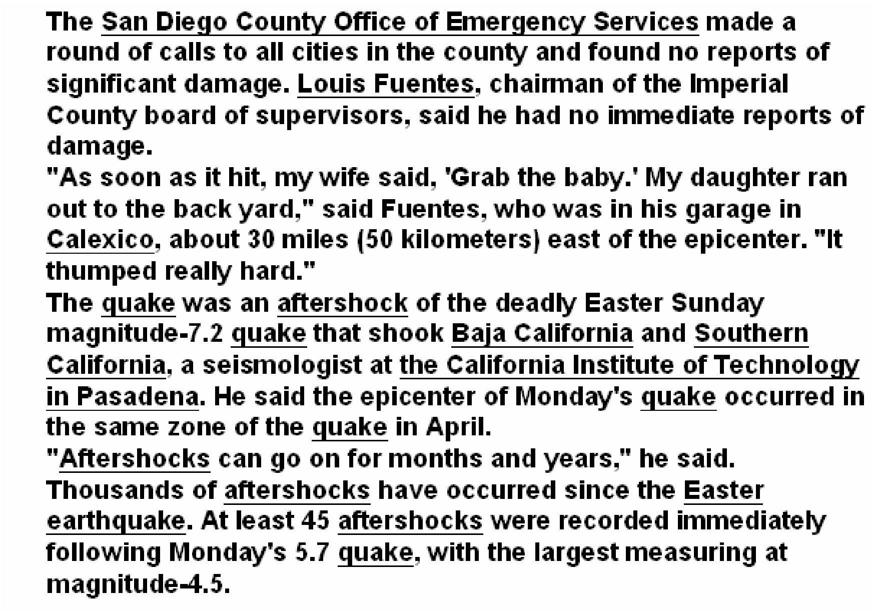


Fig. 1. The web news mining process

3.1 The Web News Terms

In this paper we perform term extraction on web news to find word sequences that are likely to have meaning in the domain, and then perform mining on the extracted terms labeling web news. As Fig.2 shows a fragment of web news with extracted terms underlined. Unlike word-based approaches, the extracted terms are fewer in number and tend to represent more meaningful concepts in the domain of the web news. Unlike keyword approaches, our term-extraction method eliminates much of the difficulties in labeling web news when faced with a new collection or new keywords.



The San Diego County Office of Emergency Services made a round of calls to all cities in the county and found no reports of significant damage. Louis Fuentes, chairman of the Imperial County board of supervisors, said he had no immediate reports of damage.

"As soon as it hit, my wife said, 'Grab the baby.' My daughter ran out to the back yard," said Fuentes, who was in his garage in Calexico, about 30 miles (50 kilometers) east of the epicenter. "It thumped really hard."

The quake was an aftershock of the deadly Easter Sunday magnitude-7.2 quake that shook Baja California and Southern California, a seismologist at the California Institute of Technology in Pasadena. He said the epicenter of Monday's quake occurred in the same zone of the quake in April.

"Aftershocks can go on for months and years," he said. Thousands of aftershocks have occurred since the Easter earthquake. At least 45 aftershocks were recorded immediately following Monday's 5.7 quake, with the largest measuring at magnitude-4.5.

Fig. 2. Example of the output of the term extraction module. Terms chosen to label the web news are underlined.

3.2 Term Extraction

The term extraction module is responsible for labeling web news with a set of terms extracted from the web news. Figure 2 gave an example of the results of this process on an excerpt of web news published on yahoo. Terms in this excerpt that were identified and designated as interesting by the term extraction module are underlined. Thus, for example, “quake” and “aftershock” are both extracted terms that would be used to label this web news.

3.3 Term Generation

In the term generation stage, sequences of tagged lemmas are selected as potential term candidates on the basis of relevant morpho-syntactic patterns, such as “Noun Noun”, “Noun Preposition Noun”, “Adjective Noun”, etc. The candidate combination stage is performed in several passes. In each pass, association rules are found and association coefficients between each pair of adjacent terms are calculated.

In the case of competing possibilities involving overlapping terms, such as (t1 t2) and (t2 t3) in (t1 t2 t3), the pair having the better association coefficient is replaced first. The news are then updated by converting all combined terms into a new single term and the whole procedure is then repeated until no new terms are generated.

The nature of the patterns used for candidate generation is an open research question. Daille [11] proposed specific operators to select longer terms as combinations of shorter ones. Justeson and Katz [12] suggest accepting prepositions as well as adjectives and nouns. This approach generates a much larger number of term results; Frantzi [13] only accepts Noun sequences to reduce the amount of "bad" terms.

To compute the association coefficient for combining two terms we currently use an ad hoc co-occurrence metric that computes a function of the number of times that the two terms match the possible extraction patterns. The term generation process combine two terms into a bigger term only if the value of the coefficient is over a threshold, τ . Although WNSS provides a default value for this threshold, it was designed a user can vary the threshold to affect the term generation process. For the experiments reported later a fixed value of 8 was used.

3.4 Term Filtering

The term generation stage produces a set of terms associated with web news without taking into account the relevance of these terms in the framework of the whole news collection. We therefore allow the term generation stage to create more terms than is truly desired, complementing generation with an additional filtering stage that prunes generated terms based on their frequencies of occurrence throughout the collection. For example, the following are examples of two-word terms that were identified during term generation, but were later eliminated during term filtering in one sample text-mining session: long-ago, same time, right hand…etc.

Our goal in term filtering is to identify terms that may not to be of interest in the context of the whole document collection either because they do not occur frequently enough or because they occur in a fairly constant distribution among the different news.

As in term generation, WNSS allows a user to select and combine these filtering methods if the user desires such control over the term generation process. For the experiments given later only the tf-idf based filter was used, with a fixed threshold of 4.5.

4 The Experiment Setup and Analysis

In this experiment, we adopt a 500MHz Pentium III computer, with an environment of 2048MB main memory and 576MB virtue memory, and use Visual C# 2005 to develop codes used from simulating the generation of datasets to implementing the WNSS algorithm. For most of what follows use 32,753 web news from the 125 website in the world for 2008-2009. This collection is 86M in size and contains over 150,000 unique words. Each web news contained on average 524 words. In the term generation stage, 1.2M terms were identified, 117k of them is unique.

Figure 3 givens an example of data preprocessing, when a user search a keyword with the different syntactical feature, and Figure 4 gives an example of a user requesting associations in various web news. The user constrains the left-hand side(LHS) of the association to contain extracted terms, and right-hand side (RHS).

Using the WNSS generated 12,000 frequent sets complying with restriction specified by association-rule query which uses a support threshold of 5 web news and confidence threshold of 0.1. These frequent sets generated 575 association rules.

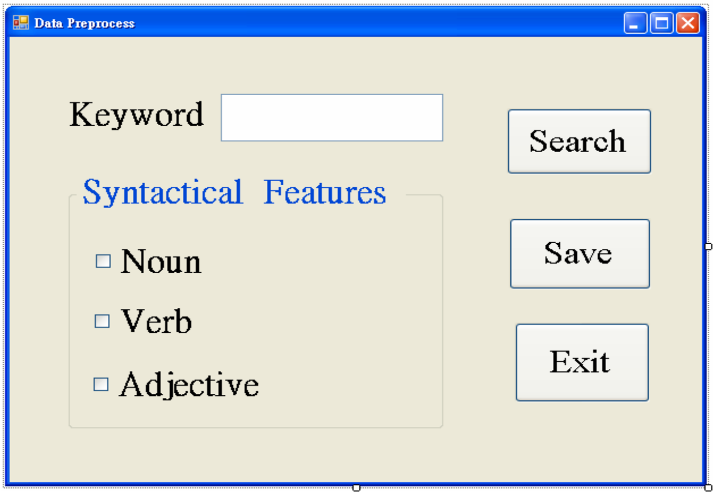


Fig. 3. Data Preprocessing of the syntactical features

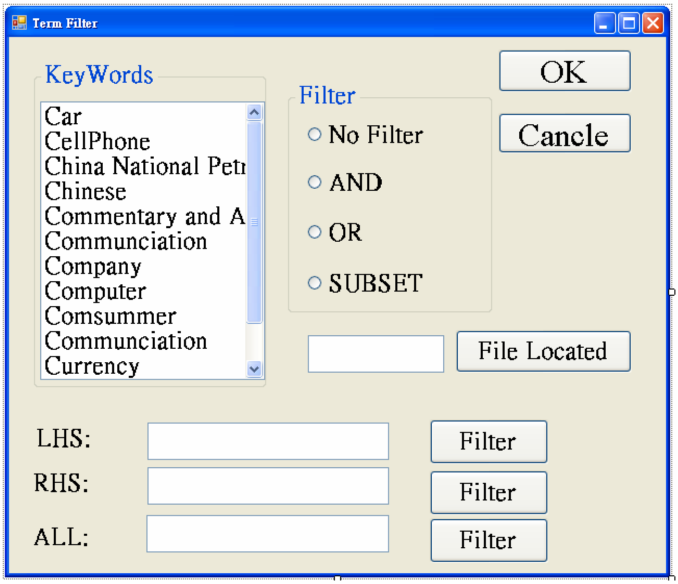


Fig. 4. Term Filter to generate association rules

5 Conclusion and Prospective Outlook

In this paper, we describe the use of data mining techniques to analyze web news collected from published on the Web news. We also proposed the Web News Search System (called WNSS) to find the useful information. These include association rules induction and clustering discovery.

Web news mining at the terms level thus hits a useful middle ground on the quest for understanding the information present in the large amount of data that is only available in textual form. Web news mining at the terms level serves as a powerful technique to manage knowledge encapsulated n large web news collections.

References

1. Feldman, R., Kloesgen, W., Zilberstein, A.: Document explorer: Discovering knowledge in document collections. In: Raś, Z.W., Skowron, A. (eds.) ISMIS 1997. Lecture Notes in Computer Science, LNAI, vol. 1325, pp. 137–146. Springer, Heidelberg (1997)
2. Brown, C.M., Danzig, B.B., Hardy, D., Manber, U., Schwartz, M.F.: The harvest information discovery and access system. In: Proc. 2nd International World Wide Web Conference (1994)
3. Konopnicki, D., Shmueli, O.: W3QS: A query system for the World Wide Web. In: Proc. of the 21th VLDB Conference, pp. 54–65 (1995)
4. Feldman, R., Dagan, I.: Kdt -knowledge discovery in texts. In: Proc. of the First Int. Conf. on Knowledge Discovery (KDD), pp. 112–117 (1995)
5. Nahm, U., Mooney, R.: Text mining with information extraction. In: i Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases (2002)
6. Gaizauskas, R.: An information extraction perspective on text mining: Tasks, technologies and prototype applications (2003), `http://www.itri.bton.ac.uk/projects/euromap/TextMiningEvent/Rob_Gaizauskas.pdf`
7. Crispdm and CRISP,: Cross industry standard process for data mining (1999), `http://www.crisp-dm.org/`
8. Hearst, M.: Untangling text data mining. In: Proc. of ACL 1999 the 37th Annual Meeting of the Association for Computational Linguistics (1999)
9. Kodratoff, Y.: Knowledge discovery in texts: A definition and applications. In: Raś, Z.W., Skowron, A. (eds.) ISMIS 1999. LNCS, vol. 1609, pp. 16–29. Springer, Heidelberg (1999)
10. Hidalgo, J.: Tutorial on text mining and internet content filtering. Tutorial Notes Online (2002), `http://ecmlpkdd.cs.helsinki.fi/pdf/hidalgo.pdf`
11. Daille, B., Gaussier, E., Lange, J.M.: Towards Automatic Extraction of Monolingual and Bilingual Terminology. In: Proceedings of International Conference on Computational Linguistics, COLING, pp. 515–521 (1994)
12. Justeson, J.S., Katz, S.M.: Technical Terminology: Some linguistic properties and an algorithm for identification in text. Natural Language Engineering 1(1), 9–27 (1995)
13. Frantzi, T.K.: Incorporating Context Information for the Extraction of Terms. In: Proceedings of ACLEACL 1997 (1997)

Inter-cloud Data Integration System Considering Privacy and Cost

Yuan Tian, Biao Song, Junyoung Park, and Eui-Nam Huh

Department of Computer Engineering
Internet Computing and Network Security Lab
KyungHee University Global Campus, South Korea
{ytian,bsong,parkhans,johnhuh}@khu.ac.kr

Abstract. In spite of all the advantages delivered by cloud computing, still many challenges are hindering the migration of customer software and data into the cloud. Whenever information is shared in the cloud, privacy and security questions may arise. Although many technologies have been proposed in order to meet users' requirements from privacy concern, however, at the same time, with the increasing number of processed data, the cost for privacy protection also increases dramatically. In this paper, we present a privacy-aware inter-cloud data integration system considering tradeoff between the privacy requirements from users and the charging for those data protection and processing. In contrast to existing data sharing techniques, our method is more practical as the cost for technical supporting privacy must be considered in the commoditized cloud computing environment.

Keywords: Cloud computing, Privacy, Service pricing.

1 Introduction

Cloud computing becomes increasingly pervasive and more and more people are starting to take advantage of the power of the cloud these days. Comparing with the traditional systems, which are masked behind firewalls and other gateway boundaries and attackers must do intensive intelligence gathering to know that they exist, cloud computing, is highly visible and are designed to be accessible to anywhere by anyone.

Similar to the real world utilities, nearly all the services provided in cloud computing are on-demand and highly commoditized, based on consumers' usage and quality of service expectations, they just need to pay for those services like water, electricity, gas and telephony [6], rather than investing heavily to maintain their own computing infrastructure.

Cloud computing fundamentally shifts the traditional "desktop as a platform" to "internet as a platform", however, at the same time, arises more problems than ever before, especially when valuable data from thousands of users in a single site are being attacked [2], which also means, whenever information is shared in the cloud, privacy and security questions may arise. When a user stores his data in a third party like cloud computing provider, those data may have fewer or weaker privacy protections than when the data just remains in the possession of this user.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 195–204, 2010.

However, there are few privacy laws apply to restrict the disclosure of customers or employees' personal information from a business to the cloud provider. Even privacy laws apply to particular categories of customer or employee information, disclosure to a cloud provider may not be restricted, as current laws that protect electronic communications may apply differently to different aspects of cloud computing [1]. We can only enjoy the full benefits of Cloud computing if we can address the very real privacy and security concerns that come along with storing sensitive personal information in databases and software scattered around the Internet [1].

In order to meet the growing requirement for privacy management, many privacy technologies have been proposed. Those technologies meet users' requirements from privacy concern, however, at the same time, with the increasing number of processed data, the cost for privacy protection also increases dramatically.

In [4], Yau et al present a method which could safeguard the privacy of data in a very secure way. They provided a privacy preserving repository to accept integration requirements from users, help data sharing services share data, collect and integrate the required data from data sharing services, and return the integration results to users. The main contribution of their work is that the processing of data is kept securely in both data sharing services and repository: data is randomized before sending to repository and encryption/decryption are used from information releasing in the repository.

However in the above method, they did not consider the practicability of uploading all the unprocessed data to the repository, as in the real scenario, the high cost for transmitting data is even more unbearable compare with keeping the privacy of data which is not that important [8, 9]. As a vital building block in many fields, privacy is possible to be a new service in the cloud environment. Consumers do not have to concern about how to protect their privacy by which technology, instead, according to consumers' requirements, the privacy services can be offered by providers to execute their applications only if consumer pay for that [3]. Both service provider and consumer would like to pay different price for those privacy services with different protection assurances according to the importance of their data.

Thus, based on the above analysis, we present a privacy-aware inter-cloud data integration system considering tradeoff between the privacy requirements from users and the charging for those data protection and processing. In general, it is up to customers to decide on a strategy of how to get a service fulfilled on the basis of their personal feeling of the importance of their data, and the cost for processing data.

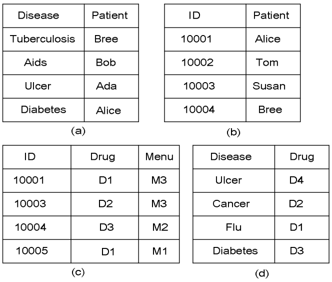
The structure of the rest of the paper is as follows. Section 2 introduces a scenario that is used as a running example throughout the paper. A component-based view of the proposed system is presented in Section 3 and the design approach is given in Section 4, and the last section of this paper presents our conclusions.

2 A Motivating Example

This section presents a scenario used throughout the paper and we use this motivating example to show how our system works. The scenario is a revised version of the case study proposed in [4, 7].

In a healthcare system there are multiple collaborated clouds which participate in processing, sharing and integrating data. We assume for the purpose of getting the menu which is proposed to ulcer suffering, the clouds may contains the data from medical research institutes, hospitals and pharmacies. For simplicity, only four databases are considered: a disease record database *T1(Disease, Patient)* which storing patient's names and corresponding diseases which they are suffering, an identification information database *T2(ID, Patient)* which stores patients' names and their IDs, a hospital database *T3(ID, Drug, Menu)* storing hospitalization information and a pharmacy database *T4(Disease, Drug)* which stores popular drugs for each disease.

Table 1. The databases which used in the motivating example



Disease	Patient
Tuberculosis	Bree
Aids	Bob
Ulcer	Ada
Diabetes	Alice

(a)

ID	Patient
10001	Alice
10002	Tom
10003	Susan
10004	Bree

(b)

ID	Drug	Menu
10001	D1	M3
10003	D2	M3
10004	D3	M2
10005	D1	M1

(c)

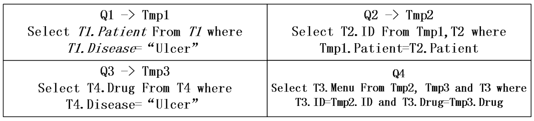
Disease	Drug
Ulcer	D4
Cancer	D2
Flu	D1
Diabetes	D3

(d)

(a) Disease records T1 (b) Identification Information T2 (c) Hospital T3 (d) Pharmacy T4

Now we are going to express our motivating example by the following four SQL queries which shown in Table 2. Query Q1, Q2 and Q3 generates three temporary tables Tmp1, Tmp2 and Tmp3 respectively, and the final results are from the last query Q4.

Table 2. The integrated query



Q1 -> Tmp1 Select *T1.Patient* From *T1* where *T1.Disease*= "Ulcer"	Q2 -> Tmp2 Select T2.ID From Tmp1,T2 where Tmp1.Patient=T2.Patient
Q3 -> Tmp3 Select T4.Drug From T4 where T4.Disease= "Ulcer"	Q4 Select T3.Menu From Tmp2, Tmp3 and T3 where T3.ID=Tmp2.ID and T3.Drug=Tmp3.Drug

As some queries may need other queries' results as inputs, the repository randomizes those results by using Hush function, which avoid the need for the repository to know that results but still keep the mapping relation between data. For example, we replace the Q1's results {Bob, Alice} by {H(Bob), H(Alice)} which protects Q1's results without disturbing Q2, as the hashed name usually remains unique, the repository can easily evaluate Q2 by comparing H(Tmp1. Patient) and H(T2.Patient). Since H(Susan) is not in the Q1's hashed result{H(Bob), H(Alice)}, the repository can find the patient Susan whose ID is 10004 is not an Aids nor diabetes patient.

3 The Proposed System

Our proposed system aims to mediate between users' privacy preferences and the cost for privacy protection. The overall architecture of our system is illustrated in Fig.1.

The user sends his integration requirements to the repository cloud and only the required data for users' integration request is collected. The query plan wrapper in the repository cloud will correctly construct a query plan for users' integrated query, decompose the query into a set of sub-queries, and discover corresponding service providers. The data which was conducted or collected in the query plan wrapper are then sent to the query allocator.

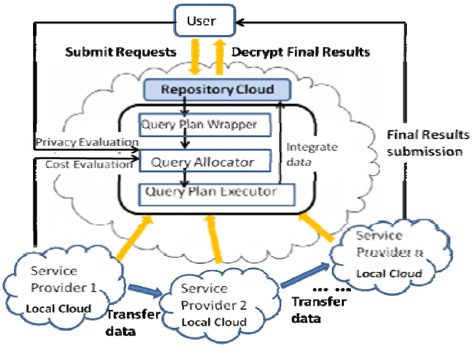


Fig. 1. System Architecture

Each sub-query could be executed in local cloud or repository cloud. For the former case, the sub-query will be executed in the service provider which stores the corresponding database. Thus, only the results for the sub-query are returned to the repository cloud or transferred to the successive service provider. The price for processing data in local cloud is relative cheaper, however, it may not be secure as both the data storing and processing are done in the same service provider. While the later one is to fetch data in local cloud, randomize all of those data and then send to the query plan executor for further processing. Executing query in query plan executor guarantees secure protection as the data is stored and processed separately. Therefore, the cost for processing data in repository cloud is high as it requires randomizing all the data in the local service provider and transfers it back to the repository cloud.

In order to help users find a balance between privacy insurance and the cost for privacy protection, the query plan executor decides at where the sub-query should be executed. Before sending those sub-queries to the query plan executor, three separating phases are proposed:

- ✓ Privacy evaluation: based on user's query, allows user to express their privacy preferences by setting risk values to decide the importance of data and the relationship between those data.
- ✓ Cost evaluation: here, the pricing mechanism is that the service provider in repository cloud should estimate a price for processing data, whereas other providers which located outside the repository cloud provides the costs for both processing data and uploading data to the repository cloud.
- ✓ Query Allocation: the query allocator chooses an optimal strategy according to user's preferences in the above two steps, and sends it to the query plan executor to enforce that method. At last, the requested data is sent back to user from the query plan executor.

4 System Design

In this section, we presented the process of our system implementation. Before discussing the functionality of our system, some basic definitions are introduced in order to establish a common ground of concepts.

Firstly, the query plan wrapper converts user's integrated query to a query plan graph G. For a integrated query, the query plan graph $G=\{V, E, C\}$ is a labeled directed acyclic graph. Among which, $V=\{v_1, v_2, \ldots, v_m,\}$ is a set of nodes where each v_i represents an intermediate search result; $E=(e_1, e_2, \ldots, e_l)$ is a set of edges where each edge $e_{ij}=(v_i, v_j)$ represents a data integration relation between v_i an v_j. $C=(c_1, c_2, \ldots, c_l)$ is a set of labels attached to each $e_{ij} \in E$ and each label $c_{i,j}=(op, attr1, attr2) \in C$ specifies that the data of v_i and v_j is integrated by the data integration operator *op* between v$_i$'s attribute attr1 and v_j's attribute attr2. Generally, the operator *op* can be any binary comparison operator chosen from $\{=, \neq, >, <\}$ or any aggregate operator chosen from $\{SUM, AVG, MAX, MIN\}$.

The motivating example in Section 2 can be represented by the query plan graph shown in Fig. 2. The query plan graph is decomposed to several sub-graphs $\{G_1, G_2, G_3, \ldots, G_n\}$ where each sub-graph represents a sub query in Table 2. The overlapping of sub-graphs shows that the immediate result from a sub-query is used as the input for another sub-query.

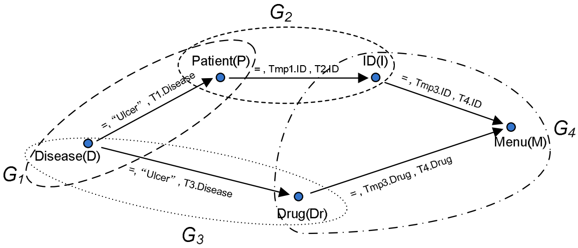


Fig. 2. The query plan graph from the motivating example

We use a matrix D_{m*m} to represent the disclosure condition of the searching results and the relationships between them. For each $d_{ii} \in D$, $d_{ii}=0$ denotes the information of v_i is not disclosed, while $d_{ii}=1$ denotes the information of v_i can be known by service provider. Similarly, $d_{ij}=0$ denotes the relationship between v_i and v_j is not disclosed while $d_{ij}=1$ shows the relationship between v_i and v_j that can be known by service provider. An instance of data disclosure condition matrix is shown in Fig.3.

As the service providers which locate in different clouds may collude each other in order to get more information that they are not supposed to get, i.e., if the relationship between disease and patient is disclosed to one service provider and the relationship between patient and ID is disclosed to another service provider, then is possible for them to find out the relationship between disease and ID. Thus, we define $d_{ij}=1$ if there exists a path P between v_i and v_j where the starting node of P is v_i, the ending node of P is v_j, and every edge in P has been disclosed.

	D	P	I	Dr	M
D	1	1	0	0	0
P	0	1	0	0	0
I	0	0	0	0	0
Dr	0	0	0	0	0
M	0	0	0	0	0

Fig. 3. An instance of a data disclosure condition matrix

As we mentioned before, each sub-query can be executed in local cloud or repository cloud, for the sub-graph G_k , we set $A(G_i) = 1$ to denote the former case and $A(G_i) = 0$ to denote the later one where $A(G_i)$ is the sub-allocation decision of sub-graph G_i. As each sub-graph contains nodes and edges, if the value of sub-graph is set as “1”, the values of all the nodes and edges within the sub-graph should be also set as “1”. For the nodes which may appear in several sub-graphs with different setting values “0” and “1”, we should consider this node (represents intermediate results) is executed in local cloud and set it’s value as “1” as it may be disclosed in the local cloud.

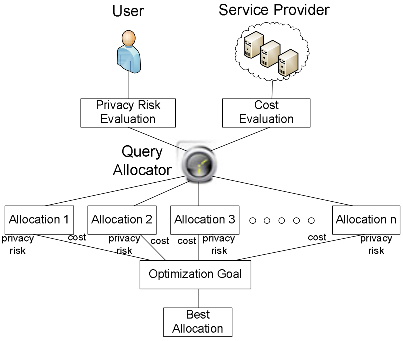


Fig. 4. Process of the proposed system

The process of our proposed approach is briefly presented in Fig.4. Based on user’s query, user and service providers first submit their evaluation for risk and cost evaluation respectively. After that, query allocator decides an optimal allocation which provides minimum trade-off value between user’s risk and service cost.

4.1 Risk Evaluation

Before submitting user’s sub-requests to the query plan executor, user could express his privacy preference by giving risk value to those intermediate nodes. Each node represents the results from the previous query and the values on the nodes show how sensitive that user care about those data or the relationships between the data.

	D	P	I	Dr	M
D	4	9	9	1	1
P	-	5	∞	-	5
I	-	-	3	-	1
Dr	-	-	-	3	1
M	-	-	-	-	0

Fig. 5. An instance of a penalty matrix

The risk values for the intermediate nodes are presented in a risk matrix R_{m*m} in Fig.5. The $r_{ii} \in R$ denotes the risk value on intermediate search result v_i and $r_{ij} \in R$ denotes the risk value on the relationship between the intermediate search result v_i and v_j. Every value in the matrix is an integer ranging from 0 to 9, or $+\infty$. For example, from Fig.5 we can see that the user gives the "Menu" information away and set the risk value to "0" as he may think this information is not that important. Whereas he cares which patients are suffer from which diseases as this kind of data disclosure may offend patients' privacy, so the risk value on the relationship between "Patient" and "Disease" is set to infinity.

$$\begin{array}{c|ccccc} & D & P & I & Dr & M \\ \hline D & 1 & 1 & 0 & 0 & 0 \\ P & 0 & 1 & 0 & 0 & 0 \\ I & 0 & 0 & 0 & 0 & 0 \\ Dr & 0 & 0 & 0 & 0 & 0 \\ M & 0 & 0 & 0 & 0 & 0 \end{array} * \begin{array}{c|ccccc} & D & P & I & Dr & M \\ \hline D & 4 & 9 & 9 & 1 & 1 \\ P & - & 5 & \infty & - & 5 \\ I & - & - & 3 & - & 1 \\ Dr & - & - & - & 3 & 1 \\ M & - & - & - & - & 0 \end{array} = 18$$

Fig. 6. Risk Penalty Value

Given disclosure condition matrix D and risk matrix R, the risk penalty value RPV can be calculated by the formula RPV=$\sum_{i=1}^{m}\sum_{j=1}^{m} d_{ij} \times r_{ij}$. An instance for showing how to calculate the risk penalty value is presented in Fig. 6.

4.2 Cost Evaluation

At the same time, the service providers should fix the service price considering different situations. Our pricing mechanism includes four kinds of costs:

1) Data randomization cost

As we mentioned before, the sub-query could be executed in the repository cloud in order to assure the data can be stored and processed separately. Thus, all the data in the local service provider needs to be randomized before transmitting to the repository cloud. Service provider evaluates this price according to the amount of data.

2) Cost for uploading all data

The user who wants to process the data in the repository cloud needs to upload all of the randomized data. This cost is for submitting those randomized data.

3) Query execution cost

The cost is for executing sub-query in local cloud or repository cloud. The query execution cost varies based on the amount of data.

4) Cost for uploading intermediate result

This cost is for transmitting the results of the sub-query after executing it locally. According to user's preference, the results can be transferred to next local server provider or to the repository cloud for further processing.

Fig.7 shows an example of our pricing model. We use notation $C_r(G_i)$ to denote the cost for processing a sub-query G_i in the repository cloud while $C_l(G_i)$ represents the cost for processing a sub-query G_i in the local cloud. If the user decides to execute sub-query in repository cloud, the service provider should firstly randomize the whole data and send it to repository cloud to execute the sub-query. So $C_r(G_i)$ includes the cost for randomization data, uploading all data, and executing query.

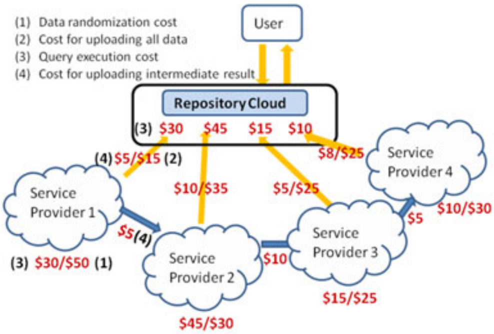


Fig. 7. Pricing Mechanism

For example, from Fig.7 we can calculate that $C_r(G_1) = \$50 + \$15 + \$30 = \$95$, which means the total cost for executing G_1 in repository cloud is \$95. On the other hand, if G_i is executed in local cloud, only the results of that sub-query will be submitted to the repository cloud or transferred to the successive service provider. So $C_l(G_i)$ includes the cost for executing query and uploading intermediate results. We can get $C_l(G_1) = \$30 + \$5 = \$35$ in the example in Fig.7, so the total cost for executing G_1 in service provider 1 is \$35.

4.3 Query Allocation

Based on the above analysis, the query allocator provides several strategies for user considering both price and privacy. User could select a service with lowest cost allocation which subjects to privacy risk constraint, or a service with lowest privacy protection which subjects to cost constraint, or a service considering trade-off between cost and privacy penalty. Based on user's choice, every possible allocation is examined and the best allocation will be selected.

Given a set of sub-allocation decision $\{A(G_1), A(G_2),..., A(G_n)\}$, the disclosure condition matrix D can be generated by combining all the sub-allocation decisions as follows:

$$\exists(v_i \in G_k \ \& \ A(G_k)=1) \xrightarrow{\forall i,k} d_{ii}=1$$

$$\exists(e_{ij} \in G_k \ \& \ A(G_k)=1) \xrightarrow{\forall i,j,k} d_{ij}=1$$

$$\exists(d_{ii}=1 \,\&\, d_{jj}=1 \,\&\, P=\{d_{ik_1}=1, d_{k_1k_2}=1, \ldots, d_{k_tj}=1\}) \xrightarrow{\forall i,j,k_1,\ldots,k_t} d_{ij}=1$$

Using RPV= $\sum_{i=1}^{m}\sum_{j=1}^{m} d_{ij} \times r_{ij}$ we can get the risk penalty value of the allocation. Then the total cost is calculated by using $\sum_{i=1}^{n} C(G_i)$ $where \begin{cases} C(G_i)=C_r(G_i) & when\ A(G_i)=0 \\ C(G_i)=C_l(G_i) & when\ A(G_i)=1 \end{cases}$

As there are four sub-queries in our motivating example, totally $2^4 = 16$ allocations are provided. In Table 3, we list all the possible allocations associated with the cost and privacy penalty. Suppose that the user prefers to select a service that produces lowest privacy penalty value while costs less than $250. From Table 3 we can see that the allocation which meets this requirement is {0,0,1,1}, which means, sub-query 1 and 2 should be executed in the repository cloud while sub-query 3 and 4 should be executed in the local cloud.

Table 3. Allocation Results

$\{A(G_1), A(G_2), A(G_3), A(G_4)\}$	Cost	Penalty	$\{A(G_1), A(G_2), A(G_3), A(G_4)\}$	Cost	Penalty
{0, 0,0,0}	335	0	{1,0,0,0}	275	18
{0,0,0,1}	288	8	{1,0,0,1}	228	26
{0,0,1,0}	290	8	{1,0,1,0}	230	22
{0,0,1,1}	205	14	{1,0,1,1}	183	28
{0,1,0,0,}	280	∞	{1,1,0,0}	220	∞
{0,1,0,1}	233	∞	{1,1,0,1}	173	∞
{0,1,1,0}	235	∞	{1,1,1,0}	175	∞
{0,1,1,1}	188	∞	{1,1,1,1}	128	∞

5 Conclusions and Future Work

In this paper, a privacy-aware inter-cloud data integration system is presented, which strikes a balance between the privacy requirements from users and the cost for data protection and processing. In contrast to existing data sharing techniques, our method is more practical as the cost for technical supporting privacy must be considered in the commoditized cloud computing environment. This work is still evolving and in the future work, we will consider more conflicted situation like how the system works under the trustable service providers.

Acknowledgment

"This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency)" (NIPA-2010-(C1090-1021-0003)) "This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MEST) (No. 2010-0016959)".

References

1. Privacy in the clouds: risks to privacy and confidentiality from cloud computing, `http://www.worldprivacyforum.org/pdf/WPF_Cloud_Privacy_Report.pdf`, (accessed on April 1, 2010)
2. Steve Mansfield-Devine, Danger in the clouds, Network security, `http://www.webvivant.com/dangers-in-the-cloud.html` (accessed on May 15, 2010)
3. Yeo, C.S., Venugopal, S., Chu, X., Buyya, R.: Autonomic Metered Pricing for a Utility Computing Service (2008)
4. Yau, S.S., et al.: A privacy preserving repository for data integration across data sharing services. EEE transactions on services computing
5. Michael Maximilien, E., Grandison, T., Sun, T., Richardson, D., Guo, S., Liu, K.:Privacy-as-a-Service: Models, Algorithms, and Results on the Facebook Platform. In: Web 2.0 Security and Privacy 2009, held in conjunction with the 2009 IEEE Symposium on Security and Privacy. Oakland, California (May 2009)
6. Foster, I., Kesselman, C. (eds.): The Grid 2: Blueprint for a new computing infrastructure. morgan Kaufmann, San Francisco (2003)
7. Canfora, G., Costante, E., Pennino, I., Visaggio, C.A.: A three-layered model to implement data privacy policies, Computer Standards & Interfaces. pp. 398–409. Elsevier Science Publishers B. V., Amsterdam (2008)
8. Yeoa, C.S., Venugopalb, S., Chua, X., Buyya, R.: Autonomic metered pricing for a utility computing service. Future Generation Computer Systems (2009)
9. Singh, T., Vara, P.K.: Smart Metering the Clouds. In: 18th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises, pp. 66–71. IEEE Computer Society, Washington

A Semantic Web Approach to Heterogeneous Metadata Integration

Shu-Hsien Liao[1], Hong-Chu Huang[2], and Ya-Ning Chen[1]

[1] Department of Management Sciences and Decision Making, Tamkang University, Taiwan
[2] Department of Information and Library Science, Tamkang University, Taiwan
{Michael,kuanin}@ mail.tku.edu.tw, arthur@sinica.edu.tw

Abstract. Heterogeneous metadata integration is an issue in digital libraries. Mapping is often used for an integrated metadata access, but the implicit knowledge and relations embedded in metadata are ignored. This paper aims to present a semantic web approach to heterogeneous metadata integration of biodiversity repositories. First, implicit knowledge and relations in metadata are extracted out and transformed into a shared ontology with expression of RDF and OWL languages. Next the shared ontology plays an inter-lingua role in harmonizing heterogeneous metadata to achieve an ontology mapping with a unified view. Then the shared ontology is expressed by SWRL for inference query to offer in-depth semantic discovery. Finally four question answering oriented queries are employed to examine the feasibility of the shared ontology for heterogeneous metadata integration.

Keywords: Information integration, heterogeneous metadata, semantic web, digital libraries.

1 Introduction

Many biodiversity heritage institutions have built up digital repositories to curate and manage their collections by digitizing biodiversity materials. Owing to various purposes and functions, digital repositories will adopt different metadata formats as data schemas or elements to describe their collections. Mapping is often-used method to achieve metadata integration. Usually, mapping just focuses on lexical equivalence of metadata elements from source format to target one, but contextual relationships between metadata elements are ignored or lost.

[5] points that data integration refers to combining data in such as way that a homogeneous and uniform view presented to users. [16] also regards that heterogeneous information integration is referred to as information synthesis of different information and data sources across disparate systems on the supply chain. One may draw a generalization that integration of data or information involves two key issues: heterogeneous and distributed. Therefore, the ideal integrated metadata access also needs to solve similar issues.

Semantic web has been proposed as the next generation of web and is composed by three components: ontologies, XML and RDF, and inference rules [3]. Actually ontologies play a core role for semantic web in offering a set of common agreed terminologies

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 205–214, 2010.

and relations to harmonize semantic heterogeneity for distributed web-based databases and systems on Internet. Ontologies can be regarded as inter-lingua either to standardize terminologies or to provide the semantic foundations for translators [15]. Therefore, it becomes an issue how to implement the semantic web into distributed and heterogeneous biodiversity repositories to achieve semantic-oriented metadata integration.

2 Methodology

This article aims to use ontologies as a harmony approach for heterogeneous metadata integration at data schema or element level. Target subjects were selected from Catalogue of Life（hereafter CoF）and Specimens Database of Native Plants in Taiwan (hereafter SDNPT), as case study to illustrate the semantic web driven approach for integration of heterogeneous metadata. The system prototype integrated Protégé 3.4 with JavaBean program language to clarify the related ontology engineering tasks for metadata integration. The rest of this article is organized as follows. First, related work based on literature analysis motivates the contribution of our proposed approach. Next, detail of ontology construction and extended applications for metadata integration is illustrated, in terms of knowledge extraction, and ontology mapping. Moreover, a method is proposed to build a connection from ontology terminologies to mashup metadata for typical relational databases. In addition, four question answering (hereafter QA) oriented queries are deployed to examine the feasibility of proposed approach. Furthermore, the discussion will be presented. Finally, concluding remarks are drawn for future research.

3 Related Work

Up to date, several approaches have been proposed to offer an integrated access to heterogeneous metadata from distributed biodiversity repositories. In terms of data value, the ID-based [10] or name-based [13] linkage approach only uses data value as a pointer to retrieve related metadata for a specific species. This approach relies on an authoritative list of unified identifiers or data value to reconcile the issue of semantic heterogeneity for biodiversity species. To this day, an official authoritative list has not agreed in biodiversity heritage. On the other hand, mapping or crosswalk is another often-used approach to provide a harmony basis for metadata integration according to a specified metadata format, such as Dublin Core（hereafter DC） and Darwin Core. One strand approach employs OAI-PHM with DC as a data aggregation mechanism to harvest and integrate various metadata formats and their elements from distributed digital repositories [2]. The other strand approach combines a networked retrieval protocol（such as Z39.50 and DiGIR） with a unified metadata format（such as EML and Darwin Core） as a federated searching service to offer integrated access to various biodiversity repositories [4][11]. However, mapping is classified as a lexical mapping which is based on lexical form, appearance or meanings [6], and contextual information embedded in metadata formats is excludes out.

Few studies have focused on how to transform metadata into ontologies in cultural heritage. One study is to illustrate in mapping DC into CIDOC/CRM ontology for metadata integration [7], the other is to mapping from DC and EAD to CIDOC/CRM [14]. These approaches are based on an existing ontology to transform metadata elements into equivalent semantic terminologies and then select and build up the required ontological concepts and relations. However, the first prerequisite of this transformation lies in that a common ontology has existed and agreed as a domain of discourse to share and exchange knowledge for a specific domain. According to the above discussion, it is worth to explore how a semantic web approach to offer a unified logical view for metadata integration in distributed biodiversity repositories.

4 A Semantic Web Approach to Integrating Metadata

CoF and SDNPT are digital repositories to manage biodiversity information relating to species and specimens respectively. CoF is a typical relational database to record the species information for 50,804 unique species in Taiwan based on Species 2000 metadata format. SDNPT is also a RDB to manage plant specimens for herbarium with 50,027 specimen's records by adoption of the HISPID 3 format. Actually implicit knowledge of CoF and SDNPT is embedded in different metadata formats. Therefore, the approach proposed by this study is to extract implicit knowledge from biodiversity repositories and then transform into machine readable and understandable with standard expression of XML-based RDF, OWL and SWRL languages. The proposed approach is illustrated in detail as follows: knowledge extraction for building a shared ontology, ontology mapping, inference query, and metadata mashup.

4.1 Knowledge Extraction from Metadata Elements for Building Ontologies

Generally metadata is defined as data about data. Furthermore, metadata can be regarded as "a materialization of domain-related knowledge that facilitates the management of data warehouses and helps in achieving good performance"[12]. Therefore one may generalize that metadata is also a kind of knowledge with shared meaning and interpretation for specific user communities. However, metadata is still a human-understandable information with implicit knowledge. Our study has to transform implicit knowledge embedded in metadata into explicit one to build up a shared ontology for harmonizing heterogeneous metadata formats and their elements.

At this stage, we first adopted approach provided by [9] to extract knowledge manually from metadata of CoF and SDNPT repositories as follows: determine the domain and scope of the ontology, consider reusing existing ontologies, enumerate important terms in the ontology, define the classes and the class hierarchy, define the properties of classes, define the facets of the slots, and create instances. Then we used RDF data model to define classes and hierarchy, properties of classes and facets of the slots, to illuminate and re-contextualize the original ontological structure and relations embedded in metadata elements. Lastly, we inputted all classes, properties and their instances into Protégé 3.4 ontology editor to build and validate a shared ontology with expression of XML-based RDF and OWL languages. During this stage, this study is successful in transforming implicit knowledge into explicit one. Moreover, we also build up a shared ontology with a unified logical view. In addition, this study also transforms human-understandable metadata into a machine readable format in RDF and OWL.

4.2 Ontological Mapping

Traditionally, mapping or crosswalks are an imperative task for metadata integration in digital libraries. In practice, crosswalk is a chart or table to represent the semantic mapping of fields of data elements in one element set to fields or data elements in another element set [1]. Once a crosswalk between two metadata formats has completed, an integrated access to heterogeneous metadata of various sources can attain. However, not all metadata formats and elements have been included into existing official crosswalks maintained by authority institutions.

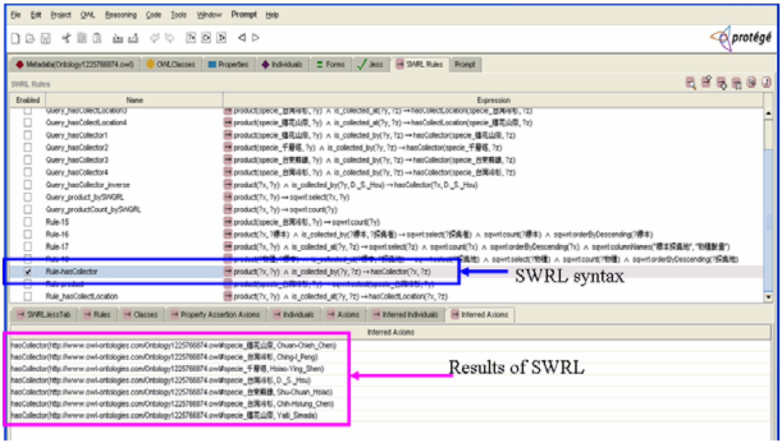
A shared ontology of this study is a set of common terminologies and relations with a unified logical view. It can be used to harmonize metadata from heterogeneous data sources in biodiversity heritage. At this stage, we adopted ontology alignment to generate ontology mapping between the shared ontology and elements of CoF and SDNPT respectively. Protégé and iPromptTab are employed to perform the semi-automatic ontology mapping, because iPromptTab can perform ontology alignment for classes according to both lexical strings and their path-based class hierarchical relations [9]. However, manual revision is still required to complete the final mapping rules. In fact, it reveals that almost elements of CoF and SDNPT can be mapped to the shared ontology, owing to the shared ontology stems from CoF and SDNPT.

4.3 Inference Query

Basically our shared ontology is a RDF data model of triples（subject, predicate, and object）with unambiguous associative relations and assertions, and stored in an XML-based RDF/OWL format. It can be extended as a knowledge representation basis to allow computer to meaningfully process the knowledge and provide semantic conclusions from our shared ontology and retrieve corresponsive metadata for answering imposed queries. Thus it can be further utilized to develop a set of semantic units of description logic（hereafter DL） such as IF-THEN rules, to draw inference query from various digital biodiversity repositories. For instance, two RDF triple statements, such as “Species has － product － Specimen” and “Specimen － is_collected_by－Collector”, can formulate an IF-THEN rule like “IF a specific plant Species has Specimen and Specimen is collected by Collector, THEN it means that Collector had collected this Species.” The SWRL syntax can be expressed as follows: product?（?x, ?y） ^ is_collected_by（?y, ?z）→hasCollector（?x, ?z）. Therefore, one can identify and combine two semantic RDF triples and statements as a basic IF-THEN rule for inference query. At this stage, we employed Protégé and SWRLTab software to manifest the deployment of SWRL language for OWL-DL based IF-THEN rules（see Fig. 1).

4.4 Mashup Metadata from Digital Repositories

How to retrieve corresponsive biodiversity metadata from CoF and SDNPT is still a problem in this study. Basically, the CoF and SDNPT are two typical relational databases, neither RDF nor networked retrieval protocol based. The proposed approach is to develop a query agent as a connection from ontological query results to mashup corresponsive metadata of CoF or SDNPT. The component of query syntax of CoF



Enabled	Name	Expression
□	Query_hasCollectLocation3	product(specie_台灣冷杉, ?y) ∧ is_collected_at(?y, ?z) → hasCollectLocation(specie_台灣冷杉, ?z)
□	Query_hasCollectLocation4	product(specie_穗花山奈, ?y) ∧ is_collected_at(?y, ?z) → hasCollectLocation(specie_穗花山奈, ?z)
□	Query_hasCollector1	product(specie_穗花山奈, ?y) ∧ is_collected_by(?y, ?z) → hasCollector(specie_穗花山奈, ?z)
□	Query_hasCollector2	product(specie_千層塔, ?y) ∧ is_collected_by(?y, ?z) → hasCollector(specie_千層塔, ?z)
□	Query_hasCollector3	product(specie_台東蘇鐵, ?y) ∧ is_collected_by(?y, ?z) → hasCollector(specie_台東蘇鐵, ?z)
□	Query_hasCollector4	product(specie_台灣冷杉, ?y) ∧ is_collected_by(?y, ?z) → hasCollector(specie_台灣冷杉, ?z)
□	Query_hasCollector_inverse	product(?x, ?y) ∧ is_collected_by(?y, D._S._Hsu) → hasCollector(?x, D._S._Hsu)
□	Query_product_bySWQRL	product(?x, ?y) → sqwrl:select(?x, ?y)
□	Query_productCount_bySWQRL	product(?x, ?y) → sqwrl:count(?y)
□	Rule-15	product(specie_台灣冷杉, ?y) → sqwrl:count(?y)
□	Rule-16	product(?x, ?標本) ∧ is_collected_by(?標本, ?採集者) → sqwrl:select(?採集者) ∧ sqwrl:count(?標本) ∧ sqwrl:orderByDescending(?標本)
□	Rule-17	product(?x, ?y) ∧ is_collected_at(?y, ?z) → sqwrl:select(?z) ∧ sqwrl:count(?x) ∧ sqwrl:orderByDescending(?x) ∧ sqwrl:columnNames("標本採集地", "物種數量")
□	Rule-18	product(?物種, ?標本) ∧ is_collected_at(?標本, ?採集地) → sqwrl:select(?採集地) ∧ sqwrl:select(?物種) ∧ sqwrl:count(?物種) ∧ sqwrl:orderByDescending(?採集地)
☑	Rule-hasCollector	product(?x, ?y) ∧ is_collected_by(?y, ?z) → hasCollector(?x, ?z)
□	Rule-product	product(specie_台灣冷杉, ?y) → sqwrl:select(specie_台灣冷杉, ?y)
□	Rule_hasCollectLocation	product(?x, ?y) ∧ is_collected_at(?y, ?z) → hasCollectLocation(?x, ?z)

SWRL syntax

Inferred Axioms

```
hasCollector(http://www.owl-ontologies.com/Ontology1225766874.owl#specie_穗花山奈, Chuan-Chieh_Chen)
hasCollector(http://www.owl-ontologies.com/Ontology1225766874.owl#specie_台灣冷杉, Ching-I_Peng)
hasCollector(http://www.owl-ontologies.com/Ontology1225766874.owl#specie_千層塔, Hsiao-Ying_Shen)
hasCollector(http://www.owl-ontologies.com/Ontology1225766874.owl#specie_台灣冷杉, D._S._Hsu)
hasCollector(http://www.owl-ontologies.com/Ontology1225766874.owl#specie_台東蘇鐵, Shu-Chuan_Hsiao)
hasCollector(http://www.owl-ontologies.com/Ontology1225766874.owl#specie_台灣冷杉, Chih-Hsiung_Chen)
hasCollector(http://www.owl-ontologies.com/Ontology1225766874.owl#specie_穗花山奈, Yati_Simada)
```

Results of SWRL

Fig. 1. An instance for SWRL query syntax and result

and SDNPT was analyzed into two parts: URL location as well as query field and string. The first part without underline is to connect the location of specified repository, and the second one with underline is to retrieve metadata records from repository based on either species' or specimen's name. Therefore several specific query syntaxes are generalized to mashup metadata as follows:

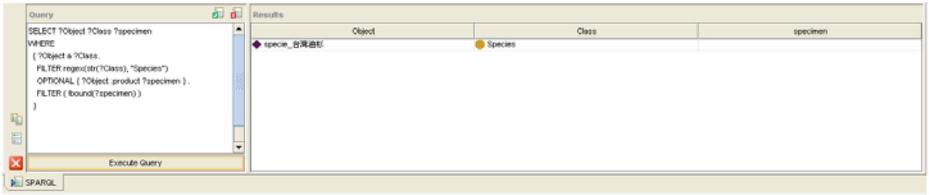
- SDNPT — http://db1n.sinica.edu.tw/textdb/hast/hast_label.php?_op=?species_m.speciesE:EngSpeciesName（query field is English species' name）
- CoF — http://taibif.org.tw/taibif_search/species_Detail.php?sc=Engspeciesname（query field is English species' name）

Second, we used JavaBean as program language to extend the function of Protégé for retrieving metadata from various digital repositories. Based on mashup connection, users can retrieve metadata from Protégé to CoF or SDNPT for reviewing detail of species or specimen records directly, no matter users select the specific term by browsing or querying ontologies. On the other hand, users can also query terms in a SPARQL syntax to access the corresponsive metadata from CoF or SDNPT respectively. Moreover, users would further use either SWRL or SQWRL language to perform OWL-DL based inference query to retrieve species' or specimen's metadata.

5 Question Answering

The design of ontological capable applications can facilitate the integrated metadata access and semantic query answering to a subject of interest from various biodiversity repositories. By means of QA-based query examples, we illustrate how biodiversity researchers not only query at various levels of ontological granularity, but also make semantically constrained queries.

Query 1: Which species has not specimen? There are two approaches may be the answer to this query. At the beginning, it can add zero into datatype attribute of specimen in our shared ontology. However, this approach is not a correct way for answering this query. Thus this study selects the second approach to find answer for this query. Essentially ontologies are assumed to be an open world. This query could convert to prove an assumption to be true or false. Therefore in this study we use SPARQL query to find the answer. The answer is Keteleeria davidiana（台灣油杉）. The syntax and query result of SPARQL are shown Fig. 2.



```
SELECT ?Object ?Class ?specimen
WHERE
 { ?Object a ?Class.
  FILTER regex(str(?Class), "Species")
  OPTIONAL { ?Object :product ?specimen } .
  FILTER ( !bound(?specimen) )
 }
```

Object	Class	specimen
◆ specie_台灣油杉	● Species	

Fig. 2. SPARQL result for query 1 from Protégé

Query 2: Which species are two different species but with synonymous name? This graph pattern for this query is illustrated in Fig. 3. It clarifies the use of OWL differentFrom property as a query restriction to assert that two species（Aralia decaisneana Hance and Aralia bipinnata Blanco) are different but with the same Chinese common name（鵲不踏).

Query 3: Which species is collected by Ching-I Peng at MIAOLI_HSIEN? This graph pattern for this query is illustrated in Fig. 4. It represents a more sophisticated query that spans over several RDF triples. This RDF graph query is a set of tuples in a sequential order, especially containing values for collector and collecting place of specimen respectively to satisfy three conditions: (i)a species has specimen, (ii) a specimen has been collected at specific place and （iii）a specimen has also been collected by specific person. The answer for this query is Abies Kawakamii（台灣冷杉） species.

Query 4: Which species has specimen, reference literature, scientific name, English common name and Chinese common name simultaneously? This graph pattern for this query is illustrated in Fig. 5. It represents another more sophisticated query that spans a greater portion of our shared ontology. The answer to this query is a set of tuples containing a species, specimen, reference and a complicated relation for name which includes scientific name, English common name and Chinese common name. The answer for this query is Hedychium coronarium Koenig（穗花山奈） with specimens（no. 101527 and 93998）, Flora of Taiwan（vol. 5, p. 717） reference literature, English common name（e.g. white ginger）, and Chinese common name（野薑花）. For this query this study uses SWRL rather than SPARQL, because SPARQL query can not de-duplicate the results.

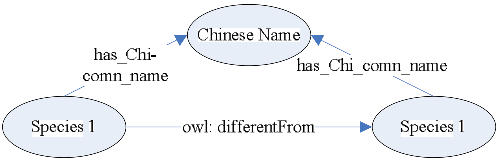


Fig. 3. RDF graph for query 2

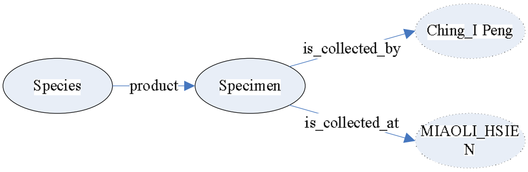


Fig. 4. RDF graph for query 3

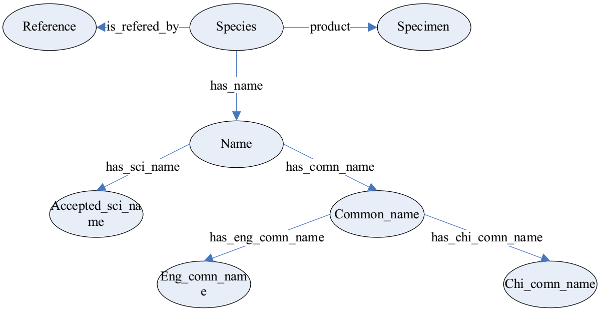


Fig. 5. RDF graph for query 4

6 Discussion

6.1 Transformation from Human-Understandable Metadata into Machine-Understandable Ontology

In fact a common ontology at element level is not available in biodiversity. In this study the proposed approach is to construct a shared ontology by transforming metadata into ontologies, rather than a mapping from metadata to ontologies. Thus this study is first to transform heterogeneous metadata into ontologies. It means that we have to extract and re-contextualize the original ontological concepts and relations embedded in metadata elements. This study has made implicit knowledge in metadata into explicit one. Therefore a human-understandable metadata expression has changed into a machine-readable RDF/OWL format. Thus the proposed approach in this study is a more practical solution than the above [7][14].

Second, it is not always feasible in heterogeneous situations for diverse user communities or domains to agree on using either the same authoritative identifiers [10] or names [13], or a specific metadata format [2][4][11]. Therefore this study provides a new and more flexibly customized approach to build up the shared ontology from metadata elements, and enrich ontological relations between elements as a domain of discourse for any communities and domains. Furthermore, the proposed approach in this study also transforms machine-readable metadata into a machine-understandable ontology in a SWRL language that can be furthering processed and inferred by semantic web software.

6.2 Manifestation of Semantic Web Technologies on Heterogeneous Metadata Integration

In this study we build up a shared ontology as a harmony mechanism to integrate metadata from heterogeneous digital biodiversity repositories at conceptual level. With addition of ontologies to metadata integration, our contribution can be drawn as follows. First, it is distinctive from simply physical or virtual data aggregation based on the same metadata element set without semantic relations [2][4][11]. The shared ontology proposed by this study is a manifestation of knowledge representation to represent associative relations of class and property. Many relations of our shared ontology are not expressed straightforward in digital repositories. It can include relations into query indexing and inference query, in addition to metadata elements. Second, the shared ontology can further support semantic query formulation across various levels of granularity, to discover any relation between two or more objects for answering in-depth questions as same as our demonstrated queries. This provides a data mining approach to discover relations between two or more resources in biodiversity. Thus, the use of associative relations among objects is an advantage of the use of ontology mapping over typical metadata mapping. Therefore semantic web driven approach is a conceptual crosswalk for heterogeneous metadata integration, rather than an element mapping without semantic relations and logic axioms.

Finally, each RDF triple can be regarded as a unit of DL to formulate into a series of IF-THEN inference rules. As shown as our demonstrated queries, the proposed approach is successful in generating IF-THEN rules compliant with SPARQL or SWRL/SQWRL languages to achieve inference query. On the other hand, biodiversity domain also needs negation, e.g. species has not specimen as same as illustrated in our query 1. The ability of using negation as failure plays an important role in QA, especially for knowledge discovery. Biodiversity researchers would like to query with the assumption that all the knowledge is available at certain point to discover new research issues or provide insight into research trends. However, “closed world” inference such as QA for negation is usable in biodiversity. Therefore, apart from representation of ontological hierarchy and relations, it can formulate inference from the shared ontology based on semantic expression of knowledge representation.

7 Conclusions

This study is successful in implementing the technology of semantic web on metadata integration for distributed digital repositories in biodiversity. First, this approach proposed is distinctive from most current studies in building up a shared ontology from bottom up, instead of adopting existing ontologies or metadata formats and elements.

In this study we also manifest how to transform metadata from implicit knowledge into explicit one. It means that this study changes metadata from human-understandable biodiversity formats and elements into a machine-understandable specification of explicit knowledge in an ontological expression. Next, in this study we employ the shared ontology to develop a set of conceptual mapping rules with logical relations and axioms. Based on conceptual mapping rules, one can integrate heterogeneous metadata from digital repositories, instead of a pure schema or element mapping table. Third, this proposed semantic web driven approach also uses the shared ontology as a knowledge representation mechanism. Thus it is allowed to include logical relations into query indexing and perform OWL-DL inference query, in order to find any possible relations among objects for in-depth QA in biodiversity heritage.

Although the approach proposed for metadata integration is semantic web driven, our target subjects are still belonging to typical relational databases without any RDFization. Furthermore, the proposed agent in this study is a tentative solution to mashup metadata from relational databases. Therefore, the RDFized normalization is needed to transform these proprietary relational databases as qualified SPARQL Endpoints for providing RDF compliant query in a distributed online environment. Moreover, the biodiversity heritage still needs to develop a common agreed ontology at data schema or element level for knowledge sharing and discovery in the long term, because this study reflects partial requirements of institutions and their digital collections.

References

1. Baca, M.: Practical Issues in Applying Metadata Schemas and Controlled Vocabularies to Cultural Heritage Information. Cataloging & Classification 36(3/4), 47–55 (2003)
2. Barros, E.G., Laender, A.H.F., Gonçalves, M.A., Barbosa, F.A.R.: A Digital Library Environment for Integrating, Disseminating and Exploring Ecological Data. Ecological Informatics 3(4-5), 295–308 (2008)
3. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American 284(5), 34–43 (2001)
4. Best, B.D., Halpin, P.N., Fujioka, E., Read, A.J., Qian, S.S., Hazen, L.J., Schick, R.R.: Geospatial Web Services within Scientific Workflow: Predicting Marine Mammal Habitats in a Dynamic Environment. Ecological Informatics 2(3), 210–223 (2007)
5. Hakimpour, F., Geppert, A.: Resolving Semantic Heterogeneity in Schema Integration: an Ontology Based Approach. In: FOIS 2001 proceedings of the international conference on formal ontology in information systems, pp. 297–308. ACM Press, New York (2001)
6. ISO/IEC JTC 1/SC32 WG2: Information Technology: Semantic Metadata Mapping Procedure: ISO/IEC WD 20943-5 (2008), http://metadata-standards.org/metadata-stds/Document-library/Documents-by-number/WG2-N1201-N1250/WG2-N1217-WD_SMMP_20081119.pdf
7. Kakali, C., Lourdi, I., Stasinopoulou, T., Bountouri, L., Papatheodorou, C., Doerr, M., Gergatsoulis, M.: Integrating Dublin Core Metadata for Cultural Heritage Collections Using Ontologies". In: Sutton, S., Chaudhry, A., Khoo, C. (eds.) Proceedings of the 2007 International Conference on Dublin Core and Metadata Applications, DCMI, Singapore, pp. 128–140 (2007), http://dcpapers.dublincore.org/ojs/pubs/article/view/877/873

8. Noy, F.N., McGuinness, D.L.: Ontology Development 101: a Guide to Creating Your First Ontology (2001), http://protege.stanford.edu/publications/ontology_development/ontology101.pdf
9. Noy, F.N., Musen, M.A.: Anchor-PROMPT: Using Non-local Context for Semantic Matching. In: Workshop on Ontologies and Iinformation Sharing at IJCAI (2001), http://www.dit.unitn.it/~accord/RelatedWork/Matching/noy.pdf
10. Page, R.: A Taxonomic Search Engine: Federating Taxonomic Databases Using Web Service. BMC Bioinformatics 6(48), 1–8 (2005)
11. Peterson, A.T., Vieglais, D.A., Sigüenza, A.G.N., Silva, M.: A Global Distributed Biodiversity Information Network: Building the World Museum. The Bulletin of The British Ornithologicsts' Club 123A, 186–196 (2003)
12. Ralaivao, J.-C., Darmont, J.: Knowledge and Metadata Integration for Warehousing Complex Data (2007), http://hal.archives-ouvertes.fr/docs/00/32/06/61/PDF/ista07-ralaivao-darmont.pdf
13. Sarkar, I.N.: Biodiversity Informatics: Organizing and Linking Information across the Spectrum of Life. Briefings in Bioinformatics 8(5), 347–357 (2007)
14. Stasinopoulou, T., Bountouri, L., Kakali, C., Lourdi, I., Papatheodorou, C., Doerr, M., Gergatsoulis, M.: Ontology-based Metadata Integration in the Cultural Heritage Domain. In: Goh, D.H.-L., Cao, T.H., Sølvberg, I.T., Rasmussen, E. (eds.) ICADL 2007. LNCS, vol. 4822, pp. 165–175. Springer, Heidelberg (2007)
15. Uschold, M., Gruninger, M.: Ontologies: Principles, Methods and Applications. Knowledge Engineering Review 11(2), 93–136 (1996)
16. Xu, Q., He, F., Qiu, R.G.: Heterogeneous Information Integration for Supply Chain. In: Proceedings of 2005 IEEE International Conference on Systems, Man and Cybernetics. pp. 97–105. IEEE Press, New Jersey (2005)

Fuzzy Multi-objective Programming Application for Time-Cost Trade-off of CPM in Project Management

Gino K. Yang

Department of Computer Science and Information Management,
Hungkuang University, Taiwan, ROC

Abstract. In pragmatic project management cases, many complex resources and large-scale related activities are involved. Moreover, when situations are under the conflict and incommensurate of time and cost, it becomes more difficult for managers to make their decisions. Hence, we construct fuzzy multi-objective programming model from CPM technique. Through emphasizing the selectable flexibility among the feasible projects, we describe the decision problem brought from uncertainty and complex in project. In our research, taking the conflict and incommensurate of time and cost as two major factors, we measure weight and priority to denote the importance degree between the objectives. Adopting Tiwari et al., we also develop the result when trading off the decision vs. the vague environment of time and cost and the past' single objective' method has become the exceptional case of multi-objective models. Secondly, we broaden the consumed time and cost of activity events when assumptions are certain to merge the practical situation. We applied fuzzy number to express estimate time and cost. It based on Lee and Li method to solve this problem. Finally, we employ numerical example to explain it with LINGO package to calculate.

Keywords: Project management, CPM, Fuzzy number, Fuzzy multi-objective programming.

1 Introduction

Managers normally must supervise a variety of activities. Some of the activities routine and repetitive, while others are not. Projects belong to the latter. Most of projects have some common factors (time, cost, performance etc.) that must be managed if the project goals are to be achieved. The management of a large-scale project requires coordinating numerous activities throughout the organization. A myriad of details must be considered. Fortunately, two closely related operation research techniques, PERT and CPM are available to assist the project manager with carrying out these responsibilities. These techniques make heavy use of networks to help plan, coordinate and display of all the necessary activities. These systems allow the project progress to be monitored and controlled by managers.

Earlier, PERT focused on the uncertainty in activities using the Three-Estimate Approach to estimate and control activity time. However, CPM presumed that the amount of consumed time and activity event coats were accurate. In this work, we focus on the relationship between consumed time and expended cost even though these two techniques

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 215–229, 2010.

have gradually merged in recent years. The purpose is to determine the project critical path to ensure that the project runs smoothly in spite of a sting of unnecessary delays.

However, achieving accurate activity time and costs in CPM is difficult. For example, activity cost cannot be estimated accurately with the complicated related resources and techniques. The pressure to achieve project completion within a certain period of time makes managers give biased time consumption estimates. Similarly, the possibilities for a modest amount of overlap in activities need not invalidate a schedule using PERT/CPM. We therefore recommend fuzzy set theory to deal with the decision-making problems resulting from uncertainty and complexity.

In recent research on fuzzy sets applied to project management, most scholars used fuzzy numbers or membership functions in the uncertain project activity environment. Some scholars employed fuzzy number model ranking for defuzzification such as Chang et al [1] McCahon [2] and Yao et al [3]. Other researchers used the α-cut technique to calculate such as Hasution [4], Rommelfanger [5], and Mon [6]. Some adopted Zadeh's [7] extension principle to produce the critical path such as Chanas et al [8]. Others applied Zimmermann's [9] fuzzy linear programming technique to trade-off time and cost such as Ignizio et al [10] and Elden [11. Others discussed how to get critical path in a fuzzy environment. In a related application, Kucha [12] utilized fuzzy numbers to evaluate the project risk. He used the decision makers' attitude and project framework to help the decision-makers analyze progress.

The traditional CPM technique involves constructing a single objective programming model from the minimum crash cost under a time constraint. The optimum solution gained from single objective programming leaves the decision maker to either accept the results unconditionally or reject them. We will construct a CPM model as a fuzzy multi-objective programming model to emphasize flexibility and explore the possible feasible substitute plans. We also describe the relative importance of weight and priority among the objectives. Through the above methods, we are able to understand the complexity of expected fuzzy situation and evaluate it accurately. We can then evaluate the minimum crash cost under the time constraint and produce the earliest finish time under the cost constraint. In the past, the single objective was a special case of multi-objective models. Under the CPM assumption (the expended time and cost are both certain), we broaden this assumption and apply fuzzy numbers to express the estimated time and cost. Via such a pragmatic project management description in which the parameters normally cannot be calculated accurately, decision makers have more flexibility to obtain a satisfactory result.

2 Time-Cost Trade-offs in CPM

The estimated project activity time is normally made under some given level of resources. In many situations, it is possible to reduce the duration of a project by injecting additional resources. This is expected to shorten the project time below a normal value under this activity for crashing, to avoid penalties resulting from the time delay, to take advantage of monetary incentives for timely or early project competition, or to free resources for employment on other projects. It is possible for project managers to deal with the restrictions of additional resources and funds to acquire the earliest finish time.

CPM cannot achieve this under its minimum crash cost and single objective model. CPM creates some bias during decision-making procedures. Hence, we establish a multi-objective programming model in CPM to make the time-cost trade-off.

It was assumed that time and cost could be predicted effectively without significant uncertainty. In searching for the appropriate crash level and injecting funds to shorten the project time to the expended time and cost, we apply a time-cost curve to determine the crash level and cost. According to Hiller and Lieberman [13] , we define two key points (normal and crash) for the curves, as shown in Figure 1.

Normal point: the normal point on the time-cost graph for an activity shows the time and cost for the activity when it is performed in the normal way.

Crash point: the crash point shows the time and cost when the activity is fully crashed, i.e. it is fully expedited with no cost spared to reduce its duration as much as possible.

In most applications, it is assumed that a partial crash of the activity at any level would produce a combination of time and cost that lies somewhere on the line segment between these two points. We will apply linear programming to make the crash decision and define the variable.

D_{ij} = normal time for activity $(i\ ,j)$

C_{Dij} = normal cost for activity (i, j)

d_{ij} = crash time for activity $(\ i\ ,j\ )$

C_{dij} = crash cost for activity $(\ i\ ,j\ )$

To obtain activity i to j to shorten activity duration time per unit, we denote the slope of the line through loaded crash cost S_{ij} by

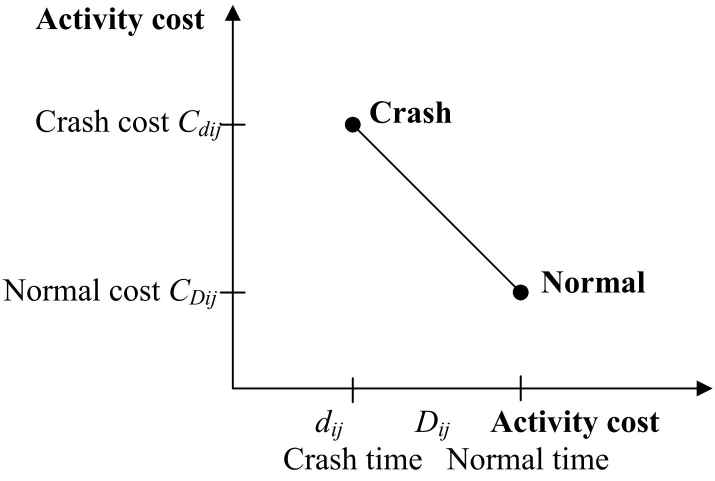


Fig. 1. Time-cost curve for activity $(\ i\ ,j\ )$

$$S_{ij} = \frac{C_{Dij} - C_{dij}}{D_{ij} - d_{ij}} \tag{1}$$

Using the AOA network diagram to show the activity event, the decision variables for the problem are shown below

x_{ij} = duration crash time for activity *i* to *j*

y_l = start time of activity *l*

CPM multi-objective linear programming clear model for cost and time analysis can be constructed as Eq. (2)

$$\begin{aligned} & Min\ Z_1 = \sum_{i,j} S_{ij} x_{ij} \\ & Min\ Z_2 = y_n - y_1 \\ s.t \quad & x_{ij} \geq 0 \\ & x_{ij} \leq D_{ij} - d_{ij} \\ & y_j + x_{ij} - y_i \geq D_{ij} \\ & y_1 = 0 \end{aligned} \tag{2}$$

y_1 is defined as the start time for the project and y_n is the end time.

3 Fuzzy Multi-objective CPM Method

The fuzzy multi-objective CPM method is a kind of mathematics programming model used under information uncertainty situations to consider multiple objectives. Its purpose is to assist the decision makers in seeking an optimum solution under the restriction of fuzzy information and objective conflict. In fuzzy set theory, the degree of the membership function for each objective represents its satisfaction level. Zimmermann [9] was the first to apply the fuzzy set theory concept with multiple objective linear programming problems. Later, many solutions for the relative problem have been developed, i.e., Lai and Hwang. [14]

We used Tiwari et al [15], Lee and Li's [16] approach to solve the fuzzy multiple objective CPM problem in this paper. Among recent research, some have applied those two approaches to solve the decision problem caused using fuzzy linear programming . In 1993, Bit et al [17] presented weights and priorities for nonequivalent objectives in transportation problems. He employed the Tiwari additive fuzzy programming model to obtain the best compromise solution. In 1998, Roy and Maiti [18] utilized multiple objective inventory models in which the coefficient was a fuzzy member to obtain the EOQ (economic order quantity) under maximum profit and minimum waste objectives together with the constraint of total average cost and warehouse space . They applied Lee and Li's fuzzy linear programming to solve these problems. In the above two applied cases, they used a fuzzy programming model to describe the decision problem generated from the uncertain environment. The results were closer to the real situation.

The fuzzy multi-objective CPM model is concerned with the problem of minimizing the time and cost. Let ($x^{(1)*}$, $x^{(2)*}$) be the ideal solution for the multiple objective CPM problem (2), i.e., $x^{(k)*}$is the optimal solution for the single objective programming problem.

$$\min_{x \in X} \; Z_k(x) \qquad (k = 1, 2)$$

The values of all of the k objective can then be calculated at all of these k optimal solutions $x^{(k)*}$(k = 1,2) to form a payoff matrix

$$\begin{array}{cc} & j \\ \begin{array}{c} Z_1 \\ Z_2 \end{array} & \begin{bmatrix} Z_1(x^{(1)*}) & Z_1(x^{(2)*}) \\ Z_2(x^{(1)*}) & Z_2(x^{(2)*}) \end{bmatrix} \end{array} \tag{3}$$

The diagonal of matrix (3) constitutes the k ideal values $Z_K^{\text{Ĝ}}\ \text{Ť}\ Z_k(x^{(k)*})$ of objectives (k = 1,2). From the payoff matrix (3), we can determine that U_{k} and L_{k} are the upper and lower bounds for the k objective, where L_{k} = aspiration level of achievement for the k objective, and U_{k} = highest acceptable level of achievement for the k objective. Hence,

$$L_{\mathrm{k}} = Z_K^{+} = Z_k(x^{(j)*})$$

$$U_{\mathrm{k}} = \max_j \{Z_k(x^{(j)*})\} \qquad (k = 1,2)$$

If CPM objective function is fuzzy in nature, the membership function is shown as Eqs. (4) and (5). This is decision maker's degree of satisfaction for each objective. The larger the membership value becomes, the more satisfaction derived.

$$\mu_1(Z_1(x)) = \begin{cases} 1 & \mathit{if}\ \ Z_1(x) \le L_1 \\ \dfrac{U_1 - Z_1(x)}{U_1 - L_1} & \mathit{if}\ \ L_1 < Z_1(x) < U_1 \\ 0 & \mathit{if}\ \ Z_1(x) \ge U_1 \end{cases} \tag{4}$$

$$\mu_2(Z_2(x)) = \begin{cases} 1 & \mathit{if}\ \ Z_2(x) \le L_2 \\ \dfrac{U_2 - Z_2(x)}{U_2 - L_2} & \mathit{if}\ \ L_2 < Z_2(x) < U_2 \\ 0 & \mathit{if}\ \ Z_2(x) \ge U_2 \end{cases} \tag{5}$$

3.1 Weighted Fuzzy Goal Additive Programming Model

In virtually every situation managers must cope with this problem when certain goals become more important than others. For example, some events could be more important than others, such as achieving a certain objective or solving a specific problem. The

Tiwari et al approach was used to assign differential weights as coefficients for the individual terms in the fuzzy membership function to reflect their relative importance, i.e., Multiplying each fuzzy goal membership value with a suitable weight and then adding them together formulated the objective function. This leads to following formulation, corresponding to (6)

$$\begin{aligned} & Max \quad w_1 \times \lambda_1 + w_2 \times \lambda_2 \\ s.t \quad & \mu_1(Z_1) = \lambda_1 \\ & \mu_2(Z_2) = \lambda_2 \\ & x_{ij} \geq 0 \\ & x_{ij} \leq D_{ij} - d_{ij} \\ & y_j + x_{ij} - y_i \geq D_{ij} \\ & y_1 = 0 \end{aligned} \tag{6}$$

Here w_1, w_2 are weight for objective function Z_1 and Z_2

3.2 Preemptive Priority Fuzzy Goal Additive Programming Model

Priority performs an important role in human decision-making processes. Recognizing this fact of life enables managers to direct their efforts to the best goals to obtain. Unless a particular objective or a subset of objectives is achieved, the other objectives should not be considered. In such cases, the weighting scheme for the previous section is not an appropriate method. The preemptive priority structure [14,17] may be stated as $P_k >>> P_{k+1}$, which means that the objectives in the k-th priority level have higher priority than the objectives in the (k +l)-th priority level.

The problem is subdivided into a subproblem. Every subproblem has a number of priority levels. In the first subproblem, only the fuzzy objective belonging to the first priority level of the membership function for the degree of satisfaction degree is considered. At other priority levels, the membership values are maximum and solved using the additive fuzzy programming model. In general the CPM becomes (7)

$$\begin{aligned} & Max \quad \lambda_{k+1} \\ s.t \quad & \mu_k(Z_k) = \lambda_k \\ & \mu_{k+1}(Z_{k+1}) = \lambda_{k+1} \\ & x_{ij} \geq 0 \\ & x_{ij} \leq D_{ij} - d_{ij} \\ & y_j + x_{ij} - y_i \geq D_{ij} \\ & y_1 = 0 \end{aligned} \tag{7}$$

Where λ_k is the achieved membership value for the DMs degree of satisfaction in the first priority level and λ_{k+1} is secondary level.

According to Tiwari et al's [15] viewpoint, the main difference between FGP (Fuzzy Goal Programming) and GP (Goal Programming) is that the GP requires the decision-maker to set definite aspiration values for each objective. Then FGP is specified in an imprecise aspiration level. Using the FGP calculation rule, the similarity between the non-fuzzy weighting method and goal programming problems can be solved. It has more flexibility and allows the decision-maker to employ different methods to describe the relative importance among the objectives.

3.3 Lee and Li Fuzzy Multi-objective Linear Programming Model

Each parameter coefficient is fuzzy, using Lee and Li's method to solve the fuzzy multiobjective-programming problem. To utilize Eq. (1) efficiently to estimate the crash cost per unit S_{ij}, the CPM coefficient is expressed as a triangle fuzzy number. When the membership value is highest, S_{ij} is calculated using the α-cut skill (6). Where $\alpha \in [0,1]$ denotes the level of possibility that all fuzzy coefficients are feasible, and $\lambda \in [0,1]$ denotes the grade of compromise to which the solution satisfies all of the fuzzy goals while the coefficients are at a feasible level α. In the synthetic possibility and satisfaction degree, when β is the overall solution satisfaction under the fuzzy objective and constraint, using Bell-Zaheh's[19] rule, β is taken as β= min(λ,α). Hence, it can be expressed as (8)

$$\begin{aligned}
&Max && \beta \\
s.t\quad & && \beta \le \lambda \\
& && \beta \le \alpha \\
& && \lambda \le [\mu_1(Z_1)]_\alpha^U \\
& && \lambda \le [\mu_2(Z_2)]_\alpha^U \\
& && x_{ij} \ge 0 \\
& && x_{ij} \le (\tilde{D}_{ij} - \tilde{d}_{ij})_\alpha^L \\
& && y_j + x_{ij} - y_i \ge (\tilde{D}_{ij})_\alpha^L \\
& && y_1 = 0 \\
& && \alpha,\ \beta,\lambda \in [0,1]
\end{aligned} \tag{8}$$

Let $(\tilde{P})_\alpha$ be the α-cut value of the fuzzy number $\tilde{P}$. $(\tilde{P})_\alpha^L$ and $(\tilde{P})_\alpha^U$ are the lower and upper bounds of the α-cut for fuzzy number $\tilde{P}$. Generally, Lee and Li's fuzzy possibility linear multi-objective programming has two extreme situations. To get the optimum solution suitable for the current situation from an infinite number of linear programming parameter combinations. [20] In this paper we make flexible adjustments for the parameters, indicating the actual meaning and the decision-maker's degree of optimism. For example, $\tilde{D}_{ij} - \tilde{d}$ has the maximum crash time, which is the upper bound. The lower bound is used to show the activity duration crash time for variable x_{ij} .

4 Numerical Example

Some project programming is shown in Figure 2. The AOA network project diagram time and cost data are shown in Table 1. When all activities are performed under normal situations, the entire project finish time is 22 days and the direct cost is $1070. The critical path is A-C-F according to the multi-objective programming model shown using Eq.(9).

$$\begin{aligned}
&Min\ Z_1 = 10x_a+10x_b+30x_c+5x_e+20x_f+20x_g \\
&Min\ Z_2 = y_6 - y_1 \\
&s.t \\
&y_2 + x_a \geq 9\ ,\quad y_3 +x_b \geq 7\ ,\quad y_4 + y_3 + x_e \geq 8\ , \\
&y_5 - y_2 + x_c \geq 6,\quad y_6- y_5 +x_f \geq 7\ ,\quad y_6 - y_4+ x_g \geq 5 \qquad (9) \\
&x_a \leq 6,\ x_b \leq 3\ ,\ x_c \leq 1,\ x_e \leq 4\ ,\ x_f \leq 3\ ,\ x_g \leq 1 \\
&x_a \geq 0\ ,\ x_b \geq 0\ ,\ x_c \geq 0\ ,\ x_e \geq 0\ ,\ x_f \geq 0\ ,\ x_g \geq 0 \\
&y_1=0
\end{aligned}$$

Table 1. Project performance time and cost data

Activity	Normal duration time	Maximum Reduction time	Standard cost	Incremental crash cost
A	9	6	140	10
B	7	3	120	10
C	6	1	200	30
D	5	0	180	0
E	8	4	140	5
F	7	3	80	20
G	5	1	210	20

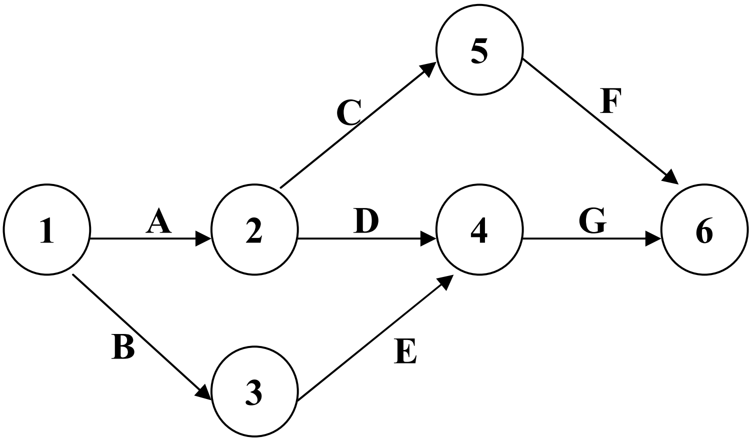


Fig. 2. AOA network diagram of project

4.1 Weighted Fuzzy Goal Additive Programming Model

Tiwari's weight additive model was used to evaluate the relative degree of importance between the objective weights. The ideal solution payoff matrix is (10). The lower bound ideal solution for objective is L_k= (0, 12) and the upper bound ideal solution is U_k = (220, 22).

$$\begin{array}{c} Z_1 \\ Z_2 \end{array}\begin{bmatrix} 0 & 220 \\ 22 & 12 \end{bmatrix} \tag{10}$$

The membership objective function can then be constructed as (11) and (12). We use (6) to calculate W_1 and W_2 of objectives shown in Table 2 and Figure 3.

$$\mu_1(Z_1(x)) = \begin{cases} 1 & if\ \ Z_1(x) \le L_1 \\ \dfrac{220 - Z_1(x)}{220 - 0} & if\ \ L_1 < Z_1(x) < U_1 \\ 0 & if\ \ Z_1(x) \ge U_1 \end{cases} \tag{11}$$

$$\mu_2(Z_2(y)) = \begin{cases} 1 & if\ \ Z_2(y) \le L_2 \\ \dfrac{22 - Z_2(y)}{22 - 12} & if\ \ L_2 < Z_2(y) < U_2 \\ 0 & if\ \ Z_2(y) \ge U_2 \end{cases} \tag{12}$$

Table 2. The finish time for the objective relative weight

W_1	W_2	Z_1	Z_2	Crash activity	Reduce time	Critical path
1.0	0	0	22	—	—	A-C-F
0.9	0.1	0	22	—	—	A-C-F
0.8	0.2	0	22	—	—	A-C-F
0.7	0.3	0	22	—	—	A-C-F
0.6	0.4	20	20	A	2	A-C-F B-E-G
0.5	0.5	80	16	A , E	6, 4	A-C-F B-E-G
0.4	0.6	170	13	A,B,E,F	6,3,4,3	A-C-F B-E-G A-D-E
0.3	0.7	220	12	A,B,C,E,F,G	6,3,1,4,3,1	A-C-F B-E-G A-D-E
0.2	0.8	220	12	A,B,C,E,F,G	6,3,1,4,3,1	A-C-F B-E-G A-D-E
0.1	0.9	220	12	A,B,C,E,F,G	6,3,1,4,3,1	A-C-F B-E-G A-D-E
0	1	220	12	A,B,C,E,F,G	6,3,1,4,3,1	A-C-F B-E-G A-D-E

We use the weight to denote the relative importance of the objective function. It differs from traditional decision-making under the time constraint and offers a distinct decision-thinking method. Many scholars have devoted research to these kinds of weight determinations.

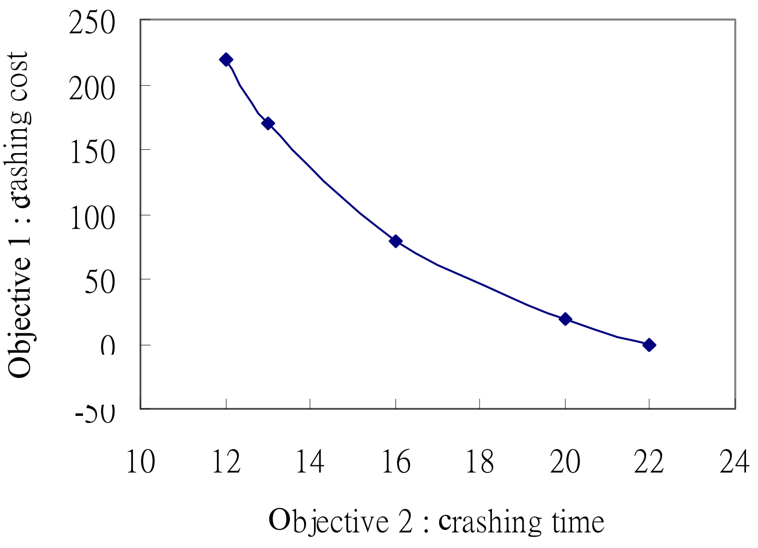


Fig. 3. Time-cost trade-off point for each objective weight

4.2 Priority Fuzzy Goal Additive Programming Model

To complete the project within the required period of time it is possible to reduce the length of the project by injecting additional resources. The time objective is a priority for completion. It is not feasible to employ a weight additive model. The project completion time is an approximate situation. The finish time cannot be estimated accurately. The decision-maker's degree of satisfaction with the projected completion time is used to indicate the membership value of the time priority objective. Equation (7) is used to obtain the objective value of the minimum crash cost. The calculation result is shown in Table 3 and Figure 4.

Table 3. Finish time under time priority objective

First priority Time membership value level	First priority Time level	Subordinate Cost level	Crash activity	Reduce time	Critical path
0	22	0	—	–	A-C-F
0.1	21	10	A	1	A-C-F
0.2	20	20	A	2	A-C-F
0.3	19	35	A, E	3, 1	A-C-F, B-E-G
0.4	18	50	A, E	4, 2	A-C-F, B-E-G
0.5	17	65	A ,E	5, 3	A-C-F, B-E-G
0.6	16	80	A, E	6, 4	A-C-F, B-E-G
0.7	15	110	A,B,E,F	6,1,4,1	A-C-F, B-E-G
0.8	14	140	A,B,E,F	6,2,4,2	A-C-F, B-E-G
0.9	13	170	A,B,E,F	6,3,4,3	A-C-F, B-E-G, A-D-E
1	12	220	A,B,C,E,F,G	6,3,1,4,3,1	A-C-F, B-E-G, A-D-E

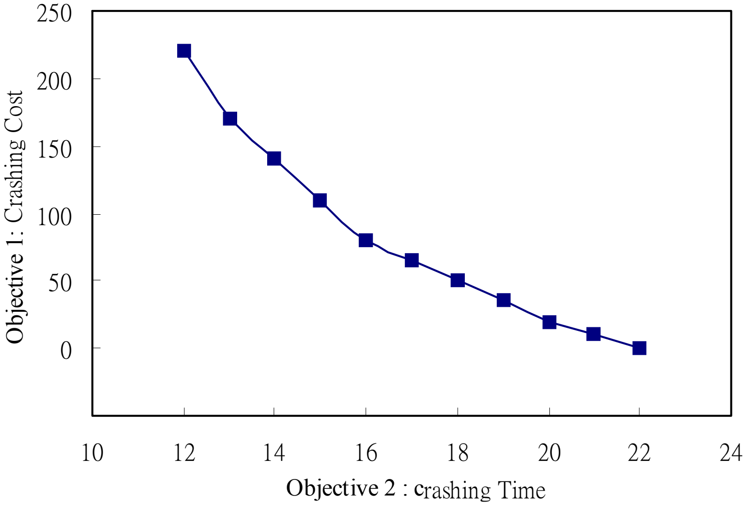


Fig. 4. Trade-off under priority time objective

Table 4. Finish time under cost priority objective

First priority Cost membership value level (λ_k)	First priority Cost level (λ_k)	Secondary Time level (λ_{k+1})	Crash activity	Reduce time	Critical path
0	220	12	A,B,C,E,F,G	6,3,1,4,3,1	A-C-F, B-E-G, A-D-E
0.1	198	12.4	A,B,C,E,F,G	6,3, 0.6, 4,3, 0.6	A-C-F, B-E-G, A-D-E
0.2	176	12.88	A,B,C,E,F,G	6,3,0.12,4,3,0.12	A-C-F, B-E-G, A-D-E
0.3	154	13.53	A,B,E,F	6 ,2.47, 4,2.47	A-C-F, B-E-G
0.4	132	14.27	A,B,E,F	6 , 1.73 , 4, 1.73	A-C-F, B-E-G
0.5	110	15	A,B,E,F	6,1,4,1	A-C-F, B-E-G
0.6	88	15.73	A,B,E,F	6 ,0.27 , 4, 0.27	A-C-F, B-E-G
0.7	66	16.93	A,E	5.07,3.07	A-C-F, B-E-G
0.8	44	18.4	A,E	3.60,1.60	A-C-F, B-E-G
0.9	22	19.87	A,E	2.13,0.13	A-C-F, B-E-G
1	0	22	—	—	A-C-F

The decision-makers must face the limits of additional resource and budget and then consider the cost priority to gain the earliest finish time under the cost restrictions. According to that degree of satisfaction the membership value of the cost priority is indicated. Equation (7) is then used to obtain the objective value of the earliest finish time. The calculation result is shown in Table 4 and Figure 5.

We applied the fuzzy multi-objective programming approach to construct a CPM model. It evaluates the minimum crash cost and predicts the earliest finish time under the cost constraint. We can consider the single objective as a special case of the multi-objective method to avoid creating bias in the decision.

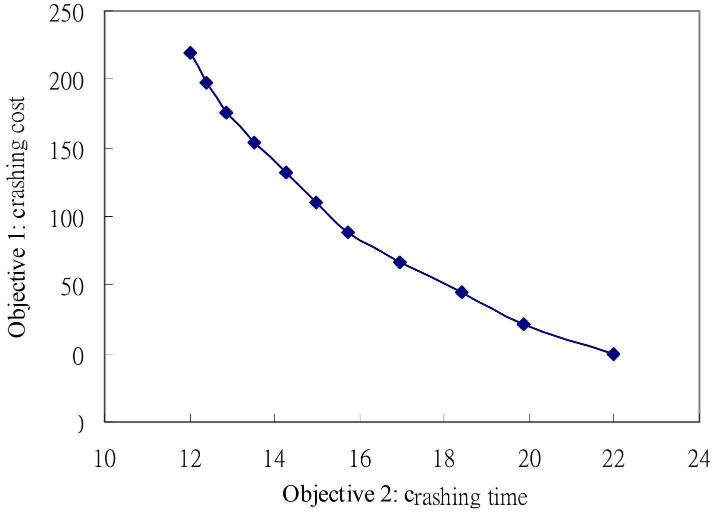


Fig. 5. Trade-off under cost priority objective

4.3 Lee and Li Fuzzy Multi-objective Linear Programming Model

The activity event time and cost is relaxed when the assumptions are ensured to merge with the practical situation. Fuzzy numbers are applied to denote the evaluated time and cost. In this paper, the time and cost for the objective were assumed to have the same weight to acquire the solution. Because the unit crash cost calculated from the other information in the objective function and normal crash time based on the constraint are the decision-maker's approximate time, both coefficients would probably create more bias. Henceforth, the slacker triangle fuzzy number is denoted by $\tilde{x} = (x-2,\ x,\ \mathrm{x}+2)$. $\tilde{x} = (x-1,\ x,\ \mathrm{x}+1)$ is therefore used to denote the triangle fuzzy number. Equation(8) is shown as Eq.(13). The result is shown in Table 5.

$$\begin{aligned}
&Min && Z_1 = (8+2\alpha)x_a + (8+2\alpha)x_b + (28+2\alpha)x_c + (3+2\alpha)x_e + (18+2\alpha)x_f \\
& && \quad + (18+2\alpha)x_g \\
&Min && Z_2 = y_6 - y_1 \\
&s.t && y_2 + x_a \ge (7+2\alpha) \qquad y_3 + x_b \ge (5+2\alpha) \\
& && y_4 - y_3 + x_e \ge (6+2\alpha) \qquad y_5 - y_2 + x_c \ge (4+2\alpha) \\
& && y_6 - y_5 + x_f \ge (5+2\alpha) \qquad y_6 - y_4 + x_g \ge (3+2\alpha) \\
& && x_a \le 5+\alpha \qquad x_b \le 2+\alpha \qquad x_c \le \alpha \\
& && x_e \le 3+\alpha \qquad x_f \le 2+\alpha \qquad x_g \le \alpha \\
& && \alpha \in [0,1] \quad x_a \ge 0 \quad x_b \ge 0 \quad x_c \ge 0 \quad x_e \ge 0 \quad x_f \ge 0 \quad x_g \ge 0
\end{aligned} \tag{13}$$

Table 5. The result of Lee and Li fuzzy linear programming

α	$Z_1^{\ L}$	$Z_1^{\ U}$	$Z_2^{\ L}$	$Z_2^{\ U}$	β	$Z_1(x)$	$Z_2(y)$
0	0	101.00	9	16	0.6011	40.2865	11.7921
0.1	0	118.82	9.3	16.6	0.6078	45.1780	12.0756
0.2	0	122.88	9.6	17.2	0.6101	47.9126	12.5633
0.3	0	134.18	9.9	17.8	0.6132	51.9005	12.9557
0.4	0	145.72	10.2	18.4	0.6156	56.0084	13.3517
0.5	0	157.50	10.5	19.0	0.6175	60.2379	13.7509
0.6	0	169.52	10.8	19.6	0.6190	64.5903	14.1530
0.6192	0	171.8553	10.8576	19.7152	0.6192	65.4401	14.2305
0.7	0	181.78	11.1	20.2	0.6201	69.0665	14.5575
0.8	0	194.28	11.4	20.8	0.6217	73.8891	14.9750
0.9	0	207.02	11.7	21.4	0.6235	79.1900	15.4105
1	0	220	12	22	0.6254	84.6154	15.8462

Table 5 shows the optimum solution and α* = β* = 0.6192，$Z_1(x) = 65.4401$, $Z_2(y) = 14.2305$. The objective values for the decision-maker are $Z_1^{\ L} = 0$, $Z_1^{U} = 171.8553$ and $Z_2^{L} = 10.8576$, $Z_2^{\ U} = 19.7152$. Under synthetic possibility and satisfaction degree, the reduced time for each activity is x_a= 5.485 and x_e =3.485. The total compromise crash time using the same cost and time weight is 14.2305. The expected cost increase is 65.4401.

In the above example, the parameter has been considered to be more optimistic situation, making the outcome consist of less activity time and cost. This can create some adjustment in practical projects. For example, we assumed certain activities in this paper to be overlapped, giving slack in the activity duration and higher result values than the normal activity time. The maximum crash time has an upper bound concept. We denote the α-cut of the fuzzy number with an extreme lower bound. Having a small amount of slack in the activity time compensates for the unexpected delays that inevitably seems to slip into a schedule. In Lee and Li's application the objective weights were unequal.

5 Conclusion

CPM considers the activity time and the activity cost. When simplifying a complex project it provides considerable information value. In this paper we used fuzzy set theory as a bridge to show the decision problems generated in a project from uncertainty and complexity. We constructed a multi-objective programming model to trade off the time and cost. Through this model construction and example, the following conclusions were obtained.

(1) Fuzzy additive programming was used to denote the relative importance of the objective functions. This differs from the traditional decision-making under a time constraint and offers a distinct decision-thinking method.

(2) Not only can the priority fuzzy goal additive programming model evaluate the minimum crash cost under the time constraint, it can also determine the earliest finish time under the cost constraint. Thus, we are able to consider the traditional single objective method as a special case of the multiple objectives to avoid creating bias in the decision process.

(3) We relaxed the time consumed and activity event costs when the assumptions merged with a pragmatic situation. Fuzzy numbers were applied to denote the evaluated time and cost. According to practical situations, adjustments are created to calculate compromise solutions for the time and cost.

It is natural to use the multi-objective programming model to denote the CPM trade off between time and cost. Fuzzy set theory was employed to show actual project situations. This application is suitable, thorough and flexible. It is suitable for human thinking modes and practical situations, and also able to deal with the complex and uncertain decision making problems in project management.

References

1. Chang, P.T., Lee, E.S.: Fuzzy decision networks and deconvolution. Computers and Mathematics with Applications 37, 53–63 (1999)
2. McCahon, C.S.: Using PERT as an Approximation of fuzzy project-network analysis. IEEE Transactions on Engineering Management 40, 146–153 (1993)

3. Yao, J.S., Lin, F.T.: Fuzzy critical path method based on signed distance ranking of fuzzy numbers. IEEE Transactions on System, Man and Cybernetics-Part A: Systems and Humans 30, 76–82 (2000)
4. Nasution, S.H.: Fuzzy critical path method. IEEE Transactions on System, Man and Cybernetics 24, 48–57 (1994)
5. Rommelfanger, H.J.: Network analysis and information flow in fuzzy environment. Fuzzy Sets and Systems 67, 119–128 (1994)
6. Mon, D.L., Cheng, C.H., Lu, H.C.: Application of fuzzy distributions a project management. Fuzzy Sets and Systems 73, 227–234 (1995)
7. Zadeh, L.A.: Fuzzy sets. Information and control 8, 338–353 (1965)
8. Chanas, S., Zieliński, P.: Critical path analysis in the network with fuzzy activity times. Fuzzy Sets and Systems 122, 195–204 (2001)
9. Zimmerman, H.: Fuzzy programming a linear programming with several objective functions. Fuzzy Sets and Systems 1, 45–55 (1978)
10. Ignizo, J.P., Daniels, S.C.: Fuzzy multicriteria integer programming via fuzzy generalized networks. Fuzzy Sets and Systems 10, 261–270 (1983)
11. Deporter, E.L., Ellis, K.P.: Optimization of project networks with goal programming and fuzzy linear programming. Computers and Industrial Engineering 19, 500–504 (1990)
12. Kuchta, D.: Use of number in project risk (criticality) assessment. International of Project Management 19, 305–310 (2001)
13. Hiller, F.S., Lieberman G.J.: Introduction to operations research. 6th.7th Ed, McGraw-Hill.Inc., NY (1997,2001)
14. Lai, Y.J., Hwang, C.L.: Fuzzy multiple objective decision making-method and applications. In: Algebraic Methods 1987, vol. 394, Springer, Berlin (1992)
15. Tiwari, R.N., Dharmar, S., Rao, J.R.: Fuzzy goal programming: an additive model. Fuzzy sets and systems 24, 27–34 (1987)
16. Lee, E.S., Li, R.J.: Fuzzy multiple objective programming and compromise programming with Pareto optimum. Fuzzy Sets and Systems 53, 275–288 (1993)
17. Bit, A.K., Biswal, M.P., Alam, S.S.: An additive fuzzy programming model for multi-objective transportation problem. Fuzzy Sets and Systems 57, 313–319 (1993)
18. Roy, T.K., Maiti, M.: Multi-objective inventory models of deteriorating items with some in a fuzzy environment. Computers and Operations Research 25(12), 1085–1095 (1998)
19. Bellman, R.E., Zadeh, L.A.: Decision making in a fuzzy environment. Management Science 17(4)B, 141–164 (1970)
20. Lee, E.S.: Fuzzy multi-criteria decision making theory and application. Science publisher, Peking (2002) (chinese)

A Study on Two-Stage Budget Planning by Fuzzy Quantifying Analysis

Gino K. Yang[1,*] and Wayne Chouhuang[2]

[1] Department of Computer Science and Information Management,
Hungkuang University, Taiwan, R.O.C.
[2] Department of Industrial Management, National Taiwan University of Science and Technology, Taiwan, R.O.C.

Abstract. This paper proposes fuzzy quantifying analysis method to solve the two-stage budget planning problems. First, it uses AHP (Analytical Hierarchy Process) to get the factors' weight and then applied Fuzzy Synthetic Decision to evaluate and analysis them for quantifying the budget planning problems. Finally, illustrate a practical example to explain the process of fuzzy quantifying on budget planning.

Keywords: Analytical Hierarchy Process, Fuzzy Synthetic Decision, Budget planning, Quantifying analysis.

1 Introduction

In our life have some vague, imprecise knowledge, and the property of nonquantity, including human's thinking and linguistic transmitting. Vague environments are met frequently. It is hard to be satisfied using classical mathematics to deal with this kind of things.

Fuzzy mathematics is what uses mathematical method to study and handle something with "fuzziness". In scientific age, some study branches urge to be quantified, such as biology, psychology, linguistics and sociology etc. Hence, people meet large-scale of fuzzy concept. Along with the highly scientific progress, the objects we study become more complex and most of complex things exist numeral exactness problems. It is considered to be contradictory to ordinary people. This contradictory phenomenon becomes fiercer along with the development of computers. It is considered to be contradictive to ordinary people. This contradictive phenomenon become fiercer along with the computers highly develop. ◦ A strict procedure needs high accuracy. The other hand, the computers or machines must perform more complex task that will involve large vague concept. That is the more complex it is, the less accurate it will be. That make people unable to focus on the whole, they should omit the less important parts and just catch the most important parts. At this time its clear concept becomes vague again.

[*] Corresponding author.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 230–242, 2010.

Traditionally, people considered that financial and budget planning are “accurate” numeric and figure affair. It doesn’t agree with “approximate” or “similar” these kinds of estimate figure. Even though financial and budget affairs involve huge and complex figure and data. And complex affair is hard to get exact figures. It will exit estimate, predictive and uncertain figure in the decision and calculation process. But finance and budget involve huge and complex figure and data. Complex things are hard to be quantified. It will exit uncertain figure of estimate and forecast. Hence, it brings a huge and complex fuzzy problem in finance figure is bigger increasingly. The traditional mathematics model becomes huger and more abstract and the researchers takes much time, manpower, material and financial resource attempt to construct a complex mathematical model. That model will be not easy to execute, and lost its practical value. Meanwhile we can apply fuzzy mathematics to solve that kind of fuzzy phenomenon for analyzing and determining the complex, huge, abstract and vague figure and data by fuzzy mathematics. It qualities and analyzes the vague data originally, and makes it is useful for financial and budge decision.

When the decision makers face the budget planning and allocation problem, they encounter the budget is limited and subordinate demand unlimited (demand bigger then supply). Meanwhile it is the time to trial the decision makers how to make a fair and reasonable allocation. Generally, decision makers will think or face those kinds of problems subjectively. Few people will consider about the real demand or opinion. It will result understanding inconsistent and bring decision bias. Meanwhile it involves the difference of upper and lower levels’ opinion and understanding. Under this situation, it is necessary to think about the subordinate opinion. The decision maker should consider not only himself subjective location but also subordinate demand and thinking when he makes decision. Such that he can make fair decision and avoid partial decision. Hence, how to give consideration to both of upper and lower level’s opinion and standpoint is the issue the decision makers must face it actively.

This study offer a approach based on AHP (Analytical Hierarchy Process) and (Fuzzy Synthetic Decision) to construct a set and step of process for decision reference and assisting decision maker to face the opinion imbalance of higher and lower level. And hope it is useful for decision-making. Finally, we offer a practical budget-planning example for application and reference.

2 The Basic Concept of Fuzzy Set

Fuzzy set is used to denote that all set of certain things that bound or edge is indefinite. It like normal set that fuzzy set is a subset of some universe of discourse. So, fuzzy set is named fuzzy subset, denote as $\tilde{A}$, $\tilde{B}$, $\tilde{C}$ …etc.

People often talk about the concept of “tall, short, fat, thin, old, young, middle age”. It has no clear bound and its expression is different with people condition. It is that the meanings is uncertain, unbounded and vague concept. Such as the concept of “middle age”, someone consider it is between 30 and 50 years old, others think that 30 to 50 years old is “young” and over 50 is old man. Hence, we can fine that the meanings and bound are unclear. [8, 12]

2.1 Characteristic Function and Membership Function

Characteristic function means the relationship of factor or object and set or entirety. It takes only one of either “belong to” or “not belong to”. It maps characteristic function value is 1 or0. This property can use characteristic function x_A to denote:

$$x_A : X \to \{0,1\}$$

$$x \to x_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \overline{\in} A \end{cases}$$

If only characteristic function, it can’t deal with the entire vague phenomenon in all fields. In fact, characteristic function value shows the membership degree of element to set, which is membership function can take place characteristic function. In fuzzy set, the membership degree is not limited on 0 and 1, it can be any value between 0 and 1. So, fuzzy sets can be obtained from normal sets.

[Definition 1]. Let U is universe of discourse, u is a function that its value is between close interval [0,1] take from U, that is $\mu_{\tilde{A}} : U \to [0,1]$

Fuzzy subset $\tilde{A}$ on U denotes $\mu_{\tilde{A}}(x)$, it means the degree of element u belong to fuzzy set $\tilde{A}$, $u \to \mu_{\tilde{A}}(u)$.

2.2 Fuzzy Sets Calculation

2.2.1 Definition of Calculation

Utilizing membership function to denote fuzzy set, we can apply membership function to quest some fuzzy set basic concept. [12]

[Definition 2]. Two fuzzy sets $\tilde{A}$ and $\tilde{B}$, to element x, both of their membership functions are equal, that we call $\tilde{A}$ and $\tilde{B}$ are equal。

That is $\tilde{A} = \tilde{B} \Leftrightarrow f_{\tilde{A}}(x) = f_{\tilde{B}}(x) \quad \forall_x \in U$

[Definition 3]. Two fuzzy sets $\tilde{A}$ and $\tilde{B}$ ’s union, intersection and complement, **we** utilize membership function to define:

$$(1)\ f_{\tilde{A}\cap\tilde{B}}(x) = \min\left\{f_{\tilde{A}}(x), f_{\tilde{B}}(x)\right\} = f_{\tilde{A}}(x) \wedge f_{\tilde{B}}(x)$$

$$(2)\ f_{\tilde{A}\cup\tilde{B}}(x) = \max\left\{f_{\tilde{A}}(x), f_{\tilde{B}}(x)\right\} = f_{\tilde{A}}(x) \vee f_{\tilde{B}}(x)$$

$$(3)\ f_{\tilde{A}}^{c}(x) = 1 - f_{\tilde{A}}(x)$$

2.2.2 Fuzzy Vector

Assume $X=\{x_1, x_2, \ldots, x_n\}$, fuzzy subset on X is

$$\tilde{A}=\frac{a_1}{x_1}+\frac{a_2}{x_2}+\cdots+\frac{a_n}{x_n}\ ,\ \text{where } a_i=\mu_{\tilde{A}}(x_i)$$

then $\tilde{A}$ is n-dimensional fuzzy vector is denote as

$\tilde{A}$ =(a_1 , a_2 ,…, a_n) , where n-component vector is a_1 , a_2 ,…, $a_n \in [0,1]$

The difference between fuzzy vector and normal vector are their component vector a_i is in closed interval [0,1] ; any n-dimensional fuzzy vector can be seen as a 1×*n* dimensional fuzzy matrix. [12]

2.2.3 Fuzzy Matrixes and Its Calculation

(1) Fuzzy matrix

[Definition 4]. If universe of discourse *X* has *m* factors, universe of discourse *Y* has *n* factors, then fuzzy relative matrix show below (1)

$\tilde{R}$ is defined by $\tilde{R}=[r_{ij}]_{m\times n}$

where $r_{ij}=f_{\tilde{R}}(x_i,y_j)$: $X\times Y\to[0,1]$

$$\text{that is } \tilde{R}=[r_{ij}]_{m\times n}=\begin{bmatrix} r_{11} & r_{12} & r_{13} & \cdots & r_{1n} \\ r_{21} & r_{22} & r_{23} & \cdots & r_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{m1} & r_{m2} & r_{m3} & \cdots & r_{mn} \end{bmatrix} \tag{1}$$

where $0\le r_{ij}\le 1\ \forall_i=1,2,\ldots,m\ \forall_j=1,2,\ldots,n$

Fuzzy relative matrix is called fuzzy matrix for short.

The differences between fuzzy matrix and normal matrix are their factor r_{ij} values are located in closed interval [0,1].

(2) Calculation of fuzzy matrix

Calculation of fuzzy matrix is the tool to show duality fuzzy vectors and same as normal matrix *m*×*n* fuzzy matrix can be considered *m* fuzzy vectors formed it [9]

(a) Equal : If $a_{ij}=b_{ij}$, then $\tilde{A}=\tilde{B}$ 。

(b) Comprise : If $a_{ij}\le b_{ij}$, then $\tilde{A}\subseteq\tilde{B}$ 。

(c) Intersection : Assume $c_{ij}=a_{ij}\vee b_{ij}$, then $\tilde{C}$ =(c_{ij}) is $\tilde{A}$, $\tilde{B}$'s intersection, denoted as $\tilde{C}=\tilde{A}\cup\tilde{B}$

(d) Union : Assume $c_{ij}=a_{ij}\wedge b_{ij}$, then $\tilde{C}$ =(c_{ij}) is $\tilde{A}$, $\tilde{B}$'s union, denoted as $\tilde{C}=\tilde{A}\cap\tilde{B}$

(e) Complement : $\tilde{A}^c$ =(1- a_{ij}) is a complement matrix of $\tilde{A}$ =(a_{ij})。

2.2.4 Combined Calculation of Fuzzy Matrix [12,13,14]

(1) $\tilde{B} = \tilde{A} \cdot \tilde{R}$ =(b_1 , b_2 ,…, b_n)

$$b_j = \bigvee_{i=1}^{n} (a_i \wedge r_{ij}) \quad (j=1,2,\ldots,m)$$

where " · " take " ∧ ", " ∨ " to operate, which is Zadeh operator. It is denoted as M(∧,∨) °

Zadeh operator is " ∧ " and " ∨ ", the former is min-operation, the latter is max-operation.

$a \vee b$=Max(a,b) $a \wedge b$=Min(a,b)

(2) $\tilde{B} = \tilde{A} \circ \tilde{R}$ = (b_1 , b_2 ,…, b_m)

$$b_j = \sum_{i-1}^{n} (a_i , r_{ij}) = (a_1 \cdot r_{1j}) \oplus (a_2 \cdot r_{2j}) \oplus \ldots \oplus (a_n \cdot r_{nj})$$

Where " ∘ " takes " · " and " ⊕ " operators; $a \cdot b = a \cdot b$ is product operator; $a \oplus b = (a+b) \wedge 1$ is called Bounded sum. $\sum_{i=1}^{n}$ means to get sum on ⊕ from some numbers. It is denote as $M(\cdot , \oplus)$ °

The calculation result of $M(\wedge, \vee)$ and $M(\cdot , \oplus)$ is different, it is because their operations are different. Obviously, when $\tilde{A}$'s factors are more average, the result is not true to the original, that we use $M(\wedge, \vee)$ to calculation. When we apply $M(\cdot , \oplus)$, it can supplement the shortage of $M(\wedge, \vee)$ calculation. So, in practical application we should notice the choice of operators.

The matrix product is

$$\tilde{B} = \tilde{A} \circ \tilde{R} = (b_1 , b_2 ,\ldots, b_m) \qquad b_j = \sum_{i-1}^{n} a_i r_{ij}$$

Where " ∘ " is $M(\cdot , +)$'s operator, that is normal matrix product.

3 Research Approach

3.1 Analytical Hierarchy Process (AHP)

The Analytic Hierarchy Process (AHP) is that integrates qualitative and quantitative multicriteria decision-making model. It has systematic, flexible and simple features, and can divide complex question into several factors and uses hierarchy structure to express the relationship among factors. It employs Pairwise comparison of preferences to ensure relative importance on the same hierarchy and synthesize the judgment of decision maker to decide the importance ranking of alternatives. This decision process shows the essential characteristics of human thinking mode that means resolving, judgment and integration. We use numeral type to express and solve the complex problems by human experience and judgment. AHP technique is popular and practical, and used on the multicriteria decision making problem in many research fields. Its main steps show below: [4,5,10]

(1) To assure the goal and decision factor sets U
(2) To construct a decision matrix
(3) Calculating the importance ranking

According to pairwise matrix, we can get largest eigenvalue mapping eigenvector, This characteristic vector is the importance ranking of all decision factors. That is called weight allocation.

(4) Test [11]

Above characteristic vector is the weight we want. If weight allocation is reasonable? It need a consistence test, and showed below:(2)

$$CR=CI／RI \tag{2}$$

where

CR is the random consistency ratio of judgment matrix
CI is the normal consistency indicator of judgment matrix, is given from

$$CI=\frac{1}{n-1}(\lambda_{\max}-n)$$

RI is the average random consistency indicator of judgment matrix, its value shows below:

n	1	2	3	4	5	6	7	8	9
RI	0.00	0.00	0.58	0.90	1.12	1.24	1.32	1.41	1.45

3.2 Fuzzy Synthetic Decision

A thing has multiple attributes and influenced by many factors, so in decision procedure, we must make synthetic consideration and decision to multiple factors or phenomenon. This is called synthetic decision. When this decision procedure involves vague factor, it must apply fuzzy mathematics to make synthetic decision. It is called fuzzy synthetic decision.

In behavioral science research, many issues map to each factor that it can't just use a simple number to evaluate it. It ought to survey to assure each factor's weight in advance, then can get appropriate synthetic decision. The basic method and steps of fuzzy synthetic decision show below:[1,2,3,6]

(1) To construct the factor set

Factor set is a normal set made by all factors that influence decision targets.

(2) To construct the decision set

Decision set is a set made up by all decision result that decision maker evaluates the decision target.

(3) Constructing the weight set

In the factor, each factor's importance is different. We give each mapping weight $a_i(i=1,2,\ldots,m)$ to each factor $u_i(i=1,2,\ldots,m)$ to map each factor's importance. $\tilde{A}=(a_1,a_2,\cdots,a_m)$ is a set made by each weight. It is called factor weight set or weight set.

(4) Single factor fuzzy decision

From single factor to make decision, to ensure the membership degree of decision target to decision factors. It is called single factor fuzzy decision.

(5) Fuzzy synthetic decision

Single factor decision only reacts the influence of each factor to each decision target. It is not enough obviously. Our purpose is to consider all of the influential factors and get a scientific decision result. It is called fuzzy synthetic decision.

$$\tilde{B}=\tilde{A}\cdot\tilde{R}$$

Weight set $\tilde{A}$ is considered a m row fuzzy matrix. Then $\tilde{B}$ is called fuzzy synthetic decision set.

(6) Manage the decision indicators

After getting the decision indicators $b_j(j=1,2,\ldots,n)$, we according to below methods to ensure the result of decision target.

(a) Maximum membership degree method max membership degree method

Take maximum decision indicator $\max_j b_j$, its mapping decision factor v_L is the decision result.

(b) The weighted averages method

Use weight b_j to make weighted average at all decision factors, the value is the decision result.

(c) Fuzzy probability distribution method

Take decision indicator to be its decision result directly or normalize decision indicator, then use normalized indicator as the decision result.

All of the decision indicators are mapping the distribution condition of decision targets' characteristic. It makes the decision maker understand decision targets more deeply and can manage it more flexibly.

4 Calculation Steps

This calculation steps utilize concept and calculation process of AHP and fuzzy synthetic decision. We make a conclusion showed below for the reference in making budget planning decision.

【**Step 1**】Determine the decision factor of upper to lower level, and utilize expert group decision-making to evaluate the factor, then make fuzzy decision to lower level, and get a fuzzy matrix $\tilde{R}_1$ (3)

$$\tilde{R}_1 = \left[\begin{array}{ccccc|c} X_1 & \ldots & X_j & \ldots & X_n & \\ \mu_{11} & \Lambda & \mu_{1j} & \Lambda & \mu_{1n} & U_1 \\ \mathrm{M} & \mathrm{O} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} \\ \mu_{i1} & \Lambda & \mu_{ij} & \Lambda & \mu_{in} & U_i \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} & \mathrm{M} \\ \mu_{m1} & \Lambda & \mu_{mj} & \Lambda & \mu_{mn} & U_m \end{array}\right] \qquad (3)$$

X_j : is jth lower level unit ; j=1,2,…,n。

U_i : is ith decision factor of lower level unit ; i=1,2,…,m 。

μ_{ij} : is ith lower level unit, jth decision factor quantifying value, $\mu_{ij} \in [0,1]$。

【**Step 2**】 To evaluate lower level offered factor $\tilde{R}_2$ (4)

$$\tilde{R}_2 = \left[\begin{array}{ccccc|c} X_1 & \ldots & X_j & \ldots & X_n & \\ \mu_{11} & \Lambda & \mu_{1j} & \Lambda & \mu_{1n} & U_1 \\ \mathrm{M} & \mathrm{O} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} \\ \mu_{i1} & \Lambda & \mu_{ij} & \Lambda & \mu_{in} & U_i \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} & \mathrm{M} \\ \mu_{m1} & \Lambda & \mu_{mj} & \Lambda & \mu_{mn} & U_m \end{array}\right] \qquad (4)$$

X_j : is jth lower level ; j=1,2,…,n 。

U_i : is ith decision factor that lower level offer ; i=1,2,…,m 。

μ_{ij} : is ith lower level offered, jth decision factor quantifying value.

【**Step 3**】 Ask experts for ranking the upper and lower levels' factors, to construct a Pairwise comparison of preferences matrix and utilize AHP sum-product method to obtain upper and lower factors' weights $\tilde{W}$ (5)

$$\tilde{A} = \left[\begin{array}{cccc} u_1 & u_2 & \ldots & u_{\mathrm{n}} \\ u_{11} & u_{12} & \Lambda & u_{1n} \\ u_{21} & u_{22} & \Lambda & u_{2n} \\ \mathrm{M} & \mathrm{M} & & \mathrm{M} \\ u_{n1} & u_{n2} & \mathrm{M} & u_{nn} \end{array}\right] \qquad (5)$$

$$\tilde{W} = [w_1, w_2, \Lambda\ , w_n]$$

u_i : is evaluating factor, (i=1,2,…,n)

u_{ij} : is u_i to u_j relative importance value, (j=1,2,…,n) 。

【**Step 4**】 Take the upper and lower level combination factor fuzzy set $\tilde{R}$ and factor weight fuzzy set $\tilde{W}$ to make fuzzy synthetic decision. [6]

$$\tilde{R} = \begin{bmatrix} \tilde{R}_1 \\ \tilde{R}_2 \end{bmatrix} \qquad \tilde{B} = \tilde{W} \cdot \tilde{R} \tag{6}$$

【**Step 5**】To allocate after normalizing the decision result, and test the lower level 's offered demand if satisfied. If exceed, satisfy it and then reject it. Relocate the left budget and back to step 4 make evaluation again till get the most suitable allocation.

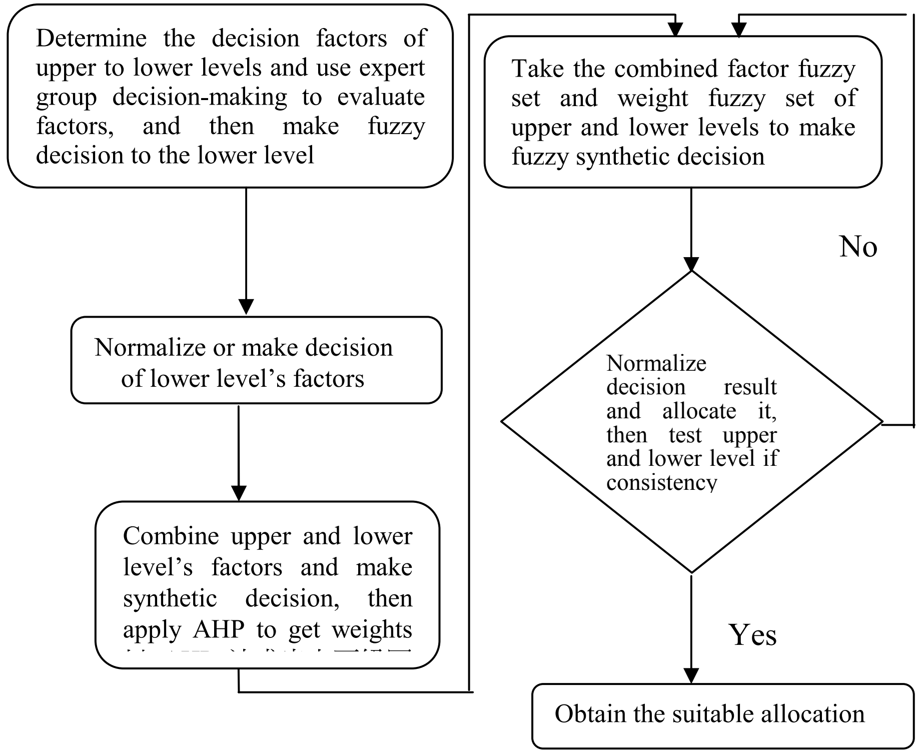

Determine the decision factors of upper to lower levels and use expert group decision-making to evaluate factors, and then make fuzzy decision to the lower level
Normalize or make decision of lower level's factors
Combine upper and lower level's factors and make synthetic decision, then apply AHP to get weights
Take the combined factor fuzzy set and weight fuzzy set of upper and lower levels to make fuzzy synthetic decision
No
Normalize decision result and allocate it, then test upper and lower level if consistency
Yes
Obtain the suitable allocation

Fig. 1. A flow chart of calculation steps

5 Budget Planning by Fuzzy Quantifying Analysis

On study the budget problems; we generally emphasize budget scale estimate, competition and repellence, and budget striving and reasonable, fair allocation. It is rare to consider the upper and lower level's standpoint or different attitude of mind. Hence, it brought some problems in upper and lower level communication. Most of the previous scholars stressed on applying Statistics methods, regression analysis or system dynamics…etc. to research the budget problems. In this paper, we try to utilize Analytical Hierarchy Process (AHP) and Fuzzy Synthetic Decision to construct a set of two-stage budget planning, allocation procedure. It can offer the decision makers of reference and application, and make reasonable and fair decision. Let the limited resource and finance bring maximum efficacy.

5.1 Problem Description

The government finance is getting tough recently, ever it have under a unbalance situation. It is not easy to satisfy the demand of all branches and subordinates. They are in urgent demand of budget, how to satisfy each branch and unit is a issue to test the authorities.

Assumed has a military organization, it has ten subordinate units. When make budget planning, it must satisfy their basic demand first. Then their present other demand show as Table 1. Total available funds are only 50 billion show as table 2 and column 2 and 3 are the importance and urgency for expert group fuzzy decision result. According to this, the decision maker how to make a decision, then let the allocation is fair and reasonable? Here the urgency means warfare situation, facilities situation, mission importance…etc. The importance means budget performance, organization size, strategy…etc.

Table 1. The budget demand of unit A to J

	A	B	C	D	E	F	G	H	I	J
Subordinate unit budget demand (billion)	9	8	7	9	5	4	8	6	3	6

Table 2. Importance and urgency get from expert group decision

	A	B	C	D	E	F	G	H	I	J
Importance	0.9	0.7	0.6	0.8	0.5	0.3	0.4	0.8	0.7	0.2
Urgency	0.5	0.4	0.9	0.3	0.4	0.6	0.9	0.2	0.5	0.5

5.2 Apply Fuzzy Quantifying Analysis to Make Budget Planning

In this question, according to the calculation steps offered by this paper. The calculation, analysis, and result show below:

【Step 1】Determine the decision factors that upper level to lower level, use expert group decision model to evaluate factor and make fuzzy decision to lower level, then get $\tilde{R}_1$

$$\tilde{R}_1 = \begin{bmatrix} 0.9 & 0.7 & 0.6 & 0.8 & 0.5 & 0.3 & 0.4 & 0.8 & 0.7 & 0.2 \\ 0.5 & 0.4 & 0.9 & 0.3 & 0.4 & 0.6 & 0.9 & 0.2 & 0.5 & 0.5 \end{bmatrix}$$

【Step 2】Evaluate subordinate offered factors $\tilde{R}_2$

$$\tilde{R}_2 = [9,8,7,9,5,4,8,6,3,6]$$

Normalized result show below:

[0.138, 0.123, 0.108, 0.138, 0.077, 0.062, 0.123, 0.092, 0.046, 0.092]

【Step 3】Take upper and lower level's factors to ask expert group for ranking, construct a Pairwise comparison of preferences matrix $\tilde{A}$ and use AHP sum-product method to get weights $\tilde{W}$ of upper and lower levels

$$\begin{array}{cccc} U_1 & U_2 & U_3 & \\ \end{array}$$

$$\tilde{A} = \begin{bmatrix} 1 & 1/3 & 3 \\ 3 & 1 & 5 \\ 1/3 & 1/5 & 1 \end{bmatrix} \begin{array}{c} U_1 \\ U_2 \\ U_3 \end{array}$$

U$_1$: is importance ; U$_2$: is urgency ; U$_3$: is the budget present from lower level

In the paper, we apply "sum-product method" gets the value show below:

$$\tilde{W} = [0.276, 0.617, 0.107]$$

【Step 4】Take the combined factor fuzzy set and factor weight set of upper and lower levels to make fuzzy synthetic decision

$$\tilde{B} = \tilde{W} \cdot \tilde{R}$$

$$\tilde{B} = [0.276 \quad 0.617 \quad 0.107] \cdot \begin{bmatrix} 0.9 & 0.7 & 0.6 & 0.8 & 0.5 & 0.3 & 0.4 & 0.8 & 0.7 & 0.2 \\ 0.5 & 0.4 & 0.9 & 0.3 & 0.4 & 0.6 & 0.9 & 0.2 & 0.5 & 0.5 \\ 0.138 & 0.123 & 0.108 & 0.138 & 0.077 & 0.062 & 0.123 & 0.092 & 0.046 & 0.092 \end{bmatrix}$$

The fuzzy synthetic decision result show below:

$$\tilde{B} = [0.572, 0.453, 0.723, 0.421, 0.393, 0.46, 0.679, 0.354, 0.507, 0.374]$$

【Step 5】Normalize the decision result, then make allocation. Test lower level's offered demand if enough, if it is then satisfy them first and reject it. Reallocate left budget, then back to step 4 for making decision till get the suitable solution.

Normalizing result is:

[0.116, 0.092, 0.148, 0.085, 0.079, 0.093, 0.137, 0.072, 0.102, 0.076]

Allocated result is:

[5.782, 4.583, 7.407, 4.255, 3.975, 4.648, 6.866, 3.581, 5.124, 3.778]

where C, F, I over their required demand, satisfy their demand first and reject it, then reallocate the left budget.

The final allocation result:

[6.342, 5.032, 4.679, 4.354, 7.518, 3.93, 4.145]

Finally we get the most suitable allocation solution shows as table 5-3。

Table 3. Result of most suitable budget allocation

	A	B	C	D	E	F	G	H	I	J
Most suitable budget allocation (billion)	6.342	5.032	7	4.679	4.354	4	7.518	3.93	3	4.145

5.3 Result Discussion and Analysis

(1) Decision factors are the most importance for influencing the research result. It involves the decision maker's subjective understanding. Hence, it is very important to analysis and judges the decision factors.

(2) The most important reason for the units satisfied first is that their importance and urgency are higher.

(3) For the supply is limited, some unit's demand can't be satisfied completely. In practical case, these units must consider the importance and urgency of themselves carefully or consider the others influencing factors and differentiate them in order of importance and urgency. When the budget was cut off, it can self-adjusted. Postpone less important affair to next time. It is unreasonable for cut off the budget in fixed proportion. Under this situation, it should have the capability to deal with it.

6 Conclusions and Suggestion

We can make a conclusion from about calculation process and a practical case for test:

6.1 Research Contributions

(1) It is rare for people apply fuzzy theory to research the budget or financial issue before. This study attempts to solve some budget-planning question by fuzzy mathematics theory. This study approach can apply on military organizations or government for budget allocation, budget planning, or resource allocation.

(2) Consider upper and lower levels opinion and standpoint at the same time, to decide the budget allocation and avoid bias.

(3) Offer a resource allocation model and procedure, to solve the resource allocation conflict between government, corporation, and ordinary branches. Under the supply and demand unbalance, can be used to solve this problem, let budget and resource be able to be allocated fairly and give the best efficacy.

(4) Apply fuzzy quantifying calculation steps to offer a reliable and effective decision approach for dealing with budget planning question. It is a reliable and effective decision approach. To urgent budget decision problems it is a feasible decision approach and for decision reference.

6.2 Suggestions

(1). Utilize fuzzy Delphi method to fine the decision factors of upper and lower levels and make the budget planning problems get better result.
(2). Apply statistics method to request weights and test decision factors, if it is independent, then compare with the result of fuzzy mathematics method.
(3). Use fuzzy mathematics theory, concept to quest, research budget and financial issues continually, let the fuzzy mathematics concept extend to the application of budget and financial relative issue, and hope more scholars utilize fuzzy mathematics concept to study on defense budget issue and make a good breakthrough.

References

1. Xiaohong, C., Mai, L., Yasuhiko, T.: Fuzzy Decision Making for Selection of Information System Development approach. In: Chinese Society Aeronautics Astronautics, New York, NY, USA, vol. 8, pp. 172–176 (1998)
2. Chen, S.M.: Evaluating Weapon Systems using fuzzy arithmetic operations. In: Fuzzy Sets and Systems, 77th edn., pp. 265–276 (1996)
3. Cheng, C.H.: Evaluating Weapon Systems Using Ranking Fuzzy Numbers, vol. 107, pp. 25–35. Elsevier Science Publishers B.V, Amsterdam (1999)
4. Cheng, C.H.: Evaluating Naval Tactical Missile System by AHP Based on the grade of membership function. European Journal of Operational Research 96, 343–350 (1997)
5. Cheng, C.H., Mon, D.L.: Evaluating Weapon System by Analytic Hierarchy Process Based on fuzzy scales. In: Fuzzy Sets and Systems, 63th edn., pp. 1–10 (1994)
6. Cheng, J., Yang, Z.Y.: Fuzzy Synthetic Decision-Making System in Ferrographic Analysis, vol. 222, pp. 1–20. Elsevier Science S.A, Lausanne (1998)
7. Rabin, J., Lynch, T.D.: Handbook on Public Budgeting and Financial Management. Marcel Dekker,Inc., New York (1983)
8. Kaufmann, A., Gupta, M.M.: Introduction to Fuzzy Arithmetic Theory and Applications. Van Nostrand Reinhold, New York (1991)
9. Li, H.X., Yen, V.C.: Fuzzy sets and Fuzzy Decision-Making. CRC Press, Inc., New York (1995)
10. Mon, D.L., Cheng, C.H., Lin, J.C.: Evaluating Weapon System Using Fuzzy Analytic Hierarchy Process Based on Entropy Weight. In: Fuzzy Sets and Systems, vol. 62, pp. 127–134 (1994)
11. Saaty, T.L.: The Analytic Hierarchy Process. McGraw-Hill International Book Company, New York (1980)
12. Zadeh, L.A.: Fuzzy sets. Information and control 8, 338–353 (1965)
13. Zimmermann, H.J.: Fuzzy Set Theory and its Applications, 2nd edn. Kluwer Academic Publishers, Boston (1991)
14. Zimmermann, H.J.: Fuzzy Set, Decision Making, and Expert systems. Kluwer Academic Publishers, Boston (1993)

Hybrid Support Vector Regression and GA/TS for Radio-Wave Path-Loss Prediction

Kuo-Chen Hung[1], Kuo-Ping Lin[2], Gino K. Yang[3], and Y.-C. Tsai[4]

[1] Department of Logistics Management, National Defense University,
Beitou Taipei 112, Taiwan
kuochen.hung@msa.hinet.net
[2] Department of Information Management, Lunghwa University of Science and Technology,
Taiwan, R.O.C.
kplin@mail.lhu.edu.tw
[3] Department of Computer Science and Information Management, Hungkuang University,
Taiwan, R.O.C.
yangklung@yahoo.com.tw
[4] Department of Marketing and Distribution Management,
Overseas Chinese University, Taiwan, R.O.C.
sultan@ocit.edu.tw

Abstract. This paper presents support vector regression with hybrid genetic algorithms and tabu search (GA/TS) algorithms (SVRGA/TS) models for the prediction of radio-wave path-loss in suburban environment. The support vector regression (SVR) model is a novel forecasting approach and has been successfully used to solve time series problems. However, the application of SVR model in a radio-wave path-loss forecasting has not been widely investigated. This study aims at developing a SVRGA/TS model to forecast radio-wave path-loss data. Furthermore, the genetic algorithm and tabu search techniques have be applied to select important parameters for SVR model. In this study, four forecasting models, Egli, Walfisch and Bertoni (W&B), generalized regression neural networks (GRNN) and SVRGA/TS models are employed for forecasting the same data sets. Empirical results indicate that the SVRGA/TS outperforms other models in terms of forecasting accuracy. Thus, the SVRGA/TS model is an effective method for radio-wave path-loss forecasting in suburban environment.

Keywords: Support vector regression, genetic algorithms and tabu search, radio-wave path-loss.

1 Introduction

The prediction of propagation path loss is an important step in planning a wireless communication system and accurate prediction methods are needed in order to determine the parameters of a radio system that will provide efficient and reliable coverage of a specified area. However, several global and local parameters (e.g., relief, objects at the propagation path, climate zones, foliage characteristics, reflecting index in

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 243–251, 2010.

atmosphere, propagation along many paths, etc.) will affect path-loss prediction model. Previous existing models mainly focus on experimental models. Traditional path-loss prediction model mainly used experimental models such as Egli 1, W&B 11 models, which briefly be introduced in following sections, 6, and 3. The problem of these models is that expressions are based on the qualitative propagation environments such as urban, suburban and open areas. Recent years the artificial neural networks (ANNs) has been adopted to forecast path-loss data. Milovanovic *et al.* 5 hybrid Okumura-Hata with neural networks to estimate path-loss measured result. However, the ANNs model has not been widely investigated in path-loss prediction. In this study, a SVRGA model is presented to forecast radio-wave path-loss data in suburban environment. The remainder of this paper is organized as follows. Introduction of the theory of SVRGA/TS is given in Section 2. Numerical example of radio-wave path-loss predictions and empirical results are presented in Section 3. Some concluding remarks are provided in Section 4.

1.1 Egli Model

Egli prediction model is an empirical model which has been proposed by 1. The Egli model is a simplistic model to approach radio-wave path-loss of irregular topography. Based on real data the path-loss approaching can be formulated as following:

$$L = 117 + 40\log d + 20\log f - 20\log(h_b h_m) \tag{1}$$

where d (Km) is total distance in meter, f (MHz) is frequency of transmit, h_b and h_m are the transmitting and receiving antenna heights, respectively.

Thus the free spaces path-loss model is better considered. The free spaces model (L_{f_s}), which includes total distance in meter and is independent of frequency, is formulated as following

$$L_{f_s} = 32.27 + 20\log d + 20\log f \tag{2}$$

The Egli model considers the free spaces path-loss, hence the Egli prediction model (L_E) can be formulated as following:

$$L_E = L - L_{f_s} = 84.73 + 20\log(d) - 20\log(h_b h_m) \tag{3}$$

1.2 Walfisch-Bertoni Model

Walfisch and Bertoni 11 proposed the half empirical prediction model. The model considers the impact of rooftops and building height by using diffraction to predict average signal strength at street level. The method describes urban propagation loss as a sum of three terms: (1) free space losses (2) rooftop-to-street losses and (3) multiple diffraction losses. Hence, the Walfisch and Bertoni prediction model can be formulated as following:

$$L_{WB} = L_0 + L_{msd} + L_{rts} \tag{4}$$

where L_0 is free spaces path-loss, L_{msd} means multiple diffraction losses and L_{rts} means rooftop-to-street losses. Furthermore, the Walfisch and Bertoni prediction model can be derived as following:

$$\begin{aligned} L_{WB} &= 89.55 + 21\log f + 38\log d - 38\log H + (5\log((b/2)^2 + (h_b - h_m)^2) \\ &\quad - 9\log b + 20\log(\tan^{-1}(2(h_b - h_m)/b))) - 18\log(1 - (d^2/17H)) \end{aligned} \tag{5}$$

where H is height of transmit, b is distance between buildings.

2 Support Vector Regression with GA/TS

Support vector machine model has been successfully extended for dealing with nonlinear regression problems (8, 9 and 10). The support vector regression model is based on the idea of mapping the original data x nonlinearly into a higher dimensional feature space. The SVR approach is to approximate an unknown function by a training data set $\{(x_i, Y_i), i=1, \ldots, N\}$. The regression function can be formulated as follows:

$$F = w\phi(x_i) + b \tag{6}$$

where $\phi(x_i)$ denotes the feature of the inputs, and w and b indicate coefficients. The coefficients (w_i and b) are estimated by minimizing the following regularized risk function.

$$R(F) = C\frac{1}{N}\sum_{i=1}^{N} L_\varepsilon(Y_i, F_i) + \frac{1}{2}\|w\|^2 \tag{7}$$

where

$$L_\varepsilon(Y_i, F_i) = \begin{cases} 0 & if \quad |Y_i - F_i| \le \varepsilon \\ |Y_i - F_i| - \varepsilon & otherwise \end{cases} \tag{8}$$

where C and ε are user-defined parameters. The parameter ε is the difference between actual values and values calculated from the regression function. This difference can be viewed as a tube around the regression function. The points outside the tube are regarded as training errors. In Eq. (7), $L_\varepsilon(Y_i, F_i)$ is called an ε-insensitive loss function, and can be illustrated as Fig. 1.

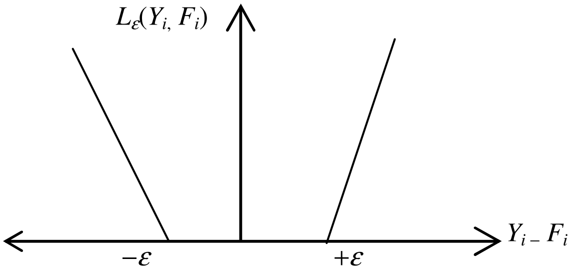


Fig. 1. The ε-insensitive loss function

The loss equals zero if the approximate value is within the ε-tube. Additionally, the second item of Eq. (7), $\frac{1}{2}\|w\|^2$, is adopted to estimate the flatness of a function which can avoid overfitting. Therefore, *C* indicates a parameter determining the trade-off between the empirical risk and the model flatness. Two positive slack variables (ξ_i and $\xi^*{}_i$), representing the distance from actual values to the corresponding boundary values of the ε-tube, are then introduced. These two slack variables equal zero when the data points fall within the ε-tube. Eq. (7) is then reformulated into the following constrained form:

$$\begin{aligned}
&Min\, f(w,\xi,\xi^*) = \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{N}(\xi_i+\xi_i^*)\right) \\
&subjective\ to \\
&w\phi(x_i)+b-Y_i \le \varepsilon+\xi_i^*\ ,\ i=1,2,\cdots,\ N \\
&Y_i-w\phi(x_i)-b \le \varepsilon+\xi_i\ ,\ i=1,2,\cdots,\ N \\
&\xi_i,\xi_i^* \ge 0,\ i=1,2,\cdots,\ N
\end{aligned} \tag{9}$$

This constrained optimization problem can be solved using the following primal Lagrangian form:

$$\begin{aligned}
&Min\frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{N}(\xi_i+\xi_i^*)\right) - \sum_{i=1}^{N}\beta_i\left[w\phi(x_i)+b-Y_i+\varepsilon+\xi_i\right] \\
&-\sum_{i=1}^{N}\beta_i^*\left[Y_i-w\phi(x_i)-b+\varepsilon+\xi_i^*\right] \\
&-\sum_{i=1}^{N}(\alpha_i\xi_i+\alpha_i^*\xi_i^*)
\end{aligned} \tag{10}$$

Eq. (10) is minimized with respect to primal variables *w*, *b*, ξ, and ξ^*, and is maximized with regard to non-negative Lagrangian multipliers α_i, α_i^* , β_i, and β_i^* . Finally, Karush-Kuhn-Tucker conditions are applied to Eq. (9), and the dual Lagrangian form given by Eq. (6).

$$\begin{aligned}
&Max\sum_{i=1}^{N}Y_i(\beta_i-\beta_i^*) - \varepsilon\sum_{i=1}^{N}(\beta_i+\beta_i^*) - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\beta_i-\beta_i^*)(\beta_j-\beta_j^*)K(x_i,x_j) \\
&subjective\ to \\
&\sum_{i=1}^{N}(\beta_i-\beta_i^*)=0 \\
&0\le\beta_i\le C,\quad i=1,2,\cdots,N \\
&0\le\beta_i^*\le C,\quad i=1,2,\cdots,N
\end{aligned} \tag{11}$$

The Lagrange multipliers in Eq. (11) satisfy the equality$\beta_i * \beta_i^*$ =0. The Lagrange multipliers, β_i and β_i^* , are determined, and an optimal weight vector of the regression hyper plane is written by Eq. (12).

$$w^* = \sum_{i=1}^{N} (\beta_i - \beta_i^*) K(x, x_i) \tag{12}$$

Thus, the regression function is given by:

$$F(x, \beta, \beta^*) = \sum_{i=1}^{N} (\beta_i - \beta_i^*) K(x, x_i) + b \tag{13}$$

Herein, $K(x_i, x_j)$ denotes a Kernel function whose value equals the inner product of two vectors, x_i and x_j , in the feature space $\phi(x_{\mathrm{i}})$ and $\phi(x_j)$, meaning that $K(x_i, x_j) = \phi(x_i) * \phi(x_j)$. Any function that satisfies Mercer's condition 4 can act as the Kernel function. This work uses the Gaussian function. Hence, in the case of Gaussian kernels, the SVR has three tuning parameters (C, ε, σ).

In this study, the GA/TS 2 algorithm with binary coding was employed to determine parameters of SVR models. The procedure of GA/TS is illustrated as follows.

Step 1 (Initialization):

Establish randomly an initial population of chromosomes. Three parameters, σ, C and ε, are expressed in a binary format; and represented by a chromosome.

Step 2 (Evaluating fitness):

Evaluate the fitness of each chromosome. In this study, the negative MAPE was used as the fitness function which can see Eq. (14).

Step 3 (Selection):

Based on the fitness functions, chromosomes with higher fitness values are more likely to yield offspring in the next generation. The roulette wheel selection principle (Holland, 1975) is applied to select chromosomes for reproduction.

Step 4 (Crossover and mutation):

Create new offspring by performing crossover and mutation operations. Mutations were performed randomly by converting a "1" bit into a "0" bit or a "0" bit into a "1" bit. In this study, the single-point-crossover principle was employed. Segments of paired chromosomes between two determined breakpoints are exchanged. The probabilities of crossover and mutation were set at 0.5 and 0.1, respectively.

Step 5 (Perform TS on each chromosome):

Evaluating neighbor chromosome and adjusting the tabu list. The tabu list size is 20 in this study. The chromosome with the smallest MAPE value and not having been recorded in the tabu list was placed in the tabu list. In this study, a first-in-first-out policy is performed for operating the tabu list. If the best neighbor chromosome is the same as one of the chromosome in the tabu list, then it generates the next set of neighbor chromosomes and calculates the fitness value of chromosome. The next set of neighbor chromosome is generated from the best neighbor chromosome in the current iteration.

Step 6 (Current chromosome selection by TS):

If the best neighbor chromosome is better then the current chromosome, then the current chromosome is replaced by the best neighbor chromosome. Otherwise, keep the current chromosome.

Step 7 (Next generation):

Form a population for the next generation. The size of the population was set to 50.

Step 8 (Stop criterion).

If the number of epochs equals a given scale, then the best chromosomes are presented as a solution; otherwise go back to Step 2. The number of epochs was set to 1000.

Fig. 2 shows the flowchart of the proposed SVRGA/TS model. The SVR model was then employed to forecast wave-radio path-loss data; and the GA/TS was performed to determine parameters of SVR models. The three parameters resulting in the smallest validation error are then used to develop a appropriate SVR model. Finalized forecasting vales are obtained by SVR. Finally, after comparison with the testing data, the testing errors are obtained.

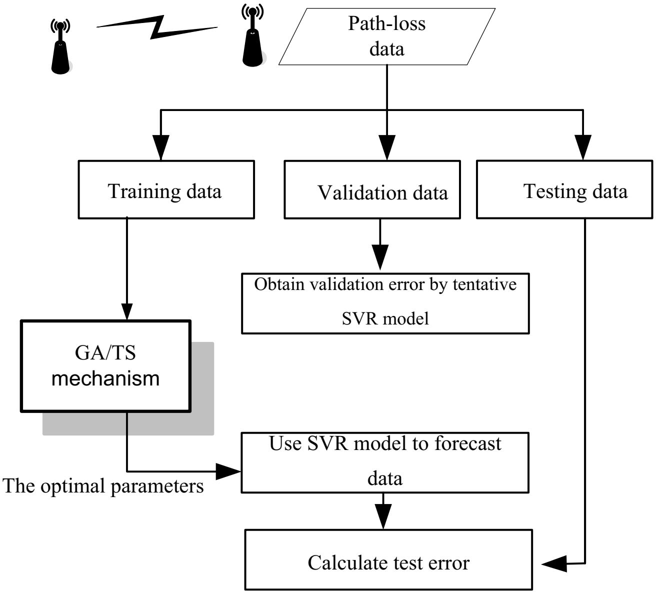


Fig. 2. A flowchart of the SVRGA/TS

3 Numerical Example and Discussions

In this research numerical example, which is actual wave-radio path-loss data in suburban environment, is examined. Table1 shows the actual wave-radio path-loss data with measurement distance between the 0.5 and 5 Km. Moreover, the experimental parameters value should be known, which frequency of transmit (f) is 937 MHz, the transmitting antenna height (h_b) is 30 m, and receiving antenna height (h_m) is 1.5 m, respectively. In addition, MAPE and the root mean square error (RMSE) are used to measure the forecasting accuracy of three models. Eqs. (14) and (15) show the expression of MAPE and RMSE respectively:

$$\text{MAPE}(\%) = -\frac{100}{M}\sum_{l=1}^{M}\left|\frac{g_l - z_l}{g_l}\right| \tag{14}$$

$$\text{RMSE} = \left\{\frac{1}{M}\sum_{l=1}^{M}\left(g_l - z_l\right)^2\right\}^{0.5} \tag{15}$$

where M is the number of forecasting periods; g_l is the actual value at period l; and z_l is the forecasting value at period l.

Fig. 3 shows the estimated results of Egli and W&B models, which have been introduced in previous sections, for comparison of performance. Moreover, this research also compares popular ANN model, which is GRNN 7 model, with SVRGA/TS.

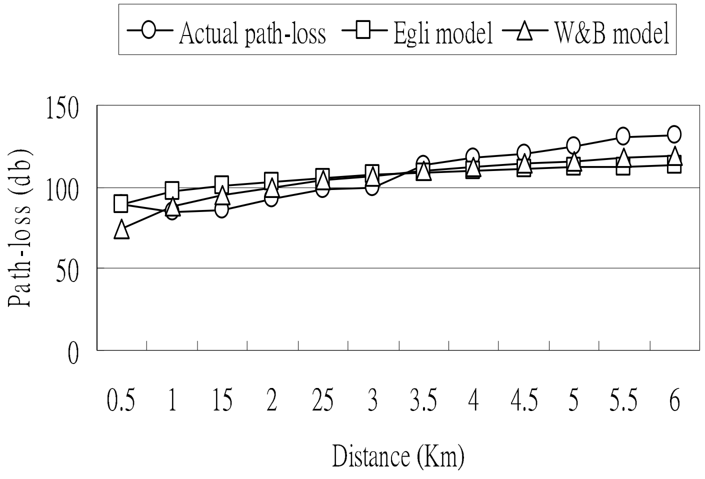


Fig. 3. Illustration of actual values and forecasting values of traditional models

For GRNN and SVRGA/TS models, experimental data are divided into training, validation and testing data sets. Table 1 depicts the actual path-loss data employed in this work for ANN modes. The training data set was used to determine the forecasting model; the validation data set was for the purpose of preventing over-fitting of forecasting models; and the testing data set is employed to investigate the performance of different forecasting models. For both examples, the training data set and the validation set are used to design SVRGA/TS models. For comparing the forecasting accuracy, the same testing data set is examined for ANN forecasting models. In the research we test the path-loss data between 2.5 and 6 Km. Table 1 show the forecasting performances and preferred parameters of two ANN models for path-loss data. SVRGA/TS model outperformed the GRNN model in terms of forecasting accuracy. In Table 1 also shows the optimal parameters for SVRGA/TS model. Figs. 3 and 4 make point-to-point comparisons of actual values and predicted values of path-loss data. As shown in Fig. 3, the traditional Egli and W&B models cannot efficiently capture the trend of data. In Fig. 4 the ANN models can find that efficiently captures the trend of data, and the SVRGA/TS and GRNN are able to follow the data trend. Table 2 shows and ranks the performances of four models. The outcome of the SVRGA/TS is better than other models in terms of forecasting accuracy.

Table 1. Comparison of the forecasting results and parameters from ANN model

Distance	GRNN	SVRGA/TS	C/ε/σ
2.5	92.19	92.95	98.06/ 0.05/0.61
3.0	98.29	98.95	96.80/ 0.11/0.55
3.5	99.86	99.97	97.19/0.03/ 0.70
4.0	111.51	112.20	96.13/0.05/0.64
4.5	117.38	117.62	97.77/0.38/0.85
5.0	119.75	119.46	98.16/ 0.54/ 0.83
5.5	124.42	124.55	110.88/0.05/0.91
6.0	129.41	129.91	130.51/ 0.04/0.88

Table 2. Forecasting performances of four models

Forecasting models	Egli model	W&B model	GRNN	SVRGA/TS
MAPE(%)	9.5	7.4	4.7	4.4
RMSE	11.42	8.82	6.48	6.20
Rank	(4)	(3)	(2)	(1)

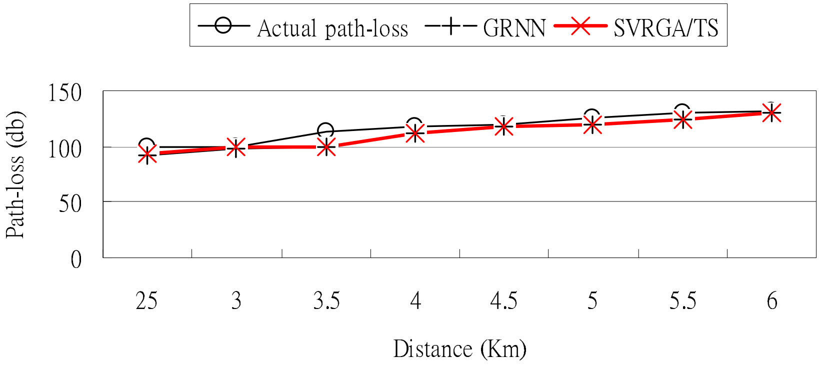


Fig. 4. Illustration of actual values and forecasting values of ANN models

4 Conclusions

This study proposed a hybrid SVRGA/TS model for forecasting path-loss, with experimental results being valid and satisfied. The superior performance of the SVRGA/TS model can be ascribed to two causes. First, the SVRGA can efficiently capital trend of nonlinear data and estimate precisely. Second, based on GA/TS searching, the SVRGA/TS model can provide optimal parameters for radio-wave path-loss predicting. For future work, forecasting other environments of path-loss data by a SVR-related model is a challenging issue for study.

Acknowledgments. The authors would like to thank the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC-98-2410-H-606-004 and NSC 98-2410-H-262 -002.

References

1. Egli, J.J.: Radio Propagation above 40 Mc over Irregular Terrain. In: IRE 45, pp. 1383–1391 (1957)
2. Glover, F., Kelly, J.P., Laguna, M.: Genetic algorithms and tabu search: hybrids for optimization. Computers and Operations Research 22, 111–134 (1995)
3. Joshi, G.G., Dietrich Jr., C.B., Anderson, C.R., Newhall, W.G., Davis, W.A., Isaacs, J., Barnett, G.: Near-ground Channel Measurements over Line-of-sight and Forested Paths. IEEE Proc.-Microw. Antennas Propag 152, 589–596 (2005)
4. Mercer, J.: Function of Positive and Negative Type and Their Connection with the Theory of Integral Equations. Philosophical Transactions of the Royal Society A209, 415–446 (1909)
5. Milovanovic, B., Stankovic, Z., Milijić, M., Sarevska, M.: Near-Earth Propagation Loss Prediction in Open Rural Environment using Hybrid Empirical Neural Model. In: TELSIKS 2007, pp. 423–426. IEEE Press, Serbia (2007)
6. Rama Rao, T., Vijaya Bhaskara Rao, S., Prasad, M.V.S.N., Sain, M., Iqbal, A., Lakshmi, D.R.: Mobile Radio Propagation Path Loss Studies at VHF/UHF Bands in Southern India. IEEE Transactions on Broadcasting 46, 158–164 (2000)
7. Specht, D.F.: A General Regression Neural Network. IEEE Transactions on Neural Networks 2, 568–576 (1991)
8. Vapnic, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)
9. Vapnik, V., Golowich, S., Smola, A.: Support Vector Machine for Function Approximation Regression Estimation, and Signal Processing. Advances in Neural Information Processing Systems 9, 281–287 (1996)
10. Vapnic, V.: Statistical Learning Theory. Wiley, New York (1998)
11. Walfisch, J., Bertoni, H.L.: A Theoretical Model of UHF Propagation in Urban environments. IEEE Transactions on Antennas and Propagation 36, 1788–1796 (1988)

Inventory Model with Fractional Brownian Motion Demand

Jennifer Shu-Jen Lin

Department of Transportation Logistics & Marketing Management, Toko University
jennifer1592001@yahoo.com.tw

Abstract. This paper studies the inventory model where the demand satisfies fractional Brownian motion. It is an extension of the Master thesis of Buffy Yang of the Department of Industrial Management in the National Taiwan University of Science and Technology in 2009. The purpose of this paper is threefold. First, we point out why the standard derivation of the total demand during the lead time is proportional to the lead time with Hurst exponent. Second, we analytically prove that our proposed inventory model has a unique minimum solution. Third, we demonstrate by simulation for the comparison between our proposed models with the traditional models to indicate that our average saving is about 6%. Our model will provide a better managerial policy for the demand that meets the behavior of fractional Brownian motion.

Keywords: inventory model, fractional Brownian motion, crashable lead time.

1 Introduction

The Brownian motion (Bm), especially the fractional Brownian motion (fBm), has seldom been employed in the field of inventory although the investigation of inventory system and related applications have been undertaking for many years. As a matter of fact, through sensitive observation and analysis, it can be found that the characteristic of dependence among a series of data of fractional Brownian motion may pave a way for the breath-taking development of inventory management. With the study of fractional Brownian motion, Hurst [6] found an analogous situation: a simple relation among a set of long-term time-series data for the natural events. For properly identifying the so-called persistent and anti-persistent phenomenon and gauging the intensity of long-range dependence of the data in the time series generated by a number of natural events, Hurst derived a parameter named as Hurst exponent and denoted by H. Afterwards, some academicians have managed to effectively link the statistical applications with fractional Brownian motion, for examples, Piterbarg [9], and Rao [10]. Queuing theory is a statistics-related area that has captured a big chunk of studies on the linkage to the fractional Brownian motion. The relevant papers that have been published include Brichet *et al.* [2], Chang *et al.* [3], Kim *et al.* [7], and Awad and Glynn [1].

For capturing the randomness of lead time demand and high dependence among lead time demand, and shortage cost, Chou [4] applied Hurst exponent [6] and fractional Brownian motion to find the optimal order quantity while achieving the minimum average total cost for a (Q, r) inventory system with multi-products, backorders, and

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 252–259, 2010.

investment constraint. Yang [11] developed an inventory model where the demand during the fixed lead time follows fractional Brownian motion. She employed Hurst exponent to diagnose and modify the (Q, r) inventory model. It turned out that the results obtained are better than those derived from the demand-independent models.

Mainly motivated by Yang [11], this paper expands the investigation to the inventory system where the market demand satisfies fractional Brownian motion. The authors validly justify the proportional relationship between the standard derivation of the total demand during the lead time and the lead time itself with Hurst exponent. The verification of the existence of a unique minimum solution of the inventory model proposed by authors is also provided in the study. At last, through the demonstration of simulation, the comparative results indicate the marked cost saving of 6% contributed by the proposed models over the traditional models.

2 Notation and Assumptions

In order to describe the proposed inventory model, the following notation and assumptions are used.

2.1 Notation

X	stochastic demand
D	average demand per year
μ	average demand per day, with $D = 365\mu$
Q	order quantity
A	fixed ordering cost per order
h	inventory holding cost per item per year
t	fixed replenishment lead time
X_t	the demand during lead time
r	reorder point
k	safety factor
β	fraction of the demand backordered during the stockout period, $0 \le \beta \le 1$
q	allowable stockout probability during lead time, with $\Pr\left(X_t > r\right) = q$
π	fixed penalty cost per unit short
π_0	marginal profit per unit
H	Hurst exponent, $0 < H < 1$
C_H	the unit time variance from fractional Brownian motion
$B(r)$	the expected shortage at the end of the cycle period

2.2 Assumptions

1. The demand is a discrete random variable, that is characteristic by fractional Brownian motion.
2. The standard derivation during lead time, t, is denoted as $t^H \sqrt{C_H}$.
3. The inventory model is a continuously reviewed. Replenishments are made

whenever the inventory level drops to the reorder point, r.

4. Symbol β represents the fraction of the demand backordered during the stockout period. So 1-β is the ratio of the lost sales from the stockout.
5. The reorder point r consists of mean and safety stock, so $r = t\mu + kt^H\sqrt{C_H}$, where k is the safety factor.

3 Proposed Inventory Model

We consider the inventory model with stochastic demand so that we do not know the distribution of the demand but from the past record, we have its mean and variance and the demand followed fractional Brownian motion. According to previous research Yang [11], the expected annual cost is expressed as

$$EAC(Q,r) = A\frac{D}{Q} + h[\frac{Q}{2} + k\sqrt{C_H}t^H + (1-\beta)B(r)] + \frac{D}{Q}[\pi + \pi_0(1-\beta)]B(r) \tag{1}$$

In order to resolve the problems derived from unknowing distribution and only knowing the mean and variance of the lead time demand. We must use the following proposition was asserted by Gallego *et al.* [5] that provided a tight upper bound for the unknown shortage during lead time.

Proposition 1: Gallego *et al.* [5],

$$B(r) \le \frac{1}{2}[\sqrt{C_H t^{2H} + (r-\mu t)^2} - (r-\mu t)], \tag{2}$$

with $r = \mu t + k\sqrt{C_H}t^H$, then we rewrite equation (2),

$$B(r) \le \frac{1}{2}\sqrt{C_H}t^H(\sqrt{1+k^2} - k). \tag{3}$$

Since the reorder point is decided by the safety factor, to simplify the expression, we treat the safety factor as a new variable to instead the reorder point, then we study an approximated expected annual cost,

$$\begin{aligned} EAC(Q,k) &= A\frac{D}{Q} + h\left[\frac{Q}{2} + k\sqrt{C_H}t^H\right] \\ &+ \frac{t^H}{2}\sqrt{C_H}\left(\sqrt{1+k^2} - k\right)\left\{h(1-\beta) + \frac{D}{Q}\left[\pi + \pi_0(1-\beta)\right]\right\}. \end{aligned} \tag{4}$$

Next, we consider the first order partial derivatives,

$$\frac{\partial EAC(Q,k)}{\partial Q} = \frac{-AD}{Q^2} + \frac{h}{2} - \frac{D}{2Q^2}[\pi + \pi_0(1-\beta)]\sqrt{C_H}t^H(\sqrt{1+k^2} - k), \tag{5}$$

and

$$\frac{\partial EAC(Q,k)}{\partial k} = h\sqrt{C_H}t^H + \frac{1}{2}\{h(1-\beta) + \frac{D}{Q}[\pi + \pi_0(1-\beta)]\}\sqrt{C_H}t^H(\frac{k}{\sqrt{1+k^2}} - 1). \tag{6}$$

And then the second order partial derivative,

$$\frac{\partial^2 EAC(Q,k)}{\partial Q^2}=\frac{2AD}{Q^3}+\frac{D}{Q^3}[\pi+\pi_0(1-\beta)]\sqrt{C_H}t^H(\sqrt{1+k^2}-k)>0\,, \tag{7}$$

$$\frac{\partial^2 EAC(Q,k)}{\partial k\partial Q}=\frac{\partial^2 EAC\left(Q,k\right)}{\partial Q\partial k}=\frac{-D}{2Q^2}\left(\pi+\pi_0(1-\beta)\right)\sqrt{C_H}t^H(\frac{k}{\sqrt{1+k^2}}-1)\,, \tag{8}$$

and

$$\frac{\partial^2 EAC(Q,k)}{\partial k^2}=\frac{1}{2}\{h(1-\beta)+\frac{D}{Q}[\pi+\pi_0(1-\beta)]\}\sqrt{C_H}t^H[(1+k^2)^{\frac{-3}{2}}]>0\,. \tag{9}$$

According to

$$\frac{\partial^2 EAC(Q,k)}{\partial Q^2}>\frac{D}{Q^3}[\pi+\pi_0(1-\beta)]\sqrt{C_H}t^H(\sqrt{1+k^2}-k)\,, \tag{10}$$

and

$$\frac{\partial^2 EAC(Q,k)}{\partial k^2}>\frac{D}{Q}[\pi+\pi_0(1-\beta)]\}\sqrt{C_H}t^H[(1+k^2)^{\frac{-3}{2}}]\,, \tag{11}$$

we know that

$$\begin{aligned}&\frac{\partial^2 EAC\left(Q,k\right)}{\partial Q^2}\frac{\partial^2 EAC\left(Q,k\right)}{\partial k^2}-\left[\frac{\partial^2 EAC(Q,k)}{\partial k\partial Q}\right]^2\\&>\frac{D^2}{4Q^4}\left(\frac{\pi+\pi_0\left(1-\beta\right)}{1+k^2}\right)^2 C_H t^{2H}\left\{1-k^2+2k^3\left(\sqrt{1+k^2}-k\right)\right\}.\end{aligned} \tag{12}$$

Motivated by equation (12), we assume an auxiliary function, say $h\left(k\right)=1-k^2+2k^3\left(\sqrt{1+k^2}-k\right)$ for $k>0$. We may rewrite $h\left(k\right)$ as follows,

$$\begin{aligned}h\left(k\right)&=1-k^2+\frac{2k^3}{\sqrt{1+k^2}+k}=1+k^2\left(\frac{k-\sqrt{1+k^2}}{k+\sqrt{1+k^2}}\right)\\&=1-\frac{k^2}{\left(k+\sqrt{1+k^2}\right)^2}=\sqrt{1+k^2}\,\frac{2k+\sqrt{1+k^2}}{\left(k+\sqrt{1+k^2}\right)^2}>0.\end{aligned} \tag{13}$$

From equations (7), (12) and (13), it yields that the Hessian matrix of $EAC\left(Q,k\right)$ is positive definite so that if $\left(Q^{\#},k^{\#}\right)$ is a solution satisfies the system of the first partial derivatives, equations (5) and (6), then $\left(Q^{\#},k^{\#}\right)$ is the absolute minimum solution. From equations (5) and (6) to solve $\frac{\partial EAC(Q,k)}{\partial Q}=0$ and $\frac{\partial EAC(Q,k)}{\partial k}=0$, then we need to solve the following two equations simultaneously,

$$Q^2 = \frac{D}{h}\left\{2A + \left[\pi + \pi_0(1-\beta)\right]\frac{\sqrt{C_H}\, t^H}{\sqrt{1+k^2}+k}\right\}, \tag{14}$$

and

$$\frac{k^*}{\sqrt{1+k^{*2}}} = 1 - \frac{2h}{h(1-\beta) + \frac{D}{Q}\left[\pi + \pi_0(1-\beta)\right]}. \tag{15}$$

We combine equations (14) and (15) to cancel out Q, then it yields that

$$\begin{gathered}\frac{D}{h}\left[\pi + \pi_0(1-\beta)\right]^2 = \\ \left(1 + 2k\left(k + \sqrt{1+k^2}\right) + \beta\right)^2 \left\{2A + \left[\pi + \pi_0(1-\beta)\right]\frac{\sqrt{C_H}\, t^H}{\sqrt{1+k^2}+k}\right\}.\end{gathered} \tag{16}$$

Motivated by equations (16), we assume an auxiliary function, say $f(k)$, where

$$f(k) = \left(1 + 2k\left(k + \sqrt{1+k^2}\right) + \beta\right)^2 \left\{2A + \left[\pi + \pi_0(1-\beta)\right]\frac{\sqrt{C_H}\, t^H}{\sqrt{1+k^2}+k}\right\}, \tag{17}$$

and then we rewrite equations (17) as follows

$$\begin{aligned} f(k) &= 2A\left(1 + 2k\left(k + \sqrt{1+k^2}\right) + \beta\right)^2 \\ &+ \left(1 + 2k\left(k + \sqrt{1+k^2}\right) + \beta\right)\left[\pi + \pi_0(1-\beta)\right]\sqrt{C_H}\; t^H g(k), \end{aligned} \tag{18}$$

with $g(k) = 2k + \frac{1+\beta}{k + \sqrt{1+k^2}}$. It shows that

$$\begin{aligned} g'(k) &= 2 - \frac{1+\beta}{\sqrt{1+k^2}\left(\sqrt{1+k^2}+k\right)} \\ &= 2 - (1+\beta)\frac{\sqrt{1+k^2}-k}{\sqrt{1+k^2}} = 1 - \beta + (1+\beta)\frac{k}{\sqrt{1+k^2}} > 0. \end{aligned} \tag{19}$$

It is trivial that $1 + 2k\left(k + \sqrt{1+k^2}\right) + \beta$ is an increasing function of k. According to equations (18), we know that $f(k)$ is an increasing function of k from $f(0) = (1+\beta)^2\left(2A + \left[\pi + \pi_0(1-\beta)\right]\sqrt{C_H}\, t^H\right)$ to $\lim_{k\to\infty} f(k) = \infty$.

We divide into two cases: Case 1, $\frac{D}{h}\left[\pi + \pi_0(1-\beta)\right]^2 \ge f(0)$, and Case (2), $\frac{D}{h}\left[\pi + \pi_0(1-\beta)\right]^2 < f(0)$.

When Case 1 happens, then there is a unique k that is the solution of equation (16), say k^*, which is the optimal safety factor. Consequently, using equations (14), we derive the optimal order quantity.

When Case 2 happens, from equation (9), it shows that $\frac{\partial EAC(Q,k)}{\partial k}$ is an increasing function in variable k and $\lim_{k\to\infty} \frac{\partial EAC(Q,k)}{\partial k} = h\sqrt{C_H} t^H > 0$. It motivates us to prove that $\frac{\partial EAC(Q,k)}{\partial k}$ is positive when $k = 0$, consequently $\frac{\partial EAC(Q,k)}{\partial k} > 0$ for $k > 0$ so that the minimum value occurs at $k = 0$.

We compute $\frac{\partial EAC(Q,k)}{\partial k}$ at $k = 0$, then

$$\frac{\partial EAC(Q,k)}{\partial k}\Big|_{k=0} = \frac{\sqrt{C_H} t^H}{2}\left\{h(1+\beta) - \frac{D}{Q}\left(\pi + \pi_0(1-\beta)\right)\right\}. \tag{20}$$

From equation (10), $EAC(Q,k)$ is a convex function in Q, so when k is fixed, for example $k = 0$ for the time being, the best policy of Q is equation (14) with $k = 0$. Hence, to verify $\frac{\partial EAC(Q,k)}{\partial k}\Big|_{k=0} > 0$ is equivalent to show that

$$h^2(1+\beta)^2 Q^2 > D^2\left[\pi + \pi_0(1-\beta)\right]^2. \tag{21}$$

where Q satisfies equation (14) with $k = 0$, that is the condition of Case (2).

4 Numerical Example

We have applied the Mandelbrot and John method [8] to generate a daily record of discrete demand that satisfies fractional Brownian motion with $H = 0.9$. We consider two inventory models as mentioned in equation (9) with different values of Hurst exponent, H, with $H = 0.9$ and $H = 0.5$ so the expect average costs are expressed as $EAC_{H=0.9}(Q,k)$ and $EAC_{H=0.5}(Q,k)$ with the following parameters: A = \$200 per order, h = \$15 per item per year, π =10 per unit short, π_0 =3 marginal profit per unit, C_H =4, fixed replenishment lead time t = 16 days.

After we generate a 365 days data, we compute its sum as D, demand per year. For $EAC_{H=0.9}(Q,k)$, we use equation (16) to find the safety factor, $k = 1.285$, and then apply equation (14) to locate the order quantity $Q = 312.9806$ with reorder point, $r = 192.4408$. We use a computer program to evaluate the daily holding coat, when inventory level drop to reorder point, $r = 192.4408$, then we replenish $Q = 312.9806$, after 16 days, the items will arrive. During the lead time period, we still compute the daily holding cost and shortage cost, if necessary. There are 13 replenishments during one year, with

$$TC_{H=0.9}\left(Q=312.9806,k=1.285\right)=5708.8127\,.$$

Similarly, we consider if we overlook the data is satisfied the fractional Brownian motion with $H=0.9$ and still apply the traditional inventory model with $H=0.5$, then we find that

$$TC_{H=0.5}\left(Q=313.4417,k=1.2838\right)=6058.0781,$$

with reorder point $r=171.5468$.

The relative saving for this example is computed by

$$\frac{TC_{H=0.5}-TC_{H=0.9}}{TC_{H=0.9}}=6.13\%\,.$$

Next, we have run 100 tests, 87 of them the $TC_{H=0.9}$ is cheaper than $TC_{H=0.5}$ and the rest owing to shortage dose not happen, then $TC_{H=0.9}$ is more expansive than $TC_{H=0.5}$.

If we consider the 100 experiments to discover that our proposed inventory model is better with an average saving of 5.77%. It points out that with the demand satisfies the fractional Brownian motion, then our proposed invent model provides a better managerial replenishment policy.

5 Conclusion

We developed a new inventory model under fractional Brownian demand. Our contribution is to explain that the derivation of the total demand during the lead time is proportional to the lead time with Hurst exponent. We have analytically divided the problem into two cases and then proved that our model has a unique minimum solution for each case. Our approach will be useful for the future research to guarantee the uniqueness of the minimum solution. From the simulation experiment, our proposed inventory model has an average saving of 6% over the traditional inventory model to signify the contribution of our paper.

References

1. Awad, H., Glynn, P.: Conditional Limit Theorems for Regulated Fractional Brownian Motion. Annual Application of Probability 19(6), 2102–2136 (2009)
2. Brichet, F., Roberts, J., Simonian, A., Veitch, D.: Heavy Traffic Analysis of a Storage Model with Long Range Dependent On/Off Sources. Queuing Systems 23(1-4), 197–215 (1996)
3. Chang, C.S., Yao, D.D., Zajic, T.: Large Deviations and Moderate Deviation, and Queues with Long-range Dependent Input. Advanced Applied Probability 31, 254–277 (1999)
4. Chou, S.H.: The fBm Demand Model in a Multi-product (Q, r) Inventory System with Backorders and Investment Constraint. Master Thesis, Department of Industrial Management, Taipei Technology University (2006)
5. Gallego, G., Yao, D.D., Moon, I.: The Distribution Free Newsboy Problem. The Journal of Operational Research Society 44(8), 825–834 (1993)

6. Hurst, H.E.: Long-term Storage Capacity of Reservoirs. American Society of Civil Engineers 116, 770–808 (1956)
7. Kim, S., Nam, S.Y., Sung, D.K.: Effective Bandwidth for a Single Server Queuing System with Fractional Brownian Input. Performance Evaluation 61(2-3), 203–223 (2005)
8. Mandelbrot, B.B., John, W.V.: Fractional Brownian Motions, Fractional Noises and Applications. SIAM Review 10, 422–437 (1968)
9. Piterbarg, V.I.: Large Deviations of a Storage Process with Fractional Brownian Motion as Input. Extremes 4, 147–164 (2001)
10. Rao, B.L.S.P.: Identification for Linear Stochastic Systems Driven by Fractional Brownian Motion. Stochastic Analysis and Applications 22(6), 1487–1509 (2005)
11. Yang, S.T.: Inventory Management for Dependent Demand Characterized by Fractional Brownian Motion. Master Thesis, Department of Industrial Management, Taipei Technology University (2009)

Note on Inventory Model with a Stochastic Demand

Yu-Wen Wou

Department of Finance, Chihlee Institute of Technology
yuwen@mail.chihlee.edu.tw

Abstract. This paper is a response to Mandal and Pal (1998), Wu and Ouyang (2000) and Deng et al. (2007). We study their paper to point out an interesting phenomenon. In the future research, after researchers provide a reasonable explanation for this phenomenon that may provide an insightful understanding for inventory models.

Keywords: Inventory model; deteriorating item; ramp type demand.

1 Introduction

There has been a trend towards examining inventory models with ramp type demand both in depth and breadth since Hill [5] launched the study of that kind. Some examples include Mandal and Pal [6] with deterioration items; Wu et al. [8] assuming that the backlogging rate is proportional to the waiting time; Wu and Ouyang [9] with two different strategies beginning with stock or shortage; Wu [10] with the Weibull distributed deterioration; Giri et al. [4] with a more generalized Weibull distributed deterioration; Deng [2] to revise the work of Wu et al. [8]; Manna and Chaudhuri [7] to extend the inventory model a with time dependent deterioration rate; and Deng et al. [3] who modify Mandal and Pal [6] and Wu and Ouyang [9].

However, their studies and corresponding results are all under one assumption that the demand pattern must be ramp type. This paper points out that there is a general property effectively shared for every kind of demand. That is, the optimal solution of an inventory system is independent of the demand types. Our findings show that the lengthy discussion and derivation of Deng et al. [3] in the consideration of two different expressions of ramp type demand are in fact superfluous. An inventory model is presented in the study where each item has two stages: the first stage with stock and the second stage with shortage. The results indicate that different inventory levels of different items with different demands are able to reach zero simultaneously.

2 Assumptions and Notation

We try to generalize the inventory model of Mandal and Pal [6], Wu and Ouyang [9] and Deng et al. [3] with the following assumptions and notations for the deterministic inventory replenishment policy with a general demand.

(1) The replenishment rate is infinite; thus, replenishments are instantaneous.
(2) The lead time is zero.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 260–264, 2010.

(3) T is the finite time horizon under consideration. We follow the assumption of Deng et al. [3] to set $T=1$ so that the length of the inventory model equals to the unit time.

(4) C_h is the inventory holding cost per unit per unit of time.

(5) C_s is the shortage cost per unit per unit of time.

(6) C_d is the cost of each deteriorated item.

(7) θ is the constant fraction of the on-hand inventory deterioration per unit of time.

(8) $I(t)$ is the on-hand inventory level at time t over the ordering cycle $[0,T]$.

(9) Shortage is allowed and fully backordered.

(10) S is the maximum inventory level for each ordering cycle, that means $S=I(0)$.

(11) The demand rate $R(t)$ is assumed to be any nonnegative function.

(12) t_1 is the time when the inventory level reaches zero.

(13) t_1^* is the optimal solution for t_1.

(14) $f(t_1)$ is an auxiliary function defined as $\left(C_d+\frac{C_h}{\theta}\right)\left(e^{\theta t_1}-1\right)-C_s(T-t_1)$.

(15) $C(t_1)$ is the total cost that consistent of holding cost, deterioration cost, and shortage cost.

3 Review of Previous Results

Previous papers of Mandal and Pal [6], Wu and Ouyang [9] and Deng et al. [3] must divide the inventory model into several cases that is depending on (a) the ramp type demand have different expression, and (b) the relation of the turning point of the ramp type demand and the time the inventory level reaches to zero. Therefore, their derivation became very complicate such that sometimes some cases are overlooked by some researchers. Deng et al. [3] had revised the results of Mandal and Pal [6] and Wu and Ouyang [9]. Recently, Cheng and Wang [1] further extended the ramp type demand to a trapezoidal type demand such that the demand classification is three phases and then their analytical work becomes more complex. We will abstract the demand to a general nonnegative function and then to show that there is a unique optimal that is independent of the demand such that those discussion of different type demands can be uniformly solved by our approach.

4 Our Proposed Inventory Model

We consider an inventory model that starts with stock. This model was first proposed by Hill [5], and then further investigated by Mandal and Pal [6], Wu and Ouyang [9] and Deng et al. [3].

Therefore, the total cost is expressed as

$$C(t_1)=C_d\int_0^{t_1}R(x)\left(e^{\theta x}-1\right)dx+C_h\int_0^{t_1}R(x)\frac{e^{\theta x}-1}{\theta}dx+C_s\int_{t_1}^{T}(T-x)R(x)dx. \quad (1)$$

From equation (1), it follows that

$$C'(t_1) = R(t_1)\left[\left(C_d + \frac{C_h}{\theta}\right)\left(e^{\theta t_1} - 1\right) - C_s\left(T - t_1\right)\right]. \tag{2}$$

Motivated by equation (2), we assume an auxiliary function, say $f(t_1)$.

$$f(t_1) = \left(C_d + \frac{C_h}{\theta}\right)\left(e^{\theta t_1} - 1\right) - C_s\left(T - t_1\right) \tag{3}$$

As a matter of fact, it is nothing but the same one proposed by Deng et al. [3]. It point out that after complicated computation, Deng et al. [3] can not realize that the solution of equation (1) is independent of the demand.

By taking derivative of $f(t_1)$, i.e., $f'(t_1) = (\theta C_d + C_h)e^{\theta t_1} + C_s > 0$, it is easy to find that $f(t_1)$ increases from $f(0) = -C_s T < 0$ to $f(T) = \left(C_d + \frac{C_h}{\theta}\right)\left(e^{\theta T} - 1\right) > 0$. Hence, obviously there exists a unique point, say $t_1^{\#}$, that satisfies $f\left(t_1^{\#}\right) = 0$ and the following equation holds.

$$\left(C_d + \frac{C_h}{\theta}\right)\left(e^{\theta t_1^{\#}} - 1\right) = C_s\left(T - t_1^{\#}\right). \tag{4}$$

Since $C'(t_1) \le 0$ for $0 \le t_1 \le t_1^{\#}$ and $C'(t_1) \ge 0$ for $t_1^{\#} \le t_1 \le T$ such that $t_1^{\#}$ is the minimum solution for $C(t_1)$. We summarize our findings in the next theorem.

Theorem 1. For the inventory model beginning with stock, the minimum solution satisfies $f\left(t_1^{\#}\right) = 0$ and is independent of the demand.

5 Explanation of Our Findings

We may assume that there are products $P_1,...,P_m$ with the same holding cost C_h, deterioration cost C_d and shortage cost C_d. However, they have a different demand, $R_1(t),...,R_m(t)$. In the beginning, the initial inventory levels are replenished with $I_1(0),...,I_m(0)$. Along with the demand and deterioration, the stock consumes and inventory levels finally drop to zero, at $t_1,...,t_m$ respectively.

In general, we may predict that there are no connections among $t_1,...,t_m$. However, by our findings that points out

$$t_1 = ... = t_m = t_1^{\#} \tag{5}$$

which is actually independent of the demand. To consolidate the application and significance of this finding, let's quote the business model from investment banks such as Merrill Lynch, JP Morgan, Morgan Stanley, Bear Stearns, and so on. In light of market needs at different time points, those investment banks design and issue several financial commodities including for instance federal bond fund, high income fund, balanced fund, and diverse underlying-linked structured commodities. Each of them has different maturity, return, risk, service fee, and contract conditions in order to meet various requirements of investors. However, for the purposes of process simplification, management centralization, and cost reduction, the issuers always set up a fixed selling period for salesmen, say two weeks for example. During which all promoted financial commodities are planned to be sold out, i.e., zero inventory, by the end of designated fixed selling period. And after which another cluster of targeted customer-driven financial commodities will be engineered and stood-by.

6 Different View of Our Findings

We intend to provide an explanation from microscope point of view to discuss our findings. If there is an item with demand quantity Q that takes place at time t, then there are two replenishment policies: (a) fulfill the demand from the stock, or (b) satisfy the demand from backorder. If we decide to fulfill the demand from the stock, we need to store $Qe^{\theta t}$ at time $t=0$. Note that the solution of $\frac{d}{dt}I(t)+\theta\, I(t)=0$ is $I(t)=I(0)e^{-\theta t}$ so that the beginning stock is $I(0)=Qe^{\theta t}$. It follows that at time is t after deteriorated items are removed, the remaining stock Q is just enough to meet the demand. The amount of deteriorated items is $Q\left(e^{\theta t}-1\right)$. From the inventory level $Qe^{\theta t}e^{-\theta x}$ for $x\in[0,t]$, the holding cost can be calculated by

$$C_h\int_0^t Qe^{\theta t}e^{-\theta x}dx = C_h\frac{Q}{\theta}\left(e^{\theta t}-1\right). \tag{6}$$

Hence, the total cost for demand Q fulfilled from the stock is $Q\left(\frac{C_h}{\theta}+C_d\right)\left(e^{\theta t}-1\right)$.

On the other hand, if the policy of backlog is adopted then the shortage cost is $QC_s(T-t)$. We find that if

$$Q\left(\frac{C_h}{\theta}+C_d\right)\left(e^{\theta t}-1\right) < QC_s(T-t) \tag{7}$$

then the better policy is to satisfy demand from the stock. Otherwise, if

$$Q\left(\frac{C_h}{\theta}+C_d\right)\left(e^{\theta t}-1\right) > QC_s(T-t) \tag{8}$$

then the better policy is to hold demand as backlog until the replenishment takes place. Recall from equation (3), we have shown that $f(t)=\left(C_d+\frac{C_h}{\theta}\right)\left(e^{\theta\, t}-1\right)-C_s(T-t)$ has an unique root, say $t_1^{\#}$. It means that for those demands occurring during $\left[0,t_1^{\#}\right)$, they should be fulfilled from the stock And for those demands occurring during

$\left(t_1^{\#}, T\right]$, they should be replenished by backorder. Our alternative approach comes up with the same results as our previous presentation by analytical method.

7 Conclusion

In the paper we discover an interesting phenomenon for finite time horizon inventory model that the single item at diverse segmenting markets or different items with the same holding cost, deterioration cost and shortage cost can be deliberately scheduled to achieve the same in-stock period for the purpose of reaching minimum inventory cost. The generalized form of inventory model is developed and corresponding optimal solution is derived. We believe that out findings provide an essential benchmark for those researchers who have the motive of pursuing different optimal inventory systems along with the changes of demand types. This study gives the solid evidence that the optimal solution is independent of the demand no matter ramp type, trapezoid type, fixed type or any other kinds of it are targeted.

References

1. Cheng, M., Wang, G.: A Note on the Inventory Model for Deteriorating Items with Tapezoidal Type Demand rate. Computers & Industrial Engineering 56, 1296–1300 (2009)
2. Deng, P.S.: Improved Inventory Models with Ramp Type Demand and Weibull Deterioration. International Journal of Information and Management Sciences 16(4), 79–86 (2005)
3. Deng, P.S., Lin, R., Chu, P.: A Note on the Inventory Models for Deteriorating Items with Ramp Type Demand Rate. European Journal of Operational Research 178, 112–120 (2007)
4. Giri, B.C., Jalan, A.K., Chaudhuri, K.S.: Economic Order Quantity Model with Weibull Deterioration Distribution, Shortage and Ramp-type Demand. International Journal of Systems Science 34(4), 237–243 (2003)
5. Hill, R.M.: Inventory Models for Increasing Demand Followed by Level Demand. Journal of the Operational Research Society 46(10), 1250–1259 (1995)
6. Mandal, B., Pal, A.K.: Order Level Inventory System with Ramp Type Demand Rate for Deteriorating Items. Journal of Interdisciplinary Mathematics 1, 49–66 (1998)
7. Manna, S.K., Chaudhuri, K.S.: An EOQ Model with Ramp Type Demand Rate, Time Dependent Deterioration Rate, Unit Production Cost and Shortages. European Journal of Operational Research 171(2), 557–566 (2006)
8. Wu, J.W., Lin, C., Tan, B., Lee, W.C.: An EOQ Model with Ramp Type Demand Rate for Items with Weibull Deterioration. International Journal of Information and Management Sciences 10, 41–51 (1999)
9. Wu, K.S., Ouyang, L.Y.: A replenishment policy for deteriorating items with ramp type demand rate. Proceeding of National Science Council ROC (A) 24, 279–286 (2000)
10. Wu, K.S.: An EOQ Inventory Model for Items with Weibull Distribution Deterioration, Ramp Type Demand Rate and Partial Backlogging. Production Planning & Control 12, 787–793 (2001)

A GA-Based Support Vector Machine Diagnosis Model for Business Crisis

Ming-Fen Yang[1] and Huey-Der Hsiao[2]

[1] Department of Leisure and Sports Management, Far East University, Taiwan, R.O.C
[2] Department of Business Administration, Far East University, Taiwan, R.O.C
{arthur,shiao322}@cc.feu.edu.tw

Abstract. This research proposes a diagnosis model for business crisis integrated a real-valued genetic algorithm and support vector machine. A series of learning and testing processes with real business data show that the diagnosis model has a crisis prediction accuracy of up to 95.56%, demonstrating the applicability of the proposed method. Six features, including five financial and one intellectual capital indices, are used for the diagnosis. These features are common and easily accessible from publicly available information. The proposed GA-SVM diagnosis model can be used by firms for self-diagnosis and evaluation.

Keywords: Business crisis, Diagnosis model, Genetic algorithm, Support vector machine.

1 Introduction

A business crisis has significant adverse effects on individuals and households, and consequently on the national and social economy. The avoidance of such crises can enhance the relationship between society and individuals, leading to greater general prosperity. Therefore, it is necessary to establish a diagnostic model for business crises to aid business managers in monitoring business performance.

This research attempt to integrate the real-valued genetic algorithm (GA) and support vector machine (SVM), i.e., the GA-SVM model, to diagnose business crises. The use of real-valued GA is to find the optimal settings of parameters in SVM. In addition to finance features, the features of intellectual capital (IC) are also included to clarify whether they are helpful for the task. Furthermore, for improving the model's performance, feature selection is undertaken by employing discriminant analysis (DA) to find the critical features for the input variables.

The rest of this paper is organized as follows. Section 2 briefly introduces various diagnosis methods relevant to business crises and research methodologies adopted by this research. The structure and procedure of the GA-SVM are described in Section 3. In Section 4 and 5, the empirical study and their analytical results of the diagnosis model are presented. The conclusion is provided in the final section.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 265–273, 2010.

2 Paper Review

2.1 Various Diagnosis Methods of Business Crisis

For more than 40 years scholars have conducted research into business bankruptcy, and in the process have developed a number of diagnosis techniques. Among these, traditional statistical methods are the most common, such as logistic regression (logit) and discriminant analysis [1,2,3,4,5].

Non-parametric statistical analysis and artificial intelligence (AI) methods, such as those that use artificial neural networks (ANNs), have garnered the most attention in recent years [3,6,7,8]. The classification and regression trees (CART) approach is another common technique that has been used as a business crisis indicator [9,10].

In addition to the aforementioned classification methods, the support vector machine method has received attention for its ability to separate hyperplanes to separate data for pattern identification [11]. Recently, there has been increased interest in the application of SVM for the prediction of credit ratings, stock prices, bankruptcy, and insurance claim fraud detection [8,12,13,14,15,16,17].

2.2 Support Vector Machine

Support vector machine, introduced by Vapnik [18], is based on a learning system that uses linear functions in a high dimensional feature space. SVM was mainly used to find out a separating hyperplane to separate two classes of data from the given data set. SVM is trained with a learning algorithm based on optimization theory to implement a learning bias derived from statistical learning theory. The learning strategy and its extension [19] have been widely adopted by many researchers. SVMs have been shown to outperform most other systems in a wide variety of applications [20].

2.3 Genetic Algorithm

A genetic algorithm is used to solve global optimization problems. The procedure starts from a set of randomly created or selected possible solutions, referred to as the population. Every individual in the population means a possible solution, referred to as a chromosome. Within every generation, a fitness function should be used to evaluate the quality of every chromosome to determine the probability of it surviving to the next generation; usually, the chromosomes with larger fitness have a higher survival probability. Thus, GA should select the chromosomes with larger fitness for reproduction by using operations like selection, crossover and mutation in order to form a new group of chromosomes which are more likely to reach the goal. This reproduction goes through one generation to another, until it converges on the individual generation with the most fitness for goal functions or the required number of generations was reached. The optimal solution is then determined.

GA coding strategies mainly include two sectors; one sector recommends the least digits for coding usage, such as binary codes; another one recommends using the real valued coding based on calculation convenience and accuracy [21].

2.4 Discriminant Analysis

Discriminant analysis is a conventional statistical method that is widely applied in various fields. The implementation procedures lie in the previous validation to inspect whether the group gravity center comes with the variance, thereafter, it continues to find the predicting variables with the maximum separating capabilities that could be grouped into separable functions. Finally, based on the separable function, test data can be evaluated and then assigned to a certain group. Aside from the fact that it is used to predict different groups, it is also able to screen out the critical attributes for groups from multi-attribute groups [22].

3 GA-SVM Model

The proposed diagnosis model is shown in Fig. 1. SVM is used to train the dataset and to perform the diagnostic task with the aid of real-valued GA and DA.. Real-valued GA is employed for searching for the optimum parameter settings in SVM, and DA is used for feature selection. The operating procedure is described as follows:

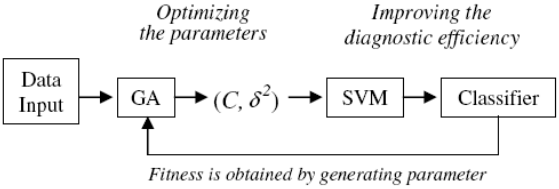

Optimizing the parameters
Improving the diagnostic efficiency
Data Input
GA
(C, δ^2)
SVM
Classifier
Fitness is obtained by generating parameter

Fig. 1. Schema of the proposed GA-SVM model

Step 1. Make settings for initial value ranges of (C, δ^2) and construct an initial population.
Step 2. Randomize initial population.
Step 3. Training SVM model (5-fold cross-validation) using each pair of (C, δ^2).
Step 4. Calculate fitness values.
Step 5. If the stopping condition (five rounds) is satisfied, go to Step 8.
If 50 generations are carried out, go to Step 7.
Step6. Perform selection, reproduce, crossover, and mutation operations to create a new population. Go to Step 3 for the next generation.
Step 7. Repeat steps 2–7.
Step 8. Find out the optimal (C*, δ^2*).
Step 9. Train the SVM model with (C*, δ^2*).
Step 10. Implement the diagnosis model against business crises.

4 An Empirical Study

4.1 Variables and Samples

A business crisis can be defined as a situation in which firms on the public stock market have become full-cash delivery or suspended stock transactions, require dramatic restructuring, declare bankruptcy, or withdraw from the market [15,23,24,25,26,27,28,29].

The diagnosis of a business crisis should not only consider financial information, but must also include intellectual capital. 27 financial features and 7 intellectual capital attributes were selected (seen Table 1). For training, every feature must include at least four quarters of data before the business crisis (excluding data from the current quarter). Each quarter of data of a feature is defined as a variable in the diagnosis model; there are thus a total of 136 variables.

Table 1. Features of business crisis

Attribute	Relevant features	Quantity
Finance	Efficiency	13
	Operating profit capability,	7
	Liability fulfillment capability	7
Intellectual capital	Manpower	3
	Structure	2
	Customer	2

The sampled data for this study were obtained from the Taiwan Economic Journal Data Bank (TEJ) based on two criteria. Firstly, as mentioned above, the sampled firms had at least four quarters of complete public information before their business crisis occurred. Secondly, sufficiently similar companies were selected to act as the opposing samples in an industry. Based on these principles, 186 firms from roughly all the industries listed in the TEJ were selected with relevant data from June 1999 to March 2006.

4.2 Parameters and Their Ranges

To establish an SVM model, a kernel function should first be determined. In this research, based on Smola's proposal [30], radial basis function (RBF) is selected as the kernel function to establish an SVM, since it produces better classification outcomes when data attributes are unknown. For the RBF functions, the values of parameter δ^2 and cost parameter C, which serve as the upper bounds of the Lagrange coefficients, should be carefully determined in order to reach satisfactory outcomes for prediction.

Studies have shown that real-valued GA performs well over a wide search range. The present study defines the search ranges of C and δ^2 as [1, 1000] and [0.5, 1000], respectively, instead of [10, 100] and [1, 100], as recommended by Tay and Cao [16]. The initial population is composed of 50 sets of chromosomes; five-fold cross-validation is carried out in the training process; a standard roulette wheel selection is adopted; and the crossover and mutation possibilities are rated at 0.8 and 0.05, respectively. Finally, every round has 50 generations, with a total of five rounds implemented.

4.3 Feature Selection

To build up a simple and robust diagnosis model, 15 data sets are established. Each dataset consists of 186 sampled firms in a random sequence to break the original firm parity orders.

Among 15 data sets, 10 datasets are randomly selected. Each contains two types of data. Data sets with finance attribute only could be denoted with R_f(i.e., R_{f1}, . . ., R_{f10}). If there were features of IC attributes added, then it could be denoted with R_{fi}(i.e., R_{fi1}, . . ., R_{fi10}). Therefore, 20 datasets are actually used. R_{fi1} and R_{f1} mean the same group of random arrangement.

Firstly, the comparisons of the diagnosis results between *Rf* and *Rfi* datasets are made by conducting the classification training and testing all variables with the GA-SVM model. In addition, DA is used to find the significant features from each dataset. The significant features from each dataset are then adopted to act as the diagnosis variables of GA-SVM*d* to the corresponding dataset. The comparisons of the diagnosis results from the above models can achieve two goals; one is to indicate the necessity of adding features of the IC attributes, and another is for confirming the sufficiency of only the significant features adopted in the GA-SVM*d*. We named the above processes as the phase-1 feature selection. Furthermore, since DA is also a classification approach, the diagnosis results from DA can be compared with those from the GA-SVM model to prove the effectiveness of GA-SVM model. Fig. 2 shows the comparisons in the phase-1 processes.

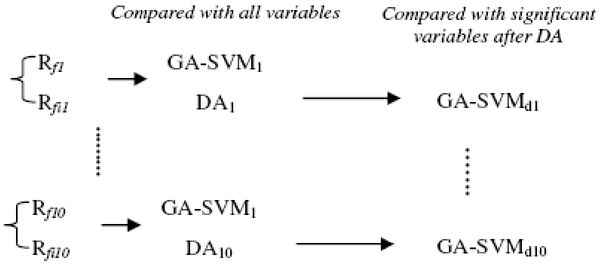

Compared with all variables
Compared with significant variables after DA
R_f1
R_fi1
GA-SVM1
DA1
GA-SVMd1
R_f10
R_fi10
GA-SVM1
DA10
GA-SVMd10

Fig. 2. Comparison of R_f and R_{fi}

Based on the outcomes in the first phase, the features that are always significant in the four quarters in DA (DA1, . . ., DA10) are adopted for the remaining five datasets. Obviously, the quantity of features will be reduced, and the model could be adopted for the future diagnosis when the diagnosis results are proved effective. This stage is called the phase-2 process.

5 Empirical Analysis

5.1 Experiments in the 1st Phase

The experiments were conducted following the GA-SVM procedure in Section 3. Before implementations, the value of every data point should be scaled within [0, 1] to avoid the dominance among different data points. For each dataset, the first 150 entries in a sequence were used for the SVM learning and testing operations through the general 5-fold cross-validation. The best and average testing accuracy rates of each generation from the cross-validation can be used for further analysis.

Fig. 3 shows the accuracy analysis of cross-validation for R_{fi1} experienced with the first round GA-SVM. The figure shows the convergence happens from the 13th generation for both the best and average testing accuracy, indicating that the parameters found by GA result in a good performance of SVM in training for R_{fi1} in this round. The accuracy and stability from the other rounds could not be superior to those of the first round, so the outcome in the first round could represent the cross-validation results of R_{fi1}. Furthermore, from the converged generations in the first round, the parameters (C^* = 879, δ^{2*} = 3.8851) with the best accuracy could be used to conduct the training of 150 entries and then perform the testing of the remaining 36 entries, reaching a training and testing accuracy of 97.333% and 88.889%, respectively, shown as random group 1 in Fig. 4.

Following the above processes, the training and testing accuracy for R_{f1} were 92.667% and 88.889%, respectively. Through the comparisons for testing accuracy between R_{f1}~R_{f10} and R_{fi1}~R_{fi10}, the trends in the figure indicate that GA-SVM is more stable with regard to testing accuracy with the entries that are IC attributes. The average testing accuracy rated at 83.68% was slightly better than that produced via the entries with only features of finance attributes, rated at 83.42%, as illustrated in Table 2.

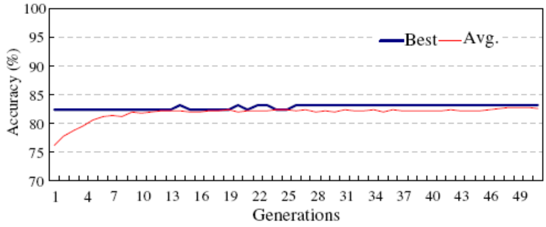


Fig. 3. Cross-validation for R_{fi1} at 1st round

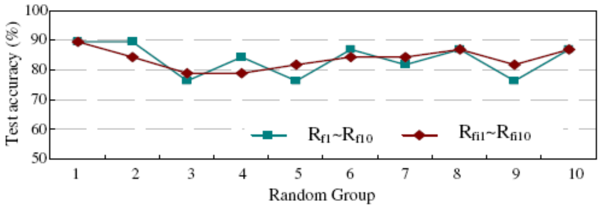


Fig. 4. Comparison of accuracy for GA-SVM using all variables

In addition, using the same data sets and variables as those in Fig. 4, the conventional DA was applied to select the significant variables with Wilks' Lambda value ≦ 0.01 in order to conduct classification for the testing data sets. As an additional comparison, the average testing accuracy (75.53%) for the entries added with features of IC attributes was slightly more stable than that (73.16%) for those with only features of finance attributes; It also shows that the diagnosis effect of GA-SVM was better than that of DA.

Variables with high significance after performing DA act as the diagnosis variables for the phase-1 GA-SVM model. 63 and 67 variables were adopted for R_f and R_{fi}, as indicated in Table 2. The quantity of variables was reduced by roughly 50%~60%, however the average testing accuracy increased to 87.11% for R_{fi}.

Table 2. Summarized comparison of R_f and R_{fi}

	GA-SVM using all variables		DA		GA-SVMd after DA	
	R_{f1}~R_{f10}	R_{fi1}~R_{fi10}	R_{f1}~R_{f10}	R_{fi1}~R_{fi10}	R_{f1}~R_{f10}	R_{fi1}~R_{fi10}
Quantity of training variables	108	136	108	136	63	67
Average training accuracy (%)	91.96	94.46	94.53	94.05	91.55	95.54
Quantity of testing variables	108	136	63	67	63	67
Average test accuracy (%)	83.42	83.68	73.16	75.53	83.68	87.11

5.2 Diagnosis Models for Business Crises

After applying DA to the R_{fi} data sets, there are 67 significant variables, belonging to 14 features, as listed in Table 3. For further decreasing the quantity of significant features, only those containing four significant subordinate variables, are adopted as the phase-2 diagnosis variables, namely liabilities ratio (LBR), debt adhere degree (DAD), return of assets (ROA), earnings per share (EPS), cash flow ratio (CFR) and operating benefit margin per capita (OBMC). The above six features contained a total of 24 variables that acted as the diagnosis variables for five data sets (R_{fi11}~R_{fi15}). Fig. 5 shows the cross-validation results from R_{fi11} in the first round. In Fig. 5, the best and average

Table 3. Features with significance in the first and second phases

Attribute	Feature
Finance	Current ratio, Quick ratio, Liability ratio, Debt adhere degree, Total asset turnover, Return on asset, Net value return after tax, Operating profit margin, EPS, Cash flow per share, Net value growth rate, Cash flow ratio
Intellectual Capital	Operating profit margin per capita, Return rate on manpower asset

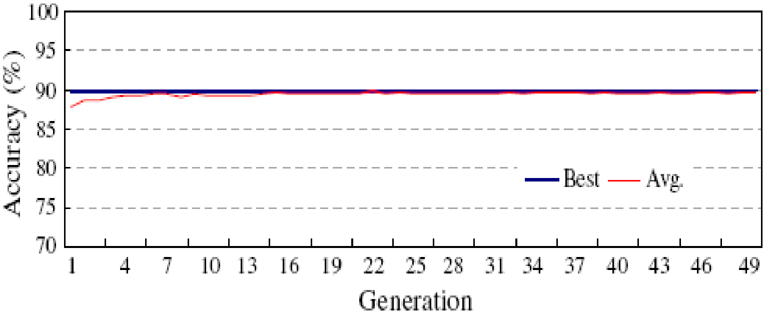


Fig. 5. Testing accuracy on cross validation for R_{fi11} in the first round

testing accuracy of the cross validations are almost converged from the fifth generation. This means that the six selected features can achieve a satisfactory diagnosis performance. Adopting the parameter values ($C^* = 483$, $\delta^2{}^* = 2.0412$) at the 44th generation, the 150 entries in R_{fi11} are the training set and the remaining 36 entries are the testing set, reaching the accuracy 97.33% and 97.22%, respectively. Similar processes were applied to R_{fi12}~R_{fi15}, and the associated outcomes are shown in Table 4. The diagnosis results for business crises had an average rate of 95.56%, demonstrating a quite high identification level.

Table 4. The results of business crisis diagnosis

	R_{fi11}	R_{fi12}	R_{fi13}	R_{fi14}	R_{fi15}	Average
Round	1	5	2	4	3	
Generation	44	50	50	46	49	
Learning accuracy (%) (150 entries)	97.33	97.33	96.0	94.67	96.33	96.13
Testing accuracy (%) (36 entries)	97.22	100	94.44	91.67	94.44	95.56
C^*	483	470	416	369	487	
$\delta^2{}^*$	2.0412	5.1653	1.0242	1.3581	4.7133	

6 Conclusion

This research integrated GA and SVM to establish the diagnosis model for business crises by using their traits in parameter evolution as well as data training and classification. After training the data sets from the real data base containing the financial and intellectual capital data of business units in Taiwan, the proposed GA-SVM diagnosis model reached an average testing accuracy of 95.56% with only six features. These features are common and easily accessible from publicly available information, making the proposed method very practical for managers to conduct a real-time investigation of the potential for a business crisis.

References

1. Altman, E.I., Marco, G., Varetto, F.: Corporate distress diagnosis comparisons using linear discriminant analysis and neural networks. Journal of Banking and Finance 18(3), 505–529 (1994)
2. Falbo, P.: Credit-scoring by enlarged discriminant models. Omega 19(4), 275–289 (1992)
3. Jo, H., Han, I., Lee, H.: Bankruptcy prediction using case-based reasoning, neural networks and discriminant analysis. Expert Systems With Applications 13(2), 97–108 (1997)
4. Martin, D.: Early warning of bank failure: A logit regression approach. Journal of Banking and Finance 1, 249–276 (1997)
5. Ohlson, J.A.: Financial ratios and probabilistic prediction of bankruptcy. Journal of Accounting Research 18(1), 109–131 (1980)
6. Boritz, J., Kennedy, D.: Effectiveness of neural networks types for prediction of business failure. Expert Systems with Applications 9, 503–512 (1995)
7. Malhotra, R., Malhotra, D.K.: Differentiating between good credits and bad credits using neuro- fuzzy systems. European Journal of Operational Research 136(2), 190–211 (2002)

8. Zurada, J.M., Malinowski, A., Usui, S.: Perturbation method for deleting redundant inputs of perceptron networks. Neurocomputing 14, 177–193 (1997)
9. Frydman, H.E., Altman, E.I., Kao, D.: Introducing recursive partitioning for financial classification: The case of financial distress. The Journal of Finance 40(1), 269–291 (1985)
10. Srinivasan, V., Ruparel, B.: CGX: An expert support system for credit granting. European Journal of Operational Research 45, 293–308 (1990)
11. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998)
12. Fan, A., Palaniswami, M.: Selecting bankruptcy predictors using a support vector machine approach. In: Proc. of the International Joint Conference on Neural Networks, vol. 6, pp. 354–359 (2000)
13. Huang, C.L., Wang, C.J.: A GA-based feature selection and parameters optimization for support vector machines. Expert Systems with Applications 31, 231–240 (2006)
14. Min, J.H., Lee, Y.C.: Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Systems with Applications 28, 603–614 (2005)
15. Shin, K.S., Lee, T.S., Kim, H.J.: An application of support vector machines in a bankruptcy prediction model. Expert Systems with Applications 28, 127–135 (2005)
16. Tay, F.E.H., Cao, L.J.: Application of support vector machines in financial time series forecasting. Omega 29(4), 309–317 (2001)
17. Tay, F.E.H., Cao, L.J.: Modified support vector machines in financial time series forecasting. Neurocompuitng 48, 847–861 (2002)
18. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, New York (1995)
19. Zhong, P., Wang, L.: Support vector regression with input data uncertainty. International Journal of Innovative Computing, Information and Control 4(9), 2325–2332 (2008)
20. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines. Cambridge University Press, Cambridge (2000)
21. Haupt, R.L., Haupt, S.E.: Practical genetic algorithms. John Wiley and Sons, New York (1998)
22. Johnson, R.A., Wichern, D.W.: Applied multivariate statistical analysis. Prentice-Hall Inc., New York (1998)
23. Ahn, B.S., Cho, S.S., Kim, C.Y.: The integrated methodology of rough set theory and artificial neural network for business failure prediction. Expert Systems with Applications 18, 65–74 (2000)
24. Altman, E.I.: Corporate Financial Distress: A Complete Guide to Predicting, Avoiding and Dealing with Bankruptcy. John Wiley and Sons, New York (1983)
25. Coat, P.K., Fant, L.F.: Recognizing financial distress patterns using a neural network tool. Financial Management 12(3), 142–155 (1993)
26. Deakin, E.: A discriminant analysis of predictors of business failure. Journal of Accounting Research 10, 167–179 (1972)
27. Gilson, S.C.: Management turnover and financial distress. Journal of Financial Economics 25, 241–262 (1989)
28. McGurr, P.T., DeVaney, S.A.: Predicting business failure of retail firms: An analysis using a mixed industry model. Journal of Business Research 43, 169–176 (1998)
29. Min, S.H., Lee, J., Han, I.: Hybrid genetic algorithms and support vector machines for bankruptcy prediction. Expert Systems with Applications 31, 652–660 (2006)
30. Smola, A.J.: Learning with Kernels. Ph.D.Thesis, Technical University Berlin, Germany (1998)

A Book-Agent System for Library Marketing

Toshiro Minami

Kyushu Institute of Information Sciences, 6-3-1 Saifu, Dazaifu,
Fukuoka 818-0117 Japan
Kyushu University Library, 6-10-1 Hakozaki, Higashi,
Fukuoka 812-8581 Japan
minami@kiis.ac.jp, minami@lib.kyushu-u.ac.jp
http://www.kiis.ac.jp/~minami/

Abstract. Libraries have been adapting themselves to the patrons' changing requests. Library marketing with data analysis will inspire the improvement and development of patron services in such a situation. In this paper we will propose a framework for library marketing based on the multi agent system (MAS), where not only the library but also the books, services, and patrons are represented as agents. Agents make a social network and develop new knowledge that is useful for library marketing. Library marketing system based on the MAS framework will provide novel inspirations to the libraries for better patron services and better library management. We put special interest to MAS of books. We propose an idea for extracting knowledge about various relationships of books based on the co-usage data.

1 Introduction

Libraries and their relating participants form a multi agent system (MAS) in nature. Each library organization has its own management policy and independent from other library organizations. A university library is run by the university, a public library is run by the federal or a local government in general, and one in a company is run by the company. The main target people as patrons are students, professors, and staff for a university library, people who live and/or work in the area for public libraries, and the workers for the company libraries.

At the same time, libraries also rely on each other by joining library associations and provide with cooperative patron services. ILL (Inter-Library Loan) is a typical example of such a service. A patron, or a user, of a library can ask the library to get a book for him/her even if it does not be found in the library. The library will help him/her by finding a library that owns the requested book, asking the library for borrowing or making a copy of the book, and getting the book for the patron. The most important mission of libraries is to help with their patrons with providing information material, e.g. books, magazines, newspapers, and with supporting them for strengthening their information literacy skills. ILL is a win-win system for such libraries in this point of view.

The mission of library is well described in the well-known "The five laws of library science" [13], which was advocated a half century ago: (1) Books are

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 274–283, 2010.

for use, (2) Every reader his book, (3) Every book his reader, (4) Save the time of the reader, (5) The library is a growing organism. In order to carry out such missions libraries have been introducing many tools and machines. They have installed computers and provided OPAC (Online Public Access Catalog) service for book retrieval in 1980s and connected them to the Internet so that their patrons are able to access their services like Web-OPAC, online referencing and other library services in 1990s. These laws are still applicable by replacing "book" with "information," "material," or "service." Now libraries have started in introducing the Web 2.0 services such as blogs, SNSs, and others as new patron services.

The major aim of this paper is to propose a multi-agent library system model that is very appropriate to construct such systems. More precisely: (i) to present a model for describing and realizing the library services based on the MAS framework, (ii) to give a grand design for constructing such a MAS system, and (iii) to propose an idea how to create useful knowledge by integrating the "experiences" of agents.

The rest of this paper is organized as follows: In Section 2, we will describe the concept of library marketing (LM). LM is quite an important concept for the future libraries as an information and educational service organization in the network-based ubiquitous society. In Section 3, we will present a concept model for library agent system, where not only libraries but also patrons, books, librarians, and library services are all considered as agents. Then in Section 4, we will show the intelligent bookshelf system based on the radio frequency identification technology and the virtual bookshelf system. We will demonstrate their usefulness as the tools for automatically collecting the book usage data. In Section 5, we will propose a method for extracting closeness relationships between books and demonstrate its usefulness in obtaining the social structure of books and patrons, and using them for new services. Finally in Section 6, we will summarize what we have discussed in this paper and show the prospect of the agent-based library system.

2 Library Marketing

According to the American Marketing Association (AMA) [2] the concept of marketing is defined as "the activity, set of institutions, and processes for creating, communicating, delivering, and exchanging offerings that have value for customers, clients, partners, and society at large." They changed the more profit-oriented old definition to this new one so that it is applicable to non-profit organizations like libraries. Based on this new definition we define library marketing as "the activity that aims for better services to library users, or patrons, so that the libraries can get better reputations, be considered more reliable organizations, and get more customer satisfaction (patron satisfaction)." In addition to this aim, it is preferable to perform their jobs more efficiently, and with less cost; which can be another important aim of library marketing.

We focus on the library marketing based on the analysis methods of the objective data and extracting useful information and knowledge for both libraries

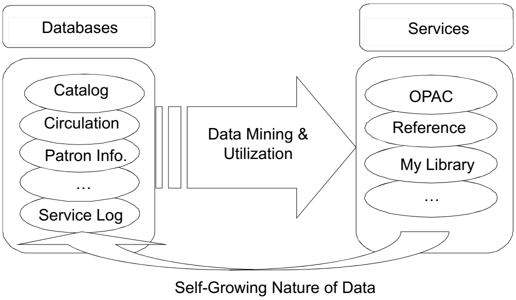


Fig. 1. Library Marketing System based on Data

and their patrons. Figure 1 illustrates the concept of library marketing, with utilizing the data together with the data analysis/mining tools.

It is the mutual depending structure between databases and services. In the right part of the figure are library services provided by the library. Reference service is a consulting service to the patrons who want to know how to find appropriate material, or need some advices. Many libraries provide this service online these days. "My Library" is a service that provides patrons with personalized information. It can be used also for providing the patron with more sophisticated services including SNS, blog, personalized bookshelf in the cyber space, or virtual bookshelf (VBS), and so on.

These services depend on the databases that consist of catalog data of materials, circulation data, patrons' profile data, etc. These data are collected at a circulation counter, at a self checkout machine, and in other ways. The log data, which are collected for recording how the system provides services to patrons, are stored in the service log database and used for various services.

The collected data are supposed to be analyzed, or data mined. The results obtained by analysis should be reflected to the services in order to improve the quality of patron services. If the quality of services gets higher, then more patrons will use the library and the system. If many patrons use them, then the system gets more service log data. If the system is able to get more data, then they can be used for better services. As a result, the customer satisfaction, or patron satisfaction, will rise. Such a natural self-growing mechanism is very important for library marketing with databases and services.

3 Multi-Agent System for a Library

The concept model is the one we consider all the participants of library system as agents. The participants consist of patrons, librarians, books, bookshelves, libraries, services, etc. In this model, we can see the whole library system more clearly and easily than other models that come from other points of view.

Figure 2 illustrates the relationships between agents. The left area indicates the group of agents of the library that is represented as the left-most library in

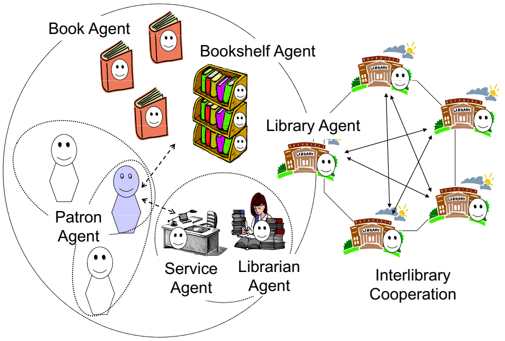


Fig. 2. A Concept Model for Library Agent System

the figure. There are several kinds of agents; patron, book, bookshelf, service, and librarian. Each agent is a representation of one of them. The agents may form a group. In the figure the patron agents have two groups. The groups are formed according to the patron's interests, attributes, or whatever he/she wants. Books and bookshelves also make a group. It is a group based on the relationship between books and a tier of bookshelves; i.e. each book belongs to one of the bookshelf in the tier.

The upper dashed arrow indicates the interactions between the gray patron agent and the group of books and bookshelves. An interaction may happen when the patron takes out the book from a bookshelf. Other possible relationships happen when the patron reserves the book, the patron gives comments in a Web page of the library, or whatever action the patron takes about the book.

The service agent and the librarian agent form a group in Figure 2. It is formed when a librarian is assigned to the service. In other words, the service agent is the representation of the service. Suppose, for example, the job is the reference. Then the service agent represents the reference service. It supports the librarian working as the reference librarian. This group of agents represents a team for this service/job. The lower dashed arrow shows that the patron agent gets the service.

Such groups and interactions are in a library. The libraries make a network for collaboration. A typical example is the ILL (Interlibrary Loan) activity. A patron in a library may want to read a book which is not in the library he/she is using. Then the "home library" of the patron asks another library for borrowing the book. With this ILL system, a patron can borrow whichever book that is purchased by a library in the library group, where the member libraries agree with such collaboration. One possible model for matchmaking the request and provision of information is the word-of-mouth network [12].

The library system consisting as a MAS system has the following advantages: (1) All participants, actions, services, etc. are dealt with in a unified framework. (2) The system becomes scalable and flexible in nature. It is easy to extend the library services and its organization in MAS. (3) As was pointed out, the library system itself is a MAS. It is easy to design and implement the system.

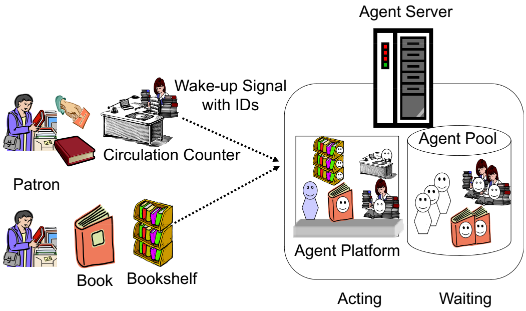


Fig. 3. Wake-Up Mechanism of Agent

Figure 3 illustrates the wake-up mechanism of agents. An ID card of a patron represents the patron himself/herself in the system. It has its corresponding patron agent in the agent server. The ID card itself has no capability of information processing for the patron. Similarly neither a book nor the ID label for the book has the ability of information processing. The book has its corresponding agent, i.e. book agent as a delegate of the book that can perform the information processing for the book or the book ID itself.

With delegate agent mechanism, passive objects like patron's ID card and book ID label have the ability of information processing. A delegate agent is an agent program in a computer called agent server and it does the necessary information processing and calculations for the passive ID objects that exist physically. The agent programs, or agents, are saved in a storage device of the server, or agent pool. We will call that they are "sleeping/waiting" and waiting for been waken-up when they are stored in the agent pool. A sleeping agent will be woken-up when a triggering event occurs. The agent starts running in the agent platform after it is waken-up.

In the upper-left part of Figure 3, a patron arrives at a library circulation counter and asks the librarian for borrowing a book. Then the librarian uses a barcode reader or something and reads the patron and book IDs. Then the library system sends the wake-up signal to the agent server together with the IDs of patron, book, and circulation agent ID. Then the agent server let the corresponding agents run and let them do their jobs. In the lower-left part of Figure 3, a patron puts a book from the bookshelf. Then the corresponding book agent and the bookshelf agent are waken-up if they are sleeping. The event information will be transmitted to the agents and the acting agents will do their necessary jobs.

Now we are developing a library agent system using the JADE platform [6]. The wake-up mechanism is introduced to the book agents (BAs) so that an intelligent bookshelf, as will be explained in the next section, will transmit information to the book manager agent (BMA) about the books as they are either put on the shelf or taken out of the shelf. The BMA will check if the corresponding BAs are acting or waiting. It will wake-up the waiting agents and let them know that their corresponding books get status changes. The BAs will check the catalog and usage records of their corresponding books. They will revise the data about the relationships to other BAs, to the bookshelf agents, to the patron agents, and so on. We will be back to this issue in the next section.

4 Usage Data of Books

This section deals with the two tools for collecting usage data of books; IBS and VBS. IBS (Intelligent Bookshelf) [9] is a bookshelf equipped with antennas for RFID (Radio Frequency Identification) [3] tag system. With IBSs the system can automatically collect the usage data for the books on the bookshelf. VBS (Virtual Bookshelf) [10] is a bookshelf realized on a server and the patron will use it via network with an interface software; e.g. a browser. The image similar to the real bookshelf will be displayed on the screen and the patron can see and manipulate it as if it is a real bookshelf. A VBS is also good equipment for collecting usage data of books and other information sources.

RFID and IBS

For the services using the network, it is easy to collect not only providing reading and other materials for the patrons but also to collect log data. Everything can be carried out automatically. For physical media such as printed books, magazines, audio visual materials like CDs and DVDs, it is a difficult task to do the job and collect the usage data, because these jobs are supposed to be processed by the library staff.

Automatic identification and data capture (AIDC) technology [1] is a great help in such jobs. So far the barcode system has been used in libraries. RFID is getting to be more and more popularly used by libraries in these years [4,5,11] because it is faster and easier to use for library jobs. Further there are more advantages in RFID such as the circulation processing becomes faster and easier, the patrons are happy to use self-checkout machines and do the processing themselves, inventory of library materials becomes faster. Considering these benefits it is worth putting the costs on equipments and RFID tags.

From the library marketing standpoint IBS (Figure 4) is very useful for collecting data. An IBS is a bookshelf which is equipped with RFID antennas so that the IDs of books that are shelved in it are detected and recorded automatically. The antennas and their controller, i.e. R/W, transmit the data to the manager PC and the data will be analyzed either real time or later on.

By using such equipments we can collect the usage data of books. By analyzing the usage data we may extract useful knowledge. For example, we can get the information how often a specific book is used and the differences of usage patterns according to the day of the week, time zone in a day, etc. We can provide such information to the library patrons. It may be used by library staff for planning improved services as well. Such technology should be very important for libraries in the ubiquitous environment that will come in the near future.

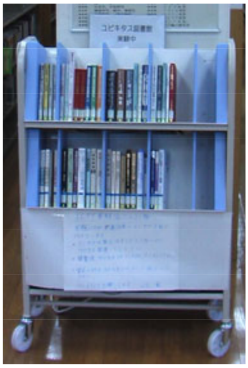


Fig. 4. IBS (Intelligent Bookshelf)

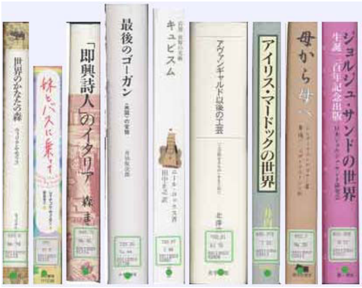


Fig. 5. Virtual Bookshelf

VBS

A virtual bookshelf (VBS) (Figure 5) is an image of bookshelf on a computer screen, with which users can see a list of images of books as if they were in a real bookshelf (See also [15,16]). The VBS not only displays the bookshelf image but also the users can rearrange the books on it just in the same way of real bookshelf [10]. Further, users can put comments and index data, or meta-data, on the shelved books. The VBS system provided by the libraries are basically owned by each patron and at the same time they are somewhat shared. VBS is another good source for usage data of books.

5 Knowledge Discovery from Book Usage Data

In this section we propose a method of extracting knowledge about relationships between books, books and patrons, and patrons. First, we describe how to measure the closeness between two books. Then we show how to make the relationship graph for it. We will define the "best friend" network based on the graph. Then we can apply various methods such as the collaborative filtering for recommendation of reading books to patrons.

5.1 Closeness between Books

The method proposed in this paper for measuring the closeness between books is based on the co-occurrence or co-appearance in the data from the IBSs, VBSs, circulation records, etc. In our MAS model, the corresponding agents meet in a same place as they appear in these data. For example, we suppose that a book A is placed on a bookshelf S. Through the wake-up mechanism, it means that the agent A enters a room organized by the corresponding bookshelf agent S. In this room every agent can recognize other agents staying in the same room via S. If the book B is already in the room S, the agents A and B will be able to

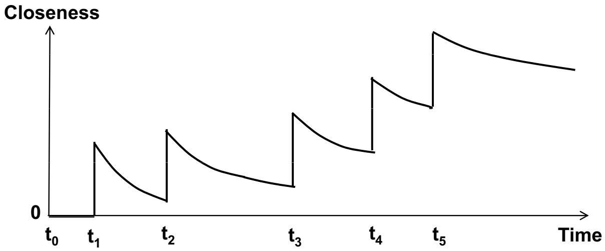


Fig. 6. Closeness between Books

meet each other; unless it happens that either of them does not want to know other agents. This is an event of meeting between books A and B.

Our idea of measuring the closeness between the agent A and the agent B, or between the book A and the book B, is based on the meeting events. Figure 6 illustrates the measurement process. The horizontal axis is the time and the vertical axis is the closeness value. The closeness 0 means that the books A and B have no closeness between them. We suppose the closeness value between A and B starts from 0. It can be different if they have something in common in advance, for example in such a case as they have the common DDC (Dewey Decimal Classification) number, for example.

In the figure, let t_0, t_1, ... be the meeting event times. A meeting event time is the time when the book A is put on the same bookshelf as the book B, the time when they are borrowed together by a patron at the circulation counter, or the time something like these happens. The closeness value rises up at the meeting event time. The rising value depends on the impact of the meeting event. If the event is the one that they are borrowed at the same time, the value might be higher than the one that they are stored in the same bookshelf. The closeness value decreases as time passes until the next meeting event. As is illustrated in Figure 6, the closeness value will increase in the long run if the two books meet occasionally.

One possible function that satisfies such requirements is:

$$f(t) = \begin{cases} 0 & \text{if } t = t_0 = 0 \\ f(t_i)e^{-\frac{1}{f(t_i)}(t-t_i)} & \text{if } t_i \leq t < t_{i+1} \\ f(t_i)e^{-\frac{1}{f(t_i)}(t_{i+1}-t_i)} + I & \text{if } t = t_{i+1} \end{cases}$$

where $i = 0, 1, 2, ...$ and I is the impact value of the meeting event.

5.2 Possible Analysis Methods Based on the Relationships between Books

Based on the closeness values between book agents we can define social networking relationship. For each book agent, say A, he/she can make a list of relating book agents according to the order of closeness values defined in the previous subsection. The book agent A can choose the "best friends" according to a criterion; say the top 10% closest agents, the agents that have the closeness value more than a pre-determined value, or something like these.

We can draw the best friend graph by making an arrow from A to B if B is a best friend of A. Then we can analyze the structure of book agents' relationships just like the link structure analysis of Web pages [7]. For example, a book having many arrows directing to it might be a good book probably in the sense that it can be used as something like a dictionary or a reference book, a textbook, something that provides with fundamental knowledge, or on the other hand it gives a good application tips. In order to know more clearly about the book, we would need to check other properties, for example, the DDCs of the book and of the books that direct to the book.

By analyzing the circulation data, probably in the same method as was described in the previous subsection, we can get the closeness relationship between books and patrons. In this case also we can extract the best friend graph. A best friend of a book is a patron that borrows the book more often than other patrons. A best friend of a patron is a book that the patron borrows very frequently and thus should be important for the patron. By combining this heterogeneous best friend relationship with the previous book-to-book best friend relationship, we can get another closeness relationship values between books and books, and patrons and patrons. These are somewhat similar to the well-known collaborative filtering [14] method and thus can be used for recommendation of reading books to patrons.

6 Concluding Remarks

In this paper we have investigated the potentiality of book agent system for library marketing. The participants of a library and the library network are multi agent system in nature. In our model the books, librarians, bookshelves, and even the library services are considered as agents. The bookshelf agents and service agents are organizers of rooms, or fields, where the participating book agents and other participants are able to get services from the organizing agents. One of the services is the meeting service where each book agent has the opportunity to meet other book agents.

A book agent will revise its closeness values to other agents as they meet in an event like sharing the same bookshelf, being borrowed at the same time by a patron, and in other co-occurrence events. The closeness value will decrease as the time passes without meeting, and it will go up as the book agents meet again in the future.

Based on the closeness value, we can make a best friend network of book agents. By analyzing the network we can find the key book agents who are very popular in the network, and other characteristic agents. Such knowledge extracted from the structurally characteristics will be useful in book recommendation and other library services.

The study presented in this paper is just a beginning of the research toward this direction of analysis methods that might be valuable to investigate multi agent based distributed knowledge extraction and integration mechanism. However by considering that the libraries are independent in nature they are managed according to their own policy, this approach should be an essential one for collective intelligence research for the library marketing. We would pursue our research in this direction.

References

1. AIM Global, http://www.aimglobal.org/
2. American Marketing Association (AMA), http://www.marketingpower.com/
3. Finkenzeller, K.: RFID Handbook, 2nd edn. John Wiley & Sons, Chichester (2003)
4. Fukuoka City Library, http://www.cinela.com/english/index.htm
5. Gwacheon City Library, http://www.gclib.net/
6. JADE (Java Agent DEvelopment Framework), http://jade.tilab.com/
7. Kosala, R., Blockeel, H.: Web mining research: A survey. In: SIGKDD Explorations, vol. 2, pp. 1–15 (2000)
8. Minami, T.: Library Services as Multi Agent System. In: Nguyen, N.T., Jo, G.-S., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2008. LNCS (LNAI), vol. 4953, pp. 222–231. Springer, Heidelberg (2008)
9. Minami, T.: A Library Marketing System for Decision Making. In: Nakamatsu, K., et al. (eds.) Proc. First KES Symposium on Intelligent Decision Technologies (KES-IDT 2009), New Advances in Intelligent Decision Technologies, pp. 97–106. Springer SCI 199, Heidelberg (2009)
10. Minami, T.: Towards Patron-Oriented Library Service Improvement with Data Analysis for Library Marketing. Information Journal 13(3) (2010)
11. National Library Board Singapore, http://www.nlb.gov.sg/
12. Ohtani, T., Minami, T.: The Word-of-Mouth Agent System for Finding Useful Web Documents. In: Proc. Asia Pacific Web Conference (APWeb 1998), pp. 369–374 (1998)
13. Ranganathan, S.R.: The Five Laws of Library Science. Asia Publishing House, Bombay (1963)
14. Resnick, P., Iacovou, N., et al.: GroupLens: an open architecture for collaborative filtering of netnews. In: Proc. 1994 ACM conference on Computer supported cooperative work, pp. 175–186 (1994)
15. Rauber, A., Bina, H.: Visualizing Electronic Document Repositories: Drawing Books and Papers in a Digital Library. In: Proc. 5. IFIP 2.6 Working Conference on Visual Database Systems (VDB5), pp. 95–114 (2000)
16. Sugimoto, S., et al.: Enhancing usability of network-based library information system - experimental studies of a user interface for OPAC and of a collaboration tool for library services. In: Proc. Digital Libraries 1995, pp. 115–122 (1995)

Adaptive Modelling of Social Decision Making by Agents Integrating Simulated Behaviour and Perception Chains

Alexei Sharpanskykh and Jan Treur

Vrije Universiteit Amsterdam, Department of Artificial Intelligence
De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
{sharp,treur}@few.vu.nl
http://www.few.vu.nl/~{sharp,treur}

Abstract. It is widely recognized that both cognitive and affective aspects play an important role in human decision making. In most recent approaches for computational modelling of affective agents emotions have a cognitive origin. In contrast to these approaches, the computational social decision making model proposed in this paper is grounded on neurological principles and theories, thus providing a deeper level of understanding of decision making. The proposed approach integrates existing neurological and cognitive theories of emotion in a decision model based on evaluation of simulated behaviour chains. The application of the proposed model is demonstrated in the context of an emergency scenario.

Keywords: Social decision making, affective, cognitive, simulated behavioural chains, neurological modelling.

1 Introduction

Traditionally, human decision making has been modelled as the problem of rational choice from a number of options using economic utility-based theories [17, 18]. In the last decades such approaches were criticized by many authors [18] for the lack of realism and limited applicability. In particular, it is imputed to the traditional decision making modelling methods that the role of human cognitive heuristics and biases, and affective states is totally neglected. Much evidence exist [1, 4, 5, 10] that affective states have a significant impact on a human's decision making. However, computational models to explain this evidence are rare. In this paper the focus is on the integration of cognitive and affective aspects in a computational social decision making model which is grounded in neurological theories.

In the areas of Artificial Intelligence and Cognitive Science a number of computational models of decision making with emotions have been developed [9, 21, 22], which use variants of the OCC model developed by Ortony, Clore and Collins [19] as a basis. The OCC model postulates that emotions are valenced reactions to events, agents, and objects, where valuations are based on similarities between achieved states and goal states; thus emotions in this model have a cognitive origin.

The model proposed in this paper exploits some of the principles underlying the OCC model but embeds them in a neurological context that includes theories that

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 284–295, 2010.

cover other aspects as well, thus providing a deeper and wider level of understanding of social decision making. More specifically, options in decision making involving sequences of actions are modelled using the neurological theory of simulated behaviour (and perception) chains proposed by Hesslow [13]. Moreover, the emergence of emotional states in these behavioural chains is modelled using emotion generation principles described by Damasio [4-8]. Evaluation of sensory consequences of actions in behavioural chains, also uses elements borrowed from the OCC model. Different types of emotions can be distinguished and their roles in the decision making clarified. The social aspect comes in by processes of emotion and intention contagion between different persons.

Evaluation and the emotions involved in it usually have a strong impact from the human's earlier experiences. In the proposed model for social decision making, this form of adaptivity to past experiences is also incorporated based on neurological principles. In such a way elements from neurological, affective and cognitive theories were integrated in the adaptive agent model proposed.

The paper is organised as follows. The general modelling principles on which the proposed computational model is based are described in Section 2. A detailed formalisation of the proposed model is provided in Section 3. In Section 4 it is demonstrated how the proposed social decision making model is applied in an emergency scenario. Finally, Section 5 concludes the paper.

2 Neurological Principles Adopted

Considering options and evaluating them is viewed as a central process in human decision making. In this paper options are not single actions but sequences of actions, as in planning. To model considering such sequences, from the neurological literature the *simulation hypothesis* proposed by Hesslow [13] was adopted. Based on this hypothesis, chains of behaviour can be simulated as follows: some situation elicits activation of s1 in the sensory cortex that leads to preparation for action r1. Then, associations are used such that r1 will generate s2, which is the most connected sensory consequence of the action for which r1 was generated. This sensory state serves as a stimulus for a new response, and so on. In such a way long chains of simulated responses and perceptions representing plans of action can be formed. These chains are simulated by an agent internally as follows:

> 'An anticipation mechanism will enable an organism to simulate the behavioural chain by performing covert responses and the perceptual activity elicited by the anticipation mechanism. Even if no overt movements and no sensory consequences occur, a large part of what goes on inside the organism will resemble the events arising during actual interaction with the environment.' ([13])

As reported in [13], behavioural experiments have demonstrated a number of striking similarities between simulated and actual behaviour.

Hesslow argues in [13] that the simulated sensory states elicit emotions, which can guide future behaviour either by reinforcing or punishing simulated actions. However, specific mechanisms for emotion elicitation are not provided. This gap can be filled by combining the simulation hypothesis with a second source of knowledge from the neurological area: Damasio's emotion generation principles based on *(as-if) body* *loops*, and the principle of *somatic marking* [1, 6]. These principles were adopted to model evaluation of options.

Damasio [4, 5, 7] argues that sensory or other representation states of a person often induce emotions felt within this person, according to a *body loop* described by the following causal chain:

sensory state → preparation for the induced bodily response → induced bodily response → sensing the bodily response → sensory representation of the bodily response → induced feeling

As a variation, an *as if body loop* uses a direct causal relation as a shortcut in the causal chain: preparation for the induced bodily response → sensory representation of the induced bodily response. The body loop (or 'as if body loop') is extended to a recursive body loop (or recursive 'as if body loop') by assuming that the preparation of the bodily response is also affected by the state of feeling the emotion as an additional causal relation: feeling → preparation for the bodily response. Thus, agent emotions are modelled based on reciprocal causation relations between emotion felt and body states. Following these emotion generation principles, an 'as if body' loop can be incorporated in a simulated behavioural chain as shown in Fig.1 (left). Note that based on the sensory states different types of emotions may be generated.

In the *OCC model* [19] a number of cognitive structures for different types of emotions are described. By evaluating sensory consequences of actions s1, s2, …, sn from Fig. 1 using cognitive structures from the OCC model, different types of emotions can be distinguished. More specifically, the emergence of hope and fear in agent decision making in an emergency scenario will be considered in Section 4.

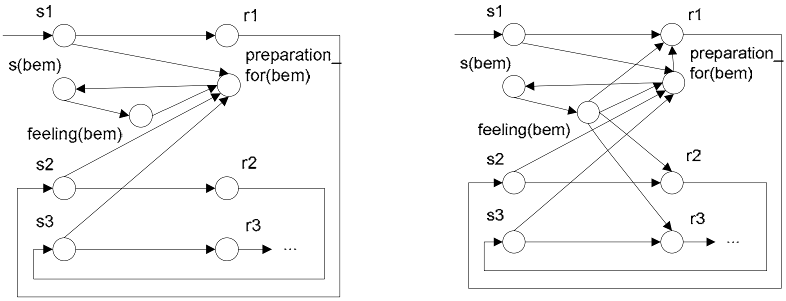


Fig. 1. Simulation of a behavioural chain extended with an 'as if body' loop with emotional state bem (left) and with emotional influences on preparation states (right)

Hesslow argues in [13] that emotions may reinforce or punish simulated actions, which may transfer to overt actions, or serve as discriminative stimuli. Again, specific mechanisms are not provided. To fill this gap the Damasio's *Somatic Marker Hypothesis* was adopted. This hypothesis provides a central role in decision making to emotions felt. Within a given context, each represented decision option induces (via an emotional response) a feeling which is used to mark the option. For example, a strongly negative somatic marker linked to a particular option occurs as a strongly negative feeling for that option. Similarly, a positive somatic marker occurs as a positive feeling for that option. Damasio describes the use of somatic markers in the following way:

> 'the somatic marker (..) forces attention on the negative outcome to which a given action may lead, and functions as an automated alarm signal which says: beware of danger ahead if you choose the option which leads to this outcome. The signal may lead you to reject, *immediately*, the negative course of action and thus make you choose among other alternatives. (…) When a positive somatic marker is juxtaposed instead, it becomes a beacon of incentive.' ([7], pp. 173-174)

To realise the somatic marker hypothesis in behavioural chains, emotional influences on the preparation state for an action are defined as shown in Fig. 1 (right). Through these connections emotions influence the agent's readiness to choose the option. From a neurological perspective, the impact of a sensory state to an action preparation state via the connection between them in a behavioural chain will depend on how the consequences of the action are felt emotionally.

As neurons involved in these states and in the associated 'as if body' loop will often be activated simultaneously, such a connection from the sensory state to the preparation to action state may be strengthened based on a general *Hebbian learning* principle ([2, 11, 12]) that was adopted as well. It describes how connections between neurons that are activated simultaneously are strengthened, similar to what has been proposed for the emergence of mirror neurons; e.g., [8, 16, 20].

Thus, by these processes an agent differentiates options to act based on the strength of the connection between the sensory state of an option and the corresponding preparation to an action state, influenced by its emotions. The option with the highest activation of preparation is chosen to be performed by the agent.

As also used as an inspiration in [14], in a social context, the idea of somatic marking can be combined with recent neurological findings on the *mirroring function* of certain neurons (e.g., [8, 16, 20]). Mirror neurons are neurons which, in the context of the neural circuits in which they are embedded, show both a function to prepare for certain actions or bodily changes and a function to mirror similar states of other persons. They are active not only when a person intends to perform a specific action or body change, but also when the person observes somebody else intending or performing this action or body change. This includes expressing emotions in body states, such as facial expressions. The mirroring function relates to decision making in two different ways. In the first place *mirroring of emotions* indicates how emotions felt in different individuals about a certain considered decision option mutually affect each other, and, assuming a context of somatic marking, in this way affect how by individuals decision options are valuated in relation to how they feel about them. A second way in which a mirroring function relates to decision making is by applying it to the *mirroring of intentions or action tendencies* of individuals (i.e., preparation states for an action) for the respective decision options. This may work when by verbal and/or nonverbal behaviour individuals show in how far they tend to choose for a certain option. In the computational model introduced below in Section 3 both of these (emotion and preparation) mirroring effects are incorporated.

3 A Computational Model for Decision Making with Emotions

First, in Section 3.1 a modelling language is described used for formalisation of the decision making model proposed. Then, the formal model is provided in Section 3.2.

3.1 Modeling Language

To specify dynamic properties of a system, the order-sorted predicate logic-based language called LEADSTO is used [3]. Dynamics in LEADSTO is represented as evolution of states over time. A state is characterized by a set of properties that do or do not hold at a certain point in time. To specify state properties for system components, ontologies are used which are defined by a number of sorts, sorted constants, variables, functions and predicates (i.e., a signature). For every system component A a number of ontologies can be distinguished: the ontologies IntOnt(A), InOnt(A), OutOnt(A), and ExtOnt(A) are used to express respectively internal, input, output and external state properties of the component A. Input ontologies contain elements for describing perceptions of an agent from the external world, whereas output ontologies describe actions and communications of agents. For a given ontology Ont, the propositional language signature consisting of all state ground atoms based on Ont is denoted by APROP(Ont). State properties are specified based on such ontology by propositions that can be made (using conjunction, negation, disjunction, implication) from the ground atoms. Then, a *state* S is an indication of which atomic state properties are true and which are false: S: APROP(Ont) → {true, false}.

LEADSTO enables modeling of direct temporal dependencies between two state properties in successive states, also called *dynamic properties*. The format is defined as follows. Let α and β be state properties of the form 'conjunction of atoms or negations of atoms', and e, f, g, h non-negative real numbers. In the LEADSTO language the notation $\alpha \twoheadrightarrow_{e, f, g, h} \beta$ means: if state property α holds for a certain time interval with duration g, then after some delay (between e and f) state property β will hold for a certain time interval of length h. When e = f = 0 and g = h = 1, called standard time parameters, we shall write $\alpha \twoheadrightarrow \beta$. To indicate the type of a state property in a LEADSTO property we shall use prefixes input(c), internal(c) and output(c), where c is the name of a component. Consider an example dynamic property:

input(A)|observation_result(fire) ↠ output(A)| performed(runs_away_from_fire)

Informally, this example expresses that if agent A observes fire during some time unit, then after that A will run away from the fire during the following time unit.

3.2 The Computational Model

Depending on a situational context an agent determines a set of applicable options to satisfy a goal at hand. In the model proposed here the applicable options are generated via connections from activated sensory states reflecting this situational context to preparation states for the relevant actions related to an option. The issue of how precisely the strengths of these connections from a particular context to relevant action preparations have come into existence is out of scope of this paper; some related research can be found in [18]. An option is represented by a (partially) ordered sequence of actions (i.e., a plan) to satisfy the agent's goals. Computationally, alternative options considered by an agent are being generated and evaluated in parallel. The evaluation of options is based on the simulation of a behavioural chain as described in Section 2 (see Fig. 2). The social context in which decision making is performed is represented by a group of agents interacting (verbally, nonverbally) on the relevant options. It is assumed that the preparation states of an agent for the actions constituting options and for emotional responses for the options are body states that can be observed with a certain intensity or strength by other agents from the group. The *contagion strength* of the interaction from agent A_2 to agent A_2 for a preparation state p is defined as follows:

$$\gamma_{pA_2A_1} = \varepsilon_{pA_2} \cdot \alpha_{pA_2A_1} \cdot \delta_{pA_1} \tag{1}$$

Here ε_{pA_2} is the personal characteristic expressiveness of the sender (agent A_2) for p, δ_{pA_1} is the personal characteristic openness of the receiver (agent A_1) for p, and $\alpha_{pA_2A_1}$ is the interaction characteristic channel strength for p from sender A_2 to receiver A_1.

By aggregating such input, an agent A_i perceives the group's joint attitude towards each option, which comprises the following dynamic properties. Note that for the sake of simplicity no intermediate states for this process have been included, such as effector states, body states proper, or sensor states; the process from internal states to external expression, transfer and receipt is characterised at once by using parameters such as ε_{pA_j}, $\alpha_{pA_jA_i}$ and δ_{pA_i} introduced above.

(a) the aggregated group preparation to (i.e., the externally observable intention to perform) each action p constituting the option. This is expressed by the following dynamic property:

$$\wedge_{j\neq i}\ \text{internal}(A_j)\ |\ \text{preparation\_for}(p, V_j) \twoheadrightarrow \text{internal}(A_i)\ |\ \text{srs}(G(p), \Sigma_{j\neq i}\ \gamma_{pA_jA_i} V_j / \Sigma_{j\neq i}\ \gamma_{pA_jA_i}\varepsilon_{pA_j}) \tag{2}$$

(b) the aggregated group preparation to an emotional response (body state) be for each option. In general an option may induce different types of emotions (e.g., fear, hope, joy). For each of them a separate preparation state is introduced. Formally:

$$\wedge_{j\neq i}\ \text{internal}(A_j)\ |\text{preparation\_for}(be, V_j) \twoheadrightarrow \text{internal}(A_i)\ |\ \text{srs}(G(be), \Sigma_{j\neq l}\ \gamma_{beA_jA_i} V_j / \Sigma_{j\neq i}\ \gamma_{beA_jA_i}\varepsilon_{beA_j}) \tag{3}$$

Furthermore, the emotional responses induced by options affect preparation states for the actions from options via 'as-if body' loops as described in Section 2. Thus, the preparation state for the first action from an option is affected by the sensory representations of the option, of the perceived group preparation for the action and of the emotion felt towards the option. Formally:

$$\begin{aligned} &\text{srs}(O, V1)\ \&\ \text{srs}(be, V2)\ \&\ \text{srs}(G(a1), V3)\ \&\ \text{preparation\_for}(a1, V4) \\ &\twoheadrightarrow \text{preparation\_for}(a1, V4 + \gamma(h(V1, V2, V3) - V4)\Delta t), \end{aligned} \tag{4}$$

where O is an option, be is an emotional response state, $Ga1$ is the aggregated group preparation to action $a1$, $h(V1, V2, V3)$ is a combination function:

$$h(V1, V2, V3) = \beta\ (1\text{-}(1\text{-}\ V1)(1\text{-}\ V2)(1\text{-}\ V3)) + (1\text{-}\beta)\ V1\ V2\ V3.$$

The simulated perception of the effect of an action from a plan in a simulated behavioural chain is modelled by the following property:

$$\text{preparation\_for}(\text{a}, \text{V}) \twoheadrightarrow \text{srs}(\text{effect}(\text{a}), \text{V}) \tag{5}$$

The confidence that an action will result in a particular effect is specified as the strength of the link between the preparation for the action state and the sensory representation of the corresponding effect state.

The preparation state for each following action a from the behavioural chain is specified by:

$$\begin{aligned}&\text{srs}(\text{effect}(\text{a}), \text{V1})\ \&\ \text{srs}(\text{be}, \text{V2})\ \&\ \text{srs}(\text{G}(\text{a}), \text{V3})\ \&\ \text{preparation\_for}(\text{a}, \text{V4})\\ &\twoheadrightarrow \text{preparation\_for}(\text{a}, \text{V4} + \gamma(\text{h}(\text{V1}, \text{V2}, \text{V3}) - \text{V4})\ \Delta t),\end{aligned} \tag{6}$$

An emotional response is generated based on an evaluation of the effects of each action of the option. In such an evaluation the effect state for each action is compared to a goal state(s) of the agent.

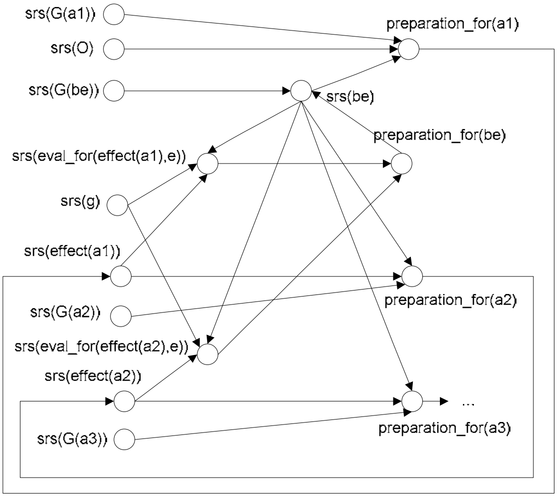


Fig. 2. A graphical representation of the emotional decision making model in the social context

Note that for different types of emotions different aspects of a goal state or different types of goals may be used. In [19] a number of cognitive structures eliciting particular types of emotions are described. As a simulated behavioural chain is a kind of a behavioural projection, cognitive structures of prospect-based emotions (e.g., fear, hope, satisfaction, disappointment) from [19] are particularly relevant for the evaluation process. Such structures can be represented formally as evaluation properties. Examples of such properties for the emotions fear and hope are provided in the following section 4. As indicated in [19], the intensity of prospect-based emotions depends on the likelihood (confidence) that a prospect state will occur. Thus, the strength of the link between the preparation state for an action and the sensory representation of its effect state is taken into account as a factor in the evaluation property. The generic evaluation property of the effect of the action a compared with the goal state g is specified formally as:

$$\begin{array}{l}\text{srs(g, V1) \& srs(effect(a), V2) \& srs(be, V3) \& connection\_between\_strength(preparation\_for(a),}\\ \text{srs(effect(a)), V4) \& srs(eval\_for(effect(a), be), V5)}\\ \twoheadrightarrow \text{srs(eval\_for(effect(a), be), V5 + }\gamma\text{(h(V4*f(g, effect(a)), V3) – V5) }\Delta\text{t),}\end{array} \quad (7)$$

where f(g, effect(a)) is an evaluation function depending on the cognitive structure used for the evaluation.

The evaluation of the effects of the actions for a particular emotional response to an option together with the aggregated group preparation to the emotional response determine the intensity of the emotional response:

$$\begin{array}{l}\wedge_{i=1..n}\ \text{srs(eval\_for(effect(}a_i\text{), be), }V_i\text{) \& srs(G(be), V3)}\\ \twoheadrightarrow \text{preparation\_for(be, f(}V_1\text{, ..., }V_n\text{)),}\end{array} \quad (8)$$

where bem is a particular type of the emotional response.

The agent perceives its own emotional response and creates the sensory representation state for it:

$$\text{preparation\_for(be, V)} \twoheadrightarrow \text{srs(be, V)} \quad (9)$$

The Hebbian learning principle for links connecting the sensory representation of options, and effects of the actions from these options, with preparation states for subsequent actions in the simulation of a behavioural chain is formalised as follows (cf. [11]):

$$\begin{array}{l}\text{connection\_between\_strength(O, preparation\_for(a1), V1) \& srs(O, V2) \& preparation\_for(a1, V3)}\\ \twoheadrightarrow \text{connection\_between\_strength(O, preparation\_for(a1), V1 + (}\eta\text{ V2 V3 (1 – V1) - }\xi\text{V1)}\Delta\text{t),}\end{array} \quad (10)$$

where η is a learning rate and ξ is an extinction rate.

$$\begin{array}{l}\text{connection\_between\_strength(srs(effect(}a_i\text{)), preparation\_for(}a_{i+1}\text{), V1) \& srs(effect(}a_i\text{), V2) \&}\\ \text{preparation\_for(}a_{i+1}\text{, V3)}\\ \twoheadrightarrow \text{connection\_between\_strength(srs(effect(}a_i\text{)),}\\ \qquad\qquad \text{preparation\_for(}a_{i+1}\text{), V1 + (}\eta\text{ V2 V3 (1 – V1) - }\xi\text{ V1)}\Delta\text{t)}\end{array} \quad (11)$$

4 Decision Making in Emergency Situations: A Case Study

In this section it is demonstrated how decision making in an evacuation scenario can be modelled using the proposed approach (see Fig. 3). In this scenario a group of agents considers different options (paths) to move outside of a burning building. Each option is generated based on the agent's beliefs about the accessibility of locations in the building. Each option is represented by a sequence of locations with an exit as the last location, specified by follows_after(move_from_to(p1, p2), move_from_to(p2, p3)). The strength of a link between a preparation for a movement action and a sensory representation of the effect of the action is used to represent confidence values of the agent's beliefs about the accessibility of locations. For example, if the agent's confidence of the belief that location p1 is accessible from location p2 is ω, then the strength of the link between the states preparation_for(move_from_to(p2, p1)) and srs(is_at_location(p1)) is also ω.

Considered options (i.e., activation of the preparations for the actions involved) evoke two types of emotions - fear and hope, which are often considered in the emergency context [17]. According to [19], the intensity of fear induced by an event depends on the degree to which the event is undesirable and on the likelihood of the event. The intensity of hope induced by an event depends on the degree to which the event is desirable and on the likelihood of the event. Thus, both emotions are generated based on the evaluation of a distance between the effect states for the actions from an option and the agent's goal states. In this example each agent in the group has two goal states '*be outside*' and '*be safe*'. The evaluation functions for both emotions include two aspects: (1) how far is the agent's location from the nearest reachable exit; (2) how dangerous is the agent's location (i.e., the amount of smoke and fire). Formally these two aspects are combined in the evaluation function from (7) using the formula

$$\omega V1 + (1-\omega)/(1+\lambda e^{-\varphi V2}), \qquad (12)$$

where V1 is the degree of danger of the location, V2 is the distance in number of actions that need to be executed to reach the nearest accessible exit, λ and φ are parameters of the threshold function, ω is a weight.

The goal value in (7) is obtained by setting V1=0 and V2=0 in (12): $(1-\omega)/(1+\lambda)$.

According to the two emotions are considered in the example, (7) is refined into two specialized evaluation properties – one for fear and one for hope:

srs(g, V1) & srs(effect(a), V2) & srs(bfear, V3) &
connection_between_strength(preparation_for(a), srs(effect(a)), V4) &
srs(eval_for(effect(a), bfear), V5)
↠ srs(eval_for(effect(a1), bfear), V5 + γ(h(V4*f(g, effect(a)), V3) – V5) Δt),
where f(g, effect(a)) = |V1-V6|, and V6 is calculated by (12) for state effect(a).

srs(g, V1) & srs(effect(a), V2) & srs(bhope, V3) &
connection_between_strength(preparation_for(a), srs(effect(a)), V4) &
srs(eval_for(effect(a), bhope), V5)
↠ srs(eval_for(effect(a), bhope), V5 + γ(h(V4* f(g, effect(a)), V3) – V5) Δt),
where f(g, effect(a))=1-|V1-V6|, and V6 is calculated by (12) for state effect(a).

Also specialized versions of other generic properties 3-9 are defined by replacing the generic state bem in them by specific emotional response states bfear and bhope.

Using the developed model a number of simulations have been performed. In particular, social decision making in a group of 6 agents with 3 agents of type 1 (see Table 1) and 3 agents of type 2 has been modelled.

Table 1. Two types of agents used in the simulation

Agent type	ε for all states to all agents	δ for all states from all agents	α	β	γ	η	ξ
Type 1: Extravert with a positive thinking attitude	*0.8*	*0.8*	*1*	*0.7*	*0.7*	*0.6*	*0.1*
Type 2: Introvert with a negative thinking attitude	*0.4*	*0.4*	*1*	*0.3*	*0.7*	*0.6*	*0.1*

The agents in the group are making choice among two options to move out of the building. It is assumed that all agents have the same beliefs about the availability of locations in the building and the degree of danger of each location. Path 1 considered by each agent is short, but also is more dangerous; whereas the alternative path 2 is much longer, but is considered to be more safe. The dynamics of spread of fire and smoke is taken into account in the internal processing of the agents.

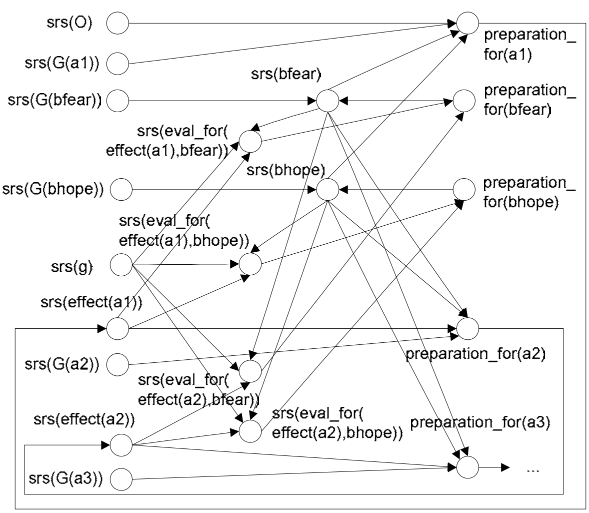


Fig. 3. A graphical representation of the emotional decision making model for an emergency scenario

In Fig. 4 the change of the strength of the links over time between the sensory representations of both options and the corresponding preparation states to start the option execution is depicted. As one can see from the graphs all agents are more inclined to choose the second option. Furthermore, as the group reaches the consensus, the difference in the strength of the link for option 2 for both types of agents decreases over time. The agents of type 2 are consistently lower in their estimation of the options than the agents of type 1. This can be explained by their personal characteristics from Table 1.

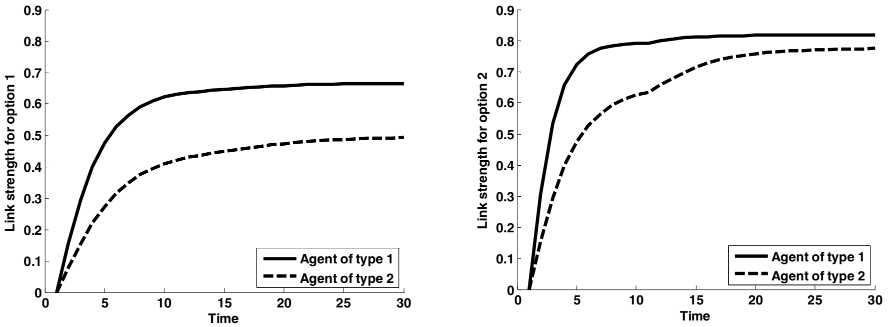


Fig. 4. The change of the strength of the link over time between the sensory representations of option1 (left) and option 2 (right) and their preparation states to start the option execution

In Fig. 5 the change of fear and hope over time for option 1 for both types of agents is depicted.

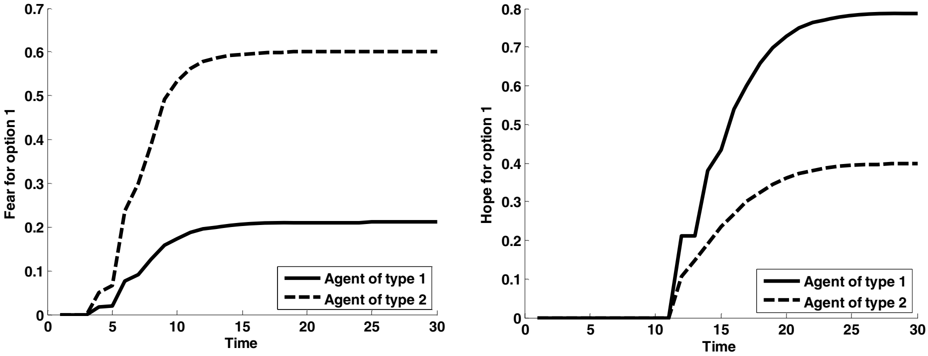


Fig. 5. The change of fear (left) and hope (right) for option 1 for both types of agents

As can be seen from the graphs, the agents of type 1 have much more hope and much less fear than the agents of type 2, even though option 1 is not the most promising option. Such dynamics is largely accounted for by the settings of the individual parameters of agents from Table 1.

5 Conclusion

In this paper a computational approach for modelling adaptive decision making of individuals in a group is proposed. The approach is based on a number of neurological theories and principles supplementing each other in a consistent manner. By taking a neurological perspective and incorporating cognitive and affective elements in one integrated model, a more realistic and deeper and wider understanding of the internal processing underlying human decision making in social situations has been achieved. This gives a richer type of model than models purely at the cognitive level (and ignoring affective aspects), or diffusion or contagion models at the social level abstracting from internal processing, for example, as addressed in [14].

Although the neurological theories and principles used as a basis for the model proposed have been validated to a certain extent, in the future a large-scale validation study for the model in the frames of the EU-project SOCIONICAL is planned (http://www.socionical.eu).

Previously, a number of computational models for human decision making including different types of cognitive biases and heuristics have been developed, also in LEADSTO language [15]. Such models can be readily integrated with the model proposed in this paper. More specifically, models of cognitive biases can be used for the generation of effect of action states and for the evaluation of these states for the generation of emotions.

References

1. Bechara, A., Damasio, A.: The Somatic Marker Hypothesis: a neural theory of economic decision. Games and Economic Behavior 52, 336–372 (2004)
2. Bi, G.Q., Poo, M.M.: Synaptic Modifications by Correlated Activity: Hebb's Postulate Revisited. Ann. Rev. Neurosci. 24, 139–166 (2001)

3. Bosse, T., Jonker, C.M., van der Meij, L., Treur, J.: A Language and Environment for Analysis of Dynamics by Simulation. Int. J. of AI Tools 16, 435–464 (2007)
4. Damasio, A.: Looking for Spinoza. Vintage books, London (2004)
5. Damasio, A.: The Feeling of What Happens. In: Body and Emotion in the Making of Consciousness, Harcourt Brace, New York (1999)
6. Damasio, A.: The Somatic Marker Hypothesis and the Possible Functions of the Prefrontal Cortex. Philosophical Transactions of the Royal Society: Biological Sciences 351, 1413–1420 (1996)
7. Damasio, A.: Descartes' Error: Emotion, Reason and the Human Brain, Papermac, London (1994)
8. Damasio, A., Meyer, K.: Behind the looking-glass. Nature 454, 167–168 (2008)
9. de Byl, P., Lukose, D.: An Affective Decision Making Agent Architecture Using Emotion Appraisals. In: Proceedings of PRICAI 2002, pp. 581–590 (2002)
10. Eich, E., Kihlstrom, J.F., Bower, G.H., Forgas, J.P., Niedenthal, P.M.: Cognition and Emotion. Oxford University Press, New York (2000)
11. Gerstner, W., Kistler, W.M.: Mathematical formulations of Hebbian learning. Biol. Cybern. 87, 404–415 (2002)
12. Hebb, D.O.: The Organisation of Behavior. Wiley, New York (1949)
13. Hesslow, G.: Conscious thought as simulation of behaviour and perception. Trends in Cog. Sci. 6, 242–247 (2002)
14. Hoogendoorn, M., Treur, J., van der Wal, C.N., van Wissen, A.: Modelling the Emergence of Group Decisions Based on Mirroring and Somatic Marking. Technical Report, Vrije Universiteit Amsterdam (2010)
15. Heuvelink, A., Klein, M. C. A., Treur, J.: An Agent Memory Model Enabling Rational and Biased Reasoning. In: IAT 2008, 193–199 (2008)
16. Iacoboni, M.: Understanding others: imitation, language, empathy. In: Hurley, S., Chater, N. (eds.) Perspectives on imitation: from cognitive neuroscience to social science, vol. 1, pp. 77–100. MIT Press, Cambridge (2005)
17. Janis, I., Mann, L.: Decision making: A psychological analysis of conflict, choice, and commitment. The Free Press, New York (1977)
18. Kahneman, D., Slovic, P., Tversky, A.: Judgement under uncertainty - Heuristics and biases. Cambridge University Press, Cambridge (1981)
19. Ortony, A., Clore, G.L., Collins, A.: The Cognitive Structure of Emotions. Cambridge University Press, Cambridge (1988)
20. Rizzolatti, G., Craighero, L.: The mirror-neuron system. Annu. Rev. Neurosci. 27, 69–92 (2004)
21. Santos, R., Marreiros, G., Ramos, C., Neves, J., Bulas-Cruz, J.: Multi-agent Approach for Ubiquitous Group Decision Support Involving Emotions. In: Ma, J., Jin, H., Yang, L.T., Tsai, J.J.-P. (eds.) UIC 2006. LNCS, vol. 4159, pp. 1174–1185. Springer, Heidelberg (2006)
22. Steunebrink, B.R., Dastani, M., Meyer, J.-J.C.: A logic of emotions for intelligent agents. In: Proceedings of the 22nd Conference on Artificial Intelligence (AAAI 2007), AAAI Press, Menlo Park (2007)

Modeling Agent-Based Collaborative Process

Moamin Ahmed, Mohd Sharifuddin Ahmad, and Mohd Zaliman M. Yusoff

Universiti Tenaga Nasional, Jalan IKRAM-UNITEN,
43009 Kajang, Selangor, Malaysia
moamin84@yahoo.com, {sharif,zaliman}@uniten.edu.my

Abstract. In this paper, we develop a model for a collaborative process using software agents to resolve issues and problems with the process. The structure, workflow, and entities of the process are analyzed to develop a model in which software agents take over the communicating tasks and motivating their human counterparts by reminding, alerting, and evaluating them to comply with scheduled deadlines. The model is used as a blueprint to develop a multi-agent system for the collaborative process.

Keywords: Intelligent Software Agents, Multiagent Systems, Workflow, Collaboration.

1 Introduction

In most organizations, the effective use of resources that spans across geographical space presents a major challenge. While knowledge workers perform tasks in different locations, they need to work collaboratively to achieve common organizational goals. The tasks they perform are often non-trivial and require specialized expertise and resources that are distributed across the organization. The ability to utilize this distributed knowledge and resources in achieving a common goal requires smooth and effective coordination.

The diversity of tasks imposed on humans poses a major challenge in keeping and monitoring the time to implement scheduled tasks. One way of overcoming this problem is to use a scheduler or a time management system which keeps track of deadlines and provides reminders for time-critical tasks. However, such systems do not always provide the needed assistance to perform mundane follow-up tasks in a collaborative process. A more motivational approach is required to ensure that everyone in a team plays his/her role effectively and diligently.

In this paper, we demonstrate the modeling and application of software agents to assist humans in complying with the schedule of the collaborative work of Examination Paper Preparation and Moderation Process (EPMP) in our academic faculty. We identify the structure, workflow, and entities of the EPMP and develop a model for each entity. We define each entity, analyze, and formalize the underlying theories that are applicable for the entity. While doing so, we concoct possible solutions to human- and system-related problems which we discovered in the EPMP.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 296–305, 2010.

2 Issues and Problems in the EPMP

The EPMP is the standard process of our faculty for examination paper preparation and moderation. The process starts when an Examination Committee (EC) sends out an instruction to start prepare examination papers. A Lecturer then prepares the examination paper, together with the solutions and the marking scheme (Set A). Upon completion, he/she then submits the set to be checked by an appointed Moderator.

The Moderator checks the set and returns them with a moderation report (Set B) to the Lecturer. If there are no corrections, the Lecturer submits the set to the Examination Committee for further actions. Otherwise, the Lecturer needs to correct the paper and resubmit the corrected paper to the Moderator for further inspection. If corrections have been made, the Moderator returns the paper to the Lecturer. Finally, the Lecturer submits the paper to the Committee for further processing. The process continues over a period of four weeks in two 2-week preparation-moderation-correction cycles. Figure 1 shows the process flow of the EPMP.

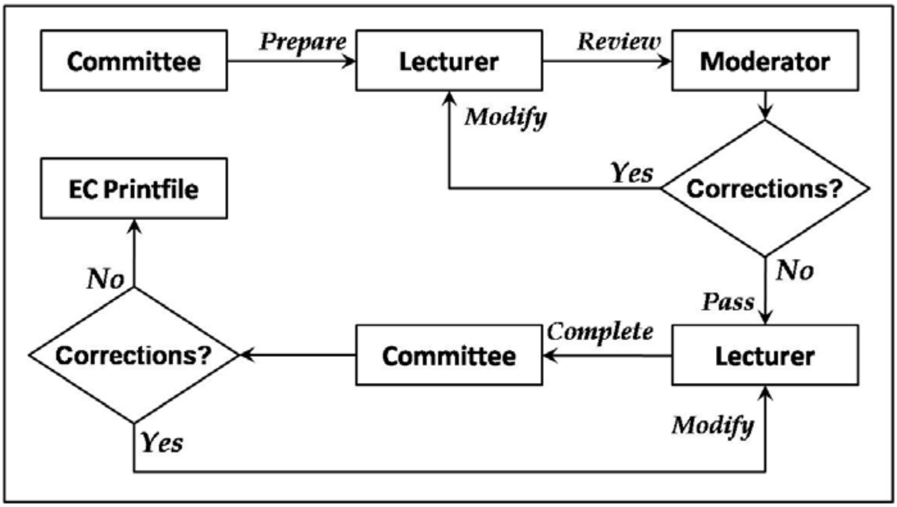


Fig. 1. The EPMP Process Flow

Lack of enforcement and the diverse tasks of lecturers and moderators cause the EPMP to suffer from delays in actions by the academicians. Our observation and investigation reveal that the lecturers and moderators do not observe the deadlines of the two 2-week moderation cycles. Some lecturers wait until the last few days of the second moderation cycle to submit their examination papers, which leaves insufficient time for the moderators to qualitatively scrutinize the papers.

Due to the manual nature of the process, (e.g. lecturers personally deliver the documents to the moderators), there is no mechanism that records the adherence to deadlines and tracks the activities of defaulters. The absence of such mechanism makes it impossible for the Committee to identify the perpetrators who fail to submit their documents by the deadlines.

Notwithstanding, our observation reveals that all examination papers reach the Committee by the last few days of the second moderation cycle albeit with numerous errors. In some cases, the percentage of papers returned for corrections (in content and format) is as high as 90%. Such high percentage has seriously burdened the Committee with keeping and maintaining the quality of examination papers. We can distinguish two problems in EPMP:

1. Human-related problems: The lack of coordination between the Committee, moderators, and lecturers causes delays in document submission. There is also a lack of enforcement on the deadlines especially in the first moderation cycle.

   Due to the diversity of tasks imposed on Lecturers and Moderators, their need to remember and perform those other tasks is compelling. The EPMP's tasks and schedule add significant additional stress on their cognitive load, because they have to depend on themselves to follow through the process until the end.

2. System-related problems: The main collaboration problem from the system's perspective is due to the manual nature of the system and its inability to provide:

   - consistent schedule for document submission, which could take two days depending on the circumstances affecting humans.
   - tracking of documents, in which case, the Committee could not know the current location of the documents without querying the Lecturers or Moderators.
   - reminders, which are very important in urging humans to speed up the process.
   - recording of submission date, which could assist the Committee in identifying the defaulters.
   - evaluation of individual performance, to identify Lecturers or Moderators who are diligently fulfilling their obligations and those who are not.
   - enforcement on the two-cycle moderation schedule, which is not strictly implemented. It seems that, effectively, only one moderation cycle is implemented since the process takes the whole four weeks.

To resolve some of these problems, we resort to the use of software agents to take over the communication and document submission tasks by the agents and the reminding and alerting tasks directed to humans.

3 Related Work

The development of our model is inspired by the work of many researchers in agent-based, collaboration, and workflow systems [2, 4, 5, 11, 12, 14, 15]. Wooldridge [15] and Ferber [5] present a comprehensive introduction to the field of agent-based systems, encompassing the theory and practice of multiagent systems. Labrou and Finin [11] propose an agent communication language for multiagent coordination which has been developed further by the Foundation for Intelligent Physical Agents (FIPA) [7, 8, 9].

Other researchers attempt to resolve issues in collaboration involving individuals and groups [2, 14], which closely relate to research in workflow systems [6, 12, 13]. For example, Chen et al. [2] present DIAMS, a system of distributed, collaborative agents to help users access, manage, share and exchange information. Steinfield et al. [14] develop TeamSCOPE, which is a collaborative system specifically designed to address problems faced by distributed teams. It is an integrative framework that focuses on facilitation of group members' awareness of group activities, communications, and resources.

Many business processes use workflow systems to exploit their known benefits such as automation, coordination, and collaboration between entities. A workflow describes the order of a set of tasks performed by various software and human to complete a given procedure within an organization [1]. Repetitive workflows are often automated, particularly in organizations that handle high volumes of forms or documents according to fixed procedures. Savarimuthu et al. [13] and Fluerke et al. [6] describe the advantages of their agent-based framework JBees, such as distribution, flexibility, and ability to dynamically incorporate a new process model. Researches have also been made on the monitoring and controlling of workflow [12].

4 Development of the EPMP Model

To model the EPMP, we use Win-Prolog and its extended module Chimera, which has the ability to handle multi-agent systems [3]. Chimera provides the module to implement peer-to-peer communication via the use of TCP/IP. Each agent is identified by a port number and an IP address. Agents send and receive messages through such configurations.

To develop the model, we first identify the entities that are involved in the domain's dynamics. The entities of the EPMP which form the underlying theory of the model include humans, agents, environment, agent actions, schedule, documents used in the process, and a set of expressions and terminologies used by humans and agents to advance the workflow. We first define each of these entities as follows:

Definition 4.1: An agent, α, is defined as an entity which performs a set of actions, a_i, based on the states of the environment, e_j, to achieve its goal.

If A is a set of actions, i.e. $a_i \in A$, and E is the environment, i.e. $e_j \in E$, then an agent maps to a function of its actions and environment, i.e.,

$$\alpha \rightarrow f(a_i, e_j) \tag{1}$$

where $i, j \geq 1$.

Definition 4.2: An environment, E, is a set of states, e_j, that influences and is influenced by the agent actions, i.e. if e_j is a state of an environment E, then the environment is the union of all the states.

$$E = \bigcup_{j=1}^{N} e_j \tag{2}$$

where j = 1... N.

Definition 4.3: An agent action, a_i, is a discrete operation performed by the agent which contributes to the achievement of its goal. If A is a set of agent actions, then,

$$A = \bigcup_{i=1}^{M} a_i \tag{3}$$

where i = 1... M.

We distinguish two types of actions: *tasks* and *message exchanges*. A task, $t_{m\alpha}$, is an action that an agent performs as a consequence of a message it receives from other agents and a message exchange, $x_{n\alpha}$, is a communicative action which the agent performs to advance the workflow, where m is the task number, n is the message exchange number and α refers to agents. Therefore,

$$A = \bigcup_{m,n=1}^{M} t_{m\alpha}, x_{n\alpha} \tag{4}$$

From Definition 4.1, $\alpha \rightarrow f(a_i, e_j)$, but $a_i = \{t_{m\alpha}, x_{n\alpha}\}$, thus we redefine an agent as follows:

$$\alpha \rightarrow f(\bigcup_{m,n=1}^{M} t_{m\alpha}, x_{n\alpha}, \bigcup_{j=1}^{N} e_j) \tag{5}$$

where j, m, n = 1... M, N.

Definition 4.4: A schedule of a process is a time slot defined by the beginning and the end of the process. In this model, we define four schedules: the main EPMP schedule, the examination paper preparation, and the first and the second moderation cycles. For example, if δ_i is the beginning of a moderation cycle and δ_j is the end, then the schedule of a moderation cycle,

$$\sigma = \delta_j - \delta_i \tag{6}$$

where δ_j is the deadline, $\delta_i < \delta_j$, δ_i and δ_j has a date structure (dd-mm-yy) and σ is the number of days between two dates.

Definition 4.5: A document, $d_i \in D$, $i \geq 1$, and D is a set of documents used in a process, is an object which agents exchange on behalf of their human counterparts to advance the workflow.

In our model, the internal states of a document such as updates made by humans are opaque to the agents. However, agents monitor the spatial and temporal states of the document in folders (i.e., exist or absence and when) to detect whether humans have indeed taken some actions on the document (e.g. submitted before the deadline).

Definition 4.6: An ontology term, $o_t \in O$, $t \geq 1$, and O is a set on ontology terms, is a keyword used by agents to communicate the semantics of their beliefs, desires, and intentions and other domain parameters.

In this model, we distinguish several types of ontology terms: *tasks*, *performatives*, and *object names*. We have defined a task, $t_{m\alpha}$, in Definition 4.3 above. A performative, p_i, is a keyword that defines a speech act, which an agent may use e.g. Prepare, while an object name, n_j, is an atom that identifies an object, which agents exchange and to be worked upon by humans, e.g. a moderation form. Therefore,

$$o_t = \{t_{m\alpha}, p_i, n_j\} \in O, \tag{7}$$

where i, j, m, t ≥ 1.

Definition 4.7: An ontology, Ω, is a structured construct of rules that defines the semantics of an ontology term and has the prolog clausal form,

$$\Omega \leftarrow \mathtt{T}_1, \ldots, \mathtt{T}_u. \tag{8}$$

where $u \geq 1$, Ω is the goal or head of the clause, and the conjuncts $\mathtt{T}_1, \ldots, \mathtt{T}_u$ are the sub-goals making up the body of the clause. Instantiations of these sub-goals are represented by o_t.

5 Modeling Actions

As defined in Definition 4.3, a task is an action that an agent performs as a consequence of a message it receives from other agents. In modeling the EPMP, we analyze what tasks are required based on the model's logical and architectural requirements. For example, our model uses the peer-to-peer architecture to simulate the manual system. Consequently, some tasks are specifically designed to enable agents in performing the peer-to-peer communication.

5.1 Tasks, $t_{m\alpha}$

Upon analysis of the domain, we concoct the following agent tasks, $t_{m\alpha}$:

a) Connect, $t_{1\alpha}$: When an agent needs to send a message or when it senses a message, it makes a connection with a remote agent (peer-to-peer).

b) Open a document, $t_{2\alpha}$: An agent opens a (new) document for the human Lecturer, Moderator, or Committee. The document could be a moderation form, committee form or examination paper. This task is required to automatically open the documents to encourage humans to update the documents.

c) Disconnect, $t_{3\alpha}$: When an agent has sent or received a message, this task disconnects the remote agent.

d) Display a message window on the screen, $t_{4\alpha}$: When an agent receives a message, it displays the message window on the screen. This message helps humans to know his/her task, the number of days left to submit or any other information about the state of examination paper.

e) Log action, $t_{5\alpha}$: An agent logs a message and its date with the details of the port number and IP address in a log file. The Head of Department is able to open the log file at anytime.

f) Remind action, $t_{6\alpha}$: The Committee agent sends a remind message to an agent that currently holds the examination documents.

g) Advertise, $t_{7\alpha}$: When a deadline is breached and a human Lecturer or Moderator who holds the examination documents has not completed the required task, the corresponding agent informs this state of affair to all other agents that its human counterpart is delaying the process. This action motivates the human to submit before the deadline.

h) Track documents, $t_{8\alpha}$: The Committee agent tracks the documents to alert the agent which holds them. An agent which submits the documents writes a message in a track file. This track file belongs to the Committee agent and it has all the information about the documents' spatial and temporal states.

i) Record action, $t_{9\alpha}$: When an agent performs an action, it records the action in a subprogram to facilitate the monitoring of paper preparation and moderation even if the system is turned off.
j) Calculate merit/demerit points, $t_{10\alpha}$: An agent evaluates its human counterpart's compliance with deadlines by calculating merit/demerit points. This action could motivate humans to work much more diligently to avoid breaching the deadline. The merit/demerit points are updated in the log file (see (e)).

5.2 Message Exchanges, $\mathbf{x}_{m\alpha}$

A message exchange is another type of action performed by agents to advance the workflow. We use the FIPA ACL [7] to implement message exchanges between agents. The message consists of a number of parameters, which include the mandatory performative, p_i, and other optional parameters, π_j, i.e.,

$$x_{n\alpha} \rightarrow (p_i, (\pi_j, f(\pi_j)), \ldots). \quad (9)$$

where $i, j, n \geq 1$, and $f(\pi_j)$ is a function which evaluates and returns a value for π_j.

The use of performatives enables agents to recognize the intent of a requesting agent for a specific service. We use most of the parameters π_j defined by FIPA and our own performatives. We define our own nine additional performatives from the analysis of the EPMP process which are Prepare, Check, Remind, Review, Complete, Modify, Advertise, Inform_all and Acknowledge. For example, in the Prepare performative, the sender advises the receiver to start prepare examination paper by performing some actions to enable its human counterpart to do so. The content of the message is a description of the action to be performed. The receiver understands the message and is capable of performing the action.

6 Modeling Environment, Ontology, Schedules, and Documents

6.1 Environment

Based on our analysis of the environment parameters required for the EPMP, we model the environment, E, to consist of four parts:

a) Status of uploaded files, e_1: The agent checks its human counterpart if he/she has uploaded Set A or Set B to a specified folder. If he/she has done so, the agent checks the next step. This state assists the agent in monitoring the submission of documents.
b) Status of deadlines, e_2: The agent checks the system's date everyday and compare it with the deadline. This state helps the agent to perform appropriate actions e.g. to send a remind message to an agent.
c) Status of subprograms, e_3: When an agent performs a task, it writes the action in a subprogram, e.g. when the Committee agent sends the Prepare message, it writes this event in the subprogram which it uses later for sending a remind message.
d) Status of Message Signal, e_4: The agent opens a port and makes a connection when it senses a message coming from a remote agent.

6.2 Ontology

In our model, the ontology defines the meaning of tasks, performatives and object names. The task ontology defines the meaning of tasks in agent actions. For example, in the task "Display a message window on the screen, $\mathtt{t}_{4\alpha}$," an agent processes appropriate information and create a window to show the message on the screen for its human counterpart. When an agent sends a message to a remote agent, the remote agent understands the meaning of the message by checking the performative ontology.

Object names like examination paper, lecturer form, moderation form, and committee form are ontology terms used in the EPMP model. When an agent receives a message, it checks these object names and process the objects based on the performative. For example, when the Committee agent sends a `Prepare` message to the Lecturer agent, it checks the `Prepare` ontology by which it opens the examination paper and the lecturer form. In other words, object names subsume in performatives.

6.3 Schedules

We model the schedules as discrete timeslots, $\sigma_{\mathtt{i}}$. A schedule can be nested in another schedule. For example, the main schedule is the start of the EPMP until all the examination documents have been collected. The second schedule is the examination paper preparation, the third schedule is the first moderation cycle and the fourth schedule is the second moderation cycle. The examination paper preparation, first moderation and second moderation schedules are nested in the main EPMP schedule. We represent such structure as follows:

$\sigma_1 = \delta_{\mathtt{start}}$ to $\delta_{\mathtt{end;}}$	EPMP process duration,
$\sigma_2 = \delta_{\mathtt{start}}$ to $\delta_{\mathtt{1start;}}$	Examination paper preparation,
$\sigma_3 = \delta_{\mathtt{1start}}$ to $\delta_{\mathtt{1end;}}$	First moderation cycle,
$\sigma_4 = \delta_{\mathtt{2start}}$ to $\delta_{\mathtt{2end;}}$	Second moderation cycle,

where $\delta_{\mathtt{2start}} = \delta_{\mathtt{1end}}$, and $\delta_{\mathtt{2end}} = \delta_{\mathtt{end}}$.

Two forms of reasoning are performed on these schedules: reminding/alerting humans to complete the scheduled task before the deadlines, and calculating the merit/demerit points based on the deadlines and actual submission dates. Figure 2 shows the schedules and the reasoning based on the schedules' parameters.

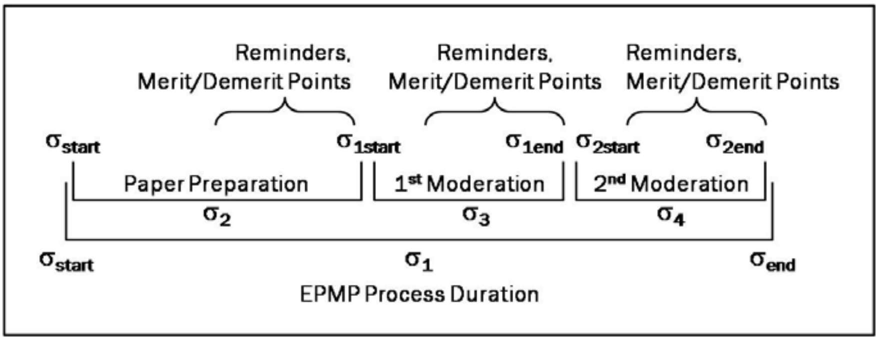


Fig. 2. The EPMP Schedules

6.4 Documents

Agents exchange documents by uploading the documents from the sender's to the receiver's folders. The agents always use the updated documents and upon receiving them, the receiver agent opens the documents for its human counterpart. The human counterparts can always open the documents at any time using some agent interface.

While the internal states of the documents are opaque to the agents, agents reason on the spatial and the temporal states of the documents. The agents then perform the autonomous action of submitting the documents to the next recipient.

In reasoning the spatial state of the documents, agents check the status of the uploaded files in a specified folder. The state of the uploaded files is true when its human counterpart clicks a Submit button in the agent interface. The agents then check other states to make the right decision. For reasoning on the temporal state of the documents, agents refer to the schedules (i.e. deadlines, $\delta_{\mathtt{1end}}$ and $\delta_{\mathtt{2end}}$). The agents use the deadlines to evaluate mathematical equations in performing appropriate actions. For example, if $\delta_{\mathtt{start}}$ is true, the committee agent sends the `Prepare` message.

7 The EPMP Model

An overview of the complete EPMP model is shown in Figure 3. The diagram shows the entities involved and the proposed solutions to resolve some of the human- and system-related problems.

Two essential premises of this model are the adoption and application of agent-based technology and the need to motivate or urge humans to abide by the scheduled deadlines while using the system. Based on these premises, we will analyze and study the issues pertaining to development methodologies, communication and semantics to design and manifest an effective agent-based collaboration and workflow system.

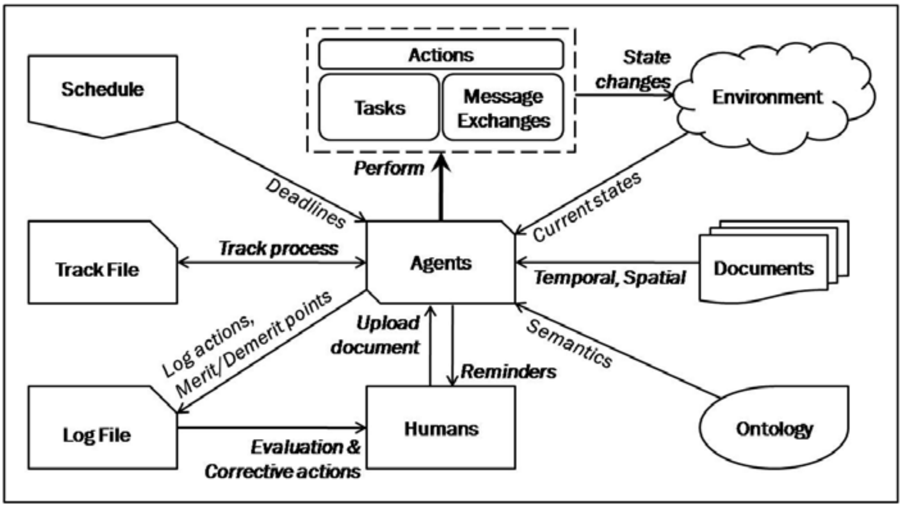


Fig. 3. An Overview of the EPMP Model

8 Conclusions

In this paper, we developed a model of the EPMP by considering the structure, process flow and entities of the domain. These attributes form the underlying theory of the model that includes agents, actions, environment, ontology, schedule, and documents. The model represents a blueprint for the subsequent development of a multi-agent system for the EPMP.

In our future work, we will look at the various agent development methodologies and use them or our own methodology to develop and implement an agent-based EPMP system.

References

[1] Bramley I.: SOA Takes Off – New WebSphere SOA Foundation Extends IBM's Lead with New System z9 Mainframes as the Hub of the Enterprise, 2nd (ed.) (2005)

[2] Chen, J.R., Wolfe, S.R., McLean, S.D.W.: A Distributed Multiagent System for Collaborative Information Management and Sharing, Virginia, US, pp. 382–388 (2000) ISBN:1-58113-320-0

[3] Chimera Agents for WIN-Prolog, `http://www.lpa.co.uk/chi.htm`

[4] DeLoach, S.A.: Multiagent Systems Engineering - A Methodology and Language for Designing Agent Systems. In: Proc. of Agent Oriented Information Systems, pp. 45–57 (1999)

[5] Ferber, J.: Multi-agent Systems. Addison-Wesley, Reading (1999)

[6] Fleurke, M., Ehrler, L., Purvis, M., Bees, J.: An Adaptive and Distributed Framework for Workflow Systems. In: Proc. IEEE/WIC International Conference on Intelligent Agent Technology, Halifax, Canada (2003)

[7] FIPA ACL Message Structure Specification: SC00061G (2002)

[8] FIPA Ontology Service Specification: XC00086D (2001)

[9] FIPA Communicative Act Library Specification SC00037J (2002)

[10] Kremer, R.: Artificial Intelligence Tutorial 2009 (AIT 2009), Agent Communication Paradigms (2009)

[11] Labrou, Y., Finin, T.: State of the Art and Challenges for Agent Communication Languages, Department of Computer Science and Electrical Engineering University of Maryland, Baltimore County Baltimore, MD 21250 (1994)

[12] Muehlen, Z., Rosemann, M.: Workflow-based Process Monitoring and Controlling - Technical and Organizational Issues. In: Sprague Jr., R. (ed.) Proceedings of the 33rd Hawaii' International Conference on Systems Sciences, Wailea, HI (2000)

[13] Savarimuthu, B.T.R., Purvis, M., Fleurke, M.: Monitoring and Controlling of a Multiagent based Workflow System. In: Purvis, M. (ed.) Proc. Australasian Workshop on Data Mining and Web Intelligence (DMWI 2004), Dunedin, New Zealand, CRPIT, pp. 127–132. ACS (2004)

[14] Steinfield, C., Jang, C., Pfaff, B.: Supporting Virtual Team Collaboration: The TeamSCOPE System, Phoenix, Arizona, pp. 81-90 (1999)

[15] Wooldridge, M.: An Introduction to Multi-agent Systems. John Wiley & Sons, Ltd., Chichester (2002)

A Three-Dimensional Abstraction Framework to Compare Multi-Agent System Models

Tibor Bosse, Mark Hoogendoorn, Michel C.A. Klein, and Jan Treur

VU University Amsterdam, Department of Artificial Intelligence,
De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
{tbosse,mhoogen,mcaklein,treur}@cs.vu.nl
http://www.cs.vu.nl/~{tbosse,mhoogen,mcaklein,treur}

Abstract. Models of agents or multiagent systems in a certain application area can be made at different levels of abstraction. The aim of this paper is to clarify different dimensions of abstraction. The three dimensions of considered are: the process abstraction dimension, the temporal dimension, and the agent cluster dimension. Thus a three-dimensional framework is obtained in which different types of multi-agent system models can be classified. For a number of multi-agent system models in different application areas from the literature it is discussed how they fit in the framework.

1 Introduction

One of the important choices made in (multi)agent system modelling in a certain application area is the grain-size or abstraction level of the model being developed. Models which are too coarse-grained (too abstract) may miss details of the domain modelled that are essential to the aim of the model, and models which are too fine-grained (too detailed) for the aim of the model may become difficult to handle because they are too complex and/or intransparent. Therefore choosing the right level of abstraction may be crucial. However, upon further consideration the notion of abstraction level itself is not so self-evident. For example, does more abstract mean that the models for the internal processes within agents are described at a more abstract level (abstracting from internal physiological, cognitive and affective detail)? Or should it be interpreted such that the models describe relationships over larger time intervals (abstracting from the smaller time steps)? Or does more abstract mean that within the multiagent system model higher level structures are used that aggregate individual agents (considering groups or clusters of agents as entities, as in organisation models)? In this paper the viewpoint is taken that to clarify the notion of abstraction level the point of departure should be that different dimensions of abstraction have to be considered.

The focus of the paper is on the following dimensions of abstraction: the *process abstraction*, *temporal*, and *agent cluster dimension*. For each of these dimensions a multiagent system application can be modeled from a more local-level perspective and from a more global-level perspective. Thus a framework is obtained to distinguish abstraction levels for different types of models (see the cube in Figure 1). It will be

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 306–319, 2010.

shown that this framework is able to distinguish and position different types of multi-agent system models known from the literature, including, for example, population-based vs agent-based models, behavioural vs cognitive agent models, and executable models vs requirements models. Note that the distinctions made by the framework are purely semantic, independent of any representation format of a model. In the paper, first the three dimensions are introduced in Sections 2, 3 and 4. Next, in Section 5 it is shown how the three-dimensional framework can be used to position models and approaches available in the literature. Finally, Section 6 is a discussion.

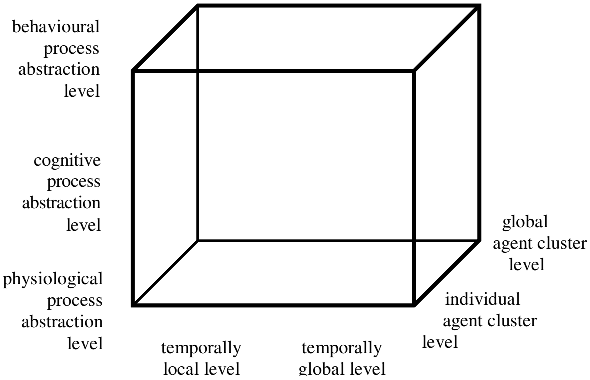

behavioural process abstraction level
cognitive process abstraction level
physiological process abstraction level
temporally local level
temporally global level
global agent cluster level
individual agent cluster level

Fig. 1. Process abstraction, temporal and agent cluster dimension

2 The Process Abstraction Dimension

For the process abstraction level dimension, a multi-agent process can be conceptualised (from more abstract, higher to less abstract, lower levels) at the behavioural level, the cognitive level, or the physiological level; see Fig. 1.

Behavioural Level. At the *behavioural level* relations are described between input and output states of agents, for example:

- direct reactive behaviour relating received input to an immediate response
- an avoidance reaction after three times a negative experience with something or somebody
- spread of information over a population by gossipping

At the behavioural process abstraction level, models abstract from internal (e.g., cognitive or physiological) concepts, such as beliefs, desires, intentions or activation states, and subprocesses involving them. Behavioural specifications only make use of concepts related to input states or output states (e.g., observations, actions, outgoing and incoming communication) of an agent.

The different dimensions will be illustrated by a toy case study involving an instant dating service: an Internet agent ID that can be contacted when somebody wants a date. This is to be considered a fictitious example just for the purpose of illustration. In principle many activities during a date are possible, such as hiking, going out for a dance club, attending a concert, et cetera. For the case study three characteristics make up a type of an activity: active or not, involving eating or not, and romantic or not. Moreover, it is assumed that dating is steered by an identity described by specific body states: blood sugar level, adrenaline level, testosterone level, and dopamine lack. Each person involved has interaction in three phases:

- Based on a chosen identity the person requests ID for a date
- After a candidate has been proposed by ID the person proposes ID a desired activity type
- When a received activity type is desired, a date with this candidate is performed

These have been specified in the form of temporally local behavioural properties LBP1, LBP2, and LBP3, as follows. Here ↠ is used to denote a causal relation.

LBP1 Requesting for a date

If *P* observed that the dopamine lack is high, the adrenaline level is *X1,* the blood sugar lack is *X2,* and the testosterone level is *X3,*
then *P* will request *ID* to look for a date with identity *id(X1, X2, X3)*

```
observed(P, dopaminelack, high) & observed(P, adrenaline, X1) &
observed(P, bloodsugarlack, X2) & observed(P, testosterone, X3)
   ↠ communication(P, date_with_id(id(X1, X2, X3)), ID)
```

LBP2 Proposing an activity

If it was communicated by *ID* to *P* that *Q* with identity *id(X1, X2, X3)* agrees to have a date,
and *P* observed for adrenaline *X1*, blood sugar lack *X2*, testosterone *X3*,
then *P* will propose *ID* to perform activity *act(X1, X2, X3).*

```
communication(ID, agrees(Q, id(X1, X2, X3)), P) &
observed(P, adrenaline, X1) & observed(P, bloodsugarlack, X2)  & observed(P, testosterone, X3)
↠ communication(P, wants(P, act(X1, X2, X3)), ID)
```

LBP3 Performing an agreed activity

If it was communicated by *ID* to *P* that *Q* with identity *id(X1, X2, X3)* agrees to have a date,
and *P* observed for adrenaline *X1*, blood sugar lack *X2*, testosterone *X3*,
and it was communicated that *Q* wants to do activity *act(X1, X2, X3)*,
then *P* will perform *act(X1, X2, X3)* with *Q*.

```
communication(ID, agrees(Q, id(X1, X2, X3)), P) & observed(P, adrenaline, X1) &
observed(P, bloodsugar, X2)  & observed(P, testosterone, X3) &
communication(ID, wants(Q, act(X1, X2, X3)), P)
↠ performed(P, date(act(X1, X2, X3), Q))
```

Cognitive Level. At the *cognitive level*, both informational concepts (e.g., observations, communications, beliefs), and affective or motivational concepts (e.g. emotions, feelings, desires) are considered. For the case study it is assumed that persons aim at maintaining homeostasis for the specific body states. By monitoring the body, specific desires are generated which motivate actions to fulfill them (e.g., communicating a request for a date or a proposal for an activity). For these actions, first intentions are generated, and from an intention, the actual performance of the action is generated. As an example, a desire to find a date is directly related to the state of the body, in particular the dopamine, adrenaline, blood sugar and testosterone levels**.** For the sake of simplicity it is assumed that the desire leads to a corresponding intention, and in turn this leads to the related communication.

LCP1 Generating a desire to date using some identity

If *P* observed that the dopamine lack is high, adrenaline level is *X1,*
blood sugar lack level *X2*, and testosterone level *X3*
then *P* will have the desire to obtain a date with identity *id(X1, X2, X3)*

```
observed(P, dopaminelack, high) & observed(P, adrenaline, X1) &
```

observed(P, bloodsugarlack, X2) & observed(P, testosterone, X3)
→ desire(P, date_with_id(id(X1, X2, X3)))

LCP2 Generating an intention to date based on a desire
If *P* has the desire to obtain a date with identity *I*,
then *P* will have the intention to obtain a date with identity *I*.
desire(P, date_with_id(I)) → intention(P, date_with_id(I))

LCP3 Requesting to look for a date based on the intention
If *P* has an intention for a date with identity *I*,
then *P* will request *ID* to look for a date with identity *I*
intention(P, date_with_id(I)) → communication(P, date_with_id(I), ID)

Physiological Level. At the physiological level, processes are described in terms of body and brain states and relationships between them. For example, this may concern muscle activity, levels of (bio)chemical substances in the blood, sensor and effector states, or activation states of neurons, or groups of neurons. Within the case study the four body aspects are considered as before, and in addition sensor and effector states, and activations of (groups of) neurons. For the physiological process abstraction level, a neural network structure was specified, together with connection strengths and threshold values. Threshold functions were used, defined as $th(\sigma, \tau, V) = 1/(1+e^{-\sigma(V-\tau)})$ with *steepness* σ and *threshold* τ. This specification assumes that specific values are given for each of the occurrences of connection weight ω, steepness σ, and threshold τ, in dependence of variables such as *B*, *I* and *A*. The assumed sensor system for body states maintains socalled *neural body maps* using sensory neurons (e.g., [22], [23]).

LPP1 Generating a sensor state for a body state
If body state property *B* occurs for person *P* with level *V*
and the connection strength from body state property *B* to a sensor state for *B* is ω
and the steepness and threshold value for this sensor state are σ and τ
then the sensor state for *B* will have activation level $th(\sigma, \tau, \omega V)$.
activation(BS(P, B), V) & connection_strength(BS(P, B), S(P, B), ω) &
steepness(S(P, B), σ) & threshold(S(P, B), τ)
→ activation(S(P, B), th(σ, τ, ωV))

LPP2 Activating sensory neurons for a body map
If the sensor state for body state *B* has activation level *V*
and the connection strength from sensor state for *B* to the sensory neuron for *B* is ω
and the steepness and threshold value for this sensor state are σ and τ
then the sensory neuron for *B* will have activation level $th(\sigma, \tau, \omega V)$.
activation(S(P, B), V) & connection_strength(S(P, B), SN(P, B),ω) &
steepness (SN(P, B), σ) & threshold(SN(P, B), τ)
→ activation(SN(P, B), th(σ, τ, ωV))

Depending on these neural body maps certain preparation neurons become activated. In particular, the sensory neurons for lack of dopamine (indicating 'being bored') have a positive effect on the activation for preparing to request a date; here the body maps for adrenaline, blood sugar lack and testosterone have effect on which identity is chosen for such a date. How the effects of the activation levels of the four connected sensory neurons that provide input for the preparation neurons are combined, depends on the respective connection strengths and on the threshold value of the preparation neuron, using a threshold function as described above. It is configured in a way so that for each combination of the other three body states (adrenaline, blood sugar lack and testosterone level) a suitable identity is available, (specified by strengths of *-1* or *1* of the connections ω_i to the preparation for this identity). This determines how a preparation neuron for an initiative to get a date fitting to the body map is activated.

3 The Temporal Dimension

For the temporal dimension a process can be conceptualised by descriptions over longer time periods (e.g., *emerging patterns*), or by descriptions relating states over smaller time steps (e.g., *mechanisms*). Temporal perspectives on the dynamics of processes are addressed in a wide variety of literature, from different disciplines, including Cognitive Science , [3], [42], [56], and Artificial Intelligence and Computer Science [5], [12], [40], [13], [21], [31], [50], [59], [4]. Temporally local properties were used in Section 2 to describe the (physiological, cognitive, or behavioural) mechanisms of a process considered in a step-by-step manner. Usually descriptions specified as difference or differential equations, transition systems, or immediate causal relationships, are examples of temporally local descriptions. Patterns that occur as emergent phenomena over time can be described in a temporally more global manner, by relations over longer time periods. For specification of emerging patterns a more expressive temporal modelling language is needed, for example, linear time or branching time modal temporal logics (such as LTL or CTL), or temporal predicate logics (such as the Temporal Trace Language TTL; cf. [12], [40]). Below, different types of often considered global temporal properties are described.

Achievement properties express that *eventually a certain state is reached* in which some (desired) state property holds. They relate to what in the literature in Computer Science are sometimes called *liveness* properties, and what in agent literature is referred to as *achievement goals* (in contrast to, for example, maintenance goals). An example of an achievement property expresses that whenever the dopamine lack of a person is high, eventually the dopamine lack will be low, and, the adrenaline, testosterone level and the blood sugar lack will be low:

GBP1 Body state achieved with low levels
If at some point in time T the dopamine lack in the body is high
then at a later time point T1 in the body the dopamine lack, adrenaline,
blood sugar lack, and testosterone all will be low.

$$\begin{aligned}
&\forall T, P, V\ \ [\,[\,\text{state}(\gamma, T) \models \text{body\_state}(P, \text{dopaminelack}, V)\ \&\ V \geq 0.5\,]\\
&\quad \Rightarrow \exists T1, V1, V2, V3, V4\ [\,T1 \geq T\ \&\\
&\qquad \text{state}(\gamma, T1) \models \text{body\_state}(P, \text{adrenaline}, V1)\ \&\ V1 < 0.5\ \&\\
&\qquad \text{state}(\gamma, T1) \models \text{body\_state}(P, \text{bloodsugarlack}, V2)\ \&\ V2 < 0.5\ \&\\
&\qquad \text{state}(\gamma, T1) \models \text{body\_state}(P, \text{testosterone}, V3)\ \&\ V3 < 0.5\ \&\\
&\qquad \text{state}(\gamma, T1) \models \text{body\_state}(P, \text{dopaminelack}, V4)\ \&\ V4 < 0.5\,]\,]
\end{aligned}$$

Here $\text{state}(\gamma, T) \models \text{body\_state}(P. \text{dopaminelack}, V)$ denotes that within the state $\text{state}(\gamma, T)$ at time point T in trace $\gamma$the state property $\text{body\_state}(P. \text{dopaminelack}, V)$ holds, denoted by the (infix) predicate $\models$ for the satisfaction relation. A specific type of achievement properties are *equilibrium properties*, expressing that after some time an equilibrium state is reached; for example:

GBP6 Reaching an equilibrium while ID functions properly

Eventually, after some point in time *T* person *P*’s body states are constant with dopamine lack <0.5, and no dates or communications with ID occur.

$$\begin{array}{l}
\forall P\ \exists T\ \forall T1 \geq T\ [\ \text{state}(\gamma, T1) \models \exists V1\ [\ \text{body\_state}(P, \text{dopaminelack}, V1)\ \&\ V1<0.5\ ]\ \& \\
\quad \forall B, V\ [\ \text{state}(\gamma, T) \models \text{body\_state}(P, B, V)\ \Rightarrow \text{state}(\gamma, T1) \models \text{body\_state}(P, B, V)\ ]\ \& \\
\qquad \text{state}(\gamma, T1) \models \neg\ \exists A, Q\ \text{performed}(P, \text{date}(A, Q))\ \& \\
\qquad \text{state}(\gamma, T1) \models \neg\ \exists C\ \text{communication}(C, P, ID)\ \&\ \neg\exists C\ \text{communication}(C, ID, P)\ ]
\end{array}$$

Milestone properties express that under certain conditions some *intermediate state* will be reached. These properties can be used, for example, to decompose the overall process into a number of subprocesses that, possibly depending on other milestones already achieved, each result in the achievement of some intermediate state (milestone). For the case study, to achieve GBP1, a person has interaction in three phases, specified as behavioural properties GBP2, GBP3, and GBP4:

GBP2 Requesting for a date

If *P* observed that the dopamine lack is high, the adrenaline level is *X1,*
the blood sugar lack is *X2,* and the testosterone level is *X3*
then *P* will request *ID* to look for a date with identity *id(X1, X2, X3)*

$$\begin{array}{l}
\forall T, P, X1, X2, X3 \\
[\ [\ \text{state}(\gamma, T) \models \text{observed}(P, \text{dopaminelack}, \text{high})\ \&\ \text{state}(\gamma, T) \models \text{observed}(P, \text{adrenaline}, X1)\ \& \\
\quad \text{state}(\gamma, T) \models \text{observed}(P, \text{bloodsugarlack}, X2)\ \&\ \ \text{state}(\gamma, T) \models \text{observed}(P, \text{testosterone}, X3)\ ] \\
\quad \Rightarrow\ \exists T1\ \ [\ T1 \geq T\ \&\ \text{state}(\gamma, T1) \models \text{communication}(P, \text{date\_with\_id}(\text{id}(X1, X2, X3)), ID)\ ]
\end{array}$$

GBP3 Proposing an activity

If it was communicated by *ID* to *P* that *Q* with identity *id(X1, X2, X3*) agrees to have a date,
and *P* observed for adrenaline *X1*, blood sugar lack *X2*, testosterone *X3*,
then *P* will propose *ID* to perform activity *act(X1, X2, X3).*

$$\begin{array}{l}
\forall T, P, X1, X2, X3 \\
\quad [\ [\ \text{state}(\gamma, T) \models \text{communication}(ID, \text{agrees}(Q, \text{id}(X1, X2, X3)), P)\ \&\ \ \text{state}(\gamma, T) \models \text{observed}(P, \text{adrenaline}, \\
X1)\ \&\ \text{state}(\gamma, T) \models \text{observed}(P, \text{bloodsugarlack}, X2)\ \&\ \text{state}(\gamma, T) \models \text{observed}(P, \text{testosterone}, X3)\ ] \\
\quad \Rightarrow\ \exists T1\ \ [\ T1 \geq T\ \&\ \text{state}(\gamma, T1) \models \text{communication}(P, \text{wants}(P, \text{act}(X1, X2, X3)), ID)\ ]
\end{array}$$

GBP4 Performing an agreed activity

If it was communicated by *ID* to *P* that *Q* with identity *id(X1, X2, X3*) agrees to have a date,
and *P* observed for adrenaline *X1*, blood sugar lack *X2*, testosterone *X3*,
and it was communicated that *Q* wants to do activity *act(X1, X2, X3)*,
then *P* will will perform *act(X1, X2, X3)* with *Q*.

$$\begin{array}{l}
\forall T, P, X1, X2, X3 \\
[\ [\ \text{state}(\gamma, T) \models \text{communication}(ID, \text{agrees}(Q, \text{id}(X1, X2, X3)), P)\ \& \\
\text{state}(\gamma, T) \models \text{observed}(P, \text{adrenaline}, X1)\ \&\ \text{state}(\gamma, T) \models \text{observed}(P, \text{bloodsugarlack}, X2)\ \& \\
\text{state}(\gamma, T) \models \text{observed}(P, \text{testosterone}, X3)\ \&\ \text{state}(\gamma, T) \models \text{communication}(ID, \text{wants}(Q, \text{act}(X1, X2, X3)), P) \\
\quad \Rightarrow\ \exists T1\ \ [\ T1 \geq T\ \&\ \text{state}(\gamma, T1) \models \text{performed}(P, \text{date}(\text{act}(X1, X2, X3), Q))\ ]
\end{array}$$

Maintenance properties express that certain state properties do not (or not much) change during a certain time period: *maintenance properties*. For example, specifying that a form of homeostasis is maintained: a state in which variables stay within certain fixed bounds. This class of properties relates to *safety* properties in computer science and what in agent literature is referred to as *maintenance goals*.

Representation properties express how an internal state relates to external states (or input and/or output states) in past (*backward*) or future times (*forward*); e.g., [8], [44], [15], [16], [39]. For example, when a sensory neuron has a high activation level when a stimulus occurred (or has been sensed) before, this is a backward representation relation for this neuron. Moreover, if a high activation level of some preparation neuron later on leads to the effect of the corresponding action, this is an example of a forward representation relation for this neuron. A forward representation property for desires describes, for example, the behaviours or actions that are driven or motivated by these desires, under certain conditions. For the behaviour of a person to make sense, these actions are not arbitrary, but are usually assumed to achieve fulfillment of the desires. The following example of a temporally global property at the cognitive process abstraction level is a forward representation property for a desire to have a date with a certain identity.

GCP5 Forward representation relation for a desire to date

If at some point in time *T* person *P* desires to have a date with identity *I*
then at a time point *T1* within time duration *D* after *T* person *P* will
communicate to ID a date request with identity *I.*

$\forall$T, P, I state(γ, T) |= desire(P, date_with_id(I)) &
$\Rightarrow \exists$T1 [T $\leq$ T1 & T1 $\leq$ T + D & state(γ, T1) |= communication(date_with_id(I), P, ID)]

Looking forward in time some steps further, desires can be viewed as representing a future state in which they are satisfied. As after a desire has been satisfied, in principle it is not there anymore, this may sound a bit circular and paradoxal: desires that exist in the present represent their own future nonexistence:

GCP6 Fulfillment of a desire to date

If at some point in time *T* person *P* desires to have a date with identity *I*
then at a time *T1* within duration *D* after *T* person *P* will not desire this anymore for any *I1.*

$\forall$T, P, I state(γ, T) |= desire(P, date_with_id(I)) &
$\Rightarrow \exists$T1 [T $\leq$ T1 & T1 $\leq$ T + D & $\neg$ $\exists$I1 state(γ, T1) |= desire(P, date_with_id(I1))]

Comparison properties describe how certain state properties at different time points are compared, and properties in which certain state properties in different traces are compared. *Time comparison properties* are properties expressing, for example, that under certain conditions (e.g., after some time point) the value of a variable is *monotonically increasing* or *decreasing* over time. *Trace comparison properties* express, for example, that if in one trace the values of certain variables are *lower* than in a second trace, then the value of some other variable will also be *lower* in that trace. In the case study dates have an effect on body states in that they decrease the values. As there are no other effects on body states described in the model, this implies that over time the values can never become higher, only lower. This holds for the body states themselves, but also for the related sensor states and sensory neurons. Such monotonic patterns can be expressed by time comparison properties, in particular monotonicity properties, such as the following:

GPP5 Monotonically decreasing values in body maps

If at some point in time *T1* in the body map *B* has value *V1*
and at some point in time *T2* in the body map *B* has value *V2*
and *T1* $\leq$ *T2*, then *V1* $\geq$ *V2.*

$\forall$T1, T2, P, B, V1, V2 [state(γ, T1) |= activation(SN(P, B), V1) & state(γ, T2) |= activation(SN(P, B), V2) &
T1 $\leq$ T2 $\Rightarrow$ V1 $\geq$ V2]

An example of a trace comparison property describes traces where at some point in time the dopamine lack is high in comparison to other traces. In the former type of traces eventually the values of all body states will be lower than or equal to their values in the latter type of traces:

GBP5 Trace comparison for body states

When in trace γ1 at some time *T* dopamine lack ≥*0.5* occurs, and trace γ2 is any trace, then there is a point in time *T1*≥*T* such that at each time point *T2*≥*T1* all body state values in trace γ1 are at most as high as the body state values in trace γ2 at *T2*.

∀γ1, γ2 ∀T, P, V [state(γ1, T) |= body_state(P, dopaminelack, V) & V≥0.5 ⇒
∃T1≥T ∀T2≥T1 [∀B,V1,V2 [state(γ1, T2) |= body_state(P, B, V1) & state(γ2, T2) |= body_state(P, B, V2)]
⇒ V1≤V2]]

4 The Agent Cluster Dimension

Domains where numbers of agents are quite large can be modelled in a more manageable manner by not introducing conceptual entities for each individual agent separately, but by taking *groups*, *(sub)populations* or *clusters* of agents as basic conceptual entities. Examples of such clusters are: divisions or departments within an organisation, age-related groups in a population, subcommunities within a society, or sports teams in a competition. The agent cluster dimension describes how a process can be conceptualised from an individual agent perspective to a more global perspective where a number of clusters, each consisting of multiple agents, are considered as basic entities. For the dating case, the following clusters are considered:

- persons in the process of obtaining a date: who requested but did not yet start a date (R)
- persons performing a date (D)
- persons not in a process of dating: not in a date nor in a process of obtaining a date (N)

By temporally local descriptions it can be specified how the states of the clusters (e.g., numbers of agents *N(t)* in cluster N) at a given time point relate to the states at a next time point. For example, within the time interval from *t* to *t* + Δ*t* a number of agents *NtoR(t)* per time unit will leave cluster N to join cluster R (entering the process to obtain a date), and a number of agents *DtoN(t)* per time unit will leave cluster D to join cluster N (after finishing a date). The following equation results:

$N(t+\Delta t) = N(t) + DtoN(t)\,\Delta t - NtoR(t)\,\Delta t$

This temporally local dynamic property (difference equation) can be used to perform simulations by calculating the values at the next time point from the values at the current time point. Here, *DtoN(t)* and *NtoR(t)* are expressed in the values of the states at time *t* as follows. For a date an average time duration is assumed: *dd*. By approximation from the cluster D at each time unit a fraction $\alpha = 1/dd$ will finish the date. Therefore per time unit the number making the transition from cluster D to cluster N is $DtoN(t) = \alpha D(t)$. Similarly it is estimated which part of cluster N per time unit makes the transition to cluster R. Here the number *N(do, t)* within cluster N with high dopamine lack is used: $NtoR(t) = \beta N(do, t)$. Parameter β indicates how fast the date request will be done. If this duration is on average *rd* time units, then $\beta = 1/rd$ can be taken. The following difference equation is obtained:

$N(t+\Delta t) = N(t) + \alpha D(t)\,\Delta t - \beta N(do, t)\,\Delta t$

It can also be written as

$\Delta N(t)/\Delta t = \alpha D(t) - \beta N(do, t)$

where by definition $\Delta N(t) = N(t+\Delta t) - N(t)$. In a continuous form this temporally local dynamic property can be represented by a *differential equation* as follows:

$dN(t)/dt = \alpha D(t) - \beta N(do, t)$

In a similar way temporally local dynamic properties can be obtained for the other clusters, which results in the following set of temporally *local* properties (in difference equation format) for the clusters:

$$\Delta N(t)/\Delta t = \alpha D(t) - \beta N(do, t)$$
$$\Delta R(t)/\Delta t = \beta N(do, t) - \gamma R(t)$$
$$\Delta D(t)/\Delta t = \gamma R(t) - \alpha D(t)$$
$$\Delta N(do, t)/\Delta t = \eta N(t) - \beta N(do, t)$$

Here $\gamma$is the fraction (per time unit) of the cluster R that starts a date. It is assumed that the transition from cluster D to cluster N only contributes to agents in cluster N with body states dopamine lack low. Moreover, it is assumed that always (by other processes) a certain fraction η (per time unit) of cluster N gets a high dopamine lack; this explains the fourth equation above.

For temporally *global* descriptions at the *behavioural* level an example achievement property indicates that the size of the cluster describing the persons with high body state for x in N is near *0* (using a small margin ε).

GCBP1 Achieving low body states
Eventually a state is reached in which there are no high body states $B \neq do$
$\exists T\ \forall T1 \geq T\ \ \forall B, V\ [\ B \neq do\ \&\ state(\gamma, T1) \models has\_size(N(B), V)\ \Rightarrow V < \varepsilon\ ]$

Descriptions of clusters at the *cognitive* process abstraction level address the internal cognitive structures of persons within a cluster. For example, it is expressed that a certain fraction of a cluster or population has a desire for a certain type of activity or to date, or has the intention to perform a certain activity or to request a date. In such a way mental states can be aggregated to *collective mental states* of a cluster, which occur with a certain strength. For instance, statements such as ‘it is commonly believed that the climate change is caused by human activities’ refers to *collective beliefs* present in the population with a certain strength. Yet another mental concept that can take an aggregated form, is *collective trust*, for example, in the world wide financial system. It is also possible to describe temporal relationships between collective cognitive states, for example, the collective belief that the climate change is caused by human activities, leads after some time to the collective intention to take measures. Descriptions of clusters at the *physiological* process abstraction level involve collective neural activation states. This can be addressed in a manner similar to how the cognitive level was addressed.

5 Discussion and Classification of Existing Models

As a main contribution this paper aims at clarifying different types of abstraction levels for agent system models by making explicit different dimensions for such abstraction levels. Any multi-agent system model for a certain application area can be positioned in this three-dimensional space by indicating coordinates for each of these dimensions at a scale from a more local-level (less abstract) perspective and from a more global-level (more abstract) perspective. In this section for different areas in the cube it will be indicated which types of models from the literature fit in this area; see Table 1. Here at each of the axes a distinction was made according to local vs nonlocal (or global), which results in eight different areas. For the process abstraction dimension, the internal (physiological or cognitive) level will be taken as local, in contrast to the behavioural level taken as global.

Temporally local agent-based internal and behavioural models are physiological, cognitive or behavioural agent models described at an individual agent level in small time steps. Many examples of such executable models are available, described in a logical (temporal rules), numerical (difference or differential equations), or a hybrid manner, or by transition systems, finite automata or Petri nets. For example, in [11] internal cognitive states of the agents concern the attractiveness of locations, in [17] internal trust states are maintained on which behaviour is based, and in [58] internal states related to motivation and learning are maintained. Models with internal physiological states can also be found in the area of agent-based epidemics modelling, where the internal infection states of agents are considered. Many of the models in the areas of social simulation, self-organisation, ant computing, and swarm intelligence belong to this class, and often are purely behavioural, described in a reactive manner by stimulus-response-like associations; the complexity emerges from the interaction of large numbers of such simple agents, and the environment.

Table 1. Overview of literature

dimension			*examples from the literature*
temporal	*cluster*	*process*	
temporally local	agent-based	internal	[5] (logical); [3], [42], [56] (numerical); [11], [13], [17], [1], [46], [32], [37], [51], [58], [60], [61], [36] (hybrid); [4], [21], [6], [7], [10] (transitions, automata, Petri nets)
		behavioural	[52], [53], [26], [14] (social simulation, swarm intelligence); [27] (emotion contagion); [54] (analysis)
	cluster-based	internal	[11] (crime); [20] (organisation); [38], [45] (joint goals and intentions); [2], [37], [43] (epidemics)
		behavioural	[34], [29], [41], [35], [28], [9] (organisation); [49], [57], [62] (ecological)
temporally global	agent-based	internal	[47], [24], [25], [64], [63], [48] (requirements); [31], [50], [59], [4], [21], [6], [7], [10], [12], [40] (verification)
		behavioural	
	cluster-based	internal	[24], [25], [34], [29], [30], [41], [36], [18], [55], [64], [65] (requirements, enterprise); [43], [2], [37] (epidemics); [49], [57], [62] (ecological)
		behavioural	

Temporally local cluster-based internal and behavioural models involve collective internal agent states or behaviours for clusters of agents, such as joint or shared actions, goals or beliefs. In general, the temporally local descriptions in this class of models can take the form of executable temporal logical rules, or difference or differential equations, or of a combination of both. For instance, [20] analyses the relationship between intentions and collective activity of groups of agents. In [38] problem solving within groups is investigated based upon joint intentions, and in [45] the expression of joint goals for teams.

Temporally global agent-based internal and behavioural models describe temporally more complex properties in terms of internal physiological or cognitive states or behaviours for the individual agents involved. Such descriptions, usually expressed in a richer temporal logical format (e.g., LTL, CTL, situation calculus, TTL), play an important role as a formalisation of patterns emerging from the more local mechanisms. They often occur in literature addressing requirements modelling or verification for multiagent systems.

Temporally global cluster-based internal and behavioural models concern descriptions of multiagent systems at a temporally global and global clustering level, but involve internal states of agents. Such descriptions are usually modelled as temporally complex properties expressing emerging patterns in terms of collective internal agent states or behaviours for clusters of agents such as joint actions or shared goals, emotions or beliefs. Such descriptions involve temporally complex properties expressing emerging patterns in terms of collective dynamics for clusters of agents.

References

[1] Anderson, J.R., Lebiere, C.: The atomic components of thought. Lawrence Erlbaum Associates, Mahwah (1998)

[2] Anderson, R.A., May, R.M.: Infectious Diseases of Humans: Dynamics and Control. Oxford University Press, Oxford (1992)

[3] Ashby, W.R.: Design for a Brain. Revised (edn.) Chapman & Hall, London (1960)

[4] Baier, C., Katoen, J.-P.: Principles of Model Checking. MIT Press, Cambridge (2008)

[5] Barringer, H., Fisher, M., Gabbay, D., Owens, R., Reynolds, M.: The Imperative Future: Principles of Executable Temporal Logic. John Wiley & Sons, Chichester (1996)

[6] Behrens, T.M., Dix, J.: Model checking multi-agent systems with logic based Petri nets. Ann. Math. Artif. Intell. 51, 81–121 (2007)

[7] Benerecetti, M., Giunchiglia, F., Serafini, L.: Model Checking Multiagent Systems. Journal of Logic and Computation 8(3), 401–423 (1998)

[8] Bickhard, M.H.: Representational Content in Humans and Machines. J. of Exp. and Theor. Artificial Intelligence 5, 285–333 (1993)

[9] Boella, G., van der Torre, L.: Organisations as Socially Constructed Agents in the Agent Oriented Paradigm. In: Gleizes, M.-P., Omicini, A., Zambonelli, F. (eds.) ESAW 2004. LNCS (LNAI), vol. 3451, pp. 1–13. Springer, Heidelberg (2005)

[10] Bordini, R.H., Fisher, M., Visser, W., Wooldridge, M.: Verifying multi-agent programs by model checking. Auton Agent Multi-Agent Sys. 12, 239–256 (2006)

[11] Bosse, T., Gerritsen, C., Hoogendoorn, M., Jaffry, S.W., Treur, J.: Agent-Based versus Population-Based Simulation of Displacement of Crime: A Comparative Study. Web Intelligence Agent Systems (to appear 2010)

[12] Bosse, T., Jonker, C.M., van der Meij, L., Sharpanskykh, A., Treur, J.: Specification and Verification of Dynamics in Agent Models. Int. J. of Coop. Inf. Systems 18, 167–193 (2009)

[13] Bosse, T., Jonker, C.M., van der Meij, L., Treur, J.: A Language and Environment for Analysis of Dynamics by Simulation. Intern. J. of AI Tools 16, 435–464 (2007)

[14] Bosse, T., Jonker, C.M., Schut, M.C., Treur, J.: Simulation and Analysis of Shared Extended Mind. Simulation 81, 719–732 (2005)

[15] Bosse, T., Jonker, C.M., Schut, M.C., Treur, J.: Collective Representational Content for Shared Extended Mind. Cognitive Systems Research 7, 151–174 (2006)

[16] Bosse, T., Jonker, C.M., Treur, J.: Representation for Reciprocal Agent-Environment Interaction. Cognitive Systems Research 10, 366–376 (2009)

[17] Bosse, T., Schut, M.C., Treur, J., Wendt, D.: Trust-Based Inter-Temporal Decision Making: Emergence of Altruism in a Simulated Society. In: Antunes, L., Paolucci, M., Norling, E. (eds.) MABS 2007. LNCS (LNAI), vol. 5003, pp. 96–111. Springer, Heidelberg (2008)

[18] Bresciani, P., Perini, A., Giorgini, P., Giunchiglia, F., Mylopoulos, J.: Tropos: An Agent Oriented Software Development Methodology. JAAMAS 8, 203–236 (2004)

[19] Carpenter, C., Sattenspiel, L.: The Design and Use of an Agent-Based Model to Simulate the 1918 Influenza Epidemic at Norway House, Manitoba. Amer. J. of Human Biology 21, 290–300 (2009)

[20] Castelfranchi, C.: Commitments: From Individual Intentions to Groups and Organizations. In: Proceedings of the First International Conference on Multiagent Systems, ICMAS, pp. 41–48. AAAI Press, Menlo Park (1995)

[21] Clarke, E.M., Grumberg, O., Peled, D.A.: Model Checking. MIT Press, Cambridge (2000)

[22] Damasio, A.: The Feeling of What Happens. In: Body and Emotion in the Making of Consciousness. Harcourt Brace, New York (1999)

[23] Damasio, A.: Looking for Spinoza. Vintage books (2003)

[24] Dardenne, A., van Lamsweerde, A., Fickas, S.: Goal-directed Requirements Acquisition. Science in Computer Programming 20, 3–50 (1993)

[25] Darimont, R., van Lamsweerde, A.: Formal Refinement Patterns for Goal-Driven Requirements Elaboration. In: Proc. of the Fourth ACM Symposium on the Foundation of Software Engineering (FSE4), pp. 179–190 (1996)

[26] Dorigo, M., Maniezzo, V., Colorni, A.: The Ant System: Optimization by a Colony of Cooperating Agents. IEEE Trans. Systems, Man, and Cybernetics-Part B 26(1) (1996)

[27] Duell, R., Memon, Z.A., Treur, J., van der Wal, C.N.: An Ambient Agent Model for Group Emotion Support. In: Cohn, J., Nijholt, A., Pantic, M. (eds.) Proc. of the Third Int. Conf. on Affective Computing and Intelligent Interaction, ACII 2009, pp. 550–557. IEEE Computer Society Press, Los Alamitos (2009)

[28] Esteva, M., Padget, J., Sierra, C.: Formalizing a Language for Institutions and Norms. In: Meyer, J.-J.C., Tambe, M. (eds.) ATAL 2001. LNCS (LNAI), vol. 2333, pp. 348–366. Springer, Heidelberg (2002)

[29] Ferber, J., Gutknecht, O., Jonker, C.M., Müller, J.P., Treur, J.: Organization Models and Behavioural Requirements Specification for Multi-Agent Systems. In: Demazeau, Y., Garijo, F. (eds.) Proc. of the 10th Eur. Workshop MAAMAW 2001 (2001)

[30] Fox, M.S., Barbuceanu, M., Gruniger, M.: An organisation ontology for enterprise modeling: Preliminary concepts for linking structure and behaviour. Computers in industry 29, 123–134 (1996)

[31] Goldblatt, R.: Logics of Time and Computation, 2nd edn CSLI Lecture Notes 7 (1992)

[32] Hayashida, T., Nishizaki, I., Katagiri, H.: Artificial adaptive agent model characterized by learning and fairness in the ultimatum games. J. of Telec. and Inf. Technology 4, 36–44 (2007)

[33] Herlea, D.E., Jonker, C.M., Treur, J., Wijngaards, N.J.E.: Specification of Behavioural Requirements within Compositional Multi-Agent System Design. In: Garijo, F.J., Boman, M. (eds.) MAAMAW 1999. LNCS, vol. 1647, pp. 8–27. Springer, Heidelberg (1999)

[34] Hoogendoorn, M., Jonker, C.M., Schut, M.C., Treur, J.: Modeling Centralized Organization of Organizational Change. Comp. and Math. Organization Theory 13, 147–184 (2007)

[35] Hoogendoorn, M., Schut, M.C., Treur, J.: Modeling Decentralized Organizational Change in Honeybee Societies. In: Almeida e Costa, F., Rocha, L.M., Costa, E., Harvey, I., Coutinho, A. (eds.) ECAL 2007. LNCS (LNAI), vol. 4648, pp. 615–624. Springer, Heidelberg (2007)

[36] Hoogendoorn, M., Treur, J.: An Adaptive Multi-Agent Organization Model Based on Dynamic Role Allocation. In: Nishida, T. (ed.) Proc. of the 2006 IEEE/WIC/ACM Int. Conf. on Intelligent Agent Technology (IAT 2006), pp. 474–481. IEEE CS Press, Los Alamitos (2006)

[37] Jaffry, S.W., Treur, J.: Agent-Based and Population-Based Simulation: a Comparative Case Study for Epidemics. In: Louca, L.S., Chrysanthou, Y., Oplatkova, Z., Al-Begain, K. (eds.) Proc. of the 22th European Conference on Modelling and Simulation, European Council on Modeling and Simulation, ECMS 2008, pp. 123–130 (2008)

[38] Jennings, N.R.: Controlling cooperative problem solving in industrial multi-agent systems using joint intentions. Artificial Intelligence 75, 195–240 (1995)

[39] Jonker, C.M., Treur, J.: A temporal-interactivist perspective on the dynamics of mental states. Cognitive Systems Research 4, 137–155 (2003)

[40] Jonker, C.M., Treur, J.: Compositional Verification of Multi-Agent Systems: a Formal Analysis of Pro-activeness and Reactiveness. Int. J. Coop. Inf. Sys. 11, 51–92 (2002)

[41] Jonker, C.M., Treur, J., Wijngaards, W.C.A.: Specification, Analysis and Simulation of the Dynamics Within an Organisation. J. of Applied Intelligence 27, 131–152 (2007)

[42] Kelso, J.A.S.: Dynamic Patterns: The Self-Organization of Brain and Behavior. MIT Press, Cambridge (1995)

[43] Kermack, W.O., McKendrick, W.G.: A contribution to the mathematical theory of epidemics. Proc. of the Royal Society of London, Series A 115, 700–721 (1927)

[44] Kim, J.: Philosophy of Mind. Westview Press, Boulder (1996)

[45] Kinny, D., Ljunberg, M., Rao, A., Sonenberg, L., Tidhar, G., Werner, E.: Planned team activity. In: Castelfranchi, C., Werner, E. (eds.) MAAMAW 1992. LNCS, vol. 830, pp. 226–256. Springer, Heidelberg (1994)

[46] Laird, J.E., Newell, A., Rosenbloom, P.S.: Soar: an architecture for general intelligence. Artificial Intelligence 33, 1–64 (1987)

[47] van Lamsweerde, A.: Requirements engineering: from System Goals to UML Models to Software Specifications. Wiley, Chichester (2009)

[48] Lapouchnian, A., Lespérance, Y.: Modeling Mental States in Agent-Oriented Requirements Engineering. In: Dubois, E., Pohl, K. (eds.) CAiSE 2006. LNCS, vol. 4001, pp. 480–494. Springer, Heidelberg (2006)

[49] Lotka A.J.: Elements of Physical Biology. Reprinted by Dover in 1956 as Elements of Mathematical Biology (1924)

[50] Manna, Z., Pnueli, A.: Temporal Verification of Reactive Systems: Safety. Springer, Heidelberg (1995)

[51] Oprea, M.: An adaptive negotiation model for agent-based electronic commerce. Studies in Informatics and Control 11, 271–279 (2002)

[52] Parunak, H.V.D.: Go to the Ant: Engineering Principles from Natural Multi-Agent Systems. Ann. Oper. Res. 75, 69–101 (1997)

[53] Parunak, H.V.D., Brueckner, S., Sauter, J.: Synthetic pheronome mechanisms for coordination of unmanned vehicles. In: Proceedings of AAMAS 2002, pp. 448–450. ACM Press, New York (2002)

[54] Parunak, H.V.D., Savit, R., Riolo, R.L.: Agent-Based Modeling vs. Equation-Based Modeling: A Case Study and Users' Guide. In: Sichman, J.S., Conte, R., Gilbert, N. (eds.) MABS 1998. LNCS (LNAI), vol. 1534, pp. 10–25. Springer, Heidelberg (1998)

[55] Pavon, J., Gomez-Sanz, J.: Agent Oriented Engineering with INGENIAS. In: Mařík, V., Müller, J.P., Pěchouček, M. (eds.) CEEMAS 2003. LNCS (LNAI), vol. 2691, pp. 392–403. Springer, Heidelberg (2003)

[56] Port, R.F., van Gelder, T. (eds.): Mind as Motion: Explorations in the Dynamics of Cognition. MIT Press, Cambridge (1995)

[57] Maynard Smith, J.: Models in Ecology. Cambridge University Press, Cambridge (1974)

[58] Spoelstra, M., Sklar, E.: Agent-Based Simulation of Group Learning. In: Antunes, L., Paolucci, M., Norling, E. (eds.) MABS 2007. LNCS (LNAI), vol. 5003, pp. 69–83. Springer, Heidelberg (2008)

[59] Stirling, C.: Modal and Temporal Properties of Processes. Springer, Heidelberg (2001)

[60] Sun, R.: Duality of the Mind. Lawrence Erlbaum Associates, Mahwah (2002)

[61] Sun, R.: The CLARION cognitive architecture: Extending cognitive modeling to social simulation. In: Sun, R. (ed.) Cognition and Multi-Agent Interaction. Cambridge Univ. Press, New York (2006)

[62] Volterra, V.: Variations and fluctuations of the number of individuals in animal species living together. In: Animal Ecology,McGraw-Hill, New York Translated by R. N. Chapman (1931)

[63] Wang, X., Lespérance, Y.: Agent-Oriented Requirements Engineering Using ConGolog and i*. In: Wagner, G., Karlapalem, K., Lespérance, Y., Yu, E. (eds.) Proceedings of the 3rd International Bi-Conference Workshop AOIS-2001 Agent-Oriented Information Systems 2001, pp. 59–78. iCue Publishing, Berlin (2001)

[64] Yu, E., Du Bois, P., Dubois, E., Mylopoulos, J.: From Organization Models to System Requirements: A Cooperating Agents Approach. In: Papazoglou, M.P., Schlageter, G. (eds.) Cooperative Information Systems: Trends and Directions, pp. 293–312. Academic Press, London (1997)

[65] Zambonelli, F., Jennings, N.R., Wooldridge, M.: Organisational Abstraction for the Analysis and Design of Multi-Agent Systems. In: Ciancarini, P., Wooldridge, M.J. (eds.) AOSE 2000. LNCS, vol. 1957, pp. 235–251. Springer, Heidelberg (2001)

Improving Performance of Protein Structure Similarity Searching by Distributing Computations in Hierarchical Multi-Agent System*

Alina Momot[1], Bożena Małysiak-Mrozek[1], Stanisław Kozielski[1],
Dariusz Mrozek[1], Łukasz Hera[1], Sylwia Górczyńska-Kosiorz[2],
and Michał Momot[3]

[1] Institute of Informatics, Silesian University of Technology,
Akademicka 16, 44-100 Gliwice, Poland
{alina.momot,bozena.malysiak,stanislaw.kozielski,dariusz.mrozek}@polsl.pl,
lukasz.hera@gmail.com

[2] Clinical Department of Nephrology, Diabetology and Internal Diseases,
Medical University of Silesia, Poniatowskiego 15, Katowice, Poland
sgorczynska-kosiorz@sum.edu.pl

[3] Institute of Medical Technology and Equipment,
Roosevelta 118, 41-800 Zabrze, Poland
michal.momot@itam.zabrze.pl

Abstract. Since protein structure similarity searching is very complex and time-consuming, one of the possible acceleration methods is parallelization by distributing the calculation on multiple computers. In the paper, we present a theoretical model of the hierarchical multi-agent system dedicated to the task of protein structure similarity searching. We also show results of several numerical experiments confirming a suitability of such distribution for the similarity searching performed for the Muconate Lactonizing Enzyme (PDB ID = 1MUC) from the Protein Data Bank (PDB) against the database containing almost thousand randomly chosen molecules.

Keywords: proteins, protein structure, similarity searching, multi-agent system, distributed computing.

1 Introduction

Proteins are molecules of life in all living organisms [1]. Since they act as substrates in all biochemical reactions, they are very important molecules in organisms' cells. Functions of proteins strictly depend on the internal construction of proteins, built up with hundreds of amino acids and therefore, thousands of

* Scientific research supported by the Ministry of Science and Higher Education, Poland in years 2008-2010, Grant No. N N516 265835: *Protein Structure Similarity Searching in Distributed Multi Agent System.*

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 320–329, 2010.

atoms [2,3]. Therefore, on the basis of the information regarding protein internal construction, we are able to identify functions of newly discovered proteins [4,5]. This is very important for the understanding of evolution of organisms, since we are able to observe how they differentiated during thousands of years.

The identification of proteins can be carried at different levels of protein organization - amino acid sequence (primary structure) and spatial structure (including secondary, tertiary and quaternary structure) [2] and is usually done by comparing a molecule with a database of proteins, looking for structural similarities [4]. Since protein spatial structures provide more details regarding protein construction, they are widely accepted in the analysis of protein activity in cellular reactions and in studies of their ability to interact with other molecules.

Protein structure similarity searching is a process of exploring a database of proteins in order to find molecules having an identical or related structure. In the process, we want to find certain relationships between particular regions of protein structures, regardless of the conformational changes that could appear as a result of environmental factors or binding the molecule to a ligand or other molecules. However, protein structure similarity searching is very complex and time-consuming. The main reason of this is that the construction of proteins is complex on its own. Existing methods in the area, like VAST [6], DALI [7], CE [8], LOCK2 [9], PFSC [10], FATCAT [11], FAST [12] and others, usually reduce the complexity of protein structures by representing them in much simpler form. Afterwards, they seek similarities using a pair-wise comparison of the given molecule to the subject molecule from a database.

Another significant problem is a rising number of protein structures in a world-wide repositories, like the Protein Data Bank [13], which now consists of 65 527 structures (May 25, 2010). This also slows down the process of similarity searching and causes additional delays in the identification of protein functions, since every time we have more molecules to compare in a database.

Due to these problems, the scientific community still seeks appropriate and efficient methods for protein structure similarity searching. The parallelization of the similarity searching is one of the techniques, which can be used to accelerate the search process. ProteinDBS [14] and ProCKSI [15] are examples of such systems, which make use of parallelization of the computational procedure in order to increase the speed of the search process. ProteinDBS is a real-time retrieval system for protein structure comparison. It uses Entropy Balanced Statistical k-d tree indexing and employs a resource directory to manage a collection of distributed index agents. These components handle various tasks including index organization, load balancing, database indexing and retrieval. ProCKSI is a distributed framework for protein structure comparison, which bases on the multiprocessor environment. The system is able to run both on a parallel environment using the MPI libraries and on a grid computing environment using the MPICH-G2 libraries [16].

In the paper, we show a distributed protein structure similarity searching with the use of a hierarchical multi-agent system. In the presented system, we distribute the computational procedure and data sets that are used in the search process. Using multi-agent system, we do not need any expensive hardware solutions. On the contrary, we can build a powerful framework ourselves on basis on standard PC computers. This gives us the possibility to raise the power of our framework at any moment by inviting other computers or farms of workstations to join the search process.

2 Theoretical Model

2.1 General Assumptions

In 1996 Franklin and Graesser defined the essence of being an agent as [17]: [illegible] *ff* [illegible] and thus the description of an autonomous agent should contain the environment, sensing capabilities, actions, drives and action selection architecture.

There is no universally accepted definition of the term agent, but there are two key characteristics of agents. First, agent should be autonomous, i.e. it can decide for itself what is needed to do in order to achieve the goals. Second, agent should be capable of communication with other agents and cooperation to solve problems that are beyond the capabilities of an individual agent. Multi-agent system (MAS) is composed of multiple interacting agents, which are computing entities being active objects that can migrate across machines. The major characteristics of multi-agent systems are [18]:

- each agent has just incomplete information and is restricted in its capabilities,
- system control is distributed,
- data is decentralized,
- computation is asynchronous.

Moreover MAS as distributed system offers the ability to increase computation speed (agents can operate asynchronously and in parallel) as well as reliability (failure of one or several agents does not necessarily lead to failure of achieving the goal of the whole system).

Multi-agent systems can differ in the agents themselves, the interactions among the agents and the environments in which the agents act. There are many industrial and commercial applications for MAS, for example in electronic markets, transportation systems, real-time monitoring of traffic or automated meeting scheduling, where agents act on behalf of their users to fix meeting details [19].

Multi-agent systems are often extremely complex to formally verify their properties and the only viable method to precisely study their properties is simulation. Recently, there have been developed a wide range of MAS toolkits and testbeds that can be applied in different domains. The simulation engines integrated into selected toolkits to facilitate the execution of the resulting MAS models are described in [20].

The use of such systems for analysis, design, and development for complex domains is growing rapidly [21], [22], [23]. Typical agent properties, methods and architectures are described in [24], which gives examples of historical and state-of-the-art contributions to diverse multi-agent systems. Interesting model for organizational design and a conceptual framework to classify organizational simulations is introduced in [25], which uses an ontology to describe organizational structures, environmental characteristics, and agent capabilities.

There are also multi-agent systems, which can distinguish certain classes of agents because of their functions, such as hierarchical MAS, for example hierarchical multi-agent system of taking global decisions on the basis of decision rules of local experts presented in [26], a multi-agent system for hierarchical control with self-organising database presented in [27] or a multiagent system for hierarchical control and monitoring of a complex process control system in [28].

A hierarchical multi-agent system for a distributed protein structure similarity searching will be presented below. In the presented system, the searching procedure and data sets are distributed on different types of agents.

2.2 Structure of Hierarchical MAS

Proposed structure of the hierarchical multi-agent system, which implements the protein structure similarity searching consists of three types of agents:

- **Supervisory Agent**, which is responsible for communication with a user, distribution of data for further processing over Control Agents and merging the results of the similarity searching process received from Control Agents,
- **Control Agents**, which are responsible for communication with Supervisory Agent and distribution of searching tasks over Searching Agents,
- **Searching Agents**, which evaluate the similarity measures between two given proteins.

Let C be a finite set of Control Agents $C = \{C_i : i = 1, \ldots, I\}$. Each Control Agent C_i supervises a finite set of Searching Agents $S_i = \{S_{ij} : j = 1, \ldots, J\}$. Each Searching Agent S_{ij} has access to a finite set of methods of protein structure similarity searching, such as FATCAT, CE, LOCK2, and others. Using selected method, the Searching Agent performs a single pairwise comparison $V_{S_{ij}}(QP, DP)$ for the given Query Protein (QP) and Database Protein (DP), computing similarity measures and some additional information.

Each Searching Agent S_{ij} computes the similarity between the Query Protein and some of proteins from a database, which are determined by Control Agent C_i requesting the task. The Control Agent C_i sets

- Query Protein (as PDB file or PDB ID),
- method of searching and all required parameters,
- database address (for example as IP address),
- range of proteins in the database, which should be compared in a single search process.

Based on the settings defined above each S_{ij} returns a list of similarity searching results $L_{S_{ij}}(k)$ to the Control Agent C_i, where k is a number of returned list. If a Searching Agent returns first list to the Control Agent ($L_{S_{ij}}(1)$), the Control Agent gives him the next set of data to process until the entire range of data assigned to the Control Agent is processed.

Upon receiving a list from Searching Agents S_{ij} ($j = 1, \ldots, J$) the Control Agent C_i performs merging the result list $L_{S_{ij}}(k)$ and his own list SL_{C_i} (initially empty) into the new list SL_{C_i}, which is sorted from the most to the least similar protein to the Query Protein according to a global similarity measure. The length of the sorted list SL_{C_i} should not exceed the value B, which is specified by a user at the beginning of the search process. Although in the case, when some database proteins give (in comparison to Query Protein) the same similarity measures, the length may be greater. If the length is greater, i.e. $B + G$, the following condition must be fulfilled:

$$\forall g \in \{1, \ldots, G\} \quad Sim(QP, SL_{C_i}(B)) = Sim(QP, SL_{C_i}(B + g)), \tag{1}$$

where $Sim(p_1, p_2)$ is a global similarity measure between two proteins p_1 and p_2, and $SL_{C_i}(l)$ is the lth protein in the sorted list SL_{C_i}. The global similarity measure $Sim(p_1, p_2)$ is based on the result of pairwise comparison $V_{S_{ij}}(p_1, p_2)$ performed by Searching Agent S_{ij} that computes similarity measures depending on the selected method of protein structure similarity searching.

As following from the above description, the next two conditions must be fulfilled:

$$\begin{aligned} &\forall i \in \{1, \ldots, I\}\ \forall l \in \{1, \ldots, B + G_i - 1\} \\ &\qquad\qquad Sim(QP, SL_{C_i}(l)) \geq Sim(QP, SL_{C_i}(l + 1)) \end{aligned} \tag{2}$$

and

$$\begin{aligned} &\forall i \in \{1, \ldots, I\}\ \forall l \in \{1, \ldots, B + G_i\} \\ &\qquad\qquad SL_{C_i}(l) \in \bigcup_{j \in \{1, \ldots, J\}} \bigcup_{k \in \{1, \ldots, K_{ij}\}} \left\{L_{S_{ij}}(k)\right\} . \end{aligned} \tag{3}$$

When all the data from a database (assigned to the Control Agent C_i) are analyzed, the finally updated sorted list SL_{C_i} is returned to the Supervisory Agent, which creates final list SL of sorted molecules (merging the sorted lists received from all Control Agents C_i ($i = 1, \ldots, I$) according to the same procedure as described above in the case of Control Agent C_i) and returns it to the user.

The whole process should ensure fulfilling the following condition:

$$\forall d \in D\ \exists S_{ij} \quad V_{S_{ij}}(QP, d), \tag{4}$$

where D is a set of proteins, which should be compared and QP is the query protein chosen by a user. It means that each protein from databases chosen by user is compared with the query protein by a Searching Agent.

The main idea of the hierarchical multi-agent system is presented in figure 1. The User sets the Query Protein (QP), Method of protein structure similarity searching (M) in selected Databases (D) and parameter B - length of list of the most similar database proteins to the Query Protein. The information through

Graphical User Interface (GUI) goes to the Supervisory Agent. The Supervisory Agent divides the whole range of data (N proteins) and assigns a range of data to search to each of Control Agents ($N_1, \ldots, N_I$). Control Agents pass the task of searching to the Searching Agents dividing its range into smaller parts ($N_{11}, \ldots, N_{IJ}$). Each of the Searching Agents returns a list of similarity searching results ($L_{C_{11}}(k), \ldots, L_{C_{IJ}}(k)$), these lists are merged into new sorted lists of length B by each of Control Agents ($SL_{C_1}, \ldots, SL_{C_I}$). Finally the sorted lists are returned to the Supervisory Agent, which performs the last merging of the lists from all Control Agents into one sorted list (SL), which is returned to the User.

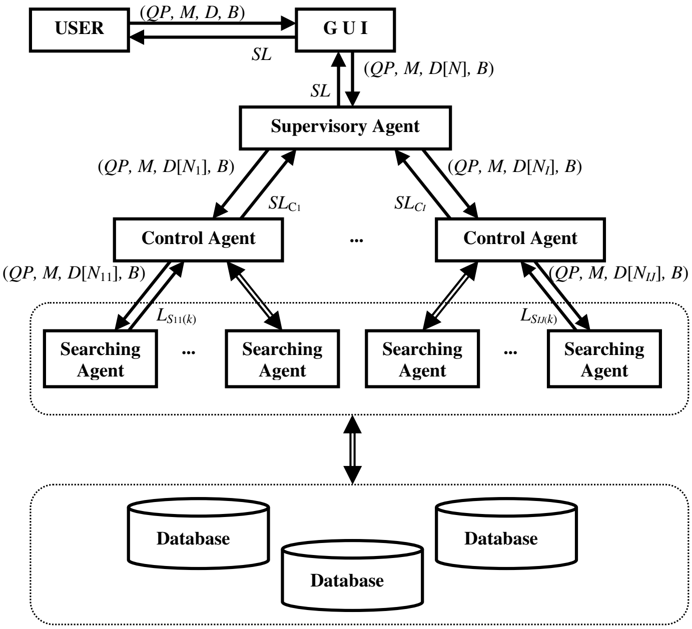


Fig. 1. Schema of structure of the hierarchical multi-agent system

In the special case, where $I = 1$, i.e. there is only one Control Agent in the system, the Supervisory Agent is not needed and its tasks are performed by the Control Agent. Different architectures may be applied depending on the situation.

The architecture containing one Supervisory Agent and many Control Agents is appropriate in each of the following cases:

- different groups of Searching agents perform different tasks, perform the same task using different methods or group of agents work on disjoint sets of data;

- a group of Searching Agents joins the ongoing calculations and will be controlled by a separate Control Agent;
- calculations involve a number of computers and the consolidation of results and communication between agents would be too burdensome for one Control Agent.

The architecture with one Control Agents merged with the Supervisory Agent is appropriate in each of the following cases:

- number of Searching Agents is relatively small, and coordination of communication is not a problem;
- we do not need any separate groups of agents due to a task performed, method used or data sets.

3 Numerical Experiments

We performed several tests in order to study the performance of the protein structure similarity searching in hierarchical multi-agent system presented in the previous section. Distributing computation accross many Searching Agents was destined to accelerate the searching process.

In our experiments, we used the JADE (Java Agent DEvelopment Framework), which is an open source platform for peer-to-peer agent based applications. The JADE simplifies the implementation of multi-agent systems through a given middle-ware and through a set of tools that supports debugging and deployment phases [29]. Searching and Control Agents run at PC computers with different, but similar, hardware components (CPU and RAM) and therefore, different computational possibilities. All computers were managed by the Windows XP Professional or Windows 7 operating system.

Each Searching Agent had an access to a finite set of methods of protein structure similarity searching, namely CE and FATCAT. However, in presented experiments we used only the CE as the method of the similarity searching M. As the Query Protein QP we used the Muconate Lactonizing Enzyme from the Protein Data Bank (PDB), which has two amino acid chains. The molecule was provided by the PDB ID (1MUC, PDB identifier). The searching process was performed on the subset of the PDB database containing 850 randomly chosen protein structures. The subset was distributed into many locations. In order to compare the QP to appropriate molecules from a PDB database, each Searching Agent established the connection to different databases on the basis of the address that was sent by the Control Agent. The parameter B - length of list of the most similar database proteins was set to 100.

In the performed experiments there was only one Control Agent (the number of Control Agents $I = 1$), therefore the Supervisory Agent was not needed and its tasks were performed by the Control Agent. The number of Searching Agents J varied from 1 to 20, i.e. they were equal: 1, 2, 4, 6, 8, 10 and 20. The numbers of proteins assigned to each of J Searching Agents in a single task were equal to each other and $N_{11} = N_{12} = \ldots = N_{1J} = 5$. The total number of proteins to search $N = N_1$ was set to 850.

Performed experiments show that the acceleration of searching as a function of the number of searching agents is close to identical function of the number Searching Agents as can be seen in the figure 2. The identical function is depicted by solid line and the particular values of acceleration are depicted by square markers.

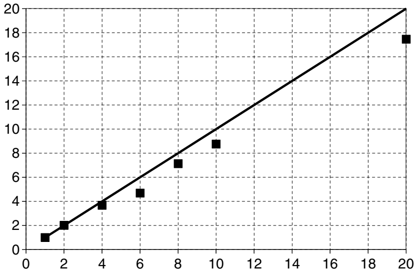


Fig. 2. Acceleration of searching as a function of the number of searching agents

For each experiment, where the number of searching agents J varied from 1 to 20, we measured the search time. The values are presented in table 1. The values in first two rows of this table are particularly interesting, because increasing agents number from 1 to 2 resulted in acceleration a little more than two times. This fact could be explained by a little better performance of the second agent than the performance of the first. In the case, where the number of searching agents increases above 4, the acceleration is still increasing, however slower than linearly. This could be explained by additional time required for communication between agents, as well as a little different computational capabilities of Searching Agents.

Table 1. Time of searching for varying Searching Agents number

Searching Agents number	time of searching [s]
1	28550
2	14217
4	7773
6	6090
8	4006
10	3260
20	1635

4 Concluding Remarks

By parallelizing the protein structure similarity searching we obtain results in shorter time, proportionally to the scaling the system horizontally by adding more working agents. Performed tests show that distributing the computational procedure across many computers results in significant acceleration of the search process. Since we do not use any other accelerating technique, like specific indexing, our method is not a real-time, like ProteinDBS [14]. However, during our tests we obtained better acceleration ratio than in the ProCKSI [15]. We also observed that with the rising number of working agents, the acceleration grows, but the dynamic of the speedup slightly decreases. Our conclusions in this regard are analogous with the conclusions presented in [16]. However, in a hierarchical multi-agent system shown in this article, the decline is not as significant and is probably the provision of the communication between agents.

Constructing the framework for our system, we used PC computers with comparable hardware parameters. However, we assume that any computer can join the similarity searching at any time, working as a Searching Agent. Even computers with a poor performance can join the framework. In the non-balanced environment, when we work with different types of computers due to hardware parameters, the Control Agent will distribute the work proportionally to the performance possibilities of particular agents. This is done by dividing the whole range of candidate proteins into small packages (eg. 5 proteins). Agents residing on faster computers return their comparison results relatively faster and obtain another package to work with, while agents residing on slower machines still process previous packages. Therefore, by assigning more work to faster computers, we optimize the time needed to carry out the exploration of protein structures.

Through the use of the multi-agent system we not only accelerate the search process, but also we raise the reliability of the system. Even disconnecting a Searching Agent from the whole environment or no response for some time will not stop the search process. Moreover, in case of having more than one Control Agent we increase the stability and reliability of the system.

References

1. Lodish, H., et al.: Molecular cell biology, 4th edn. W.H. Freeman and Company, New York (2001)
2. Allen, J.P.: Biophysical chemistry. Wiley-Blackwell, Chichester (2008)
3. Branden, C., Tooze, J.: Introduction to protein structure. Garland (1991)
4. Gibas, C., Jambeck, P.: Developing bioinformatics computer skills, 1st edn. O'Reilly, Sebastopol (2001)
5. Attwood, T.K., Parry-Smith, D.J.: Introduction to bioinformatics. Prentice Hall, Englewood Cliffs (1999)
6. Gibrat, J.F., Madej, T., Bryant, S.H.: Surprising similarities in structure comparison. Curr. Opin. Struct. Biol. 6(3), 377–385 (1996)
7. Holm, L., Sander, C.: Protein structure comparison by alignment of distance matrices. J. Mol. Biol. 233(1), 123–138 (1993)
8. Shindyalov, I.N., Bourne, P.E.: Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. 11(9), 739–747 (1998)
9. Shapiro, J., Brutlag, D.: FoldMiner and LOCK2: protein structure comparison and motif discovery on the web. Nucleic Acids Res. 32, 536–541 (2004)

10. Yang, J.: Comprehensive description of protein structures using protein folding shape code. Proteins 71(3), 1497–1518 (2008)
11. Friedberg, I., et al.: Using an alignment of fragment strings for comparing protein structures. Bioinformatics 23(2), 219–224 (2007)
12. Zhu, J.H., Weng, Z.P.: FAST: A novel protein structure algorithm. Proteins 58, 618–627 (2005)
13. Berman, H.M., et al.: The Protein Data Bank. Nucleic Acids Res. 28, 235–242 (2000)
14. Shyu, C.-R., et al.: ProteinDBS: a Real-Time Retrieval System for Protein Structure Comparison. Nucleic Acids Research 32, 572–575 (2004)
15. Barthel, D., et al.: ProCKSI: a Decision Support System for Protein (Structure) Comparison, Knowledge, Similarity and Information. BMC Bioinformatics 8(416), 1–22 (2007)
16. Folino, G., et al.: Towards a Distributed Framework for Protein (Structure) Comparison, Knowledge, Similarity and Information (ProCKSI). In: UK eScience All Hands Meeting (AHM 2008), Edinburgh, UK, pp. 1–2 (2008)
17. Franklin, S., Graesser, A.: Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. In: Jennings, N.R., Wooldridge, M.J., Müller, J.P. (eds.) ECAI-WS 1996 and ATAL 1996. LNCS, vol. 1193, pp. 21–35. Springer, Heidelberg (1997)
18. Jennings, N.R., Sycara, K., Wooldridge, M.: A roadmap of agent research and development. Autonomous Agents and Multi-Agent Systems 1, 7–38 (1998)
19. Weiss, G. (ed.): Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence. MIT Press, Cambridge (1999)
20. Theodoropoulos, G.K., Minson, R., Ewald, R., Lees, M.: Simulation Engines for Multi-Agent Systems. In: Uhrmacher, A., Weyns, D. (eds.) Multi-agent systems: simulation and applications, pp. 77–105. CRC Press, Boca Raton (2009)
21. Cavedon, L., Rao, A., Wobcke, W. (eds.): PRICAI-WS 1996. Lecture Notes in Computer Science, LNAI, vol. 1209. Springer, Heidelberg (1997)
22. Jain, L.C., Chen, Z., Ichalkaranje, N. (eds.): Intelligent agents and their applications. Physica-Verlag, Heidelberg (2002)
23. Ghose, A., Governatori, G., Sadananda, R. (eds.): Agent Computing and Multi-Agent Systems. LNCS (LNAI), vol. 5044. Springer, Heidelberg (2009)
24. Tolk, A., Uhrmacher, A.M.: Agents: Agenthood, Agent Architectures, and Agent Taxonomies. In: Yilmaz, L., et al. (eds.) Agent-Directed Simulation and Systems Engineering, pp. 75–109. Wiley-VCH, Weinheim (2009)
25. Dignum, V., Dignum, F., Sonenberg, L.: Design and Analysis of Organization Adaptation in Agent Systems. In: Yilmaz, L., et al. (eds.) Agent-Directed Simulation and Systems Engineering, pp. 237–267. Wiley-VCH, Weinheim (2009)
26. Wakulicz-Deja, A.: Przybyla-Kasperek M.: Hierarchical Multi-Agent System. In: Klopotek, M.A., et al. (eds.) Recent Advances in Intelligent Information Systems, pp. 615–628 EXIT, Warsaw (2009)
27. Choinski, D., Nocon, W., Metzger, M.: Multi-agent system for hierarchical control with self-organising database. In: Nguyen, N.T., Grzech, A., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2007. LNCS (LNAI), vol. 4496, pp. 655–664. Springer, Heidelberg (2007)
28. Tan, V.V., Yoo, D.S., Shin, J.C., Yi, M.J.: A Multiagent System for Hierarchical Control and Monitoring. J. UCS 15(13), 2485–2505 (2009)
29. Bellifemine F., et al.: JADE, A White Paper (2003), http://jade.tilab.com/papers/2003/WhitePaperJADEEXP.pdf

Modeling Super Mirroring Functionality in Action Execution, Imagination, Mirroring, and Imitation

Monique Hendriks and Jan Treur

VU University Amsterdam, Department of Artificial Intelligence
De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
M.Hendriks@student.tue.nl, treur@cs.vu.nl
http://www.few.vu.nl/~treur

Abstract. In this paper a cognitive agent model is presented that models multiple functions of preparation states: mirroring an observed action, imitation of an action, or imagining an action. The model incorporates a super mirroring function enabling the agent to keep track of the context of a mental process: mirroring another agent, imitation of another agent, action imagination, or own action performance. These cognitive functions have been adopted from informal descriptions of mirror neuron and super mirror neuron functions in neurological literature. Example simulations are presented that illustrate the model.

Keywords: Cognitive model, super mirror neurons, imitation, imagination.

1 Introduction

Within an agent's internal neurological processes, sensory representations of stimuli usually lead to preparations for responses. Recent neurological findings reveal that so called 'preparation' neurons have multiple functions; preparing for an action to be executed is only one of these functions. For example, preparation of actions may play a role in imagination, or in interpreting an observed action. In these cases, actual execution of the prepared action is suppressed. This multi-tasking aspect of preparation states, which from an evolutionary perspective can be viewed as an economic use of available resources, entails a fundamental problem concerning the representational content of such neurons. Apparently, activation of such neurons by itself has no unambiguous meaning; it is strongly context-dependent. A way out of this problem of multi-interpretability of preparation neurons is obtained when suitable forms of context can be defined. Indeed this is what is assumed to happen at the neurological level (e.g., [8], pp. 196-203, [11]), based on what are called super mirror neurons. These are neurons which were found to have a function in allowing or suppressing action execution after preparation has taken place.

This paper presents a cognitive agent model displaying the (re)use of preparation states in the processes of imagining, imitating or mirroring an action. It is shown that the use of a context provides the agent with the capability of differentiating between the different uses of the same preparation state. In Section 2 the main principles adopted are discussed. In Section 3 the cognitive agent model is described in more detail. Section 4 illustrates the functioning of the model by showing simulation results

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 330–342, 2010.

for different contexts, and Section 5 addresses a mathematical analysis. Finally, Section 6 provides a discussion.

2 Principles Adopted

This paper presents a model with four different functions for the preparation of the same action. The shared representation of action preparations in these four cases is supported in the literature by the concepts of *inner simulation* (cf. Hesslow, 2002), the as-if body loop (cf. [3, 4]) and mirroring (cf. [8, 12, 15]). Preparations for responses can lead to further mental processing via an *as-if action execution loop* from preparation state to sensory representation of the expected outcome; cf. [6]. When inner simulation is combined with the inhibition of action execution, a preparation for an action can be used to simulate the results of an action as if it were actually executed; the concept of inner simulation supports a role for preparation neurons in imagination. Such sensory representations of expected outcomes induce preparation states for a specific bodily reaction to this outcome. According to Damasio (e.g., [3, 4]) a bodily reaction can be viewed as an emotional response; sensing changes in the body state lead to feeling these emotions (*body loop*). However, the preparation for a bodily reaction can also directly lead to a sensory representation of the bodily reaction without the bodily reaction actually taking place (*as-if body loop*). For the cognitive agent model presented here, the following causal chain is assumed; see [3, 4, 6]:

> sensory representation of stimulus → preparation for response →
> sensory representation of expected outcome → preparation for body reaction →
> sensory representation of body state.

This causal chain is extended to a recursive loop by assuming that the preparation for the response is also affected by the level of feeling the emotion associated to the expected outcome of the response:

> sensory representation of body state → preparation for response

Within the agent model presented in this paper, states are assigned a quantitative (activation) level. The positive feedback loops between preparation states for responses and their associated body states, and the sensory representations of expected outcomes are triggered by a sensory representation of a stimulus and converge to a certain level of feeling and preparation.

The discovery of mirror neurons has revealed that preparation states for own actions may also be used in the recognition of actions of others (e.g., [8]). Mirror neurons are motor neurons that fire not only when an action is performed, but also when the same action is observed. So when the stimulus is an observed action, then a preparation state for that action is created and through the as-if action execution loop and the as-if body loop the expected outcome and the feelings associated with that action are recognised. The existence of mirror neurons provides support for the role of preparation neurons in action recognition through mirroring, and in imitation. When the preparation state has a mirroring function or is created through imagination, the prepared action will not be executed. Therefore in the step from preparation to actual response, in addition to the preparation level another factor has to be taken into account. Within neurological literature such as [8], pp. 196-203, and [11], such a factor is assumed to be realised by a specific type of mirror neurons, called *super mirror neurons* [8] to keep track of the context in which a mental process takes place. In single cell recording experiments with epileptic patients [11], neurons were found that are active when the person prepares an own action to be executed, but shut down when the action is only observed, which suggests that these cells may be involved in the distinction between a self-generated preparation state and a preparation state generated through observation of an action [8]. In [8], pp. 201-202, it is also described that certain cells are sensitive to a specific person, so that in the case of an observed action, this action can also be attributed to the specific person that was observed.

Within the cognitive agent model presented in the next section, the functions of super mirror neurons have been incorporated in focus states, generated by processing of available (sensory) context information. For the case modeled, this focus can refer to the person her or himself, observing another person's action, imitating an action, or imagination of performing an action. When the focus is on action execution or imitation, this has a positive effect on performing the action. When the focus is on imagination, or recognizing another person's action, it has a suppressing effect. In the latter case the preparations and related feeling generated through simulation can be used to attribute them to that person.

3 Specification of the Cognitive Agent Model

To formalise the agent model, the hybrid dynamic modeling language LEADSTO has been used; cf. [1]. Within LEADSTO the dynamic property $a \rightarrow_D b$ denotes that when a state property a occurs, then after a certain time delay (specified as any positive real number D), state property b will occur. This D will be taken as the time step Δt, and not be mentioned explicitly. Both logical and quantitative calculations can be specified, and a software environment is available to support specification and simulation.

Fig. 1 shows a graphical representation of the agent model. In this figure, the circles represent state properties. Occurrence of one or more state properties at a certain point in time can lead to the occurrence of another state property at the next point in time, as represented by the labeled arrows. The state properties in this model contain variables. The state property sensor_state(c[,a]) represents the sensing of a context. Variable c denotes the specific context and optional variable a denotes the agent to which the action preparation state to be generated should be attributed (in the case of mirroring or imitation, the action preparation should be attributed to another agent). State property srs(c[,a]) represents the sensory representation that is generated for a sensed context. Again, variable c denotes the specific context and optional variable a denotes the agent to which the to be generated action preparation state should be attributed. State property focus(c[,a]) represents a focus state for a specific context and agent (optional) that is generated in response to a sensory representation of a context. State property attribution(b,V,a) represents the attribution of a body state b that was generated with strength V to agent a. Such an attribution state is created in case of mirroring or imitation. State properties sensor_state(s,V) and srs(s,V) represent the sensing and the generation of a sensory representation of a stimulus s that is sensed with strength V. When another agent is being mirrored or imitated, the stimulus needs to be transformed before it can be translated into the preparation for an action.

For example when the agent is imitating another agent who is lifting a box, then the sensed stimulus will be another agent lifting a box. In order to be able to generate the same action, lifting a box, the agent will imagine himself being at the location of the other agent and seeing a box in front of him. This state is represented by state property transformed_stimulus(s,V), where variable s may have an instanstiation (e.g., 'box') different from s in state property srs(s,V) (e.g., 'agent a lifting a box').

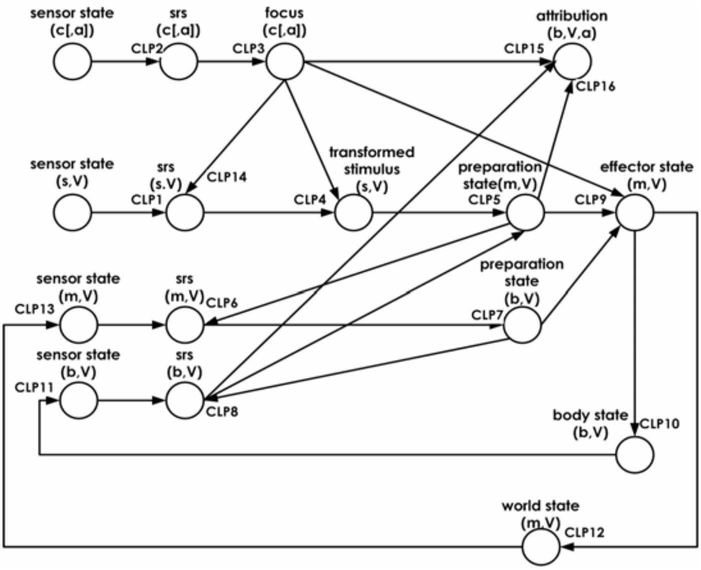


Fig. 1. Overview of the cognitive agent model

State property preparation_state(m,V) represents the state in which execution of motor plan m with strength V is prepared. State property effector_state(m,V) represents the actual execution of motor plan m with strength V. State properties sensor_state(m,V), srs(m,V), sensor_state(b,V) and srs(b,V) represent the sensing and sensory representation of execution of motor plan m with strenght V, and the sensing and sensory representation of body state b with strength V respectively. The occurrence of a sensory representation of a motor plan being executed leads to the preparation for the occurrence of a specific body state. This is denoted by state property preparation_state(b,V), with b being this specific body state and V the strength of the preparation. Finally, body_state(b,V) and world_state(m,V) represent the externally observable occurrence of a body state b with strength V and the observable execution of motor plan m with strength V respectively.

The dynamic properties, or temporal relations between these state properties are denoted by the labelled arrows in Figure 1. Dynamic properties CLP1 and CLP4 describe perception of a stimulus and its transformation.

CLP1 Generating a sensory representation for a sensed stimulus

if stimulus s is sensed with strength *V*
then a sensory representation for s with strength *V* will occur
sensor_state(s,V) → srs(s,V)

A sensory representation of stimulus s1 is transformed to a stimulus s2 when the observed stimulus is a person performing an action and the focus is to imitate or mirror this person, as was described above for the case of imitating another agent who is lifting a box. In the case of imitation or mirroring, the agent needs to imagine that he is 'standing in the shoes' of the agent that he is imitating or mirroring. Indeed neurological evidence has been found in support of this hypothesis that humans mentally displace themselves such that they literally put themselves in the shoes of the other, when they imitate or mirror the behavior of another person [10]. When the focus is to perform a self-initiated action or to imagine an action, then the stimulus is not transformed.

CLP4 Generating a transformed representation of the stimulus

if a sensory representation of stimulus s1 occurs with strength *V*
and the focus state is imitate or to mirror agent a's behavior
and transformation of stimulus s1 from the third person perspective to the first person perspective results in stimulus s2
then a transformed stimulus s2 will occur with strength *V*
srs(s1,V) & (focus(imitation,a) or focus{mirroring,a)) → transformed_stimulus(s2,V)
if a sensory representation of stimulus s occurs with strength *V*
and the focus state is to execute a self-initiated action or to imagine an action
then transformed stimulus s will occur with strength *V*
srs(s,V) & (focus(action_execution) or focus(imagination)) → transformed_stimulus(s,V)

Properties CLP5 and CLP9 define how stimulus s leads to a response in the form of the preparation and execution of a motor plan m. A preparation state for a specific motor plan is created or updated based on the transformed stimulus representation, the current state of feeling and the current state of preparation.

CLP5 Generating a preparation state

if transformed stimulus representation s occurs with strength V_1
and sensory representation for body state b occurs with strength V_2
and the current preparation state for motor plan m has strength V_3
then the updated preparation state for m has strength $V_3 + \gamma(h(\beta, \omega_1, \omega_2, V_1, V_2) - V_3)\, \Delta t$
transformed_stimulus(s,V1) & srs(b,V2) & preparation_state(m,V3)
→ preparation_state(m,V3+γ(h(β,ω1, ω2,V1,V2)-V3)Δt)

The level of preparation V_3 is updated based on a function $h(\beta, \omega_1, \omega_2, V_1, V_2)$ of the levels of strength of the stimulus and body state associated with motor plan m, defined as follows:

$$h(\beta, \omega_1, \omega_2, V_1, V_2) = \beta(1-(1-\omega_1 V_1)(1-\omega_2 V_2)) + (1-\beta)\, \omega_1 \omega_2 V_1 V_2$$

Parameter β models a person's characteristic for emotional response (from *0* as weakest response to *1* as strongest response): in the loop in which the strength of the preparation state for m is iteratively updated, β determines to what degree the strength of the preparation state for m is amplified (here $0 \leq \beta \leq 1$, which keeps the strenght within [*0*,*1*]). Parameter γ describes the rate of growth of the update function and parameters ω1 and ω2 provide an option for differential weighting of strenghts of the stimulus and the associated feeling in the update process. This function is also used in the update process of the strength of sensory representations of the outcome of an action (CLP6) and the body state resulting from an action (CLP8). Furthermore it is used in a slightly adapted manner in the update process of preparations for body states

(CLP7) and it is used to determine the strenght of an effector state of motor plan m by combining the strenghts of the preparation m and associated body state b (CLP9). The strength of the effector state for m is determined by a combination of the strengths of the preparation state for m (V_1) and the preparation state for b (V_2). An effector state only occurs when the focus is self-performed action or imitation.

CLP9 From preparation of action and preparation of body to action

if a preparation state for motor plan m occurs with strength V_1
and a preparation state for body state b occurs with strength V_2
and the focus is self-performed action or imitation
then an effector state for m occurs with strength $h(\beta, \omega_1, \omega_2, V_1, V_2)$

preparation_state(m,V1) & preparation_state(b,V2) &
(focus(action_execution) or focus(imitation,a)) → effector_state(m, h(β,ω1, ω2,V1,V2))

If motor plan m is executed, CLP10 and CLP11 describe how this leads to a feeling of the changes in the body state through the body loop [3]. The body state is first changed through property CLP10. The body state is then sensed through CLP11.

CLP10 From effector state to body state

if an effector state for motor plan m occurs with strength V
then a body state b will occur with strength V

effector_state(m,V) → body_state(b,V)

CLP11 Sensing a body state

if body state b occurs with strength V
then b will be sensed with strength V

body_state(b,V) → sensor_state(b,V)

However, even before a motor plan is executed, its predicted outcome already leads to a feeling through the as-if body loop [3] described by properties CLP6, CLP7 and CLP8. In response to a preparation state for motor plan m, a sensory representation of the predicted sensory outcome of m is formed (CLP6). This leads to a preparation state for the body state b associated with motor plan m (CLP7).

CLP6 Generating a sensory representation of a response outcome

if a preparation state for motor plan m occurs with strength V_1
and action m is sensed in the world with strength V_2
and the current sensory representation for m has strength V_3
then the updated sensory representation for m has strength $V_3 + \gamma(h(\beta, \omega_1, \omega_2, V_1, V_2) - V_3)\ \Delta t$

preparation_state(m,V1) & sensor_state(m,V2) & srs(m,V3) → srs(m, V3+γ(h(β,ω1, ω2,V1,V2)-V3)Δt)

The strength of the preparation of body state b associated with motor plan m is updated with rate γ. Again, β represents the orientation for emotional response.

CLP7 Generating a preparation of a body state

if a sensory representation of motor plan m occurs with strength V_1
and the current preparation state for body state b has strength V_2
then the updated preparation state for b has strength $V_2 + \gamma(\omega_1 V_1 - V_2)\ \Delta t$

srs(m,V1) & preparation_state(b,V2) → preparation_state(b, V2+γ(ω1V1 - V2)Δt)

Property CLP8 describes the update of the strength of a sensory representation for body state b, based on the strengths of the preparation state for b, the sensor state for b and the current strength of the sensory representation for b in a similar manner as described above for CLP5.

CLP8 Generating a sensory representation of the body state

if a preparation state for body state b occurs with strength V_1
and body state b is sensed with strength V_2
and the current sensory representation of b has strength V_3
then the updated sensory representation of b has strength $V_3 + \gamma(h(\beta, \omega_1, \omega_2, V_1, V_2) - V_3)\, \Delta t$
preparation_state(b,V1) & sensor_state(b,V2) & srs(b,V3) → srs(b, V3+γ(h(β,ω1, ω2,V1,V2)-V3)Δt)

This preparation state can directly lead to a sensory representation of the changed body state (CLP8) as opposed to the indirect path from an effector state of m resulting in a sensory representation of the changed body state through the real sensation (CLP11) of an actual change in the body state (CLP10).

The actual outcome of the execution of motor plan m is observed through CLP12 and CLP13. The execution of an action changes the world in which the agent is acting throught CLP12.

CLP12 From effector state to world state

if an effector state for motor plan m occurs with strength V
then this action m will occur in the world with strength V
effector_state(m,V) → world_state(m,V)

These changes in the world can then again be observed throught CLP13.

CLP13 Sensing outcome of a motor plan

If action m occurs in the world with strength V
then m will be sensed with strength V
world_state(m,V) → sensor_state(m,V)

Finally, properties CLP2, CLP3, CLP4, CLP9, CLP14, CLP15, and CLP16 describe how a focus state arises. First the context of the action preparation is sensed. A context can consist of a self-performed action, imagination of an action, imitation of another agent, or mirroring another agent. In case of imitation or mirroring, the context also indicates which agent is imitated or mirrored (indicated below by rectangular braces). This leads to a sensory representation of the context (CLP2). This sensory representation leads to a focus state of either performing, imagining, imitating or mirroring an action (CLP3).

CLP2 Generating a sensory representation for a sensed context

if context c is sensed [and implies agency to agent a]
then a sensory representation for c [with agency a] will occur
sensor_state(c[,a]) → srs(c[,a])

A sensory representation of a context leads to a focus state of either self-performed action, imagination, imitation or mirroring.

CLP3 From sensory representation of a context to focus state

if a sensory representation of a context c occurs [with agency a]
then a focus state for c [with agency a] will occur
srs(c[,a]) → focus(c[,a])

In case of imitating or mirroring the identity of the agent that is to be imitated or mirrorred is also added to the focus state. This focus state can have several different effects. In the case of imagining an action, the focus state is responsible for activating a sensory representation of a stimulus that is not really there (CLP14, see [6]). When the focus is imagination, then a sensory representation of a stimulus is created.

CLP14 Imagining a stimulus

if the focus is imagination
then a sensory representation of imagined stimulus s will occur with strength *1*
focus(imagination) → srs(s,1)

In the case of imitating or mirroring of another agent, the focus state is responsible for enabling transformation of a sensory representation of a stimulus from a third person perspective to a first person perspective (CLP4, see [10]). In the case of action execution or imitation, the focus state is responsible for the creation of an effector state (CLP9), while in the case of imagining or mirroring, its responsibility is to inhibit the creation of an effector state. Finally, through CLP15 and CLP16, the focus state leads to the attribution of the preparation for an action (CLP16) and the corresponding feelings (CLP15) to the self or to another agent, according to the context. In case of imitation, the action and associated feeling are attributed to the self as well as to the other agent.

CLP15 Attribution of a feeling

if a sensory representation of body state b occurs with strength *V*
and the focus is self-performed action, imagination or imitation
then the feeling of b with strength *V* is attributed to the self
srs(b,V) & (focus(action_execution) or focus(imagination) or focus(imitation,a)) → attribution(b,V,self)

if a sensory representation of body state b occurs with strength *V*
and the focus is imitation or mirroring of agent a's behavior
then the feeling of b with strength *V* is attributed to agent a
srs(b,V) & (focus(imitation,a) or focus(mirroring,a)) → attribution(b,V,a)

CLP16 Attribution of an action

if a sensory representation of action m occurs with strength *V*
and the focus is self-performed action, imagination or imitation
then action m with strength *V* is attributed to the self
srs(m,V) & (focus(action_execution) or focus(imagination) or focus(imitation,a)) → attribution(m,V,self)
if a sensory representation of action m occurs with strength *V*
and the focus is imitation or mirroring of agent a's behavior
then action m with strength *V* is attributed to the self as well as to agent a
srs(m,V) & (focus(imitation,a) or focus(mirroring,a)) → attribution(m,V,a)

4 Simulation Results

Using the supporting LEADSTO software environment, for the formally specified model a number of simulations were run. The resulting simulation traces visualise the state properties of the model over time; e.g., see Figure 2. Time is depicted at the horizontal axis. The visualisation using bars indicates of the different state properties whether they are true or false at every time point. The graphs indicate for state properties containing a quantitative variable, the development of the (real) value of this variable over time. Four example simulation results will be briefly discussed to demonstrate the ability of the agent model to display self-performed action, action imagination, imitation and mirroring. As a scenario, the lifting of a box was considered. For all scenario's shown β was set to *1* (maximal emotional response), γ was set to *0.2* (moderate rate of change), and ω_1 and ω_2 were both set to *1* (equal weighing of both factors considered in the function *h*). Furthermore, the step size Δt was set to *1*. The strength of the stimuli (box and a person lifting a box) were both set to *0.5*. Figure 2 shows simulation traces for two example simulations.

Self-performed action. In the simulation results for the self-performed action of lifting a box first the stimulus - the box - is perceived by the agent: sensor_state(box, 0.5) is true from time point 0 up to time point 100. Then a sensory representation is formed; srs(box, 0.5) becomes true after 1 time step. Since the context is a self-performed action, the stimulus does not need to be transformed, therefore the transformed stimulus is equal to the sensory representation of the observed stimulus: transformed_stim(box, 0.5). Development takes place of the strengths of feeling srs(b,V), preparation for the motor plan preparation_state(m,V), preparation for the bodily response preparation_state(b,V), the effector state for the motor plan effector_state(m,V) and the attribution of the action and the feeling to the self, attribution(m,V,self) and attribution(b,V,self) respectively. The simulation shows that not only does a preparation state for a motor plan arise, but the effector state also obtains a strenght > 0. A little later in time, the preparation state for a bodily response arises (the as-if body loop, [3, 4]). The preparation of a body state leads to the development of a feeling (srs(b,V)), the actual occurrence of the body state contributes to the strength of the feeling. The action (motor plan m) as well as the feeling (body state b) are attributed to the self.

Action imagination. The simulation results for the imagination of the action of lifting a box (Figure 2(a)) show that the sensory representation for the stimulus (box) is formed in the absence of a sensor state for the stimulus. Furthermore, even though the preparation state for m reaches the same level as that in the scenario of self-performed action, the formation of an effector state is prevented by the context of imagination; the value of variable *V* in state property effector_state(m,V) remains 0.

Imitation. The simulation results for the imitation of the action of lifting a box show that the sensed stimulus of a person lifting a box is transformed to a first person perspective: a representation of the box. State property sensor_state(person_lifting_box, 0.5) is transformed to state property transformed_stim(box, 0.5). Furthermore, it is shown that the focus state contains not only the context of imitation, but also the identity of the agent to be imitated (agent x): focus(imitation, x). Consequently, the action is attributed to the self, as well as to agent x.

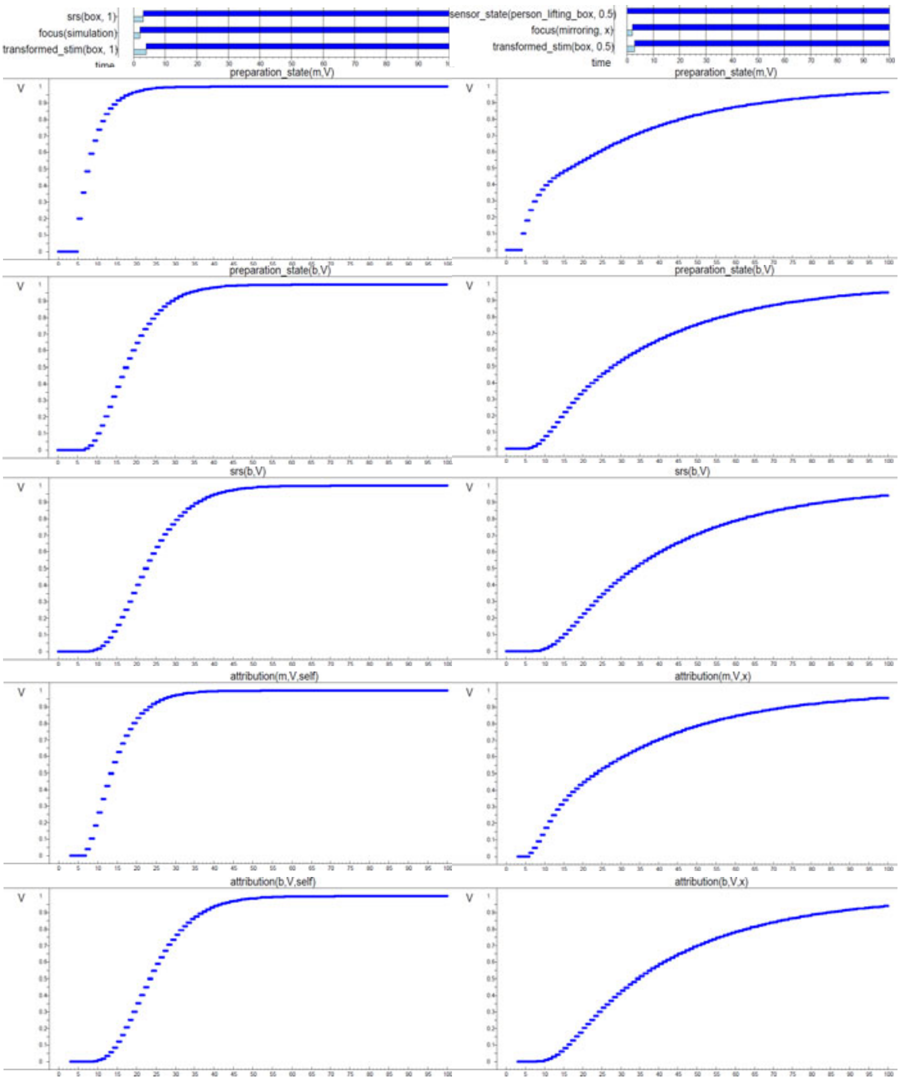


Fig. 2. Simulation traces: (a) imagination (left hand side) and (b) mirroring (right hand side)

Mirroring. The simulation results for the scenario of mirroring agent x's action of lifting a box (Figure 2(b)) show that again, the stimulus is transformed to a first person perspective and the action and feeling are attributed to agent x. Furthermore, the activation of an effector state for m is prevented by the context of mirroring. Therefore, in the case of mirroring there is no change in the body state b, since the action is not executed. Therefore, the development of the strengths of the preparation states and the associated feeling is slower. In the scenario of imagination there is also an absence of action execution.

5 Mathematical Analysis

In the example simulations discussed above it was shown that for a time period with a constant environment, the strengths of sensory representations, preparations, body states and feelings reach a stable equilibrium. By a mathematical analysis it can be addressed which types of equilibria are possible. To this end equations for equilibria were determined from the dynamical model equations. Here:

$ts(X, t)$	the level of the transformed sensory representation of X
$p(X, t)$	the level of the preparation of X
$srs(X, t)$	the level the sensory representation of X
$e(X, t)$	the level of the effector state for X
$ss(X, t)$	the level of the sensor state for X (for simplicity assumed 0 or 1)

The following equation models the transformed stimulus:

$$ts(w2, t) = 1 - (1 - ss(w1, t))\,(1 - ss(imagination, t))\,(1 - ss(action, t))$$

Note that this is 1 when one of the external factors $ss(w1, t)$, $ss(imagination, t)$ and $ss(action, t)$ is 1, and 0 otherwise. The dynamical model specifications can be expressed as differential equations as follows.

$$\begin{aligned}
\frac{dp(m,t)}{dt} &= \gamma(\beta(1\text{-}(1\text{-}ts(w2, t))(1\text{-}srs(b, t))) + (1\text{-}\beta)ts(w2, t)srs(b, t) - p(m, t))\\
\frac{de(m,t)}{dt} &= \gamma[(\beta(1\text{-}(1\text{-}p(m, t))(1\text{-}srs(w3, t)) + (1\text{-}\beta)p(m, t)srs(w3, t)]\\
&\qquad [1 - (1 - ss(imitation, t))\,(1 - ss(action, t))\,] - e(m, t))\\
\frac{dsrs(w3,t)}{dt} &= \gamma(\beta(1\text{-}(1\text{-}p(a, t))(1\text{-}e(a, t))) + (1\text{-}\beta)p(a, t)e(a, t) - srs(w3, t))\\
\frac{dsrs(b,t)}{dt} &= \gamma(\beta(1\text{-}(1\text{-}srs(w3, t))(1\text{-}e(m, t))) + (1\text{-}\beta)\, srs(w3, t)e(m, t) - srs(b, t))
\end{aligned}$$

To obtain equations for equilibria, constant values for all variables are assumed (also the ones that are used as inputs). Then in all of the equations the reference to time t can be left out, and in addition the derivatives can be replaced by 0. Assuming γ nonzero, this leads to four equations in $srs(b)$, $p(m)$, $srs(w3)$, $e(m)$, with externally determined $ss(c)$ and $ts(w2)$. For the case that $\beta = 1$, the following equations result

$$ts(w2) = 1 - (1 - ss(w1))\,(1 - ss(imagination))\,(1 - ss(action)) \tag{0}$$
$$(1\text{-}(1\text{-}ts(w2))(1\text{-}srs(b))) - p(m) = 0 \tag{1}$$
$$(1\text{-}(1\text{-}p(m)(1\text{-}srs(w3))))\ \ [1 - (1 - ss(imitation))\,(1 - ss(action))\,] - e(m) = 0 \tag{2}$$
$$(1\text{-}(1\text{-}p(m))(1\text{-}e(m))) - srs(w2) = 0 \tag{3}$$
$$(1\text{-}(1\text{-}\ srs(w3))(1\text{-}e(m))) - srs(b) = 0 \tag{4}$$

When both $\beta = 1$ and $ts(w2) = 1$, this results in the following:

$$ss(w1) = 1 \ \text{ or } \ ss(imagination) = 1 \ \text{ or } \ ss(action) = 1 \tag{0}$$
$$p(m) = 1 \tag{1}$$
$$e(m) = [1 - (1 - ss(imitation))\,(1 - ss(action))\,] \tag{2}$$
$$srs(w3) = 1 \tag{3}$$
$$srs(b) = 1 \tag{4}$$

This case is shown in the example traces in Figure 2.

6 Discussion

The cognitive agent model presented in this paper incorporates both a mirroring function and a super mirror function of preparation states. The mirroring function enables the agent to recognize actions of other agents, while experiencing feelings associated to these actions. The super mirroring function provides a control layer; it enables the agent to determine and keep track of the context of its mental processing, such as action imagination, mirroring another agent, imitation of another agent or own action execution. Depending on the context, the super mirroring function suppresses or stimulates the actual performance of a prepared action. When this super mirroring function is not well-developed this may lead to deviations in functioning, such as imperfect self-other distinction as occurs in variants of autism, or uncontrolled impulses to imitate what is observed (for example, after watching anti-social or violent actions on TV).

The cognitive functions involved have been adopted, abstracted, and formalised from mirror neuron and super mirror neuron functions informally described in neurological literature such as [6, 8, 15, 12, 11]. The cognitive agent model has been formally specified in the hybrid temporal modeling language LEADSTO [1], and has been analysed mathematically. In example simulations it was shown that preparation states for actions can be used not only for actual preparation of a to be executed action, but that such a preparation state can also be used to internally simulate the effects of the execution of an action (as in mirroring or imagination). The sensing of the context of an action and the creation of a focus state is the process that eventually leads to the differentiation between the different uses of the same preparation state. This focus state influences perception of the stimulus and execution of the action. Furthermore the focus state is also responsible for attribution of the effects of the preparation state to the correct agent.

The resulting cognitive agent model functions according to the Simulation Theory perspective on mindreading (e.g., [5]), the simulation perspective on imagination [6], and perspectives on imitation [7, 13, 14], which all assume that the own sensory representations, feelings and preparation states are used in reading another agent's actions and emotions, in imagination of actions and emotions, and in imitation. Thus it provides an economically designed generic agent model unifying different modes of mental processing. It has a high extent of flexibility, as it can easily switch between the different modes of mental processing. Due to this flexibility and its economical design, this model may provide a basis for extension to models that build on these forms of action preparation and execution, for example, to induce empathy reactions or to infer intentions of the mirrored agent.

References

1. Bosse, T., Jonker, C.M., van der Meij, L., Treur, J.: A Language and Environment for Analysis of Dynamics by Simulation. Int. J. of Artificial Intelligence Tools 16, 435–464 (2007)
2. Bosse, T., Jonker, C.M., Treur, J.: Formalisation of Damasio's Theory of Emotion, Feeling and Core Consciousness. Consciousness and Cognition 17, 94–113 (2008)
3. Damasio, A.: The Feeling of What Happens. In: Damasio, A. (ed.) Body and Emotion in the Making of Consciousness. Harcourt Brace, New York (1999)

4. Damasio, A.: Looking for Spinoza: Joy, Sorrow, and the Feeling Brain. Vintage Books, London (2003)
5. Goldman, A.I.: Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading. Oxford Univ. Press, New York (2006)
6. Hesslow, G.: Conscious thought as simulation of behavior and perception. Trends Cogn. Sci. 6, 242–247 (2002)
7. Hurley, S., Chater, N. (eds.): Perspectives on imitation: from cognitive neuroscience to social science, vol. 1. MIT Press, Cambridge (2005)
8. Iacoboni, M.: Mirroring People: the New Science of How We Connect with Others. Farrar, Straus & Giroux, New York (2008)
9. Iacoboni, M.: Understanding others: imitation, language, empathy. In: Hurley, S., Chater, N. (eds.) Perspectives on imitation: from cognitive neuroscience to social science, vol. 1, pp. 77–100. MIT Press, Cambridge (2005)
10. Jeannerod, M., Anquetil, T.: Putting oneself in the perspective of the other: A framework for self-other differentiation. Soc. Neurosc. 3, 356–367 (2008)
11. Mukamel, R., Ekstrom, A.D., Kaplan, J., Iacoboni, M., Fried, I.: Mirror properties of single cells in human medial frontal cortex. Soc. for Neuroscience (2007)
12. Pineda, J.A. (ed.): Mirror Neuron Systems: the Role of Mirroring Processes in Social Cognition. Humana Press Inc., NJ (2009)
13. Rizzolatti, G.: The mirror-neuron system and imitation. In: Hurley, S., Chater, N. (eds.) Perspectives on imitation: from cognitive neuroscience to social science, vol. 1, pp. 55–76. MIT Press, Cambridge (2005)
14. Rizzolatti, G., Fogassi, L., Gallese, V.: Neuro-physiological mechanisms underlying the understanding and imitation of action. Nature Rev. Neurosci. 2, 661–670 (2001)
15. Rizzolatti, G., Sinigaglia, C.: Mirrors in the Brain: How Our Minds Share Actions and Emotions. Oxford University Press, Oxford (2008)

Cellular GEP-Induced Classifiers

Joanna Jędrzejowicz[1] and Piotr Jędrzejowicz[2]

[1] Institute of Informatics, Gdańsk University,
Wita Stwosza 57, 80-952 Gdańsk, Poland
jj@inf.ug.edu.pl

[2] Department of Information Systems, Gdynia Maritime University,
Morska 83, 81-225 Gdynia, Poland
pj@am.gdynia.pl

Abstract. In this paper we propose integrating two collective computational intelligence techniques gene expression programming and cellular evolutionary algorithms with a view to induce expression trees, which, subsequently, serve as weak classifiers. From these classifiers stronger ensemble classifiers are constructed using majority-voting and boosting techniques. The paper includes the discussion of the validating experiment result confirming high quality of the proposed ensemble classifiers.

Keywords: gene expression programming, cellular evolutionary algorithm, ensemble classifiers.

1 Introduction

Gene expression programming introduced by Ferreira [6] is an automatic programming approach. In GEP computer programs are represented as linear character strings of fixed-length called chromosomes which, in the subsequent fitness evaluation, can be expressed as expression trees of different sizes and shapes. The approach has flexibility and power to explore the entire search space, which comes from the separation of genotype and phenotype.

Several experiments with gene expression programming for classification tasks were presented in [10], [11]. In this paper gene expression programming is strengthened by the paradigm of cellular evolutionary algorithms - the population of genes is structured by using the concept of neighborhood. Individuals can only interact with their closest neighbors in the population. For an introduction to cellular evolutionary algorithms see [1].

The paper is organized as follows. In section 2 classifiers induced by gene expression programming are proposed and discussed. In section 3 two classification algorithms based on gene expression programming and cellular evolutionary algorithms are introduced. In section 4 the results of validating experiment are shown. Finally, section 5 contains conclusions.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 343–352, 2010.

2 Using Cellular Gene Expression Programming to Induce Classifiers

Consider data classification problem. In what follows C is the set of categorical classes which are denoted $1, \dots, |C|$. We assume that the learning algorithm is provided with the training set $TD = \{< \boldsymbol{d}, c > \mid \boldsymbol{d} \in D, c \in C\} \subset D \times C$, where D is the space of attribute vectors $\boldsymbol{d} = (w_1^d, \dots, w_n^d)$ with w_i^d being symbolic or numeric values. The learning algorithm is used to find the best possible approximation $\bar{f}$ of the unknown function f such that $f(\boldsymbol{d}) = c$. Then $\bar{f}$ is a classifier which can be applied to find the class $c = \bar{f}(\boldsymbol{d})$ for any $\boldsymbol{d} \in D$.

As usual when applying GEP methodology, the algorithm uses a population of individuals (called genes), selects them according to fitness and introduces genetic variation using several genetic operators. Each individual is composed of a single gene divided into two parts as in the original head-tail method [6]. The size of the head (h) is determined by the user with the suggested size not less than the number of attributes in the dataset. The size of the tail (t) is computed as $t = h(n-1)+1$ where n is the largest arity found in the function set. In the computational experiments the functions are: logical AND, OR, XOR, NOR and NOT. Thus $n = 2$ and the size of the gene is $h + t = 2h + 1$. The terminal set contains triples $(op, attrib, const)$ where op is one of relational operators $<$, $\leq$, $>$, $\geq$, $=$, $\neq$, $attrib$ is the attribute number, and finally $const$ is a value belonging to the domain of the attribute $attrib$. As usual in GEP, the tail part of a gene always contains terminals and head can have both, terminals and functions. Observe that in this model each gene is syntactically correct and corresponds to a valid expression. Each attribute can appear once, many times or not at all. This allows to define flexible characteristics like for example $(attribute1 > 0.57)$ AND $(attribute1 < 0.80)$. On the other hand, it can also introduce inconsistencies like for example $(attribute1 > 0.57)$ AND $(attribute1 < 0.40)$. This does not cause problems since a decision subtree corresponding to such a subexpression would evaluate it to $false$. Besides, when studying the structure of the best classifiers in our experiments the above inconsistencies did not appear. Attaching an expression tree to a gene is done in exactly the same manner as in all GEP systems.

Traditionally, GEP algorithms work on a single population of genes. Here, the benefits of structuring the population by defining neighborhoods are explored. Individuals are arranged on a torus-like grid of dimension $xmax \times ymax$. Each point of the grid has a neighborhood that overlaps the neighborhood of nearby individuals; all the neighborhoods have the same size and identical shape. The boundary individuals of the grid are connected to the individuals located in the opposite borders in the same row/column, depending on the case. This results in toroidal grid and all the individuals have exactly the same number of neighbours. In the experiments the L5, or NEWS neighborhood - 4 nearest neighbours in a given axial (north, east, west, south) direction was applied.

The algorithm for learning the best classifier using cellular GEP works as follows. Suppose that a training dataset is given and each row in the dataset has a correct label representing the class. In the initial step the minimal and maximal value of each attribute is calculated and a random population of genes placed on the grid is generated.

To introduce variation in the population the following genetic operators are used: mutation, transposition of insertion sequence elements (IS transposition), root transposition (RIS transposition), one-point recombination, two-point recombination. Genetic operations may only take place in L5 neighborhood of each individual. What is more, the reproductive cycle is applied to all the individuals synchronously, that is the population for the next generation is created at the same time for all individuals on the grid.

In order to compare the quality of genes used as classifiers two measures are introduced. Suppose that the learning dataset TR and a class $cl \in C$ are fixed. The first measure counts the rows which are classified incorrectly. For a gene g

$$nF^{cl}(g) = \sum_{rw \in TR,\ g(rw)\ is\ true} sg(\text{rw is from class} \neq cl)$$

where

$$sg(\varphi) = \begin{cases} 1 & \text{if } \varphi \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

The second measure is defined as:

$$nT^{cl}(g) = \sum_{rw \in TR,\ g(rw)\ is\ true} sg(\text{rw is from class } cl)$$

Obviously, for two genes the one with the higher nT^{cl} is better and respectively for the measure nF^{cl} - the lower the better. Using the above measures, fitness of a gene g is defined as:

$$fitness^{cl}(g) = nT^{cl}(g) - nF^{cl}(g)$$

The algorithm of cGEP learning is shown as Algorithm 1.

3 Learning Algorithms

Below two learning algorithms, called cGEP-mv and cGEP-ada, are described.

3.1 Algorithm cGEP-mv

Learning takes place in two steps, shown as Algorithms 2 and Algorithm 3. Firstly, for each class (separately) the population of genes best fitting the class is found. During this step, in each generation the algorithm cGEP is applied and using the domination relation (defined below) best individuals are copied to next generation. The objective of the second step, called meta learning, is to select subsets of informative genes from the population defined in the first step in order to obtain high classification accuracy.

Algorithm 1. Algorithm cGEP learning

```
Require: class $cl$, training data, integer $NG$, integer $noInd$
create the grid with a random population $Pop$
for $i = 1$ to $NG$ do
    express genes as expression trees,
    calculate fitness of each gene,
    keep best gene
    for all $g \in Pop$ do
        nghbrs←CalculateNeighourhood($g$)
        offspring1←One-pointRecomb($g$, nghbrs)
        offspring2←Two-pointRecomb($g$, nghbrs)
        $gNew$ ← the better fitted of two offsprings
        mutation($gNew$)
        IStransposition($gNew$)
        RIStransposition($gNew$)
        Replacement($position(g)$, $AuxiliaryPop$, $gNew$)
    end for
    $Pop \leftarrow AuxiliaryPop$
end for
return $noInd$ best genes from $Pop$
```

Definition 1. [illegible] g_1 [illegible] g_2 [illegible] cl [illegible] $nT^{cl}(g_1) > nT^{cl}(g_2)$ [illegible] $nF^{cl}(g_1) < nF^{cl}(g_2)$.

The relation of domination is irreflexive, antisymmetric and transitive.

Definition 2. [illegible] g [illegible] pop [illegible] ff [illegible] pop [illegible] g

Algorithm 2. Algorithm cGEP-mv, first step

```
Require: training data with correct labels representing $|C|$ classes, integer $sizePop$,
    integer $sizeBest$
Ensure: population of $sizeBest$ genes for each class
for all $cl \in C$ do
    $population\_best^{cl} \leftarrow \phi$
    repeat
        call algorithm cGEP for class $cl$ to generate $pop$
        add to $population\_best^{cl}$ those genes from population $pop$ which are not dominated in $population\_best^{cl}$
    until $population\_best^{cl}$ contains at least $sizeBest$ elements
end for
```

In the process of meta-learning genetic algorithms are applied in order to select a subset of genes obtained in the process of learning and resulting in the population of $|C| \times sizeBest$ genes. Now individuals are defined as matrices of type $MG = \{0,1\}^{|C| \times sizeBest}$ which correspond to the distribution of genes from *population_best*. For the matrix $mg \in MG$ the set $\{i : \ mg[cl, i] = 1\}$ picks up genes from $population\_best^{cl}$ which are meaningful for the class cl. To define fitness of an individual mg, we assume the majority vote is performed to classify each data row and the number of correct answers is counted. Let $r = (\boldsymbol{x}, y)$ be a data row

$$which_i^{cl}(mg, \boldsymbol{x}) = \begin{cases} 1 & \text{if } mg[cl,i] = 1 \text{ and} \\ & population\_best^{cl}(i) \text{ is true for } \boldsymbol{x} \\ 0 & \text{otherwise} \end{cases}$$

$$mv(mg, \boldsymbol{x}) = \max_{cl}\Big(\sum_{i=1}^{sizeBest} which_i^{cl}(mg, \boldsymbol{x})\Big)$$

and, finally

$$fitness(mg) = \frac{|r : \ mv(mg, \boldsymbol{x}) = y|}{|dataset|}$$

Standard genetic operators: mutation and crossover are applied to find one matrix mg best fitting the given dataset. This matrix is then used in testing where again, the majority vote is applied. Finally, Algorithms 2 and Algorithm 3 are used to define Algorithm 4 which is majority-voting making use of cellular paradigm.

Algorithm 3. Algorithm cGEP-mv second step (meta-learning)

```
Require: training data with correct labels representing $|C|$ classes, integer $noIter$,
    population of $sizeBest$ genes for each class,
Ensure: best metagene $mg$
create random population $popMet$ of metagenes
for $i = 1$ to $noIter$ do
    calculate fitness of each metagene from $popMet$
    using roulette rule choose the metagenes for the next step,
    mutation,
    crossover,
end for
return the best metagene $mg$
```

3.2 Algorithm cGEP-ada

As suggested in [14] a weak classifier can be considerably improved by [illegible]. The general idea is to create an ensemble of classifiers by resampling the training dataset and creating a classifier for each sample. In each step the most informative training data is provided - for example those for which the previous classifier misclassified. Then an ensemble of generated classifiers together with an intelligent combination rule proves often to be a more efficient approach. Freund and Schapiro [8] suggested a refinement of a boosting algorithm called AdaBoost.

Algorithm 4. Algorithm cGEP-mv

```
Require: training data with correct labels representing $|C|$ classes, integer $sizePop$,
    integer $sizeBest$, testing data $TS$
Ensure: $qc$ quality of the majority vote classifier
apply Algorithm 2 to define $population\_best$
apply Algorithm 3 to find metagene $mg$
$qc \leftarrow 0$ {test $mg$ for testing data $TS$}
for all $(\boldsymbol{x}, y) \in TS$ do
    if $mv(mg, \boldsymbol{x}) = y$ then
        $qc \leftarrow qc + 1$
    end if
end for
$qc \leftarrow qc/|TS|$
return $qc$
```

For a predefined number of iterations T, the following procedures are performed. In the tth step, according to the current distribution - which is uniform in the first iteration, a sample is drawn from the dataset. The best classifier C_t is found for the sample and using the whole dataset the error of the current classification is calculated. The distribution is updated so that the weights of those instances that are correctly classified by the current classifier C_t are reduced by the factor depending on the error, and the weights of misclassified instances are unchanged. Once T classifiers are generated 'weighted majority voting' is used to classify the test set. The idea is to promote those classifiers that have shown good performance during training - they are rewarded with a higher weight than the others. The details are given in Algorithm 5, where cGEP is used as a weak classifier and AdaBoost methodology is applied.

It can be observed that in each iteration t the distribution weights of those instances that were correctly classified are reduced by a factor β_t and the weights of the misclassified instances stay unchanged. After the normalization the weights of instances misclassified are raised and they add up to $1/2$, and the weights of the correctly classified instances are lowered and they also add up to $1/2$. What is more, since it is required that the weak classifier has an error less than $1/2$, it is guaranteed to correctly classify at least one previously misclassified instance. In the ensemble decision those classifiers which produced small error and β_t is close to zero, have a large voting role since $1/\beta_t$ and logarithm of $1/\beta_t$ are large.

4 Computational Experiment Results

To evaluate the proposed approach computational experiment has been carried out. The experiment involved the following 2-classes datasets from the UCI Machine Learning Repository [2]: Wisconsin Breast Cancer (WBC), Diabetes,

Algorithm 5. Algorithm cGEP-ada

```
Require: training data $TD$ of size $N$, test dataset $TS$, integer $T$, integer $M \leq N$ - size of the selected dataset
Ensure: $qc$ quality of the AdaBoost classifier.
initialize the distribution $D_1(i) = \frac{1}{N}, i = 1, \ldots, N$
for $t = 1$ to $TT$ do
    for the current distribution $D_t$ select a training dataset $S_t \subset TD$ of size $M$,
    call Algorithm cGEP for the dataset $S_t$, receive the classifier $C_t$
    using the majority voting for $C_t$ calculate the error $\epsilon_t = \sum_{C_t(\mathbf{x}_i) \neq y_i} D_t(i)$
    if $\epsilon_t > 0.5$ then
        abort
    else
        $\beta_t = \epsilon_t / (1 - \epsilon_t)$
    end if
    { update the distribution}
    for $i = 1$ to $N$ do
        if $C_t(\mathbf{x}_i) = y_i$ then
            $D_t(i) \leftarrow D_t(i) \times \beta_t$
        end if
        normalize the distribution $D_{t+1}(i) = D_t(i)/Z_t,\ Z_t = \sum_i D_t(i)$
    end for
end for
{test the ensemble classifier $C_1, C_2, \ldots, C_T$ in the test dataset $TS$}
$qc \leftarrow 0$
for all $(\mathbf{x}, \mathbf{y}) \in \mathbf{TS}$ do
    $V_i = \sum_{C_t(\mathbf{x})=i} \log(1/\beta_t),\ i = 1, \ldots, |C|$
    $c \leftarrow argmax_{1 \leq j \leq |C|} V_j$
    if $c = y$ then
        $qc \leftarrow qc + 1$
    end if
end for
$qc \leftarrow qc/|TS|$
return $qc$
```

Sonar, Australian Credit (ACredit), German Credit (GCredit), Cleveland Heart (Heart), Hepatitis and Ionosphere.

In the reported experiment the following classification tools have been used: Cellular GEP with majority voting (cGEP-mv) and Cellular GEP with adaboost (cGEP-ada) described in details in the previous sections versus 16 well-known classifiers from WEKA Environment for Knowledge Analysis v. 3.7.0 [18] including Naive Bayes, Bayes Net, Logistic Regression, Radial Basis Function Network, AdaBoost, Support Vectors Machine, Ensemble Selection, Bagging, Classification via Clustering, Random Committee, Stacking, Rotation Forest, Decision Table, FT Tree, Random Forest and C4.5.

Computations involving cGEP-mv have been run with the following arbitrary parameter settings: xmax = ymax = 15; target population size 80; percentage

Table 1. Comparison of the classifier accuracy (%)

no.	classifier	WBC	Diab.	Sonar	ACr.	GCr.	Heart	Hep.	Ion.
1	Naive Bayes	95,99	76,30	67,78	77,68	75,40	83,70	84,51	82,62
2	Bayes Net	97,14	74,35	80,28	86,23	75,50	81,11	83,22	89,46
3	Logistic	96,56	77,21	73,08	85,22	75,20	83,70	82,58	88,89
4	RBF Network	95,85	75,39	72,11	79,71	74,00	84,07	85,80	92,31
5	AdaBoost M1	94,85	74,34	71,63	84,64	69,50	80,00	82,58	90,88
6	SVM	**96,99**	**77,34**	75,96	84,92	75,10	84,07	85,16	88,60
7	Ensemble Selection	94,42	74,61	75,48	84,93	73,10	80,00	81,29	90,59
8	Bagging	95,56	74,61	77,40	85,07	74,40	79,26	84,52	90,88
9	Class. via clustering	95,71	64,84	54,32	74,06	56,60	77,04	74,19	70,94
10	Random Committee	95,99	73,95	**84,13**	83,48	73,90	80,37	84,52	92,59
11	Stacking	65,52	65,10	53,36	55,51	70,00	55,56	79,35	64,10
12	Rotation Forest	96,99	76,69	84,13	87,25	74,80	80,74	82,58	**93,73**
13	Decision Table	95,28	71,22	69,23	83,48	71,00	**84,81**	76,13	89,46
14	FT	96,99	77,34	79,81	85,51	68,30	82,96	81,29	90,31
15	Random Forest	96,13	73,82	80,77	85,07	72,50	78,15	82,58	92,87
16	C4.5	94,56	73,82	71,15	86,09	70,50	76,66	83,87	91,45
17	cGEP-mv	95,58	76,99	80,79	**87,39**	76,27	80,24	86,46	91,73
18	cGEP-ada	95,86	77,21	81,24	86,52	77,37	83,84	**87,13**	91,35

of the non-dominated expression trees taking part in the majority voting 70%; number of iterations in Cellular GEP 250; probability of mutation 0.5, RIS transposition 0.2, IS transposition 0.2, 1-point recombination 0.2 and 2-point recombination 0.2. Computations involving CGEP-ada have been run with the following arbitrary parameter settings: xmax = ymax = 10; number of the Adaboost iterations 5; number of iterations in Cellular GEP 30; number of repetitions in the class learning 50. The remaining settings are identical as in case of cGEP-mv. In all WEKA classifiers the default parameter settings have been used.

Table 1 shows computation results averaged over 10 repetitions of the 10-cross-validation scheme. Performance measure is the classifier accuracy. To evaluate the performance of cGEP-mv and cGEP-ada the Friedman's non-parametric test using ranks of the data has been applied under the following hyphotheses:

- Null Hypothesis H0: All of the 18 population distribution functions are identical.
- Alternative Hypothesis H1: At least one of the populations tends to yield larger observations than at least one of the other populations.

Analyses of the experiment results shows that for the population of the classification accuracy observations the null hypothesis should be rejected at the significance level of 0,05. The average Friedmans ranks for the classification accuracies rank cGEP-ada on the first place and cGEP-mv on the third place among 18 investigated classifiers.

5 Conclusions

The paper proposes an approach based on integrating gene expression programming with cellular genetic programming to induce expression trees. The induced trees are used to construct ensemble classifiers. Main contribution of the paper can be summarized as follows:

- Class specific cellular GEP learning procedure is proposed and implemented
- Non-dominance relation between genes is used in the process of gene selection
- Two ensemble classifiers based on expression trees induced using cellular GEP learning procedure are constructed and validated.

The resulting cellular GEP-induced ensemble classifiers were validated experimentally using several datasets and the results were compared with those of other well established classification methods. Validation experiment results allow to draw the following conclusions:

- In terms of the classification accuracy and the area under the ROC curve both - cGEP-mv and cGEP-ada perform very well and are competitive in comparison with a majority of other approaches
- Both algorithms (cGEP-mv and cGEP-ada) are consistent assuring high quality classification results when applied to different datasets.

References

1. Alba, E., Dorronsoro, B.: Cellular Genetic Algorithms. Springer Science, New York (2008)
2. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository, School of Information and Computer Science. University of California (2007), http://www.ics.uci.edu/~mlearn/MLRepository.html
3. Bi, Y., Guan, J., Bell, D.: The combination of multiple classifiers using an evidentional reasoning approach. Artif. Intell. 172, 1731–1751 (2008)
4. Centeno, P., Lawrence, N.D.: Optimising Kernel Parameters and Regularisation Coefficients for Non-linear Discriminant Analysis. Journal of Machine Learning Research 7, 455–491 (2006)
5. Fawcett, T.: ROC Graphs: Notes and Practical Considerations for Researchers, HP Labs Tech Report HPL-2003-4, Palo Alto, Ca (2003)
6. Ferreira, C.: Gene Expression Programming: a New Adaptive Algorithm for Solving Problems. Complex Systems 13(2), 87–129 (2001)
7. Ferreira, C.: Gene Expression Programming. Studies in Computational Intell. 21, 337–380 (2006)
8. Freund, Y., Schapire, R.E.: Decision-theoretic generalization of on-line learning and application to boosting. Journal of Computer and System Science 55, 119–139 (1997)
9. Ishibuchi, H., Nojima, Y.: Analysis of Interpretability-accuracy Tradeoff of Fuzzy Systems by Multiobjective Fuzzy Genetics-based Machine Learning. Intern. Journal of Approximate Reasoning 44, 4–31 (2007)

10. Jedrzejowicz, J., Jedrzejowicz, P.: GEP-induced expression trees as weak classifiers. In: Perner, P. (ed.) ICDM 2008. LNCS (LNAI), vol. 5077, pp. 129–141. Springer, Heidelberg (2008)
11. Jedrzejowicz, J., Jedrzejowicz, P.: A Family of GEP-induced Ensemble Classifiers. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) ICCCI 2009. LNCS, vol. 5796, pp. 641–652. Springer, Heidelberg (2009)
12. Kretowski, M.: A Memetic Algorithm for Global Induction of Decision Trees. In: Geffert, V., et al. (eds.) SOFSEM 2008. LNCS, vol. 4910, pp. 531–540. Springer, Heidelberg (2008)
13. Polikar, R.: Ensemble Based Systems in Decision Making. IEEE Circuits and Systems Magazine 3, 22–43 (2006)
14. Schapire, R.E., Freund, Y., Bartlett, P., Lee, W.S.: Boosting the margin: A new explanation for the effectiveness of voting methods. The Annals of Statistics 26(5), 1651–1686 (1998)
15. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
16. Torre, F.: Boosting Correct Least General Generalizations, Technical Report GRApp A-0104, Grenoble (2004)
17. Weinert, W.R., Lopes, H.S.: GEPCLASS: A Classification Rule Discovery Tool Using Gene Expression Programming. In: Li, X., Zaïane, O.R., Li, Z.-h. (eds.) ADMA 2006. LNCS (LNAI), vol. 4093, pp. 871–880. Springer, Heidelberg (2006)
18. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
19. Zhou, C., Xiao, W., Tirpak, T.M., Nelson, P.C.: Evolving Accurate and Compact Classification Rules with Gene Expression Programming. IEEE Transactions on Evolutionary Computation 7(6), 519–531 (2003)

Cluster Integration for the Cluster-Based Instance Selection

Ireneusz Czarnowski and Piotr Jędrzejowicz

Department of Information Systems, Gdynia Maritime University
Morska 83, 81-225 Gdynia, Poland
{irek,pj}@am.gdynia.pl

Abstract. The problem addressed in this paper concerns data reduction through instance selection. The paper proposes an approach based on instance selection from clusters. The process of selection and learning is executed by a team of agents. The approach aims at obtaining a compact representation of the dataset, where the upper bound on the size of data is determined by the user. The basic assumption is that the instance selection is carried out after the training data have been grouped into clusters. The cluster initialization and integration strategies are proposed and experimentally evaluated.

1 Introduction

The paper focuses on learning from examples understood as the process of finding a classification model. In data mining learning from example delivers models for solving classification tasks. One of the recent research focus is development of techniques for selecting the relevant information through data reduction. Such a reduction can be achieved by selection of instances, selection of attributes/features or simultaneous reduction in both dimensions [15].

Data reduction is the process of selecting the relevant data with an objective to find patterns, also called prototypes or reference vectors, or regularities within certain attributes (see, for example [10]). Data reduction performed without losing extractable information can result in increased capabilities and generalization properties of the learning model. It is obvious that removing some instances from the training set reduces time and memory complexity of the learning process [20]. Data reduction is also considered as an approach to increasing effectiveness of learning process when the available datasets are large or distributed and when the access to data is limited and costly from the commercial point of view.

A common approach to the distributed data mining (DDM) is to build separate models at geographically distributed sites and then to combine the models using meta-learning [16]. In [18] it was, however, pointed out that such an approach offers a rather narrow view of the distributed data mining, since the data distributions in different locations are often not identical. Another approach to learning from the distributed data sets is based on moving all of the data to a central site, merging the data and building a single global learning model. Unfortunately, moving all data into a centralized location can be very time consuming, costly, or may not be feasible due to some restrictions [9, 11, 14].

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 353–362, 2010.

The selection of relevant data in distributed locations and then moving only the local patterns can eliminate or reduce the restrictions on a communication bandwidth or reduce the cost of data shipping, and speed up the distributed learning process [10]. The important distributed data mining problem is to establish a reasonable upper bound on the size of a dataset needed for an efficient analysis [19].

The aim of the paper is proposing and evaluating through computational experiment a cluster-based instance selection approach generating a representative dataset of the required size. The proposed approach is based on the assumption that prototypes are selected from clusters by the agent-based population learning algorithm. Clusters are generated at the first stage of instance selection using the similarity coefficient as the criterion procedure. Strategies for cluster initialization and integration are experimentally evaluated.

The paper is organized as follows. The agent-based instance selection algorithm and its main features are presented in Section 2. The cluster initialization procedure is described in Section 3. Section 4 provides details on the computational experiment setup and discusses its results. Finally, the last section contains conclusions and suggestions for future research.

2 A Framework for Data Reduction

This paper considers the problem of learning from examples using dataset obtained by instance selection. It was proved that the instance selection is computationally difficult combinatorial optimization problem [7]. Thus approximate algorithms including metaheuristics seem to be a practical approach to solving the instance selection problem.

In this paper the instance selection problem is solved by an agent-based population learning algorithm. The proposed approach uses the specialized team of agents (A-Team) where agents execute the improvement procedures and cooperate with a view to solve instances of the data reduction problem. Main features of the agent-based population learning algorithm and an overview of the proposed approach are included in the next subsection.

2.1 Main Features of the Agent-Based Population Learning Algorithm

The A-Team concept was originally introduced in [17]. The idea was motivated by several approaches like blackboard systems and evolutionary algorithms, which have proven to be able to successfully solve some difficult combinatorial optimization problems. Within an A-Team agents achieve an implicit cooperation by sharing a population of solutions (individuals), to the problem to be solved.

An A-Team can be also defined as a set of agents and a set of memories, forming a network in which every agent remains in a closed loop. Each agent possesses some problem-solving skills and each memory contains a population of temporary solutions to the problem at hand. All agents can work asynchronously and in parallel. During their work agents cooperate to construct, find and improve solutions which are read from the shared, common memory.

Main functionality of the agent-based population learning approach includes organizing and conducting the process of search for the best solution. It involves a sequence of the following steps:

- Generation of the initial population of solutions to be stored in the common memory.
- Activation of optimizing agents which apply solution improvement algorithms to solutions drawn from the common memory and store them back after the attempted improvement applying some user defined replacement strategy.
- Continuation of the reading-improving-replacing cycle until a stopping criterion is met. Such a criterion can be defined either or both as a predefined number of iterations or a limiting time period during which optimizing agents do not manage to improve the current best solution. After computation has been stopped the best solution achieved so far is accepted as final.

More information on the population learning algorithm with optimization procedures implemented as agents within an asynchronous team of agents (A-Team) can be found in [2].

2.2 Agent-Based Population Learning Algorithm for Instance Selection

This paper proposes an A-Team in which agents execute the improvement procedure and cooperate in a manner described in the preceding subsection with a view to solve instances of the data reduction problem.

A potential solution is represented by a string of numbers of the selected reference instances. The initial population, at the initial phase, is generated randomly – one instance from each cluster. It is clear that drawing the initial population is preceded by execution of the clustering algorithm.

Each solution from the population is evaluated and the value of its fitness is calculated. The evaluation is carried out by estimating classification accuracy of the classifier, which is constructed taking into account the instances as indicated by the solution.

In the proposed approach optimizing agents, which operate on individuals, are implementations of the local search procedures. Each optimizing agent tries to improve quality of the received solutions by applying the implemented improvement procedure. An agent, after being supplied with an individual to be improved, explores its neighborhood with the aim of finding a new, better solution in the neighborhood.

To solve the data reduction problem two kinds of optimizing agents representing two different improvement procedures have been implemented. These are: tabu search and a simple local search for instance selection. Details of both procedures can be found in [4].

The proposed A-Team uses a simple replacement strategy. Each optimizing agent receives a solution drawn at random from the population of solutions (individuals). The solution returned by optimizing agent is merged with the current population replacing the current worst solution.

3 The Cluster Initialization Procedure

It is proposed to use the clustering procedure based on the similarity coefficient. Instances are grouped into clusters according to their similarity coefficient calculated as in [4]. A cluster contains instances with identical similarity coefficients.

The main feature of the above procedure is that the number of clusters is determined by the value of the similarity coefficient. The experiment results discussed in [5] show that the instance grouping procedure based on the similarity coefficient assures a very good results, with respect to clustering quality measured by the silhouette coefficient, in comparison with other clustering methods including, for example, the *k*-means clustering or the clustering based on the stratification [8]. The detailed pseudo-code of the instance grouping procedure based on the similarity coefficient is shown below as Algorithm 1.

Algorithm 1*. The instance grouping based on the similarity coefficient values*

Input: X - the matrix containing values of all instances from the original training set T, where $X=\{x_{ij} : i=1,\ldots,n; j=1,\ldots,m+1\}$ and n denotes the number of instances, m - the number of attributes. Total length of each instance is equal to $m+1$, where element numbered $m+1$ contains the class label (the class label of each example can take any value from a finite set of decision classes $C = \{c_l : l = 1,\ldots, k\}$, which has cardinality k).
Output: clusters from which prototypes will be selected.

1. Transform data instances: each $\{x_{ij}\}$ for $i=1,\ldots,n$ and $j=1,\ldots, m$ is normalized into interval [0,1] and then rounded to the nearest integer, that is 0 or 1.
2. Calculate values:

$$s_j = \sum_{i=1}^{n} x_{ij}, \; where \;\; j = 1,\ldots,m.$$

3. For instances from X, belonging to the class c_l, calculate the value of its similarity coefficient I_i:

$$\forall_{x:x_{i,\ m+1}=c_l} \;\; I_i = \sum_{j=1}^{m} x_{ij} s_j, \; where \;\; i = 1,\ldots,n.$$

4. Map input vectors from X with the same value of similarity coefficient I_i into clusters.
5. Let $Y_i^{(l)}$ denotes the obtained clusters such that $T = \bigcup_l^k \bigcup_{i=1}^t Y_i^{(l)}$ and $\forall_{i\neq j:i,j=1,\ldots,t;\ l=1,\ldots,k} \;\; Y_i^{(l)} \cap Y_j^{(l)} = \varnothing$, and where t denotes the total number of obtained clusters.

Assuming that one prototype is selected from each cluster, the number of clusters produced has a direct influence on the final number of prototypes, i.e. on the size of the reduced dataset. Thus, the selection based on the similarity coefficient can not always assure the required data reduction rate. Hence, the approach based on the similarity coefficient is used only at the cluster initialization stage.

When the number of clusters obtained at the first stage does not assure the required data reduction rate the initially generated clusters are merged. The idea is to apply a procedure which allows at a single step to merge two "similar" clusters where similarity is measured using one of the available proximity measures. The merging is repeated several times and at each step only two clusters minimizing the proximity measure are merged. The merging is terminated when the required reduction rate is obtained. Furthermore, since it is expected that the final prototype set will identical class distribution as the original data set, the merging procedure deciding on the number of clusters is class specific. The detailed pseudo-code of the cluster merging procedure is shown as Algorithm 2.

Algorithm 2*. The merging procedure*

Input: $Y_i^{(l)}$ - initial clusters, where i=1,…,t and t denotes the number of initial clusters; t_{max} – upper bound of the number of clusters (t_{max}<t<n).

Output: clusters from which prototypes will be selected.

```
1. Set $t_x$:=$t$;
2. While ($t_x$>$t_{max}$)
3. Execute the merging procedure on clusters $Y_i^{(l)}$ containing instances belonging to
   the class $c_l$ ($l$=1,...,$k$).
4. Update $t_x$.
5. End while.
```

4 Computational Experiment

4.1 Computational Experiment Settings

To validate the proposed approach it has been decided to carry out the computational experiment. The experiment aimed at answering the following two questions:

- Does the proposed approach assure appropriate compression rate?
- Does the choice of the proximity measure, used within the merging procedure, used to measure the distance between clusters may influence the performance of the learning process?
- Does the proposed approach perform better then other data reduction algorithms?

In the reported research the following merging procedures proposed in [11] have been considered:

- *MLP – minimum linkage*. The procedure calculates the distance between two clusters as the smallest Euclidean distance between any two instances from two clusters. The clusters, where the distance is minimal, are merged.
- *CLP - complete linkage*. The procedure calculates the distance between two clusters as the largest Euclidean distance between any two instances from two clusters. The clusters, where the distance is minimal, are merged.
- *ALP - average linkage*. The procedure calculates the distance between two clusters as the average of the Euclidean distances between all the pairs of instances from two clusters. The clusters, where the distance is minimal, are merged.
- *CeLP - centroid linkage*. The procedure calculates the distance between two clusters as the Euclidean distance of the two means of the two clusters. The clusters, where the distance is minimal, are merged.

4.2 Dataset Choice and the Experiment Plan

To validate the proposed approach several benchmark classification problems have been solved. Datasets for each problem including Cleveland heart disease, credit approval, Wisconsin breast cancer and sonar problem have been obtained from the UCI Machine Learning Repository [1]. Characteristics of these datasets are shown in Table 1.

Table 1. Datasets used in the reported experiment

Dataset	Number of instances	Number of attributes	Number of classes	The best reported classification accuracy	Average number of clusters (by Algorithm 1)
heart	303	13	2	90.0% [6]	162
sonar	208	60	2	97.1% [1]	94
credit	690	15	2	86.9% [1]	184
cancer	699	9	2	97.5% [1]	133

In Table 1 the average number of clusters produced at the first stage of the cluster initialization phase by the algorithm based on the similarity coefficient (Algorithm 1) is shown.

Each benchmarking problem has been solved 30 times and the reported values of the quality measure have been averaged over all runs. The quality measure in all cases was the correct classification ratio calculated using the 10-cross-validation approach, where at first the available dataset was randomly partitioned into training and test sets. In the second step each training dataset was reduced using the proposed approach.

The above scheme has been repeated four times, with the upper bound on the number of the selected prototypes set to t_{max}={5%, 10%, 15%, 20%} of the number of instances in the original dataset. The optimization agents were running for 100 iterations. The common memory size was set to 100 individuals. The number of iterations and the size of the common memory were set arbitrary.

The proposed A-Team has been implemented using the middleware environment called JABAT [2], based on JAVA code and built using JADE (Java Agent Development Framework) [3].

4.3 Computation Experiment Results

Classification accuracy of the classifier obtained using the reduced dataset where instances are selected through applying the proposed approach, has been compared with:

- results obtained by machine classification without data reduction, i.e. on full, non-reduced dataset,
- results obtained using the set of prototypes produced through selection based on the *k*-means clustering (In this case at the first stage the *k-means* clustering has been implemented and next, from thus obtained clusters, the prototypes have been selected using the agent-based population learning algorithm).

Table 2. Accuracy of the classification results (%)

	heart	cancer	credit	sonar
C 4.5 – non-reduced dataset	77.89	94.57	84.93	74.04
MLP (t_{max}=5%)	81.13	78.43	80.43	48.46
MLP (t_{max}=10%)	82.54	77.87	84.52	48.46
MLP (t_{max}=15%)	84.06	79.02	83.32	51.32
MLP (t_{max} =20%)	87.25	78.24	81.32	56.32
CLP (t_{max}=5%)	81.43	82.13	84.43	73.21
CLP (t_{max}=10%)	79.04	85.32	83.20	72.02
CLP (t_{max}=15%)	83.55	86.90	85.32	72.59
CLP (t_{max}=20%)	83.54	89.32	87.20	73.65
ALP (t_{max}=5%)	82.43	97.53	87.03	69.43
ALP (t_{max}=10%)	84.64	96.30	86.66	78.32
ALP (t_{max}=15%)	87.84	98.20	88.14	76.62
ALP (t_{max}=20%)	86.52	97.62	89.20	74.62
CeLP (t_{max}=5%)	85.32	94.32	83.72	71.43
CeLP (t_{max}=10%)	87.32	95.43	83.26	69.73
CeLP (t_{max}=15%)	87.12	97.60	84.71	72.43
CeLP (t_{max}=20%)	89.43	96.43	86.66	73.05
k-means (t_{max}=5%)	82.00	95.71	87.12	48.46
k-means (t_{max}=10%)	85.67	94.43	88.99	54.34
k-means (t_{max} =15%)	87.67	95.09	90.14	59.48
k-means (t_{max}=20%)	88.00	96.14	90.29	72.17

Table 3. Performance comparison of different instance reduction approaches (*Acur.* and *Ret.* in %)

Problem	heart		cancer		credit		Sonar	
Approach	*Accur.*	*Ret.*	*Accur.*	*Ret.*	*Accur.*	*Ret.*	*Accur.*	*Ret.*
ALP (t_{max}=10%)	84.64	10	96.30	10	86.66	10	**78.32**	10
ALP (t_{max}=15%)	**87.84**	15	**98.20**	15	88.14	15	76.62	15
ALP (t_{max}=20%)	86.52	20	97.62	20	89.20	20	74.62	20
k-means (t_{max}=20%)	88.00	20	96.14	20	**90.29**	20	72.17	20
CNN [20]	73.95	31	95.71	7.09	77.68	24.12	74.12	32.85
SNN [20]	76.25	34	93.85	8.35	81.31	28.38	79.81	28.26
IB3 [20]	81.16	11	96.57	3.47	85.22	4.78	69.38	12.02
DROP 3 [20]	80.84	13	96.14	3.58	83.91	5.96	78.00	26.87
RMHC [15]	82.3	3	70.9	7	-	-	-	-
GA-KJ [13]	74.7	33.1	95.5	33.4	-	-	55.3	52.6

The learning tool used in the experiment was the C 4.5 algorithm [12]. Computation results are shown in Table 2. The ranking of the compared approaches is shown in Fig. 1, where horizontal axis represents the mean relative difference between the mean accuracies of the best method and the given method.

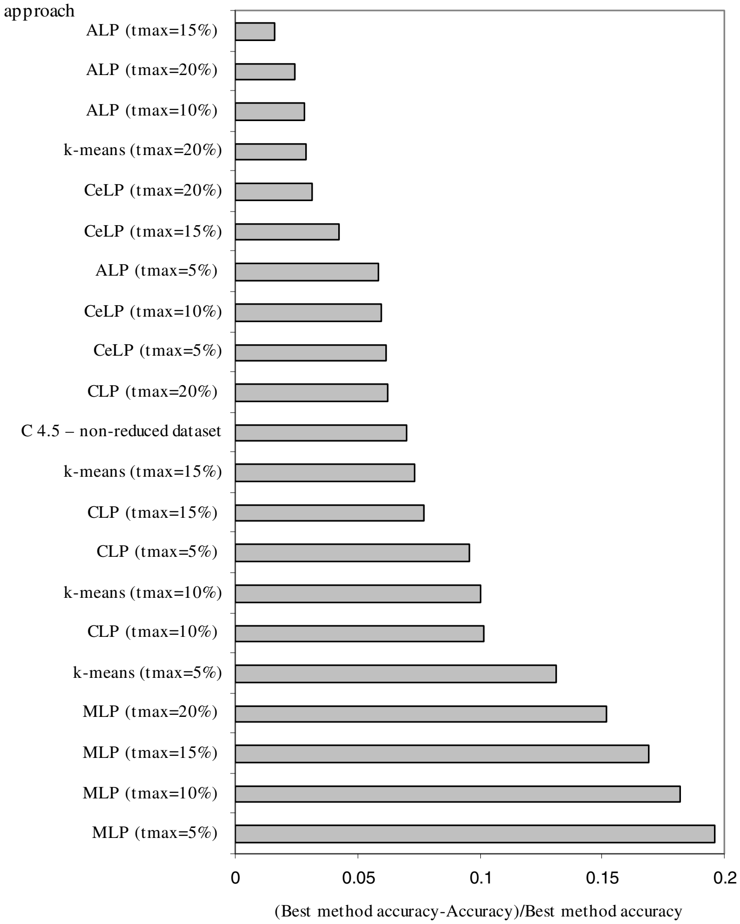


Fig. 1. Ranking of the data reduction methods

It should be noted that the proposed method produces very good results as compared with the case when data reduction is carried-out by the k-means-based approach. The cluster-based data reduction allows to induce classifiers outperforming those induced using original, non-reduced dataset. Further, from Table 1 it is also clear that the distance measure used to measure proximity of the clusters may have a direct influence on the accuracy of clustering, quality of the selected prototypes and hence, the classification accuracy. Experiment results indicate that the *ALP* cluster merging strategy assures significantly better results than other merging strategies. Moreover the *ALP* procedure is competitive, with respect to classification accuracy, in comparison with other, well known, approaches to data reduction, which can be concluded from data shown in Table 3. The column *Ret.* in Table 3 shows what percentage of instances from the original training set has been retained by the respective reduction algorithm.

5 Conclusions

The main contribution of the paper is proposing and validating a cluster-based instance selection approach assuring the required size of the reduced dataset and, at the same time, guaranteeing best possible classification accuracy of the classifier induced using the reduced dataset. To produce clusters the proposed approach uses an implementation of the agent-based population learning algorithm. Clusters are produced in two stages including cluster initialization at the first stage and cluster merging at the second.

Computational experiment results have confirmed that the cluster-based data reduction is an effective data reduction tool in data mining. The experiment shows that the clustering procedure based on the similarity coefficient is an effective cluster initialization method. The proposed approach extends the existing range of the available techniques of data reduction.

A critical factor at the stage where clusters are merged, is the distance measure used to measure the proximity of the clusters. Finding more effective merging algorithms should be the focus of further research. Effectiveness of the approach should be also studied using a wider range of the available classifiers. It should be also possible to integrate two stages of the data reduction, that is clustering and cluster merging, into a single agent-based population learning algorithm.

Acknowledgments. This research has been supported by the Polish Ministry of Science and Higher Education with grant for years 2008-2010 and grant for years 2010-2013.

References

1. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository, School of Information and Computer Science. University of California, Irvine (2007), `http://www.ics.uci.edu/~mlearn/MLRepository.html`
2. Barbucha, D., Czarnowski, I., Jędrzejowicz, P., Ratajczak-Ropel, E., Wierzbowska, I.: e-JABAT - An Implementation of the Web-Based A-Team. In: Nguyen, N.T., Jain, I.C. (eds.) Intelligent Agents in the Evolution of Web and Applications, Studies in Computational Intelligence, vol. 167, pp. 57–86. Springer, Heidelberg (2009)
3. Bellifemine, F., Caire, G., Poggi, A., Rimassa, G.: JADE. A white paper, Exp. 3(3), 6–20 (2003)
4. Czarnowski, I., Jędrzejowicz, P.: An Approach to Instance Reduction in Supervised Learning. In: Coenen, F., Preece, A., Macintosh, A. (eds.) Research and Development in Intelligent Systems XX, pp. 267–282. Springer, London (2004)

5. Czarnowski, I.: Cluster-based instance selection for machine classification. Knowledge and Information Systems (to appear, 2010)
6. Datasets used for classification: comparison of results. In. directory of data sets, `http://www.is.umk.pl/projects/datasets.html` (accessed 1 September 2009)
7. Hamo, Y., Markovitch, S.: The COMPSET Algorithm for Subset Selection. In: Proceedings of The Nineteenth International Joint Conference for Artificial Intelligence, Edinburgh, Scotland, pp. 728–733 (2005)
8. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, pp. 520–528. Springer, New York (2009)
9. Klusch, M., Lodi, S., Moro, G.-L.: Agent-Based Distributed Data Mining: The KDEC Scheme. In: Klusch, M., et al. (eds.) Intelligent Information Agents. LNCS (LNAI), vol. 2586, pp. 104–122. Springer, Heidelberg (2003)
10. Krishnaswamy, S., Zaslavsky, A., Loke, S.W.: Techniques for Estimating the Computation and Communication Costs of Distributed Data Mining. In: Sloot, P.M.A., et al. (eds.) ICCS-ComputSci 2002. LNCS, vol. 2329, pp. 603–612. Springer, Heidelberg (2002)
11. Liu, H., Lu, H., Yao, J.: Identifying Relevant Databases for Multidatabase Mining. In: Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 210–221 (1998)
12. Quinlan, J.R.: C4.5: Programs for machine learning. Morgan Kaufmann Publishers, San-Mateo (1993)
13. Rozsypal, A., Kubat, M.: Selecting Representative Examples and Attributes by a Genetic Algorithm. Intelligent Data Analysis 7(4), 291–304 (2003)
14. Silva, J., Giannella, C., Bhargava, R., Kargupta, H., Klusch, M.: Distributed Data Mining and Agents. Engineering Applications of Artificial Intelligence Journal 18, 791–807 (2005)
15. Skalak, D.B.: Prototype and Feature Selection by Sampling and Random Mutation Hill Climbing Algorithm. In: Proceedings of the International Conference on Machine Learning, pp. 293–301 (1994)
16. Stolfo, S., Prodromidis, A.L., Tselepis, S., Lee, W., Fan, D.W.: JAM: Java Agents for Meta-learning over Distributed Databases. In: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, Newport Beach, CA, pp. 74–81. AAAI Press, Menlo Park (1997)
17. Talukdar, S., Baerentzen, L., Gove, A., de Souza P.: Asynchronous Teams: Co-operation Schemes for Autonomous, Computer-Based Agents, Technical Report EDRC 18-59-96, Carnegie Mellon University, Pittsburgh (1996)
18. Tsoumakas, G., Angelis, L., Vlahavas, I.: Clustering Classifiers for Knowledge Discovery from Physical Distributed Databased. Data and Knowledge Enginering 49(3) (2004)
19. Vucetic, S., Obradovic, Z.: Performance Controlled Data Reduction for Knowledge Discovery in Distributed Databases. In: Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 29–39 (2000)
20. Wilson, D.R., Martinez, T.R.: Reduction Techniques for Instance-based Learning Algorithm. Machine Learning 33(3), 257–286 (2000)

Comparative Study of the Differential Evolution and Approximation Algorithms for Computing Optimal Mixed Strategies in Zero-Sum Games

Urszula Boryczka[1] and Przemyslaw Juszczuk[2]

[1] Institute of Computer Science, University of Silesia, ul.Bedzinska 39,
Sosnowiec, Poland
urszula.boryczka@us.edu.pl

[2] Institute of Computer Science, University of Silesia, ul.Bedzinska 39,
Sosnowiec, Poland
przemyslaw.juszczuk@us.edu.pl

Abstract. In this paper, we present the application of the Differential Evolution (DE) algorithm to the problem of finding optimal mixed strategies in zero-sum games for two players. Differential evolution (DE) is a simple and powerful optimization method, which is mainly applied to numerical optimization and many other problems (for example: neural network train, filter design or image analysis). The advantage of the DE algorithm is its capability of avoiding so-called ,,local minima" within the considered search space. Thanks to the special operator of the adaptive mutation, it is possible to direct the searching process within the solution space. Approach used in this article is based on the probability of selecting single pure strategy. In optimal mixed strategy, every strategy has some probability of being chosen. Our goal is determine this probability and maximize payoff for a single player. We also compare proposed method with well known approximation algorithm.

Keywords: Game theory, Differential evolution, zero-sum games.

1 Introduction

Competition plays very important role in our life. A basic feature of the competition is that the final outcome depends primarily on the combination of strategies selected by the adversaries. Game theory has influenced many fields, including economics (its initial focus) [1], political science [7], biology [4], and many others. Game theory is an integral part of artificial intelligence (AI) [12], e-commerce [3], and other areas of computer science. Scope of this article are the zero-sum games [11]. Every zero-sum game has at least one optimal strategy [13]. Assuming, that we consider only games, where saddle points not exist, calculating the optimal mixed strategy is a difficult task, and it may be described as a finding distribution of probability over the set of pure strategies.

Now, if the first player has a mixed strategy X^*, and the second player has a mixed strategy Y^*, such that:

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 363–372, 2010.

$$X^*My \geq v, \text{ for all } Y,$$
$$xMY^* \leq v, \text{ for all } X,$$

where M is the payoff matrix, X^*My denotes the payoff for the first player, v is the value of the the game. We refer X^* and Y^* as optimal strategies since:

- if first player uses strategy X^*, his expectation is at least v, irrespective of what the second player does;
- if second player uses strategy Y^*, he can make first player expectation at most v, irrespective, what the first player does.

This problem was successfully solved by the pivot method and by the simplex method (both methods are part of linear programming). A linear programming problem may be defined as the problem of maximizing or minimizing linear function subject to linear constraints. So matrix game for two players can be easily transformed into linear equations. Unfortunately, both methods give only one solution for the game. Sometimes player want to choose from the set of optimal mixed strategies.

Our motivation is to show a new method of finding set of optimal mixed strategies. The Differential Evolution (DE) provides at least one optimal mixed strategy in every single run of the algorithm. While increasing size of the problem, set of optimal mixed strategies also is growing. Compare this method with any other is difficult, because the DE is the nondeterministic algorithm. In this article we present preliminary study of the DE used for the game theory.

Our article is organized as follows: first we describe the problem of finding the optimal mixed strategies in large zero-sum games for two players. In section three, we recall well known approximation algorithm [13], and propose new DE based algorithm for calculating the optimal mixed strategies. Section four contains the experiments, and comparative study of two described algorithms. Finally we end with some conclusions and future works.

2 Proposed Methodology

A two-player game is a zero-sum game if the sum of the payoffs of two players is zero for any choice of the strategies. In a normal form, such game may be defined as a payoff matrix $M_{m,n}$, where every player $i \in 1, ..., n$ has a set of pure strategies (or actions) S_i, and a utility function:

$$u_i \rightarrow S_1 \times S_2 \times ... \times S_n \rightarrow R \tag{1}$$

that maps every outcome (a vector consisting of a pure strategy for each player, also known as a profile of pure strategies) to a real number.

In the theory of games a player is said to use a mixed strategy whenever he or she chooses to randomize over the set of available actions. Formally, a mixed strategy is a probability distribution that assigns to each available action a likelihood of being selected. When the player has only finite number of strategies, the mixed strategy may be described as an , - dimensional vector $S = (s_1, .., s_m)$, which also needs to meet two conditions:

$$s_i \geq 0, \tag{2}$$

which denotes, that the probability of choosing the strategy s_i is greater or equal to 0. Second condition means that the sum of probabilities must be equal to 1:

$$\sum_{i=1}^{m} s_i = 1. \tag{3}$$

Cell $M_{i,j}$ is the payoff for the first player in case, when the first player choose i pure strategy, and the second player choose j pure strategy. As it may be seen, the game on the fig. 1 is not fair, because every cell in the payoff matrix is positive. The game cost v is equal to zero if:

$$min(M_{i,j}) + max(M_{i,j}) = 0, \text{ for } i, j \in \{1, 2, ..., n\},$$

where n is number of the pure strategies for every player. If only one action has a positive probability of being selected, the player is said to use a pure strategy. So, when a mixed strategy is played, a pure strategy is chosen according to this distribution before the game starts and then followed throughout the game. General problem is to calculate values of probability of choosing each pure strategy. Complexity of this task depends on the number of the possible pure strategies. In the next section we show two methods for calculating the optimal mixed strategy.

3 Algorithms for Calculating Optimal Mixed Strategy

In this section we show two methods for calculating the optimal mixed strategy. First method is derived from the literature, and it is based on the probability distribution [13]. At every step of the algorithm, the best player strategy is chosen. The second player behaves in the same way, and chooses strategy, that is the best response. Players make their choices for a determined number of iterations. Strategies that bring the highest payoff are chosen more frequently.

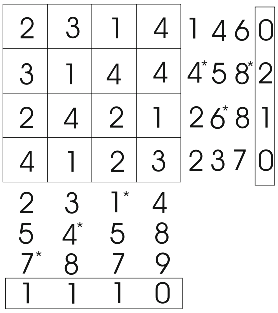
Where S_{1i} is the i-th pure strategy of the first player and M is the payoff matrix. The algorithm is very easy and brings good results - even for large games, where number of strategies for single players is greater than 10. First 6 starting steps of algorithm is shown on the fig. 1.

Separated values $1 : 1 : 1 : 0$ and $0 : 2 : 1 : 0$ are frequency of using pure strategies. Frequency vectors may be transformed into corresponding probability vectors $0.33 : 0.3 : 0.33 : 0$ and $0 : 0.66 : 0.33 : 0$. This probability may be used to calculate the game cost v.

$$v = \textstyle\sum_{i=1}^{n} S_{xi} \cdot M_{x,i}$$

Algorithm 1. Approximation algoritm

```
1 Take any row player strategy $S_{1i}$ from the matrix $M$;
2 Copy selected strategy below the matrix and mark the smallest number in it;
3 Select the column player strategy $S_{2j}$ that is above the smallest number in
  selected row $S_{1i}$;
4 Copy that strategy $S_{2j}$ at the right and mark its largest number;
5 while stop criterion is not met do
6     Add row $S_{1k}$ (pointed by the largest number from $S_{2j}$) to the row below the
      matrix and write it below that row ;
7     Select column that is above the smallest number in $S_{1k}$, add this strategy to
      the strategy $S_{2j}$ on the right and write it next to it;
8     go to step 4;
```



2	3	1	4	1	4	6	0
3	1	4	4	4*	5	8*	2
2	4	2	1	2	6*	8	1
4	1	2	3	2	3	7	0
2	3	1*	4				
5	4*	5	8				
7*	8	7	9				
1	1	1	0				

Fig. 1. Approximation algorithm

where n is length of the probability vector, S_{xi} is the $i-th$ element of the mixed strategy of the first player used against x pure strategy of the second player, $M_{x,i}$ denotes cell in the payoff matrix.

The second algorithm is based on Differential Evolution (DE). DE is a stochastic, population-based search strategy developed by Storn and Price in 1995 [9]. It has mostly been applied to optimize functions defined over continuous-valued landscapes [10]. DE has also been applied to train neural networks (NN) [6]. In this case an individual represents a complete NN. Similar approach was made for training Fuzzy Cognitive Maps [2]. Other applications of the Differential Evolution focus on clustering [8], system design [5]. The pseudocode of the general DE algorithm is presented in algorithm 2.

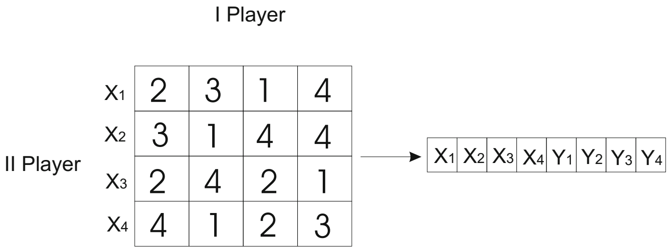
Every individual in population is created on the basis of the payoff matrix (Fig. 2), where X_i if frequency of using strategy X_i for player 1, and Y_i is frequency of using strategy Y_i (player 2). Length of the single genotype may be calculated according the formula:

$$dim = count(X) + count(Y).$$

For example, 2 player game, where each of the players has 20 pure strategies is 40-dimensional optimization problem. Every gene in genotype has a value in the

Algorithm 2. Basic DE algorithm

```
1 Create the initial population of genotypes $P_0 = \{\boldsymbol{X}_{1,0}, \boldsymbol{X}_{2,0}, ..., \boldsymbol{X}_{n,0}\}$;
2 Set the generation number $g = 0$;
3 while stop criterion is not met do
4     Compute the fitness function for every genotype in the population
      $\{f(\boldsymbol{X}_{1,g}), f(\boldsymbol{X}_{2,g}), ..., f(\boldsymbol{X}_{n,g})\}$ ;
5     Create the population of trial genotypes $V_g$ based on $P_g$;
6     Make crossover of genotypes from the population $P_g$ and $V_g$ to create
      population $U_g$;
7     Choose the genotypes with the highest fitness function from population $U_g$
      and $P_g$ for the next population;
8     $generation = generation + 1$, go to step 4;
```



X1	2	3	1	4
X2	3	1	4	4
X3	2	4	2	1
X4	4	1	2	3

X1	X2	X3	X4	Y1	Y2	Y3	Y4

Fig. 2. Creation of the single individual

range $< 0 : 1 >$ and represents the probability of choosing the pure strategy. Mutation is a primary genetic operator in the differential evolution algorithm. It is based on the few steps listened in the algorithm 3:

Algorithm 3. Mutation schema

```
Set the $F$ parameter
foreach individual $x_i \in$ population $P$ do
    Generate three different random numbers $r_1, r_2, r_3$ between range $(1, ..., n)$
    foreach gene $j \in$ genotype do
        Calculate $v_{ij} = x_{r_1 j} + F \cdot (x_{r_{2j}} - x_{r_{3j}})$,
```

An individual $v_i(t)$ represents an individual after the mutation. $(x_{r_2} - x_{r_3})$ is a differential vector created from the two random individuals x_{r_2} and x_{r_3}). The differential vector gives information about the fitness landscape and in this way the search process is directed.

The main part of the presented algorithm is a calculation of the fitness function. We can formulate following equations:

$$\bigwedge_{x_i \in X} f_1 = |Y^* M x_i - v| \tag{4}$$

$$\bigwedge_{y_i \in Y} f_2 = |X^* M y_i - v| \tag{5}$$

where x_i is a pure strategy, X is a set of pure strategies, X^* is the mixed strategy, v is a game cost (for the optimal mixed strategy, which is calculated inside the fitness function), and $X^* M y_i$ is an expected payoff for the player using the mixed strategy X^* against the pure strategy Y_i. Optimal mixed strategy X^* guarantees that game cost v is constant, while playing against every pure strategy of the second player (this rules holds also for the second player - second player plays his optimal mixed strategy Y^* against every pure strategy of first player). Second part of the fitness function is based on the Eq. 3 - sum off all probabilities over the optimal mixed strategy of single player is equal to 1.

$$f_3 = 1 - |\sum_{i=1}^{m} X_i|, \tag{6}$$

$$f_4 = 1 - |\sum_{i=1}^{m} Y_i|. \tag{7}$$

where X_i is probability of choosing strategy i for X player. Fitness function is a sum of four above equations:

$$f = f_1 + f_2 + f_3 + f_4, \tag{8}$$

Game cost is calculated as follows:

$$v = \frac{\frac{\sum_{i=1}^{n} X_{mixed} M y_i}{n} + \frac{\sum_{i=1}^{m} Y_{mixed} M x_i}{m}}{2} \tag{9}$$

where n is number of pure strategies of the second player, $X_{mixed} M y_i$ is the $i-th$ pure strategy of Y player used against the mixed strategy of the X player.

4 Experimental Results

The aim of this experiments is to compare existing approximation algorithm and the new proposed method for computing the optimal mixed strategies for two players zero-sum games. The DE algorithm has the following parameters: the binomial crossover, the crossover parameter $CR = 0.5$ [10], the mutation parameter $F = 0.7$ [10], the population size $n_X = 50$, the number of the differential vectors $n_v = 1$. Number of iteration for the DE was set to 7500 for every tested game. Number of iterations for the approximation algorithm was set to 1000000. As we can see, the number of iterations for both algorithms differ a lot. One iteration of the DE algorithm consists of mutation, crossover, fitness evaluation and selection (for 50 individuals in the population). On the other hand, the approximation algorithm iteration consists only from the add operation. We tested 10 different randomly generated games. Number of strategies for every test game differs from 10 to 20. Tested parameters are:

Table 1. The comparison of how effectively DE finds mixed optimal strategies - compared with approximation algorithm - pessimistic spacing

Number of strategies	the best pesDE spacing	average pesDE spacing	the worst pesDE spacing	median pesDE spacing	standard deviation pesDE	Approx Algorithm
10 x 10 game 1	6.530%	8.229%	9.716%	8.453%	0.992	53.186%
10 x 10 game 2	8.709%	10.156%	12.566%	10.044%	1.080	59.942%
15 x 15 game 1	7.306%	17.594%	21.209%	18.436%	3.132	59.657%
15 x 15 game 2	9.572%	17.561%	22.312%	18.010%	3.773	58.131%
15 x 15 game 3	12.446%	25.075%	32.267%	26.197%	3.366	55.277%
20 x 20 game 1	29.277%	34.064%	37.855%	34.820%	2.881	55.678%
20 x 20 game 2	21.274%	29.530%	37.881%	28.723%	4.895	57.364%
20 x 20 game 3	23.226%	31.219%	40.184%	30.499%	3.826	61.087%
20 x 20 game 4	22.032%	29.407%	38.793%	27.784%	3.913	46.075%
20 x 20 game 5	21.049%	33.774%	42.225%	35.361%	4.262	69.685%

Table 2. The comparison of how effectively DE finds mixed optimal strategies - compared with approximation algorithm - average spacing

Number of strategies	the best avgDE spacing	average avgDE spacing	the worst avgDE spacing	median avgDE spacing	standard deviation avgDE	Approx Algorithm
10 x 10 game 1	1.334%	2.184%	3.528%	1.886%	0.739	13.118%
10 x 10 game 2	0.594%	2.689%	3.816%	3.287%	0.714	18.835%
15 x 15 game 1	0.109%	2.209%	7.385%	1.638%	2.327	16.834%
15 x 15 game 2	1.387%	4.871%	10.830%	4.575%	3.025	19.570%
15 x 15 game 3	1.370%	4.530%	8.796%	4.972%	2.526	16.997%
20 x 20 game 1	1.152%	4.345%	8.644%	4.321%	2.738	15.329%
20 x 20 game 2	1.356%	2.881%	5.464%	2.692%	1.288	16.334%
20 x 20 game 3	1.528%	3.462%	7.443%	3.255%	2.150	17.439%
20 x 20 game 4	1.506%	3.656%	7.307%	3.806%	1.736	10.319%
20 x 20 game 5	1.192%	4.418%	10.006%	3.395%	3.394	13.470%

– pessimistic spacing (pesDE) : distance between minimal v calculated from set of the first player pure strategies and maximal v calculated from the second player pure strategies - expressed as a percentage.

$$pesDE = (1 - \frac{min(v_{x1}, v_{x2}, ..., v_{xn})}{max(v_{y1}, v_{y2}, ..., v_{ym}})\cdot 100 \tag{10}$$

where v_{xi} is the optimal mixed strategy of the first player used against the pure $i - th$ strategy of the second player.

– average spacing (avgDE) : average value from set of possible game costs v calculated from both players - expressed as a percentage.

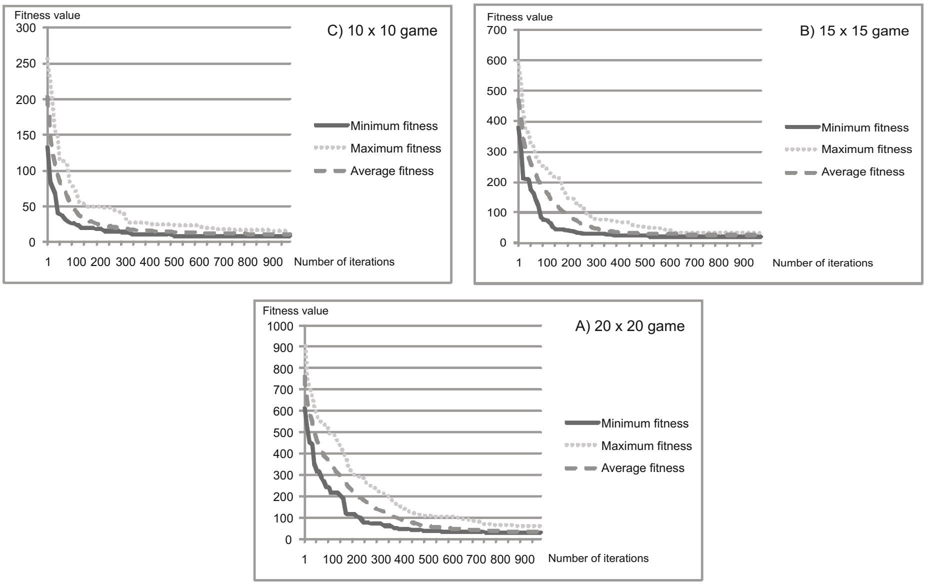


Fig. 3. Converge of population to optimum - first 1000 iterations. a) 10 *x* 10 game; b) 15 *x* 15 game; c) 20 *x* 20 game

Table 3. The 10 example DE algorithm runs - probability distribution over the set of pure strategies S for the row player

Nr.	$P(S_1)$	$P(S_2)$	$P(S_3)$	$P(S_4)$	$P(S_5)$	$P(S_6)$	$P(S_7)$	$P(S_8)$	$P(S_9)$	$P(S_{10})$
1.	0.175	0.178	0.11	0.082	0.001	0.0	0.156	0.0	0.18	0.109
2.	0.15	0.203	0.121	0.089	0.004	0.0	0.164	0.0	0.173	0.098
3.	0.06	0.238	0.09	0.2	0.0	0.0	0.16	0.111	0.138	0.0
4.	0.092	0.218	0.092	0.172	0.0	0.0	0.162	0.086	0.147	0.025
5.	0.109	0.216	0.106	0.144	0.0	0.0	0.164	0.046	0.16	0.046
6.	0.224	0.147	0.122	0.061	0.008	0.0	0.13	0.0	0.173	0.137
7.	0.177	0.191	0.113	0.078	0.0	0	0.151	0.006	0.169	0.11
8.	0.188	0.182	0.121	0.07	0.0	0.0	0.146	0.0	0.172	0.12
9.	0.15	0.2	0.118	0.089	0.001	0.0	0.163	0.003	0.175	0.096
10.	0.209	0.141	0.112	0.08	0.005	0.0	0.147	0.025	0.166	0.112

$$avgDE = \left(1 - \frac{\frac{\sum_{i=1}^{n} v_{xi}}{n}}{\frac{\sum_{j=1}^{m} v_{yj}}{m}}\right) \cdot 100 \tag{11}$$

For the differential evolution algorithm, convergence of population is also measured.

Second part of experiments is based on the convergence of population. 3 games with different difficulty were tested. Everytime the DE algorithm was set to 3000 iterations. Two interesting population features may be seen at figure 4. First, significant improvement of the fitness function is observed especially in the first

1000 iterations. At the last stage of the algorithm (> 2500 iterations) only overall improvement of the population may be seen (the best so far found solution is almost constant).

Finally, notice that the approximation algorithm allows to find only one probability distribution over the set of strategies S. In table 3 we can see 10 example runs of the DE algorithm for simple 10 x 10 game chosen from the test set. It is clear to see, that the probability distributions differ, so the DE algorithm is capable to find different optimal mixed strategies.

Results may be seen in tables 1 and 2. The DE algorithm was better in both cases. Even for very large games, evaluation error is acceptable. The approximation algorithm tends to stuck far from optimum. This situation is clearly seen especially in the table 1. It can be assumed, that significant increase the number of the iterations for the approximation algorithm should bring major improvement. We can see, that the new proposed method of computing the optimal mixed strategies in large zero-sum games is far better than the compared approximation algorithm. The large error for the pessimistic spacing, that is observed in the table 1 is caused by the calculation method. For the large games, only few strategies for each player seems to be not effective, but the pessimistic spacing method takes into account only the worst strategies. Confirmation of this observation is in the table 2, where the average spacing error is very small even for the games with 20 strategies for both players.

5 Conclusions

Differential evolution (DE) was chosen out of all evolutionary algorithms (EAs) available, to speed up the execution, since the DE algorithm is one of the simplest of all evolutionary algorithms. Proposed approach allows to generate the approximate optimal mixed strategies for the large zero-sum games. Our next goal is to describe the DE based algorithm allowing to extract pure strategies (which are part of the mixed strategy) having a big probability of choosing - which are main building blocks of the optimal mixed strategy.

References

1. Aubin, J.: Mathematical Methods of Game and Economic Theory. North-Holland Publ. CO., New York (1979)
2. Froelich, W., Juszczuk, P.: Predictive Capabilities of Adaptive and Evolutionary Fuzzy Cognitive Maps - A Comparative Study. Studies in Computational Intelligence 252, 153–174 (2009)
3. Griss, M., Letsinger, R.: Games at Work-Agent-Mediated E-Commerce Simulation. In: Proceedings of the Fourth International Conference on Autonomous Agents (2000)
4. Hammerstein, P., Selten, R.: Handbook of game theory - Chapter Game theory and evolutionary biology. University of Bonn (1994)
5. Kyprianou, A., Worden, K., Panet, M.: Identification of Hysteretic Systems using the Differential Evolution Algorithm. Journal of Sound and Vibration 248(2), 289–314 (2001)

6. Magoulas, G.D., Plagianakos, V.P., Vrahatis, M.N.: Neural Network-Based Colonoscopic Diagnosis using On-Line Learning and Differential Evolution. Applied Soft Computing 4(4), 369–379 (2004)
7. Ordeshook, P.: Game theory and political theory. Cambridge University Press, Cambridge (1986)
8. Paterlini, S., Krink, T.: High Performance Clustering with Differential Evolution. In: Proceedings of the IEEE Congress on Evolutionary Computation, vol. 2, pp. 2004–2011 (2004)
9. Storn, R., Price, K.: Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. Journal of Global Optimization 11(4), 341–359 (1997)
10. Storn, R.: On the Usage of Differential Evolution for Function Optimization. In: Proceedings of the Biennial Conference of the North American Fuzzy Information Processing Society, pp. 519–523 (1996)
11. Straffin, P.: Game Theory and Strategy. SCHOLAR (2004)
12. Tennenholtz, M.: Game Theory and Artificial Intelligence, vol. 2403, pp. 34–52. Springer, Heidelberg (2002)
13. Williams, J.D.: The Compleat Strategyst: Being a Primer on the Theory of Games of Strategy. McGraw-Hill, New York (1966)

Ant Colony Decision Trees – A New Method for Constructing Decision Trees Based on Ant Colony Optimization

Urszula Boryczka and Jan Kozak

Institute of Computer Science, University of Silesia, Będzińska 39, 41–200 Sosnowiec, Poland
Tel.:/Fax: (+48 32) 291 82 83
{urszula.boryczka,jan.kozak}@us.edu.pl

Abstract. In this paper, we would like to propose a new method for constructing decision trees based on Ant Colony Optimization (ACO). The ACO is a metaheuristic inspired by the behavior of real ants, where they search for optimal solutions by considering both local heuristic and previous knowledge, observed by pheromone changes. Good results of the ant colony algorithms for solving combinatorial optimization problems suggest an appropriate effectiveness of the approach also in the task of constructing decision trees. In order to improve the accuracy of decision trees we propose an Ant Colony algorithm for constructing Decision Trees (ACDT). A heuristic function used in a new algorithm is based on the splitting rule of the CART algorithm (Classification and Regression Trees). The proposed algorithm is evaluated on a number of well-known benchmark data sets from the UCI Machine Learning repository. What deserves particular attention is the fact that empirical results clearly show that ACDT performs very good while comparing to other techniques.

1 Introduction

One should know that data mining is a process of extracting useful knowledge from real-world data. Among several data mining tasks – such as clustering and classification – this paper focuses on classification. The aim of the classification algorithm is to discover a set of classification rules. One of algorithms for solving this task is Ant–Miner, proposed by Parpinelli and colleagues [16], which employs Ant Colony Optimization techniques [3,8] to discover classification rules. In this paper, we present a new approach of ACO in Data Mining – ant colony alghorithm for constructing decision trees. Ant Colony Optimization is a branch of a newly developed form of artificial intelligence called swarm intelligence. The swarm intelligence is a form of emergent collective intelligence of groups of simple individuals: ants, termites or bees in which indirect form of communication via pheromone was observed. Pheromone values encourage the ants following the path to build good solutions of the analyzed problem and the learning process occurring in this situation is called positive feedback or autocatalysis.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 373–382, 2010.

In the present thesis we would like to discuss a series of important sections. The First section is devoted to an introduction to the subject of this paper. Section 2 provides detailed description of the Ant Colony Optimization in data mining. Section 3 describes decision trees, especially the CART algorithm. Section 4 focuses on the presented new algorithm ACDT. Section 5 presents the experimental study that has been conducted to evaluate the performance of ACDT, taking into consideration the data sets. The last section concludes obtained results and discusses the future evolution of the presented approach.

2 Ant Colony Optimization in Data Mining

In this paper we defined an ant algorithm to be a multi–agent system inspired by the observation of real ant colony behavior exploiting the stigmergic communication paradigm. The optimization algorithm in this paper was inspired by the previous works on Ant Systems (AS) and, in general, by the term — stigmergy. This phenomenon was first introduced by P.P. Grasse [12].

An essential step in this direction was the development of Ant System by Dorigo et al. [8], a new type of heuristic inspired by analogies to the foraging behavior of real ant colonies, which has proven to work successfully in a series of experimental studies. Diverse modifications of AS have been applied to many different types of discrete optimization problems and have produced very satisfactory results [6]. Recently, the approach has been extended by Dorigo et al. [4,5,9,10] to a full discrete optimization metaheuristics, called the Ant Colony Optimization (ACO) metaheuristics.

The Ant Colony System (ACS) algorithm has been introduced by Dorigo and Gambardella to improve the performance of Ant System [7,8], which allowed to find good solutions within a reasonable time for small size problems only. The ACS is based on 3 modifications of Ant System: a different node transition rule; a different pheromone trail updating rule; the use of local and global pheromone updating rules (to favor exploration).

The node transition rule is modified to allow explicitly for exploration. An ant k in analyzed node i chooses the node j to move to following the rule:

$$j = \begin{cases} \arg\max_{u \in J_i^k} \{[\tau_{iu}(t)] \cdot [\eta_{iu}]^\beta\} & \text{if } q \leq q_0 \\ J & \text{if } q > q_0 \end{cases}$$

where q is a random variable uniformly distributed over $[0, 1]$, q_0 is a tunable parameter $(0 \leq q_0 \leq 1)$, and $J \in J_i^k$ is an analyzed node that is chosen randomly according to a probability:

$$p_{iJ}^k(t) = \begin{cases} \dfrac{\tau_{iJ}(t) \cdot [\eta_{iJ}]^\beta}{\sum\limits_{l \in J_i^k} [\tau_{il}(t)] \cdot [\eta_{il}]^\beta} \end{cases}$$

which is similar to the transition probability used by Ant System. Therefore we can easily notice that the ACS transition rule is identical to Ant System's one, when $q > q_0$, and is different when $q \leq q_0$. More precisely, $q \leq q_0$ corresponds to the exploitation of the knowledge available about the problem, that is, the heuristic knowledge about distances between nodes and the learned knowledge memorized in the form of pheromone trails, whereas $q > q_0$ favors more exploration.

An adaptation of Ant Colony Optimization to classification is a research area still not well explored and examined. The appeal of this approach similarly to the evolutionary techniques is that they provide an effective mechanism of conducting a more global search. These approaches are based on a collection of attribute–value terms, then it can be expected that these approaches will also cope better with attribute interaction than greedy induction algorithms [11].

The prototype Ant–Miner is an ant–based system [16] and it is more flexible and robust than traditional approaches. This method incorporates a simple ant system in which a heuristic value based on the entropy measure is calculated. Ant–Miner has produced good results when compared with more conventional data mining algorithms, such as C4.5 [17], ID3 and CN2 [2]. ACDT (proposes in this work) is a different approach than Ant–Miner. In Ant–Miner, the goal of the algorithm was to produce an ordered list of rules, which was then applied to the test data in order of being discovered. ACDT constructs decision trees.

One should know that there are many other characteristics of ACO which are really important in data mining applications. It is interesting to mention that ACO contrary to deterministic decision trees or rule induction algorithms, during rule induction, tries to extenuate the problem of premature convergence to local optima because of stochastic element which prefers a global search in the problem's search space. Secondly, the ACO metaheuristics is a population–based one. It permits the system to search in many independently determined points in the search space concurrently and to use the positive feedback between ants as a search mechanism [15].

3 Decision Trees

Data mining, the science and technology of exploring data sets in order to discover previously unknown patterns, is a part of the overall process of knowledge discovery in data bases. In data mining, a decision tree is a predictive model which can be used to represent both classifiers and regression models. When a decision tree is used for classification tasks, it is referred to as a classification tree. When it is used for regression tasks, it is called regression tree.

A decision tree is used in determining the optimum course of action, in situations having several possible alternatives with uncertain outcomes. The resulting chart or diagram (which looks like a cluster of tree branches) displays the structure of a particular decision, and the interrelationships and interplay between different alternatives, decisions, and possible outcomes. Decision trees are commonly used in operational research, specifically in decision analysis, for identifying an optimal strategy for reaching a goal. The evaluation function for decision trees will be calculated according to the following formula:

$$Q(T) = \phi \cdot w(T) + \psi \cdot a(T, P) \tag{1}$$

where:
$w(T)$ – the size (numer of nodes) of the decision tree T,
$a(T,P)$ – the accuracy of the classification object from a test set P by the tree T,
ϕ and ψ – constants determining the relative importance of $w(T)$ and $a(T,P)$.

Constructing optimal binary decision trees is a NP–complete problem, where an optimal tree is one which minimizes the expected number of tests required for identification the unknown objects (as shown by Hyafil and Rivest in 1976 [13]).

The problem of designing storage efficient decision trees from decision tables was examined by Murphy and McCraw [14] in 1991. They shown that for most cases, the construction of the storage optimal decision tree is a NP-complete problem, and therefore a heuristic approach to the problem is necessary. Constructing an optimal decision tree may be defined as an optimization problem in which at each stage of creating decisions we select the optimal splitting of the data [18].

In most general terms, the purpose of the analysis via tree-building algorithms is to determine a set of if–then logical (split) conditions that permit accurate prediction or classification of cases.

Classification And Regression Tree (CART) approach was developed by Breiman et al. in 1984 [1] and is characterized by the fact that it constructs binary trees, namely each internal node has exactly two outgoing edges. The splits are selected using the Twoing Criteria and Gini index. When provided, CART can consider misclassification costs in the tree induction. It also enables users to provide prior probability distribution. CART looks for splits that minimize the prediction squared error. The prediction in each leaf is based on the weighted mean for node.

A decision tree is built in accordance with splitting rule that performs the multiple splitting of learning sample into smaller parts. Data in the each node have to be divided into two parts with maximum homogeneity in the decision class.

Twoing criterion will search for two classes that will make up together more then 50% of the data. Twoing splitting rule will maximize the following change-of-impurity measure which implies the following maximization problem:

$$\underset{a_j \leq a_j^R, j=1,\dots,M}{\arg\max} \left(\frac{P_l P_r}{4} \left[\sum_{k=1}^{K} |p(k|m_l) - p(k|m_r)| \right]^2 \right). \tag{2}$$

where:
$p(k|m_l)$ – the conditional probability of the class k provided in node m_l,
P_l – the probability of transition objects into the left node m_l,
P_r – the probability of transition objects into the right node m_r,
K – the decision class,
a_j – variable j,
a_j^R– the best splitting value of variable a_j.

Although Twoing splitting rule allows us to build more balanced trees, this algorithm works slower than the Gini rule. For example, if the total number of classes is equal to K, than we will have 2^{K-1} possible splits.

Besides mentioned Twoing splitting rules, there are several other methods. Among mostly used are an Gini index, Entropy rule, χ^2 rule, a maximum deviation rule. It has been proved that the final tree is insensitive to the choice of the splitting rule. The pruning procedure is much more important than the presented rules.

4 Ant Colony Decision Trees (ACDT)

It is interesting to note that the proposed algorithm of decision tree construction is mainly based on the version of ant colony optimization. A number of slight modifications have been introduced both as a new discrete optimization algorithm for constructing decision trees and a new metaheuristics approach in data mining procedures. The presented modifications are introduced in the main transition rule and they are treated as an improvement of the quality of the classification mechanism. We have employed a classical version of ACO and simple changes concerning the main rules, dedicated to each agents–ants during the construction the tours are incorporated in the scheme. Then we have applied the classical splitting rule, firstly used in CART. Secondly, we are complied with the pheromone changes which are useful knowledge for creating a reasonable division.

In ACDT each ant chooses the appropriate attribute for splitting in each node of the constructed decision tree according to the heuristic function and pheromone values (fig. 1). The heuristic function is based on the Twoing criterion, which helps ants divide the objects into two groups, connected with the analyzed attribute values. In this way, the attribute, which well separate the objects is treated as the best condition for the analyzed node. The best splitting is observed when we classified the same number of objects in the left and right subtrees with the maximum homogenity in the decision classes. Pheromone values represent the best way (connection) from the superior to the subordinate nodes – all possible combinations in the analyzed subtrees. For each node we calculate the following values according to the objects classified using the Twoing criterion of the superior node.

The pseudo code of the proposed algorithm is presented below. Lines 2–13 describe one iteration of this algorithm. At the beginning of its work, each ant builds one decision tree (lines 4–11). At the end of the loop, the best decision tree is chosen and then the pheromone is updated according to the splits performed during the process of construction the decision tree, iteratively. While constructing the tree, agents–ants are analyzing previous structures and some modifications are performed in the single node. This process is performed till the best decision tree is obtained. The process of building the decision tree is presented in Figure 2.

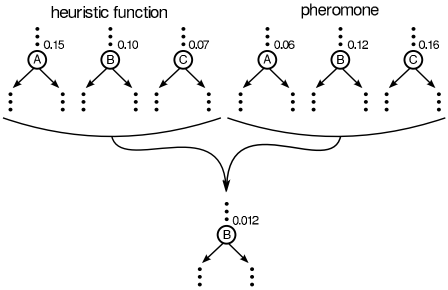


Fig. 1. Choice of splits in ACDT

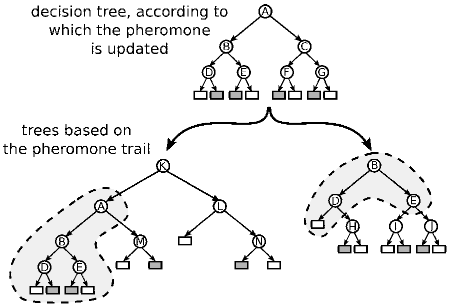

decision tree, according to which the pheromone is updated
A
B
C
D
E
F
G
trees based on the pheromone trail
K
A
L
B
M
N
D
E
B
D
E
H
I
J

Fig. 2. Building the decision tree with pheromone

Algorithm 1. Pseudo code of the proposed ACDT algorithm

```
initialization_pheromone_trail(pheromone);
for number_of_iterations do
    best_tree =      ;
    for number_of_ants do
        new_tree = build_tree(pheromone);
        pruning(new_tree);
        assessment_of_the_quality_tree(new_tree);
        if new_tree is_higher_quality_than best_tree then
            best_tree = new_tree;
        endIf
    endFor
    update_pheromone_trail(best_tree, pheromone);
endFor
result = best_constructed_tree;
```

The value of the heuristic function is determined according to the splitting rule employed in CART approach (see formula (2)), in dependence on the chosen criterion. Whereas the probability of choosing the appropriate test in the node is calculated according to a classical probability used in ACO:

$$p_{i,j} = \frac{\tau_{m,m_{L(i,j)}}(t)^{\alpha} \cdot \eta_{i,j}^{\beta}}{\sum_i^a \sum_j^{b_i} \tau_{m,m_{L(i,j)}}(t)^{\alpha} \cdot \eta_{i,j}^{\beta}} \tag{3}$$

where:

$\eta_{i,j}$ – a heuristic value for the test of the attribute i and value j,

$\tau_{m,m_{L(i,j)}}$ – an amount of pheromone currently available at time t on the connection between nodes m and m_L, (it concerns the attribute i and value j),

α, β – the relative importance with experimentally established values 1 i 3.

The initial value of the pheromone trail, similarly to the Ant–Miner approach is established in dependence on the number of attribute values. While the pheromone updates are performed (4) by increasing the previous values on each pairs of nodes (parent–child).

$$\tau_{m,m_L}(t+1) = (1-\gamma) \cdot \tau_{m,m_L}(t) + Q \tag{4}$$

where Q determines the evaluation function of decision tree (see formula (1)), and γ is a parameter represented the evaporation rate, equal to 0.1. $m_{L(i,j)}$ is the current node, in which the test of attribute i and the value j, is performed, and $m_{(p,o)}$ is the superior node, where the test of attribute p and the value o is performed.

5 Experiments

A variety of experiments were conducted to test the performance and behavior of the proposed algorithm. First we describe our experimental methodology and explain its motivation. Then we present and discuss our results. In this section we will consider an experimental study (see Table 2) performed for the following adjustments. We have performed 200 experiments for each data set. Each experiment included 25 generations with the population size of ant colony equal to 10.

Pruning can be applied in two different ways. Firstly, it can be applied as a second phase after each final decision tree has been built. Secondly, it is worthwhile to examine pruning of the lookahead trees. For this purpose, a post–pruning phase is applied to the decision trees that were generated by ACDT to form the lookahead sample, and the size of the pruned decision trees will be considered. Previous comparative studies did not find a single pruning method that is generally the best and conclude that different pruning techniques behave similarly. Therefore, during the analysis we used the Twoing criterion and Error–Based Pruning, and examine post–pruning of the final trees. We have performed the experimental study for three approaches:

ACDT – with quality coefficient calculated accordingly with the learning set, when the algorithm is evaluated based on the testing set. We also used the Twoing criterion. We evaluate the effectiveness of the approach and then update the pheromone values for the best quality decision tree.

CART (Salford Systems) – it is a standard implementation of the CART algorithm (`http://www.salfordsystems.com`). Each of the analyzed parameters' values are established in a classical way.

Ant–Miner – the classical version of the algorithm, the parameters' values are tuned as in the authors' suggestions [15].

5.1 Data Sets

Evaluation of the performance behavior of ACDT was performed using 7 public–domain data sets from the UCI (University of California at Irvine) data set repository available from: `http://archive.ics.uci.edu/ml/`. Table 1 shows the main characteristics of the data sets, which are divided into two groups in a random way: training ($\frac{2}{3}$ of all objects) and testing ($\frac{1}{3}$ of all objects) sets, appropriately. In order to obtain reliable performance estimates train–and–test were carried out to produce each of the statistics in the tables below. The experiments were carried out on the Intel Celeron 1.60 GHz Computer with 1.5 GB RAM. Possible follow-up investigations suggested by the results are also mentioned.

Table. 1. Original parameters in data sets

Dataset	N. of	N. of	Dec.
	inst.	att.	Class
zoo	101	16	7
lymphography	148	18	4
wisconsin breast cancer	699	9	2
tic-tac-toe	958	9	2
car	1728	6	4
kr-vs-kp	3196	36	3
nursery	12960	8	5

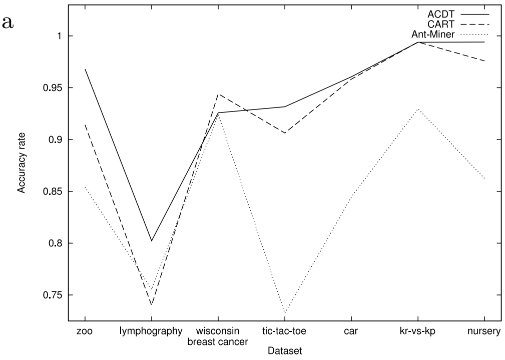


Fig. 3. Results (accuracy rate)

5.2 Results

To compare the performance of the different algorithms, we will consider three evaluation criteria over decision trees: the accuracy, measured by the ratio of the correctly classified examples in the testing set, and the tree size and run time of the analyzed approaches. Several of our experiments, which are not reported here, confirmed that relying on the tree size results in better decision trees.

The accuracy achieved by ACDT, as shown in the Table 2 and Figure 3 is better than that of CART and Ant–Miner, in some data sets. CART achieved similar results (only for two data sets better classification in the case of wisconsin breast cancer and kr-vs-kp data sets were obtained). For two analyzed data sets, the decrease in the size of the generated trees included in CART is accompanied by an increase in predictive power. The run time in the case of ACDT is decreasing in the comparison with the runtime for Ant–Miner approach. ACDT performs better than its predecessor Ant–Miner (10% better in the most analyzed data sets).

The promising outcome of the experiments shows that ACDT can deals with relatively large data sets. It should be noticed that for all data sets, good quality decision trees were found, both in terms of the classification accuracy and the tree size. It can be also observed that classification time scales almost linearly with the data sets.

Table. 2. Comparative study (standard deviations in parentheses)

Dataset	ACDT algorithm			CART (Salford Systems)			Ant–Miner algorithm		
	Acc.	Size	Time	Acc.	Size	Time	Acc.	Size	Time
zoo	**0.9680** (0.0383)	8.0	0.1	0.9142	6.0	0.0	0.8541 (0.0165)	5.0	0.3
lymphography	**0.8023** (0.0483)	15.9	0.4	0.7400	10.0	1.0	0.7554 (0.0405)	4.7	0.9
wisconsin breast cancer	0.9258 (0.0111)	21.7	0.7	**0.9442**	5.0	0.0	0.9244 (0.0094)	9.9	5.0
tic-tac-toe	**0.9316** (0.0158)	50.6	0.8	0.9063	30.0	0.0	0.7324 (0.0155)	7.7	0.9
car	**0.9606** (0.0095)	74.6	0.8	0.9583	33.0	0.0	0.8446 (0.0181)	14.8	1.3
kr-vs-kp	0.9939 (0.0011)	37.4	2.6	**0.9941**	22.0	2.0	0.9297 (0.0096)	7.2	80.8
nursery	**0.9941** (0.0018)	216.8	4.9	0.9759	148.0	4.0	0.8623 (0.0035)	17.3	10.6

Abbrev.: Acc. – accuracy rate (standard deviations in parentheses); Size – number of nodes / rules; Time – runtime of the approach in sec.

6 Conclusions

In recent years, ACO has proved its value and has been successfully applied in a variety of domains, especially in combinatorial optimization problems. However, what remains as an obstacle, ACO in data mining tasks is still opaque. The lack of transparency can be overcome through decision trees' construction, the interesting domain in data mining and what is more, the high effectiveness of the performance of ACDT – our proposition in this area.

In this paper we examined the new method for constructing decision trees – ACDT. The proposed algorithm was simulated and compared with two approaches: CART and Ant–Miner for different data sets. The results showed that the proposed modifications were similar to the classical approach and they can preserve high value of predictive accuracy.

The advantages of metaheuristics algorithms in constructing decision trees are: they work with population of ants which create candidate solutions. Secondly, they evaluate candidate solution as a whole by the objective function and they examine the pheromone values cumulated by the agents on the appropriate routes. These characteristics are in contrast with most deterministic methods, which work with a single candidate solution at a time. Moreover, they use probabilistic procedures that make them less prone to get trapped into local minima in the search space.

Empirical evidence confirms that, in general, Ant Colony Optimization in data mining copes with attribute interactions better than other algorithms. In the future work, we plan to provide a more flexible algorithm for creating a decision forest. Furthermore, we want to apply the proposed algorithm to continuous attribute values in larger data sets.

References

1. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Chapman & Hall, New York (1984)
2. Clark, P., Niblett, T.: The CN2 rule induction algorithm. Machine Learning 3(4), 261–283 (1989)
3. Corne, D., Dorigo, M., Glover, F.: New Ideas in Optimization. Mc Graw–Hill, Cambridge (1999)
4. Doerner, K.F., Merkle, D., Stützle, T.: Special issue on ant colony optimization. Swarm Intelligence 3(1), 1–2 (2009)
5. Dorigo, M., Caro, G.D.: New Ideas in Optimization. McGraw–Hill, London (1999)
6. Dorigo, M., Caro, G.D., Gambardella, L.: Ant algorithms for distributed discrete optimization. Artif. Life 5(2), 137–172 (1999)
7. Dorigo, M., Gambardella, L.M.: Ant Colony System: A cooperative learning approach to the Traveling Salesman Problem. IEEE Trans. Evol. Comp. 1, 53–66 (1997)
8. Dorigo, M., Stützle, T.: Ant Colony Optimization. MIT Press, Cambridge (2004)
9. Ant Colony Optimization and Swarm Intelligence. In: Dorigo, M., et al. (eds.) ANTS 2008. LNCS, vol. 5217, Springer, Heidelberg (2008)
10. Dorigo, M., Birattari, M., Stützle, T., Libre, U., Bruxelles, D., Roosevelt, A.F.D.: Ant colony optimization – artificial ants as a computational intelligence technique. IEEE Comput. Intell. Mag. 1, 28–39 (2006)
11. Galea, M.: Applying swarm intelligence to rule induction. Master's thesis, University of Edingbourgh (2002)
12. Grasse, P.P.: Termitologia, vol. II. Paris, Masson (1984)
13. Hyafil, L., Rivest, R.: Constructing optimal binary decision trees is NP–complete. Inf. Process. Lett. 5(1), 15–17 (1976)
14. Murphy, O., McCraw, R.: Designing Storage Efficient Decision Trees. IEEE Transactions on Computers 40, 315–320 (1991)
15. Parpinelli, R.S., Lopes, H.S., Freitas, A.A.: An ant colony algorithm for classification rule discovery. In: Abbas, H., Sarker, R., Newton, C. (eds.) Data Mining: a Heuristic Approach, Idea Group Publishing, London (2002)
16. Parpinelli, R.S., Lopes, H.S., Freitas, A.A.: Data mining with an ant colony optimization algorithm. IEEE Transactions on Evolutionary Computation, Special issue on Ant Colony Algorithms, 321–332 (2004)
17. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco (1993)
18. Rokach, L., Maimon, O.: Data Mining With Decision Trees: Theory And Applications. World Scientific Publishing, Singapore (2008)

A Cross-Entropy Based Population Learning Algorithm for Multi-mode Resource-Constrained Project Scheduling Problem with Minimum and Maximum Time Lags

Piotr Jędrzejowicz[1] and Aleksander Skakovski[2]

[1,2] Gdynia Maritime University
[1] Chair of Information Systems
[2] Department of Navigation
[1,2] ul. Morska 83F, 81-225 Gdynia, Poland
{P.Jedrzejowicz,askakow}@am.gdynia.pl

Abstract. The multi-mode resource-constrained project scheduling problem with minimum and maximum time lags is considered in the paper. An activity is performed in a mode, which determines the demand of renewable and nonrenewable resources required for its processing and minimum and maximum time lags between adjacent activities. The goal is to find a mode assignment to the activities and their start times such that all constraints are satisfied and the project duration is minimized. Because the problem is NP-hard a population-learning algorithm (PLA2) is proposed to tackle the problem. PLA2 is a population-based approach which takes advantage of the features common to the social education system rather than to the evolutionary processes. The proposed approach perfectly suits for multi-agent systems because it is based on the idea of constructing a hybrid algorithm integrating different optimization techniques complementing each other and producing a synergetic effect. Results of the experiment were compared to the results published in Project Scheduling Problem Library.

Keywords: multi-mode resource-constrained project scheduling, minimum and maximum time lags, evolutionary computation, population learning algorithm.

1 Introduction

Some real-life technological projects in civil engineering, for example, process industries (chemical industries or food industries), can be modeled as multi-mode resource-constrained project scheduling problem with minimum and maximum time lags (MRCPSP/max). An activity there requires for its processing a set of renewable (machines, manpower) and a set of nonrenewable resources (money, fuel, energy). Moreover, minimum and maximum time lags between activities are introduced, which put temporal constraints on the start times of activities. For example, between mixing of concrete and some concreting work a minimum time lag should be obeyed, since concreting cannot start earlier than the mixing has been completed. On the other hand, a

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 383–392, 2010.

maximum time lag imposed by the properties of concrete should be introduced between mixing and concreting. Concreting must be carried out within some time after mixing, otherwise concrete becomes hard. Additionally, an activity can be carried out in different ways - execution modes which differ in activity processing time, resource requirements and time lags to other activities. Execution modes imply time-resource and resource-resource tradeoffs [8]. The objectives can be the maximalization of the net present value, the minimalization of resource costs, or the minimalization of the project duration (makespan). In the paper, our goal is to determine a mode and a start time for each activity such that the time lags and resource constraints are satisfied and the project duration is minimized. In literature, the considered problem is denoted by $MPS|temp|C_{max}$ [5] and $m, 1T\ |gpr, mu|C_{max}$ [13]. It should be mentioned that MRCPSP/max is a generalization of two considered in literature problems: the resource-constrained project scheduling problem with minimum and maximum time lags (RCPSP/max) $PS|temp|C_{max}$ or $m, 1|gpr|C_{max}$ [7], where activity is carried out only in one execution mode and the multi-mode resource-constrained project scheduling problem (MRCPSP) $MPS|prec|C_{max}$ or $m, 1T\ |cpm, mu|C_{max}$ [26], [27] where only minimum time lags are considered. A thorough and explicit discussion on project scheduling problems one can find in [23]. Because the cosidered MRCPSP/max is a generalization of RCPSP/max it is NP-hard [2].

A population-learning algorithm (PLA) first proposed in [17] is used to tackle the problem since it was effective in solving other scheduling problems [14], [15], [16]. PLA is based on the idea of constructing a hybrid algorithm integrating different optimization techniques complementing each other and producing a synergetic effect. In the paper we consider PLA2 [19] where three procedures are used: cross-entropy (CE) described in Section 3.1, Tabu Search (TS) procedure and an island-based evolution algorithm (IBEA) briefly described in Section 3.2 and 3.3, and thoroughly in [18], and [19] respectively.

2 Problem Formulation

We define a multi-mode resource-constrained project scheduling problem with minimum and maximum time lags in the same way as in [12]. Namely, let a project consist of a finite set $V = \{0,1, . . . , n, n + 1\}$ of activities. Activity 0 represents the start and activity $n + 1$ the completion of the project. Both are dummy activities. For each activity $i \in V$ a set $M_i = \{1, . . . , | M_i |\}$ of (execution) modes is available. Activities 0 and $n + 1$ can be executed in only one mode: $M_0 = M_{n+1} = \{1\}$. Each activity $i \in V$ has to be performed in exactly one mode $m_i \in M_i$. The processing time of activity i executed in mode m_i is denoted by $p_i(m_i) \in Z_{\geq 0}$. The processing time of activities 0 and $n + 1$ equals 0, i.e. $p_0(1) = p_{n+1}(1) = 0$. S_i and C_i stand for the start time and the completion time (of the performance) of activity i, respectively. If we define $S_0 = 0$, S_{n+1} stands for the project duration. If activity i starts in mode m_i at time S_i, it is being executed at each point in time $t \in [S_i, S_i+p_i(m_i))$.

Between the start time S_i of activity i, which is performed in mode $m_i \in M_i$, and the start time S_j of activity j ($j \neq i$), which is performed in mode $m_j \in M_j$, a minimum time lag $d^{\min}_{i(m_i)j(m_j)} \in Z_{\geq 0}$ or a maximum time lag $d^{\max}_{i(m_i)j(m_j)} \in Z_{\geq 0}$ can be given. A time lag between activities i and j depends on mode m_i as well as on mode m_j.

Activities and time lags are represented by an activity-on-node (AoN) network $N = \langle V, E, \delta \rangle$ with node set V, arc set E, and arc weight function δ. Each element of node set V represents an activity. In the rest of the paper we will use this terms alternatively. An arc $\langle i, j \rangle \in E$ indicates that there is given a time lag between S_i and S_j. Arc weight function δ assigns to each arc $\langle i, j \rangle \in E$ a $|M_i| \times |M_j|$-matrix of arc weights as follows: for a minimum time lag $d^{\min}_{i(m_i)j(m_j)}$ we set $\delta_{i(m_i)j(m_j)} = d^{\min}_{i(m_i)j(m_j)}$, and for a maximum time lag $d^{\max}_{j(m_j)i(m_i)}$ we set $\delta_{i(m_i)j(m_j)} = -d^{\max}_{j(m_j)i(m_i)}$.

Since activity 0 represents the start of the project, it has to be ensured that no real activity can start before activity 0. For this reason, N is adapted as follows: if $\langle 0, i \rangle \notin E$ $(i \in V \setminus \{0, n+1\})$ holds, an arc $\langle 0, i \rangle$ with $\delta_{0(1)i(m_i)} = 0$ is introduced. Else, we set $\delta_{0(1)i(m_i)} = \max\{\delta_{0(1)i(m_i)}, 0\}$ $(m_i \in M_i)$. It has also to be ensured that no real activity can be completed after activity $n + 1$ since it represents the completion of the project. Hence, N can be adapted as follows: if $\langle i, n+1 \rangle \notin E$ $(i \in V \setminus \{0, n+1\})$ holds, an arc $\langle i, n + 1 \rangle$ with $\delta_{i(m_i)n+1(1)} = p_i(m_i)$, $(m_i \in M_i)$ is introduced. Otherwise, $\delta_{i(m_i)n+1(1)} = \max\{\delta_{i(m_i)n+1(1)}, p_i(m_i)\}$, $(m_i \in M_i)$.

R^{ρ} and R^{ν} denote the set of renewable resources and the set of nonrenewable resources, respectively. $R^{\rho}_k \in Z_{>0}$ stands for the capacity of renewable resource k $(k \in R^{\rho})$ which is available at each point in time. $R^{\nu}_k \in Z_{>0}$ stands for the capacity of nonrenewable resource k $(k \in R^{\nu})$ which is available in total. Provided that activity i is performed in mode m_i, $r^{\rho}_{i(m_i)k}$ units of renewable resource k $(k \in R^{\rho})$ are used at each point in time at which activity i is being executed, and $r^{\nu}_{i(m_i)k}$ units of nonrenewable resource k $(k \in R^{\nu})$ are consumed in total. For activities 0 and $n+1$ we set $r^{\rho}_{0(1)k} = r^{\rho}_{n+1(1)k} = 0$, k $(k \in R^{\rho})$ and $r^{\nu}_{0(1)k} = r^{\nu}_{n+1(1)k} = 0$, k $(k \in R^{\nu})$.

A schedule (M, S) defined in [12] consists of a mode vector M and a start time vector S. A mode vector $M = (m_i)_{i \in V}$ assigns to each activity $i \in V$ exactly one mode $m_i \in M_i$ as execution mode of i. A start time vector $S = (S_i)_{i \in V}$ assigns to each activity $i \in V$ exactly one point in time $t \geq 0$ as start time S_i with $S_0 = 0$.

With $A(M, S, t) = \{\, i \in V \mid S_i \leq t < S_i + p_i(m_i)\}$ denoting the set of activities being executed at time t for a given schedule (M, S), MRCPSP/max can be stated as follows:

$$\text{Min. } S_{n+1} \tag{1}$$

s.t.

$$S_j - S_i \geq \delta_{i(m_i)j(m_j)} \quad (\langle i, j \rangle \in E), \tag{2}$$

$$\sum_{i \in A(M,S,t)} r^{\rho}_{i(m_i)k} \le R^{\rho}_k \quad (k \in R^{\rho};\ t \ge 0), \tag{3}$$

$$\sum_{i \in V} r^{\nu}_{i(m_i)k} \le R^{\nu}_k \quad (k \in R^{\nu}), \tag{4}$$

$$m_i \in M_i \quad (i \in V), \tag{5}$$

$$S_i \ge 0 \quad (i \in V), \tag{6}$$

$$S_0 = 0. \tag{7}$$

The objective is to determine a schedule (M, S) such that the time lags (2) are observed, the constraints w.r.t. the renewable resources (3) and nonrenewable resources (4) are met, and the project duration is minimized (1). Such a schedule is called optimal. A schedule (M, S) satisfying (2) - (4) is called feasible.

3 Population Learning Algorithm

Population learning algorithm proposed in [17] has been inspired by analogies to a social phenomenon rather than to evolutionary processes. The population learning algorithm takes advantage of features that are common to social education systems:

- A generation of individuals enters the system.
- Individuals learn through organized tuition, interaction, self-study and self-improvement.
- Learning process is inherently parallel with different schools, curricula, teachers, etc.
- Learning process is divided into stages.
- More advanced and more demanding stages are entered by a diminishing number of individuals from the initial population (generation).
- At higher stages more advanced education techniques are used.
- The final stage can be reached by only a fraction of the initial population.

General idea of the present implementation of PLA2 is shown in the following pseudo code:

```
Procedure PLA2
Begin
  1. Create an initial population $P_0$ of the size $x_0$ - 1
     using procedure cross-entropy (CE).
  2. Create an individual TSI in which all tasks $i$ are to
     be executed in mode $m_i$ characterized by maximal task
     processing time.
  3. Improve the individual TSI with the tabu search (TS)
     procedure.
  4. Create population $P_1$ = $P_0$ + TSI.
  5. Improve all individuals in $P_1$ with the IBEA.
  6. Output the best solution to the problem.
End.
```

In the procedure PLA2, $x_0 = K \cdot PS$, where K – the number of islands and PS – the population size on an island defined in procedure IBEA.

All individuals (solutions) used in the PLA2 procedure can be characterized in the following manner:

- all processing modes of all activities are numbered consecutively. Thus processing mode m_b of task J_b has the number $c_b = \sum_{i=0}^{b-1} |M_i| + m_b$,
- an individual (a solution) is represented by an $(n + 2)$-element vector $S = \{c_i |\ 0 \leq i \leq n + 1\}$,
- all S representing feasible solutions are potential individuals;
- each individual can be transformed into a schedule by applying LSG, which is a specially designed list-scheduling algorithm, which schedules activities in S one by one minimizing their start times and satisfying all constraints of MRCPSP/max;
- each schedule produced by the LSG can be directly evaluated in terms of its fitness, unfeasible schedules are penalized (a description of LSG for discrete-continuous scheduling is given in [19]).

3.1 A Cross-Entropy Algorithm

In PLA2 the proposed CE procedure is perceived as the procedure preparing some solution basis for further improvement by procedure IBEA. In CE procedure a cross-entropy method first proposed in [25] is used since it was effective in solving various difficult combinatorial optimization problems [4]. Because in CE procedure a solution is viewed as a vector of n jobs, we would like to know the probability of locating job i on a particular place j in the vector. For this reason we introduce two success probability vectors $\hat{p}_j$ and $\hat{p}'_{ji}$ related to each job i and its place j in solution S. Vector $\hat{p}_j = \{p_{ji} \,|\, 1 \leq i \leq n\}$, $1 \leq j \leq n$ contains p_{ji} values, which is the probability that on place j there will be located job i. Vector $\hat{p}'_{ji} = \{p_{jil} \,|\, 1 \leq l \leq D_i\}$, $1 \leq j \leq n$, $1 \leq i \leq n$ contains p_{jil} values, which is the probability that on place j task i will be executed in mode l. A procedure CE using cross-entropy method for combinatorial optimization described in [3] and modified for solving MRCPSP/max problem is shown in the following pseudo code:

```
Procedure CE
Begin
  1. Set ic:= 1 (ic - iteration counter), $ic^{stop}$ - maximal
     number of iterations, a:= 1.
  2. Set $\hat{p}_j = \{p_{ji} = 1/n \,|\, 1 \leq i \leq n\}, 1 \leq j \leq n$.
  3. Set $\hat{p}'_{ji} = \{p_{jil} = 1/D_i \mid 1 \le l \le D_i\}, 1 \le j \le n, 1 \le i \le n.$
  4. While $ic$ • $ic^{stop}$ do
    4.1. Generate a sample $S_1$, $S_2$, … , $S_s$, … , $S_N$ of solutions with success probability vectors $\hat{p}_j$ and $\hat{p}'_{ji}$.
    4.2. Order $S_1$, $S_2$, … , $S_s$, … , $S_N$ by nondecreasing values of their fitness function.
    4.3. Set $\gamma = \lceil \rho \cdot N \rceil, \rho \in (0,1)$.
    4.4. Set $\hat{p}_j = \left\{ p_{ji} = \frac{\sum_{s=1}^{\gamma} I(S_s(j) = i)}{\gamma} \middle| 1 \le i \le n \right\}$,    (2)
         $1 \le j \le n$, $I(S_s(j) = i) = 1$, $I(S_s(j)$ • $i) = 0$, where $S_s(j)$ - number of the task located on $j$-th place in $s$-th solution $S$, $D_i$ - the number of available task modes.
    4.5. Set $\hat{p}'_{ji} = \left\{ p_{jil} = \frac{\sum_{s=1}^{\gamma} I(S_s(ji) = l)}{\gamma} \middle| 1 \le l \le D_i \right\}$,    (3)
         $1 \le j \le n$, $1 \le i \le n$, $I(S_s(ji) = l) = 1$, $I(S_s(ji)$ • $l) = 0$, where $S_s(ji)$ - an execution mode of task $i$ located on $j$-th place in $s$-th solution $S$.
    4.6. Save the first $h = \lceil K \cdot PS / ic^{stop} \rceil$ best solutions from the ordered sample into $P_0$ under address $a$. Set $a := a + h$.
    4.7. Set $ic := ic + 1$.
  EndWhile.
EndProcedure.
```

In the presented pseudo code parameters K – the number of islands and PS – the population size defined in procedure IBEA.

3.2 A Tabu Search Algorithm

Tabu search is a metaheuristic used in PLA2 (see [9]). It is a local search procedure proved by other researchers to be efficient in solving difficult computational problems [9], [21]. In the present implementation of the tabu search procedure we introduce the neighborhoods N_t and N_{md} of a solution S. N_t is a set of solutions generated from S by moving a task $J_i \in S$ from place i to the rest $n - 1$ places. Thus we yield $|N_t| = n \cdot (n - 1)$ neighbors. N_{md} is a set of solutions generated from S by assigning to task $J_i \in S$ one by one in a row all of its D_i modes. Thus we yield another $|N_{md}| = n \cdot (D_i - 1)$ neighbors. The considered tabu search procedure is shown in the following pseudo code:

```
Procedure TS
Begin
 1. Set $S_0$ = initial solution $TSI$ ($l_i$ = 1, $1 \bullet i \bullet n$).
 2. Set the best solution $S_{best}$ = $S_0$.
 3. Set Tabu List TL = {∅}.
 4. Set $N_t$ = {$S_I$} and $N_{md}$ = {∅}.
 5. Repeat the following $max\_number\_of\_iterations$ times:
  5.1. Find the best legal neighbour $S_{bln}$ of $S_0$, i.e. the
       best across $N_t$ and $N_{md}$ neighbour which is not on
       TL.
  5.2. Set $S_0$ = $S_{bln}$.
  5.3. If $S_{bln}$ is more fit than $S_{best}$ then $S_{best}$ = $S_{bln}$.
  5.4. Put $S_{bln}$ on the Tabu list.
  5.5. If the fitness of $S_0$ has not improved after $nit$
       number of iterations construct a new solution by
       moving a task $i$ in $S_0$ to one of the chosen ran-
       domly less frequently visited places on the task
       list and assigning to it one of the chosen ran-
       domly less frequently assigned execution modes.
 EndRepeat.
EndProcedure.
```

3.3 An Island-Based Evolutionary Algorithm

The island-based approach brings two benefits: a model that maps easily onto the parallel hardware and extended search area (due to multiplicity of islands) preventing from being caught in local optima [1], [2], [6], [10], [20]. The following pseudo-code shows main stages of the IBEA algorithm:

```
Procedure IBEA
Begin
 1. Set the number of islands $K$, the number of generations
    $PN$ for each island, the size $PS$ of the population on
    each island. For each island $I_k$, form an initial popu-
    lation selecting at random $PS$ individuals from $P_1$.
 2. While no stopping criteria is met do
  2.1. For each island $I_k$ do
   2.1.1. Evolve $PN$ generations using PBEA.
   2.1.2. Send the best solution to $I_{(k \bmod K) + 1}$.
   2.1.3. Incorporate the best solution from
          $I_{((K+k-2) \bmod K) + 1}$ instead of the best one.
  EndWhile.
 3. Find the best solution $S_{best}$ across all islands and
    save it as the final one.
EndProcedure.
```

PBEA (Population-Based Evolutionary Algorithm) iteratively improves a population of individuals using selection, crossover, and mutation operators selecting the best individual at the final stage. The pseudo-code of PBEA the reader can find in [19].

4 Computational Experiment

The proposed a version of cross-entropy based population learning algorithm - PLA2 for solving the multi-mode resource-constrained project scheduling problem with minimum and maximum time lags was implemented and tested. The test set of benchmark instances used for evaluation of PLA2 was downloaded from a project scheduling problem library (PSPLIB) discussed in [22] and available in [24]. The results yielded by PLA2 were compared to the results of three heuristics Prio, TS_{DR}, TS_F considered in [11], because problem instances to test them were generated, as well as for PLA2, using ProGen for MRCPSP/max with the same parameter settings. All in all, PLA2 was tested on 270 problem instances built on 100 real activities, where in each one third of the instances an activity was carried out in 3, 4 or 5 execution modes respectively. PLA2 was run on Pentium (R) 4 CPU 3.00GHz personal computer and compiled by Borland Delphi Personal v.7.0 under operating system Windows XP Home Edition.

We use the same evaluation criteria as in [11], namely, the percentage of problem instances for which a feasible solution has been determined by the respective heuristic and the average percentage deviation of the objective function value from a lower bound value for those instances for which a feasible solution has been determined by the heuristic within the given amount of time equal 100 seconds. The values for the evaluation criteria used are given in Table 1 and Table 2 respectively. The experimental results prove good, but not the best performance of PLA2 among the considered heuristics. Despite the low percentage of problem instances for which a feasible solution has been determined and high deviation of the objective function value from a lower bound value, PLA2 can achieve much higher performance, as it was shown in [19] while solving other computationally difficult problems, subject to the appropriate to MRCPSP/max tuning.

Table 1. The percentage of feasible solutions found within 100 s

m	PLA2	TS_{DR}	TS_F	Prio
3	31	53	39	100
4	35	61	46	100
5	47	67	59	100

Table 2. The average deviation from a lower bound in percents

m	PLA2	TS_{DR}	TS_F	Prio
3	124	40	105	63
4	165	91	151	113
5	241	164	215	170

5 Conclusion

In his paper we propose a hybrid population-based approach to solving the the multi-mode resource-constrained project scheduling problem with minimum and maximum time lags to minimize the project duration (makespan). Since the discussed problem is computationally difficult it has been possible to obtain only approximate solutions within a reasonable time. In such case the validation of the approach seems only possible through computational experiment and comparison of the results with those obtained by applying other approaches. The results of the experiment discussed in the paper proved that combining cross-entropy, tabu search and island-based evolutionary algorithm within the population based scheme is an effective approach. Further research should focus on further reducing computation time needed to obtain satisfactory results.

References

1. Alba, E., Troya, J.: Analysis of Synchronous and Asynchronous Parallel Distributed Genetic Algorithms with Structured and Panmictic Islands. In: Rolim, J., et al. (eds.) Proceedings of the 10th Symposium on Parallel and Distributed Processing, San Juan, Puerto Rico, USA, April 12-16, pp. 248–256 (1999)
2. Bartusch, M., Mohring, R., Radermacher, F.J.: Scheduling project networks with resource constraints and time windows. Annals of Operations Research 16, 201–240 (1988)
3. Belding, T.C.: The Distributed Genetic Algorithm Revisited. In: Eshelman, L.J. (ed.) Proceedings of the Sixth International Conference on Genetic Algorithms, pp. 114–121. Morgan Kaufmann, San Francisco (1995)
4. De Boer, P.-T., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A Tutorial on the Cross-Entropy Method. Annals of Operations Research 134(1), 19–67 (2005)
5. Brucker, P., Drexl, A., Mohring, R., Neumann, K., Pesch, E.: Resource-constrained project scheduling: notation,classification, models, and methods. European Journal of Operational Research 112, 3–41 (1999)
6. Czarnowski, I., Gutjahr, W.J., Jędrzejowicz, P., Ratajczak, E., Skakovski, A., Wierzbowska, I.: Scheduling Multiprocessor Tasks in Presence of Correlated Failures. In: Luptaćik, M., Wildburger, U.L. (eds.) Central European Journal of Operations Research, vol. 11(2), pp. 163–182. Physika-Verlag, A Springer-Verlag Company, Heidelberg (2003)
7. Dorndorf, U., Pesch, E., Toan, P.H.: A time-oriented branch-and-bound algorithm for project scheduling with generalised precedence constraints. Management Science 46, 1365–1384 (2000)
8. Elmaghraby, S.E.: Activity Networks—Project Planning and Control by Network Models. Wiley, New York (1977)
9. Glover, F.: Tabu search: a tutorial. Interfaces 20, 74–94 (1990)
10. Gordon, V.S., Whitley, D.: Serial and Parallel Genetic Algorithms as Function Optimizers. In: Forrest, S. (ed.) Proceedings of the Fifth International Conference on Genetic Algorithms, pp. 177–183. Morgan Kaufmann, San Mateo (1993)
11. Heilmann, R.: Resource-constrained project scheduling: a heuristic for the multi-mode case. OR Spektrum 23, 335–357 (2001)
12. Heilmann, R.: A branch-and-bound procedure for the multi-mode resource-constrained project scheduling problem with minimum and maximum time lags. European Journal of Operational Research 144, 348–365 (2003)
13. Herroelen, W., Demeulemeester, E., De Reyck, B.: A classification scheme for project scheduling. In: Węglarz, J. (ed.) Project Scheduling: Recent Models, Algorithms and Applications, pp. 1–26. Kluwer, Boston (1999)
14. Jędrzejowicz, J., Jędrzejowicz, P.: Population–Based Approach to Multiprocessor Task Scheduling in Multistage Hybrid Flowshops Knowledge-Based Intelligent Information and Engineering Systems. In: Palade, V., Howlett, R.J., Jain, L. (eds.) KES 2003. Lecture Notes in Computer Science, LNAI, vol. 2773, pp. 279–286. Springer, Heidelberg (2003)
15. Jędrzejowicz, J., Jędrzejowicz, P.: PLA–Based Permutation Scheduling. Foundations of Computing and Decision Sciences 28(3), 159–177 (2003)

16. Jędrzejowicz, J., Jędrzejowicz, P.: New Upper Bounds for the Permutation Flowshop Scheduling Problem. In: Ali, M., Esposito, F. (eds.) IEA/AIE 2005. LNCS (LNAI), vol. 3533, pp. 232–235. Springer, Heidelberg (2005)
17. Jędrzejowicz, P.: Social Learning Algorithm as a Tool for Solving Some Difficult Scheduling Problems. Foundation of Computing and Decision Sciences 24, 51–66 (1999)
18. Jędrzejowicz, P., Skakovski, A.: A Population Learning Algorithm for Discrete-Continuous Scheduling with Continuous Resource Discretisation. In: Chen, Y., Abraham, A. (eds.) 6th International Conference on Intelligent Systems Design and Applications, ISDA 2006 Special session: Nature Imitation Methods Theory and practice (NIM 2006), October 16-18, vol. II, pp. 1153–1158. IEEE Computer Society, Jinan (2006)
19. Jędrzejowicz, P., Skakovski, A.: A cross-entropy-based population-learning algorithm for discrete-continuous scheduling with continuous resource discretisation. Elsevier, Neurocomputing 73(4-6), 655–660 (2010)
20. Jędrzejowicz, P., Czarnowski, I., Skakovski, A., Szreder, H.: Evolution-based scheduling of multiple variant and multiple processor programs. In: Hertzberger, L.O., Sloot, P.M.A. (eds.) Future Generation Computer Systems 17, pp. 405–414. Elsevier, The Netherlands (2001)
21. Józefowska, J., Mika, M., Różycki, R., Waligóra, G., Węglarz, J.: Solving discrete-continuous scheduling problems by Tabu Search. In: 4th Metaheuristics International Conference MIC 2001, Porto, Portugal, July 16-20, pp. 667–671 (2001)
22. Kolish, R., Sprecher, A.: PSPLIB - A project scheduling problem library. European Journal of Operational Research 96, 205–216 (1996)
23. Neumann, K., Schwindt, C., Zimmermann, J.: Project Scheduling with Time Windows and Scarce Resources. Springer, Heidelberg (2003)
24. Project Scheduling Problem Library, `http://129.187.106.231/psplib/`
25. Rubinstein, R.Y.: Optimization of computer simulation models with rare events. European Journal of Operations Research 99, 89–112 (1997)
26. Słowiński, R., Soniewicki, B., Węglarz, J.: DSS for multiobjective project scheduling. European Journal of Operational Research 79, 220–229 (1994)
27. Sprecher, A.: Resource-Constrained Project Scheduling — Exact Methods for the Multi-Mode Case. Springer, Berlin (1994)

Evaluation of Agents Performance within the A-Team Solving RCPSP/Max Problem

Piotr Jedrzejowicz and Ewa Ratajczak-Ropel

Department of Information Systems
Gdynia Maritime University, Poland
{pj,ewra}@am.gdynia.pl

Abstract. In this paper the E-JABAT-based A-Team architecture dedicated for solving the RCPSP/max problem instances is proposed and experimentally validated. Computational experiment involves evaluation of agent performance within the A-Team. The paper contains the RCPSP/max problem formulation, description of the E-JABAT architecture for solving the RCPSP/max problem instances and the discussion of the computational experiment results.

1 Introduction

The paper proposes an agent-based approach to solving instances of the resource constrained project scheduling problem with minimal and maximal time lags (time windows) known in the literature as the RCPSP/max. RCPSP/max has attracted a lot of attention and many exact and heuristic algorithms have been recently proposed for solving it (see for example [4], [11]). Since the problem is known to be NP-hard the approaches proposed so far produce either approximate solutions or can be applied for solving instances of the limited size. Hence, searching for more effective algorithms and solutions to the RCPSP/max problem is still a lively field of research. One of the promising directions of such research is to take advantage of the parallel and distributed computation abilities, which are the feature of the contemporary multiple-agent systems.

The multiple-agent systems are an important and intensively expanding area of research and development. There is a number of multiple-agent approaches proposed to solve different types of optimization problems. One of them is the concept of an asynchronous team (A-Team), originally introduced in [14]. The A-Team paradigm was used to develop the JADE-based environment for solving a variety of computationally hard optimization problems called E-JABAT [2]. E-JABAT is a middleware supporting the construction of the dedicated A-Team architectures based on the population-based approach. The mobile agents used in E-JABAT allow for decentralization of computations and use of multiple hardware platforms in parallel, resulting eventually in more effective use of the available resources and reduction of the computation time.

In this paper the E-JABAT-based A-Team architecture dedicated for solving the RCPSP/max problem instances is proposed and experimentally validated.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 393–402, 2010.

To solve instances of the RCPSP/max different optimization agents are used. Optimization agents represent heuristic algorithms. The proposed architecture is an extension of the one proposed in [9]. The proposed approach has been validated experimentally.

Section 2 of the paper contains the RCPSP/max problem formulation. Section 3 gives some information on E-JABAT environment. Section 4 provides details of the E-JABAT architecture implemented for solving the RCPSP/max problem instances. Section 5 describes computational experiment carried-out with a view to validate the proposed approach. Section 6 contains conclusions and suggestions for future research.

2 Problem Formulation

In the resource-constrained project scheduling problem with minimal and maximal time lags (RCPSP/max) a set of $n+2$ activities $V = \{0, 1, \ldots, n, n+1\}$ is considered. Each activity has to be processed without interruption to complete the project. The dummy activities 0 and $n+1$ represent the beginning and the end of the project. The duration of an activity j, $j = 0, \ldots, n+1$ is denoted by d_j where $d_0 = d_{n+1} = 0$. There are r renewable resource types. The availability of each resource type k in each time period is r_k units, $k = 1, \ldots, r$. Each activity j requires $r_{j,k}$ units of resource k during each period of its duration, where $r_{0,k} = r_{n+1,k} = 0$, $k = 1, ..., r$. Each activity $j \in V$ has a start time s_j which is a decision variable. There are generalized precedence relations (temporal constraints) of the start-to-start type $s_j - s_i \geq \delta_{ij}$, $\delta_{ij} \in Z$, defined between the activities.

The structure of a project can be represented by an activity-on-node network $G = (V, A)$, where V is the set of activities and A is the set of precedence relationships. The objective is to find a schedule of activities starting times $S = [s_0, \ldots, s_{n+1}]$, where $s_0 = 0$ (project always begins at time zero) and resource constraints are satisfied, such that the schedule duration $T(S) = s_{n+1}$ is minimized. The detailed description of the problem can be found in [10].

The RCPSP/max, as an extension of the RCPSP, belongs to the class of NP-hard optimization problems ([1], [3]). The objective is to find a makespan minimal schedule that meets the constraints imposed by the precedence relations and the limited resource availabilities.

3 E-JABAT Environment

E-JABAT is a middleware allowing to design and implement A-Team architectures for solving various combinatorial optimization problems. The problem-solving paradigm on which the proposed system is based can be best defined as the population-based approach.

E-JABAT produces solutions to combinatorial optimization problems using a set of optimization agents, each representing an improvement algorithm. Each improvement (optimization) algorithm when supplied with a potential solution to the problem at hand, tries to improve this solution. An initial population of solutions (individuals) is generated or constructed. Individuals forming an initial population are, at the following computation stages, improved by independently acting agents. Main functionality of the proposed environment includes organizing and conducting the process of search for the best solution.

To perform the above described cycle two main classes of agents are used. The first class called OptiAgent is a base class for all optimization agents. The second class called SolutionManager is used to create agents or classes of agents responsible for maintenance and updating individuals in the common memory. All agents act in parallel. Each OptiAgent represents a single improvement algorithm (for example: simulated annealing, tabu search, genetic algorithm etc.).

Other important classes in E-JABAT include: Task representing an instance or a set of instances of the problem and Solution representing the solution. To initialize the agents and maintain the system the TaskManager and PlatformManager classes are used. Objects of the above classes also act as agents. Up to now the E-JABAT environment has been used to solve instances of the following problems: the resource-constrained project scheduling problem (RCPSP), the traveling salesman problem (TSP), the clustering problem (CP), the vehicle routing problem (VRP).

E-JABAT environment has been designed and implemented using JADE (Java Agent Development Framework), which is a software framework supporting the implementation of multi-agent systems. More detailed information about E-JABAT environment and its implementations can be found in [6] and [2].

4 E-JABAT Architecture for Solving the RCPSP/Max Problem

E-JABAT environment was successfully used for solving the RCPSP, MRCPSP and RCPSP/max problems (see [7], [9]). The approach proposed in this paper is the continuation of the research described in [9]. The A-Team architecture has been extended and improved.

The following E-JABAT ojects have been implemented: classes describing the problem, ontologies and optimization agents. The above sets of classes forms the package called RCPSPmax.

Classes describing the problem are responsible for reading and preprocessing the data and generating random instances of the problem. The discussed set includes the following classes:

- RCPSPmaxTask inheriting from the Task class and representing the instance of the problem,
- RCPSPmaxSolution inheriting from the Solution class and representing the solution of the problem instance,
- Activity representing the activity of the problem,
- Resource representing the renewable resource,
- PredSucc representing the predecessor or successor of the activity.

The next set includes classes allowing for definition of the vocabulary and semantics for the content of messages exchange between agents. In the proposed approach the messages include all data representing the task and solution. The discussed set includes the following classes:

- RCPSPmaxTaskOntology inheriting from the TaskOntology class,
- RCPSPmaxSolutionOntology inheriting from the SolutionOntology class,

The last set includes classes describing the optimization agents. Each of them includes the implementation of an optimization heuristic used to solve the problem. All of them are inheriting from OptiAgent class. In the considered case this set includes the following classes:

- optiLSA denoting the Local Search Algorithm (LSA),
- optiPRA denoting Path Relinking Algorithm (PRA),
- optiCA denoting Crossing Algorithm (CA),
- optiTSA denoting the Tabu Search Algorithm (TSA),
- opiGEPA denoting Gene Expression Programming Algorithm (GEPA).

The algorithms LSA, PRA and GEPA proposed in [9] has been slightly modified, however their main steps remain the same. The TSA algorithm was proposed in [13]. The shorten pseudo-code of TSA and the pseudo-code of proposed CA are presented in Figures 1 and 2 respectively.

The LSA is a simple local search algorithm which finds the local optimum by moving each activity to all possible places in the schedule. For each combination of activities the value of possible solution is calculated. The best schedule is returned.

The PRA is an implementation of the path-relinking algorithm. For a pair of solutions a path between them is constructed. The path consists of schedules obtained by carrying out a single move from the preceding schedule. The move is understood as moving one of the activities to a new position. For each schedule in the path the value of the respective solution is checked. The best schedule is remembered and finally returned.

The CA (Figure 2) is an algorithm based on the idea of one point crossover operator. For a pair of solutions one point crossover is applied. The *step* argument determines the frequency the operation is performed.

The TSA (Figure 1) is an implementation of tabu search algorithm. As the initial parameters the instance of solution ($initialSolution$) and the number of iteration in which no better solution is found ($iterationNumber$) have been used. In each iteration all possible moves for the considered solution S and the activity a_i (neighborhood $N(S, a_i)$) are checked. The move is defined as an exchange of two activities where the precedence and resource constraints are satisfied. The best move from the neighborhood which is not tabu is chosen and performed. The list of tabu moves is updated after performing the move.

The GEPA is an implementation of the gene expression program based on the gene expression programming idea proposed in [5]. The algorithm was described in details in [8]. The initial population of the GEPA consists of one individual

```
TSA($initialSolution$, $iterationNumber$)
{
   $S$=$initialSolution$;
   $tabuList$=$\emptyset$;
   while($iterationNumber$>0 || not all $a_i$ from $S$ are checked)
   {
      Choose $bestMove$ from the neighborhood $N$($S$,$a_i$) which is not tabu;
      Update $tabuList$: update tabu tenures;
      if($bestMoveValue$ is better more then 80% of the best solution known)
      {
         Make $bestMove$ on $S$;
         if($S$ is better then the best solution known)
         {
            Update $tabuList$:
               add moves from $a_i$ and $a_j$ for $0.8\cdot$iterationNumber,
               add moves to $a_i$ and $a_j$ for $0.8\cdot$iterationNumber;
            Remember the best solution;
         }
         Update $tabuList$:
            add moves from $a_i$ and $a_j$ for $0.05\cdot$iterationNumber),
            add moves from $a_i$ to $a_j$ and reverse one for $0.1\cdot$iterationNumber,
            add $bestMove$ and reverse one for $0.3\cdot$iterationNumber);
      }
      Update $iterationNumber$;
   }
}
```

Fig. 1. Pseudo code of TSA algorithm

selected from the current content of the E-JABAT common memory, and the rest generated randomly. The individual (solution) from the E-JABAT population is transformed into GEP individual representation. Next, the whole population is evolved using GEP operators: one point, two point and n-point recombination, restricted permutation, inversion, gene deletion/insertion and mutation. Each of the operators is used to produce an offspring to the fixed percent of individuals creating next generation. To calculate the fitness of an individual the function $f_x = T_g - t_x + 1$ is used, where x is the individual, t_x is the length of the schedule and T_g is the length of the largest schedule encoded in the chromosomes of the current population. The best schedule is remembered and finally returned.

All optimization agents co-operate together using the E-JABAT common memory. The initial population in the common memory is generated randomly with the exception of a few individuals which are generated by heuristics based on the priority rules [10]. Because it might be difficult to obtain feasible solution for some RCPSP/max problem instances, the random drawing of an individual could be repeated several times. If this does not produce enough feasible solutions the infeasible ones are added to the population in the common memory. All the proposed improvement algorithms can use infeasible solutions as an initial seed. Individuals in the common memory are represented as schedules of activities. The final solution is obtained from the schedule by SGSU [10].

```
CA($initialSolution1$, $initialSolution2$, $step$)
{
   for(each crossover point every $step$ activities)
   {
      Do one point crossover on $initialSolution1$ and $initialSolution2$
         and remember the result as $solution$;
      for(all priority rules used)
      {
         Use SGSU to $solution$;
         Remember the best solution;
      }
   }
}
```

Fig. 2. Pseudo code of the CA algorithm

The time and frequency an agent of each kind receives a solution or set of solutions from the common memory with a view to improve its quality is determined by the strategy. For solving the RCPSP/max problem instances the strategy where individuals forwarded to optimization agents for improvement are randomly chosen from the population stored in the common memory has been used. Such individuals are sent to optimization agents ready to start searching for a better solution. After computation the improved individual replaces the worst one stored in the common memory.

5 Computational Experiment Results

To validate the proposed approach and to evaluate the effectiveness of the optimization agents the computational experiment has been carried out using benchmark instances of RCPSP/max from PSPLIB [12] (test sets j10, j20 and j30). Each set includes 270 problem instances. The experiment involved computation with the fixed number of optimization agents representing LSA, PRA, CA, TSA, and GEPA algorithms, fixed population size, and the limited time period allowed for computation. To evaluate the effectiveness of agents they have been divided into two groups with the computational complexity as the criterion: a simple ones consisting of LSA, PRA and CA and a complex ones consisting of TSA and GEPA.

The discussed results have been obtained using 5 optimization agents: one of each kind. Population of solutions in the common memory consisted of 100 individuals. The computation has been stopped if improved solution has not been found for 60 seconds. The optimization algorithms have had a fixed parameter values described in section 4, but the number of iteration for two of them has been chosen randomly from the interval $[10, 50]$ in case of the LSA and from $[10, 200]$ in case of the TSA. The population of GEPA consisted of 50 genes which are evaluated through 10 generations.

Table 1. Performance characteristics of the group of simple agents

#Activities	Mean RE	% FS	Mean total CT [s]	Mean CT [s]
10	0.32%	100%	23.3	0.2
20	4.45%	100%	33.4	6.3
30	9.70%	100%	38.8	12.5

Table 2. Performance characteristics of the group of complex agents

#Activities	Mean RE	% FS	Mean total CT [s]	Mean CT [s]
10	0.32%	100%	23.2	0.2
20	4.24%	100%	29.7	2.7
30	9.34%	100%	35.1	9.4

Table 3. Performance characteristics of the group of all agents

#Activities	Mean RE	% FS	Mean total CT [s]	Mean CT [s]
10	0.32%	100%	23.2	0.1
20	4.16%	100%	31.0	3.9
30	9.24%	100%	35.3	9.4

Table 4. Literature reported results (ISES [4])

#Activities	Mean RE	% FS	Mean CT [s]
10	0.99%	100.00%	0.71
20	4.99%	100.00%	4.48
30	10.37%	100.00%	22.68

Table 5. C-BEST results [4]

#Activities	Mean RE	% FS	Mean CT [s]
10	0.00%	100.00%	–
20	3.97%	100.00%	–
30	8.91%	100.00%	–

The following characteristics have been calculated and recorded: mean relative error (Mean RE) calculated as the deviation from the lower bound, percent of feasible solutions (% FS), mean total computation time (Mean total CT) and mean computation time required to find the best solution (Mean CT).

Table 6. Percentage of the improved and best improved solutions obtained by the group of simple agents

#Activities	LSA	PRA	CA	Any	No one
		Percents of improved solutions			
10	38%	38%	38%	38%	62%
20	47%	47%	47%	47%	53%
30	43%	44%	44%	45%	55%
		Percents of best improved solutions			
10	02%	00%	0%	2%	98%
20	19%	06%	6%	25%	75%
30	21%	11%	7%	27%	73%

Table 7. Percentage of the improved and best improved solutions obtained by the group of complex agents

#Activities	TSA	GEPA	Any	No one
		Percents of improved solutions		
10	38%	38%	39%	61%
20	46%	47%	47%	53%
30	44%	45%	46%	54%
		Percents of best improved solutions		
10	1%	2%	3%	97%
20	8%	23%	25%	75%
30	14%	28%	33%	67%

Table 8. Percentage of the improved and best improved solutions obtained by the group of all agents

#Activities	LSA	PRA	CA	TSA	GEPA	Any	No one
				Percents of improved solutions			
10	38%	38%	38%	38%	38%	38%	62%
20	47%	46%	47%	46%	41%	47%	53%
30	41%	44%	40%	41%	32%	45%	55%
				Percents of best improved solutions			
10	0%	0%	0%	0%	2%	2%	98%
20	7%	1%	1%	4%	19%	25%	75%
30	11%	5%	4%	9%	18%	31%	69%

Each instance has been solved five times and the results have been averaged over these solutions. The results for each group of agents are presented in Tables 1-3. These results are compared with the results reported in the literature (Table 4). In Table 5 the optimal results are presented.

Additionally, the influence of each agent on the results was evaluated as the average percent of individuals which were improved by it and the average percent of the current best solutions found by it. The results in terms of the percent of solutions improved by each agent, percent of solutions improved by any agent from the group and percent of non-improved solutions by agents in the group are presented in Tables 6-8.

Experiment has been carried out using nodes of the cluster Holk of the Tricity Academic Computer Network built of 256 Intel Itanium 2 Dual Core 1.4 GHz with 12 MB L3 cache processors and with Mellanox InfiniBand interconnections with 10Gb/s bandwidth. During the computation one node per agent was used: 3 nodes for the group of simple algorithms, 2 for complex and 5 for all.

6 Conclusions

Experiment results show that the proposed E-JABAT based A-Team implementation is an effective tool for solving instances of the RCPSP/max problem. Presented results are better then the best solutions known from the literature. It can be also noticed that they are obtained in a shorter time. Time comparisons are however difficult since the discussed implementation have been using different numbers and kinds of processors. In case of the agent-based environments the significant part of the time is used for agent communication which has an influence on both - computation time and quality of the results.

It can be noticed that in a small problems of 10 activities the complexity of agents does not influence on the results obtained. In case of the problem size of 20 and 30 activities, agents of higher quality give better results but the best results are obtained by the group of simple and complex agents working together.

Evaluating particular agents it can be noticed that their contribution has been almost equal. Each agent has been producing the improved solutions for 38% to 47% of solutions from the population. Only in case of GEPA working in the group of all agents, the percent has been significantly lower for the problem sizes of 20 and 30 activities. This can be contributed to the computation time requirement of GEPA. Considering the best solutions found, differences became greater. In case of the group of simple agents the most effective one is the LSA agent but the experiment results show that the other two agents PRA and CA could find the best solutions for different individuals then LSA so co-operation of agents becomes quite effective. In case of the group of complex agents GEPA turned out much more effective and group-working also improved the results. Similarly in the group of all agents GEPA and LSA are the most effective ones.

Future research will concentrate on finding the best configuration of the heterogenous agents. Since the E-JABAT gives a possibility to run more then one copy of each agent it is interesting which agents should or should not be replicated to improve the results. Additionally, testing and adding to E-JABAT more different optimization agents could be considered.

Acknowledgments. The research has been supported by the Ministry of Science and Higher Education grant no. N N519 576438. Calculations have been performed in the Academic Computer Centre TASK in Gdansk.

References

1. Bartusch, M., Mohring, R.H., Radermacher, F.J.: Scheduling Project Networks with Resource Constraints and Time Windows. Annual Operational Research 16, 201–240 (1988)
2. Barbucha, D., Czarnowski, I., Jedrzejowicz, P., Ratajczak-Ropel, E., Wierzbowska, I.: E-JABAT An Implementation of the Web-Based A-Team. In: Nguyen, N.T., Jain, L.C. (eds.) Intelligent Agents in the Evolution of Web and Applications, pp. 57–86. Springer, Heidelberg (2009)
3. Blazewicz, J., Lenstra, J., Rinnooy, A.: Scheduling Subject to Resource Constraints: Classification and Complexity. Discrete Applied Mathematics 5, 11–24 (1983)
4. Cesta, A., Oddi, A., Smith, S.F.: A Constraint-Based Method for Project Scheduling with Time Windows. Journal of Heuristics 8, 108–136 (2002)
5. Ferreira, C.: Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence, 2nd edn. Springer, Heidelberg (2006)
6. Jedrzejowicz, P., Wierzbowska, I.: JADE-Based A-Team Environment. In: Alexandrov, V.N., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2006. LNCS, vol. 3993, pp. 719–726. Springer, Heidelberg (2006)
7. Jedrzejowicz, P., Ratajczak, E.: Agent-Based Approach to Solving the Resource Constrained Project Scheduling Problem. In: Beliczynski, B., Dzielinski, A., Iwanowski, M., Ribeiro, B. (eds.) ICANNGA 2007. LNCS, vol. 4431, pp. 480–487. Springer, Heidelberg (2007)
8. Jedrzejowicz, P., Ratajczak-Ropel, E.: Agent Based Gene Expression Programming for Solving the RCPSP/max Problem. In: Kolehmainen, M., et al. (eds.) Adaptive and Natural Computing Algorithms. LNCS, vol. 5495, pp. 203–212. Springer, Heidelberg (2009)
9. Jedrzejowicz, P., Ratajczak-Ropel, E.: Solving the RCPSP/max Problem by the Team of Agents. In: Håkansson, A., Nguyen, N.T., Hartung, R.L., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2009. LNCS, vol. 5559, pp. 734–743. Springer, Heidelberg (2009)
10. Neumann, K., Schwindt, C., Zimmermann, J.: Project Scheduling with Time Windows and Scarce Resources, 2nd edn. Springer, Heidelberg (2003)
11. Neumann, K., Schwindt, C., Zimmermann, J.: Resource-Constrained Project Scheduling with Time Windows, Recent Developments and New Applications. In: Perspectives in Modern Project Scheduling, pp. 375–407. Springer, Heidelberg (2006)
12. PSPLIB, http://129.187.106.231/psplib
13. Ratajczak-Ropel, E.: Experimental Evaluation of the A-Team Solving Instances of the RCPSP/max Problem. In: KES AMSTA 2010. LNCS (LNAI), vol. 6070, pp. 210–219. Springer, Heidelberg (2010)
14. Talukdar, S., Baerentzen, L., Gove, A. de Souza, P.: Asynchronous Teams: Cooperation Schemes for Autonomous, Computer-Based Agents. Technical Report EDRC 18-59-96. Carnegie Mellon University, Pittsburgh (1996)

Synchronous vs. Asynchronous Cooperative Approach to Solving the Vehicle Routing Problem

Dariusz Barbucha

Department of Information Systems
Gdynia Maritime University
Morska 83, 81-225 Gdynia, Poland
barbucha@am.gdynia.pl

Abstract. Cooperation as a problem-solving strategy is widely used to build methods addressing complex hard optimization problems. It involves a set of highly autonomous agents, each implementing a particular solution method. Agents cooperate during the process of solving the problem. The main goal of the paper is to evaluate to what extent a mode of cooperation (synchronous or asynchronous) between a number of optimization agents cooperating through sharing a central memory influences the quality of solutions while solving instances of the Vehicle Routing Problem.

Keywords: cooperative search, synchronous and asynchronous search, multi-agent systems, heuristics, vehicle routing problem.

1 Introduction

Among many approaches to solving different classes of real-world optimization problems the dominant position have hybrid algorithms. Typically, a [illegible] is a combination of exact or approximate algorithms used to solve the problem in hand. Often, such an approach outperforms methods based on a single optimization procedure.

The hybridization of different search algorithms, possibly running in parallel, gives the opportunity to diversify the problem-solving strategies. One of the most important and promising ones seems to be a cooperation between algorithms. According to Blum and Roli [2], [illegible] consists of a search performed by agents that exchange information about states, models, entire sub-problems, solutions or other search space characteristics.

There have been numerous cooperation mechanisms reported in the literature (see, for example [3], [8], for Vehicle Routing Problem). But although the cooperation paradigm may take different forms, they all share two features [5]:

- a set of highly [illegible], each implementing a particular solution method,

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 403–412, 2010.

- a [illegible] combining these autonomous programs into a single problem-solving strategy.

A set of autonomous programs may include known exact methods, like for example branch and bound, but in most cases various approximate algorithms (local search, evolutionary algorithms, tabu search, etc.) are engaged in finding the best solution.

A cooperation scheme, has to provide the mechanism for effective communication between autonomous programs allowing them to dynamically exchange the important pieces of information which next is used by each of them to much effective working in the process of searching.

As Crainic and Toulouse [5] suggest, the design of the information exchange mechanisms is a key factor determining a performance of the cooperative methods. Important cooperation design issues include its:

- [illegible] (what information to exchange),
- [illegible] (when to exchange it),
- [illegible] (the logical inter-processor structure),
- [illegible] (synchronous or asynchronous communications),
- [illegible] (what each autonomous program does with the received information),
- [illegible] (whether new information and knowledge is to be extracted from the exchanged data to guide the search).

The main goal of the paper is to evaluate to what extent a [illegible] ([illegible] or [illegible]) between a number of optimization agents (programs) influences the quality of the obtained results while solving instances of one of the well known combinatorial optimization problem - the Vehicle Routing Problem (VRP). To gain the required insight the computational experiment has been carried out using e-JABAT - the middleware cooperative search environment [1], based on the multi-agent paradigm.

The paper is organized as follows. Section 2 describes main features of the proposed cooperative multi-agent approach. In section 3 an implementation of the system for solving the vehicle routing problem is described. Section 4 shows the experiment plan and discusses its results. Finally, conclusions included in Section 5 end the paper.

2 A Multi-Agent Cooperative Search Environment

2.1 Overview

The natural approach to cooperative problem solving is to design and use [illegible] (CoMAS), where joint behavior of agents and interaction between them to cooperatively solve the tasks are specially emphasized [10].

The proposed cooperative search environment [1] belongs to the CoMAS class and its architecture is based on the concept of an [illegible] (A-Team), originally introduced by Talukdar [11]. A-Team is a collection of software agents which cooperate to solve a problem by dynamically evolving the population of solutions stored in the common memory. Each agent encapsulates a particular problem-solving method, which usually is inspired by some natural phenomena including, for example, evolutionary processes or particle swarm optimization, as well as local search techniques like, for example, tabu search.

The ground principle of asynchronous teams rests on combining algorithms, which alone could be inept for the task, into effective problem-solving organizations, possibly creating a synergetic effect, in which the combined effect of cooperation between agents is greater than the sum of their separate effects.

2.2 Generalized Procedure of Search for the Best Solution

Main functionality of the proposed environment is organizing and conducting the process of search for the best solution. It is organized as a sequence of steps, including initialization and improvement phases. At first the initial population of solutions is generated. Individuals forming the initial population are, at the following computation stages, improved by independently acting autonomous agents, thus increasing chances for reaching the global optimum. Finally, when the stopping criterion is met, the best solution from the population is taken as the result.

The pseudo-code of such a generalized process of search is shown as the Algorithm 1.

Algorithm 1. Generalized Search Procedure (GSP)

```
1: Generate an initial population of solutions (individuals) and store them in the
   common memory.
2: while (stopping criterion is not met) do {in parallel}
3:    Read individual from the common memory
4:    Execute improvement algorithm
5:    Store individual back in the common memory
6: end while
7: Take the best solution from the population as the result
```

The above process of search for the best solution is carried out by the two classes of agents: [illegible] (SMA) and a set of optimization agents - [illegible] ($OA = OA(1), OA(2), ..., OA(n)$).

[illegible] is an agent which manages the process of finding the best solution and is responsible for maintenance of individuals in the common memory. It generates an initial population of solutions and stores it in the common memory (step 1), reads a particular individual from common memory and sends it periodically to the [illegible], which has already announced its readiness (step 3), and update a common memory by storing in it a possibly improved solution obtained from the [illegible] (step 5).

Each [illegible] is an implementation of the improvement algorithm (local improvement heuristics, simulated annealing, tabu search, etc.) and its main task is to improve a solution obtained from [illegible] and to send it back to the [illegible] after the improvement (step 4).

2.3 Search Parameters

In order to implement the generalized search procedure, several features need to be defined. These include: method of [illegible], criteria for [illegible] including selection, acceptance, updating and stopping criterion.

Typically, [illegible] is formed randomly or using a dedicated constructive heuristic.

[illegible] criterion controls how to choose solutions from the common memory. Typically, a randomly chosen individual is forwarded to the optimizing agents for improvement.

[illegible] criterion describes whether new individual returned by the optimization agent after improvement phase is accepted or not to be added to the common memory. A common criterion of acceptance is satisfaction of the condition, that obtained solution is better than the current best solution produced so far by all optimization agents.

[illegible] criterion describes how to merge the improved solutions returned by the optimizing agents with the whole population. New individual may, for example, replace some randomly chosen solution or the worst one.

[illegible] criterion determines when to stop the process of search. It could be expressed for example, as a given number of iterations, time after which the process of searching will stop, or, for example, when calculations do not improve current best solution for some length of time, etc.

2.4 Synchronous and Asynchronous Search for the Best Solution

The proposed generalized cooperative search procedure can be executed in the [illegible] or [illegible] mode. Although in both modes, all optimization agents run autonomously and independently, the main difference between them is the way in which optimization agents are activated. In asynchronous mode, the communication between the solution manager and optimization agents takes place each time some optimization agent is reedy to act. The synchronous mode requires existence of a synchronization point in which, at the same time, all optimization agents are activated and try to improve the same solution.

Synchronous Search. In the synchronous mode, the whole process of search is divided into cycles. Each cycle starts when solution manager receives messages from all optimization agents about their readiness. Next, one individual (solution) chosen from the common memory according to a predefined selection criterion is sent to all optimization agents. They attempt to improve the received individual, and afterwards send their solutions back to the solution manager. Finally, after collecting all the returned individuals, the solution manager chooses the best one and updates the common memory, according with the updating criterion. If stopping criterion is not met the next cycle is performed.

The pseudo-code of the synchronous search procedure (SSP) including software agents with their roles is shown as the Algorithm 2.

Algorithm 2. Synchronous Search Procedure (SSP)

```
Activate the set of software agents:
- Solution Manager Agent ($SMA$) and
- Optimization Agents ($OA$), where $OA = OA(1), OA(2), ..., OA(n)$
Generate an initial population of solutions (individuals) and store them in the common memory.
while (stopping criterion is not met) do {in parallel}
    $SMA$ is waiting for messages from all $OA$ about their readiness
    After receiving messages from all $OA$, $SMA$ reads an individual from the common memory according to the predefined reading criterion and sends it to all $OA$.
    After receiving the message with an individual, each $OA$ is running own improvement procedure.
    After having completed the improvement procedure the resulting individual is returned by each $OA$ to $SMA$
    SMA holds the improved individuals in the temporal memory.
    After all individuals have been returned, $SMA$ selects the best one and updates the common in accordance with acceptance and updating criteria. Temporal memory is cleaned.
end while
```

Asynchronous Search. Opposite to the synchronous search, in the asynchronous mode, the cycle of selecting-improving-updating is defined not for the whole process and for all agents but for each optimization agent separately. Here, solution manager sends the solution selected from the common memory to an optimization agent immediately after receiving a message about its readiness to act. Of course, during the whole process, the stopping criterion is verified.

The pseudo-code of the asynchronous search procedure (ASP) including software agents with their roles is shown as the Algorithm 3.

Comparison. Both, the synchronous and the asynchronous mode of search, guarantee the exhaustive exploration of the search space, with intensification and diversification phases. In the asynchronous mode optimization agents act independently of each other and improve different individuals simultaneously. This is not the case in the synchronous mode where all optimization agents in each cycle improve the same individual. In this mode, it is also guaranteed that the knowledge about the current state of the search is identical for all optimization agents at the beginning of each cycle.

Cycle length in case of the synchronous mode is determined by the longest improvement procedure. In the parallel environment this may cause some of the

Algorithm 3. Asynchronous Search Procedure (ASP)

```
Activate the set of software agents:
- Solution Manager Agent ($SMA$) and
- Optimization Agents ($OA$), where $OA = OA(1), OA(2), ..., OA(n)$
Generate an initial population of solutions (individuals) and store them in the common memory.
while (stopping criterion is not met) do {in parallel}
    $SMA$ is waiting for a message from some $OA(j)$ about its readiness
    After receiving such message from $OA(j)$, $SMA$ reads an individual from the common memory according to predefined reading criterion and sends it to $OA(j)$.
    After receiving the message from $SMA$ including the individual, $OA(j)$ is starting the procedure of improving it using an improvement method build in it.
    After finishing the procedure of improvement, an individual (improved or not) is sent back by $OA(j)$ to $SMA$
    After receiving back a solution, $SMA$ adds it to the common memory using predefined acceptance and updating criteria.
end while
```

processors to stay idle. On the other hand in the asynchronous mode agents with improvement procedures running shorter tend to be evoked more often than the others. This can result in the increased demand for computational resources without achieving the expected improvement in the quality of solutions, especially when shorter procedures are not able to improve the received solutions.

3 Implementation of the Agent-Based Cooperative Search for Solving the Vehicle Routing Problem

The presented example has been implemented as a multi-agent system instances of the Vehicle Routing Problem. In VRP a set of customers is to be served by the fleet of vehicles in order to minimize the service cost and satisfying several customer's and vehicle's constraints: each route starts and ends at the depot, each customer is serviced exactly once by a single vehicle and the total load on any vehicle associated with a given route does not exceed vehicle capacity [7].

3.1 Agents

One [illegible] and three types of [illegible] have been implemented using the following heuristics and local search procedures:

- $OA(1)$ - an implementation of the [illegible] procedure.
- $OA(2)$ - an implementation of the λ[illegible] method ($\lambda = 2$).
- $OA(3)$ - an implementation of the [illegible] algorithm.

3.2 Search Parameters

The initial population of individuals. The initial population is constructed from randomly generated permutations which further are divided into routes satisfying the problem constraints (elements in each cluster do not exceed the capacity of the vehicle). The proposed method uses an idea originated from phase of the algorithm of Gillett and Miller [6].

Management of the population of individuals. The system offers two methods of a solution from the population: (R) - random solution is chosen from the population, and (B) - the best solution from the population is selected.

Also, two methods of the population after receiving a solution from optimization agent are proposed: (R) and (W), where a random or worst, respectively, solution from the current population is replaced by the received solution. All solutions obtained from optimization agents (locally or globally improved, or without any improvement) are and added to the common memory.

Stopping criterion. In the proposed implementation of cooperative search the system stops after a given number of seconds.

4 Computational Experiment

Computational experiment aimed at evaluating how the mode of cooperative search (synchronous or asynchronous) influences computation results, measured as mean relative error (MRE) from the best known results. Results obtained for all discussed combinations of the selecting and updating criteria are compared for both modes.

Cooperative search procedure implementation for VRP (for both modes) was tested on 5 test problems from ORLibrary benchmark set [12] containing 50-199 customers and only capacity restriction.

Each of the 5 instances of the problem was solved using each possible combination of selecting-accepting-updating strategies, in total giving 60 (5x12) test problems for each mode. Moreover, each test problem was repeatedly solved 10 times and mean results from these runs were recorded.

All computations have been carried out on the cluster Holk of the Tricity Academic Computer Network built of 256 Intel Itanium 2 Dual Core with 12 MB L3 cache processors with Mellanox InfiniBand interconnections with 10Gb/s bandwidth.

The experiment results are shown in Tab. 1, Fig. 1 and 2. Tab. 1 shows mean relative errors from the best known solution obtained by the proposed agent-based approach working in synchronous and asynchronous mode calculated for each strategy and for all considered instances. Fig. 1 and 2 present comparison of averaged results of searching for both modes. Results in Fig. 1 are averaged over all tested strategies, whereas results in Fig. 2 are averaged over all tested instances.

Table 1. The performance (measured as MRE) of the proposed agent-based approach working in synchronous and asynchronous modes calculated for each strategy and for selected instances from ORLibrary

STRATEGY	INSTANCE *vrpnc1* (50)	 *vrpnc2* (75)	 *vrpnc3* (100)	 *vrpnc4* (150)	 *vrpnc5* (199)
			Synchronous mode		
R-R	0.00%	1.81%	1.01%	3.20%	4.85%
R-W	0.00%	1.41%	0.94%	2.80%	4.35%
B-R	0.00%	3.36%	2.01%	3.83%	4.20%
B-W	0.00%	2.91%	1.88%	3.73%	4.05%
			Asynchronous mode		
R-R	0.00%	3.48%	3.25%	6.58%	7.94%
R-W	0.00%	3.30%	3.87%	6.86%	8.61%
B-R	0.00%	2.90%	1.79%	4.40%	8.62%
B-W	0.00%	2.57%	2.19%	5.26%	9.03%

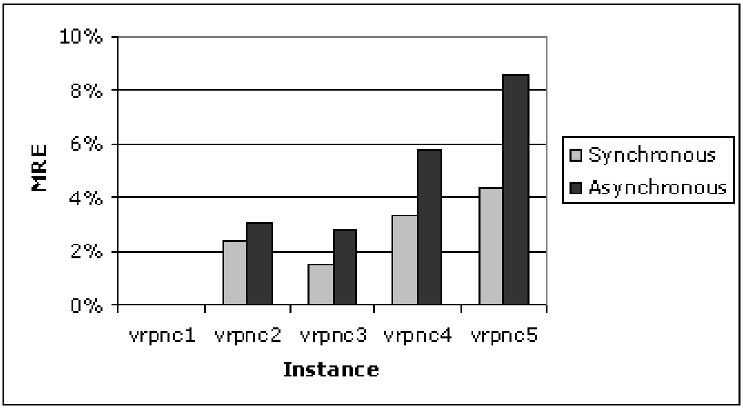


Fig. 1. Comparison of averaged results (measured as MRE) for all tested instances and for synchronous and asynchronous modes of searching

By observing results presented in Tab. 1 the influence of the mode of searching on quality of solutions is seen for almost all combinations of strategy and instance of the problem. For two strategies (R-R and R-W) synchronous organization of searching outperforms (or gives the same results) asynchronous one for all five considered instances of the VRP. Other strategies guarantee that system working in asynchronous mode produces better results only for one instance ([illegible]) in case of B-W strategy or for two instances ([illegible]) in case of B-R strategy. For instance [illegible] results obtained for all strategies are the same in both modes and are equal to the optimal solution.

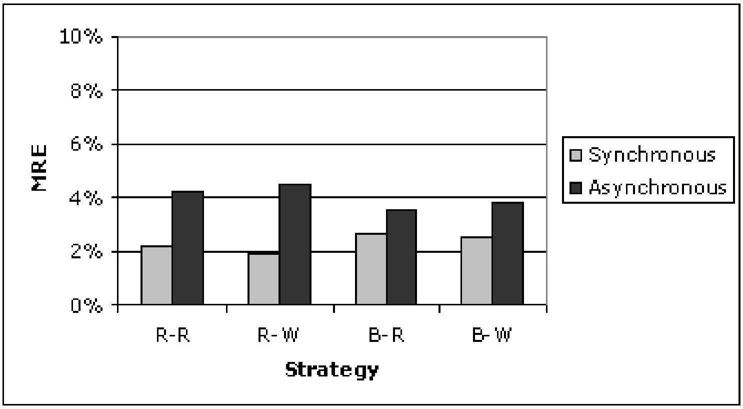


Fig. 2. Comparison of averaged results (measured as MRE) for all tested strategies and for synchronous and asynchronous modes of searching

Taking into account results averaged over all tested strategies (Fig. 1) and over all tested instances (Fig. 2), one can observe the dominance of synchronous search over asynchronous one for all cases. The differences between mean relative errors for both modes vary from one to four percent, depending on strategy and/or instance of the problem.

And finally, considering the proposed approach based on multiple agents as a tool for solving instances of the vehicle routing problem, one can conclude that its effectiveness also depends on instance of the problem. In fact, in most cases, results obtained by the system are worse than the best solutions produced by leading methods, like tabu search algorithms [7]. But on the other hand, according to the goal of the paper, a set of non-complex, representative heuristics (operating on single or many routes) for solving VRP had been chosen as optimization agents.

5 Conclusions

Cooperative search procedure implementation for VRP has been presented in the paper. During the process of solving the instances of the problem, agents exchange messages including an information about solutions obtained till now by them. In suggested approach messages between agents are exchanged indirectly and using the central memory mechanism. One of the main goal of the paper was to evaluate to what extent a mode of cooperation (synchronous or asynchronous) between a number of optimization agents influences the quality of solutions while solving instances of the VRP. The results of computational experiment have shown that mode of cooperative search for the best solution may influence the obtained results. In this context, the positive impact of synchronous search on the results has been observed.

An interesting direction of further research is to investigate the possibility of combining the approach suggested in the paper with [illegible] idea introduced by Burke et. al. [4]. The additional motivation, strongly related to this paper, is the fact that last time, Ouelhadj and Petrovic [9] also considered synchronous and asynchronous cooperative search, but in hyper-heuristic framework. In author's opinion, comparison of these approaches would bring interesting observations.

Acknowledgments. The research has been supported by the Ministry of Science and Higher Education grant no. N N519 576438. Calculations have been performed in the Academic Computer Centre TASK in Gdansk.

References

1. Barbucha, D., Czarnowski, I., Jędrzejowicz, P., Ratajczak-Ropel, E., Wierzbowska, I.: e-JABAT - An Implementation of the Web-Based A-Team. In: Intelligent Agents in the Evolution of Web and Applications, pp. 57–86 (2009)
2. Blum, C., Roli, A.: Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison. ACM Computing Surveys 35(3), 268–308 (2003)
3. Le Bouthillier, A., Crainic, T.G., Kropf, P.: A Guided Cooperative Search for the Vehicle Routing Problem with Time Windows. IEEE Intelligent Systems 20(4), 36–42 (2005)
4. Burke, E.K., Hart, E., Kendall, G., Newall, J., Ross, P., Schulenburg, S.: Hyper-heuristics: an emerging direction in modern search technology. In: Glover, F., Kochenberger, G. (eds.) Handbook of Metaheuristics, pp. 457–474. Kluwer Academic, Dordrecht (2003)
5. Crainic, T.G., Toulouse, M.: Explicit and Emergent Cooperation Schemes for Search Algorithms. In: Maniezzo, V., Battiti, R., Watson, J.-P. (eds.) LION 2007 II. LNCS, vol. 5313, pp. 95–109. Springer, Heidelberg (2008)
6. Gillett, B.E., Miller, L.R.: A heuristic algorithm for the vehicle dispatch problem. Operations Research 22, 240–349 (1974)
7. Laporte, G., Gendreau, M., Potvin, J., Semet, F.: Classical and modern heuristics for the vehicle routing problem. International Transactions in Operational Research 7, 285–300 (2000)
8. Meignan, D., Creput, J.C., Koukam, A.: A coalition-based metaheuristic for the vehicle routing problem. In: Proc. of the IEEE Congress of Evolutionary Computation (CEC 2008), Hong-Kong, pp. 1176–1182. IEEE Press, Los Alamitos (2008)
9. Ouelhadj, D., Petrovic, S.: A cooperative hyper-heuristic search framework. Journal of Heuristics (to appear, 2010)
10. Panait, L., Luke, S.: Cooperative Multi-Agent Learning: The State of the Art. Autonomous Agents and Multi-Agent Systems 11(3), 387–434 (2005)
11. Talukdar, S., Baeretzen, L., Gove, A., de Souza, P.: Asynchronous teams: Cooperation schemes for autonomous agents. Journal of Heuristics 4, 295–321 (1998)
12. OR-Library, http://people.brunel.ac.uk/~mastjjb/jeb/orlib/vrpinfo.html

Development of an Engineering Knowledge Extraction Framework

Chengter Ho

National Kaohsiung University of Applied Sciences,
80778 Kaohsiung, Taiwan
hoc@cc.kuas.edu.tw

Abstract. In this paper, an engineering knowledge extraction framework was developed based on the ontology technologies. The engineering knowledge could be extracted from unstructured engineering documents. The ontology models of the target engineering knowledge are constructed by domain experts or automatic constructed with formal concept analysis in the first stage. The engineering document is then processed to mark the relevant semantic annotations with natural language process tool and domain dictionary. These annotations will be used to populate the ontology models with the concepts and properties in the ontology models. The populated ontology models can be then used for reasoning or other tasks requiring engineering knowledge.

Keywords: Engineering Knowledge, Knowledge Extraction, Natural Language Processing, Ontology.

1 Introduction

Engineering knowledge is essential for product and process design [1]. How to preserve the engineering knowledge is the key element in knowledge management of a company with product design capability. Vast amount of engineering documents are the sheer sources of engineering knowledge. However, the knowledge stored in written form could only be processed and managed by people, because knowledge is represented in Natural and not Structured Language which is not understandable by the computer. Therefore, the possibility of utilizing this knowledge depository heavily relies on the computerization of these documents by converting them into structured knowledge. Ontology engineering is a well suited methodology to preserve engineering knowledge [2]. Sun etc. [2] has proposed an ontology-based knowledge management system to assist engineers in the design processes. The purpose of this paper is to develop an engineering knowledge extraction framework (EKEF) in order to populate this kind of ontology-based engineering knowledge base. Building ontology-based engineering knowledge involves two major phases. The ontology model of the engineering knowledge must be constructed first. These ontology models can then be populated in the second phase. In most of the systems utilizing ontology, ontology models are constructed by domain experts [1, 2]. Obitko etc. [3] introduced a methodology for constructing ontology with the aid of Formal Concept Analysis (FCA).

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 413–420, 2010.

Formal Concept Analysis (FCA) is a method of data analysis which identifies conceptual structures among data sets [3]. Therefore, it can provide help to the process of designing Ontology because Ontology design often starts with designing taxonomy. In manufacturing fields, Ontology is often used to design for representing a general product and process which includes all similar features from group of products and processes respectively. Thus, in order to facilitate the process of Ontology design in domain of engineering, we proposed using Group Technology (GT) along with the support from FCA.

After the ontology models been built, the documents can be processed with natural language processing techniques and populated to ontology models. In University of Southampton, ArtEquAKT project aims at extracting information about artists from webpage by employing Knowledge acquisition, analysis techniques and Ontology engineering [4]. Data and information involving domain of interest extracted from the webpage is used to populate Ontology. The ontology after being populated is put into Knowledge Base. ArtEquAKT makes use of that populated Ontology by accessing Knowledge Base to automatically generate artist personalized biographies. Artequakt's architecture consists of three key areas. First, knowledge extraction tools collect information items which are sentences and paragraphs from Web documents. Those information items are manually selected or obtained automatically by applying appropriate search engine technology. The tools then transfer the information fragments to the ontology server along with metadata derived from the ontology's vocabulary (Populating Ontology). Second, after populating Ontology, information is stored and consolidated in Ontology sever so that the biography generation tool can query the KB using an inference engine. Third, the Artequakt server allows user to get artist's narratives through a Web interface. Depending on the user, a particular style of biography can be selected such as a chronology or summary, or a specific focus such as the artist's style or body of work. The narrative from the KB is rendered by the server using story templates. Although EKEF system inherited some ideas from ArtEquAKT project, it however focuses on extracting knowledge within an engineering domain. Thus, it requires a dictionary of domain knowledge to be constructed and support Information Extraction (IE) process. Basically, IE technologies depend on predefined template and patterns base extraction rules to identify vocabulary that could be found on textual document. Current approaches to IE employ Natural Language Processing techniques however focus only on very restricted domains. Therefore, a dictionary of engineering domain knowledge is necessary to exploit current IE techniques. In ArtEquAKT project, XML was used to represent Ontology. The primary goal of using XML is to facilitate the sharing of data across different information systems via Internet. However, each system has it own predefined set of tags. The problem arises when trying to share data between different systems. Thus, using XML to represent Ontology limits the ability of sharing data, information and knowledge across different information systems. In OBKE system, OWL was selected to represent Ontology because OWL has been widely accepted as a standard language for Ontology representation. Therefore, the above problems of sharing data, information and knowledge across different information systems obviously were eliminated.

2 Methodology

The knowledge extraction process involves activities of retrieving explicit or tacit knowledge that resides within people, artifacts, or organizational entities [5]. In this paper, the engineering knowledge will be resided in ontology models with populated data. The proposed extraction methodology is shown in Figure 1. The engineering ontology models are constructed by experts or by engineers with software assistance. Then, the unstructured engineering document can be processed to have semantic annotations. After annotations being generated, the document with annotations can then be analyzed by matching these annotations with ontology classes resided in ontology models constructed previously. These matched annotations will be populated into the ontology-based knowledge base.

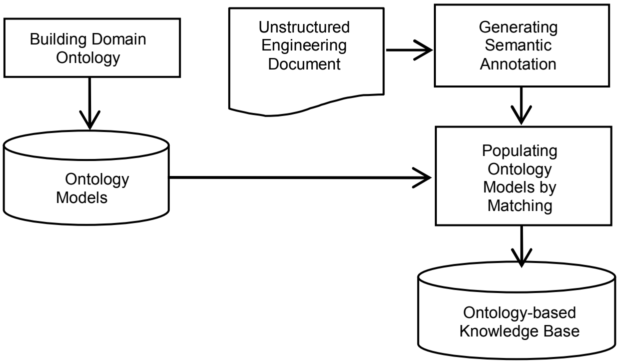

Building Domain Ontology
Unstructured Engineering Document
Generating Semantic Annotation
Ontology Models
Populating Ontology Models by Matching
Ontology-based Knowledge Base

Fig. 1. Basic Engineering Knowledge Extraction Framework

2.1 Building Domain Ontology Models

As discussed before, ontology is a high-level abstract and a simplified view of the world. It also allows both human and machine to understand the explicit knowledge that was stored on it. Hence, tacit knowledge about domain resided in an expert can be elicited through the process of ontology design. However, it may take too much effort of experts to learn how to construct ontology model. Formal Concept Analysis (FCA) was used to assist the construction of ontology model of interested engineering area. FCA is a method of data analysis that takes a matrix including a set of objects and the properties as an input, and finds all the "natural" clusters of properties and all the "natural" clusters of objects in the input data [6]. Tools developed in FCA, such as ConExp [7], can assist the user to draw a ‘draft’ ontology model from a large number of objects (classes) and properties presented in documents. A ‘draft’ ontology is created by the “natural" clusters of properties and the "natural" clusters of objects. It will be reviewed by experts to produce a complete ontology model for a family of products.

Finally, ontology alignment is needed to check for duplication or one ontology class is a subclass of another class. This task can be done with the aid of DL reasoner, for example RACER DL reasoner [8]. After the ontology of the interested domain knowledge being constructed, the ontology models can be used as the key template of knowledge extraction from unstructured documents.

2.2 Generating Semantic Annotations

Most of the information extraction (IE) tools only support processes for detecting terms, phrases and terminologies related to pre-specific terms such as person, organization and place names. Therefore, to apply IE technology on any specific domain knowledge, a dictionary of the domain knowledge must be established manually as the first component. Dictionary of Domain knowledge will be established through a process of constructing Gazetteer lists in GATE [9]. When GATE runs Gazetteer on the textual document, it will create a type of Lookup annotation for each matching string. This Lookup annotations and Token annotations are used for the purpose of generating Semantic annotations.

First, syntactic analysis of every sentence in the input document is carried out by a parser tool to provide syntactical annotation for a word found on that sentence. Basically, syntactic analysis deals with analyzing the arrangement of words in sentences, clauses, phrases and the relationships between words to form that sentence, clauses and phrases. In English, a typical sentence is formed by a standard Subject-Verb-Object word order. Any attempt to change the word order of such a sentence would change the meaning or even make that sentence meaningless. We only count sentence's syntax analysis because only a complete sentence can provide at least a Subject-Verb-Object word order which is equivalent to Subject-Relationship-Object statement from Ontology. Sentence's syntax analysis and syntactical annotations also plays an important role on later matching operation which can be found on Matching Ontology with Semantic Annotation part. Secondly, when facing several words referencing to one meaning or nearly the same meaning, the system needs an ability to understand that situation. WordNet, an online lexical reference system, would be used to solve the problems with the variation of vocabulary.

2.3 Populating Ontology

With the document with semantic annotations from previous processes and ontology models constructed by experts, matching ontology classes with semantic annotations is the first step in populating ontology into knowledge base. This process of matching semantic annotations with ontology entities is shown in Figure 2. The ontology objects/data-type properties from ontology models are accessed one by one. Then, the system will get corresponding domain and range of these objects/data-types. The ontology annotations of these objects/data-types are used to match the semantic annotations. If the content of ontology annotation property matches semantic annotation, information about ontology entities will be added to the semantic annotations. Information in ontology model will be added to semantic annotations as an annotation feature. In case the content of ontology annotation property does not match the semantic annotation, the process will be repeated from the first step with another ontology object/data-type property.

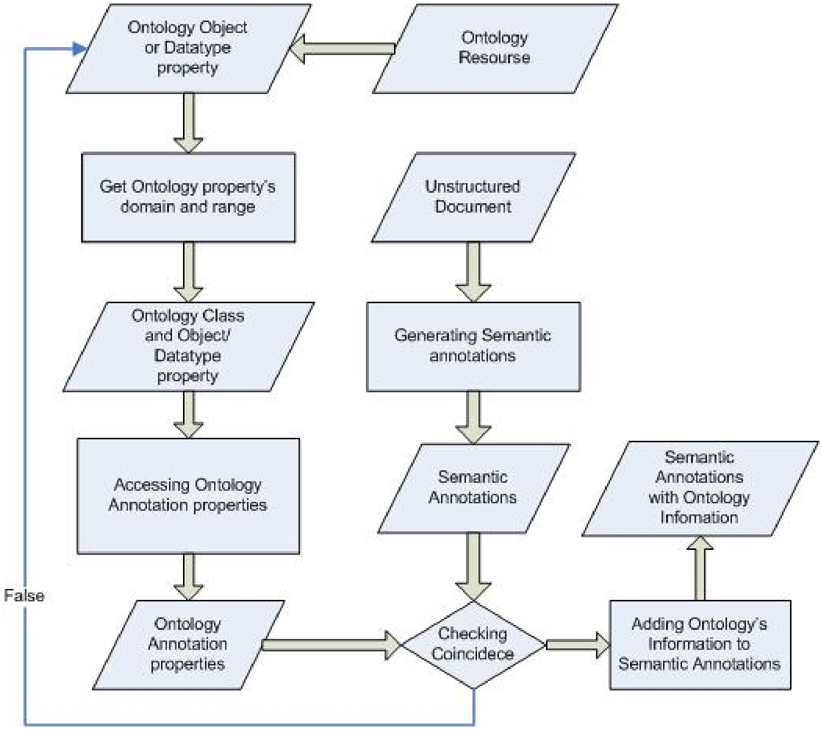

Ontology Object or Datatype property
Ontology Resourse
Get Ontology property's domain and range
Unstructured Document
Ontology Class and Object/ Datatype property
Generating Semantic annotations
Accessing Ontology Annotation properties
Semantic Annotations
Semantic Annotations with Ontology Infomation
False
Ontology Annotation properties
Checking Coincidece
Adding Ontology's Information to Semantic Annotations

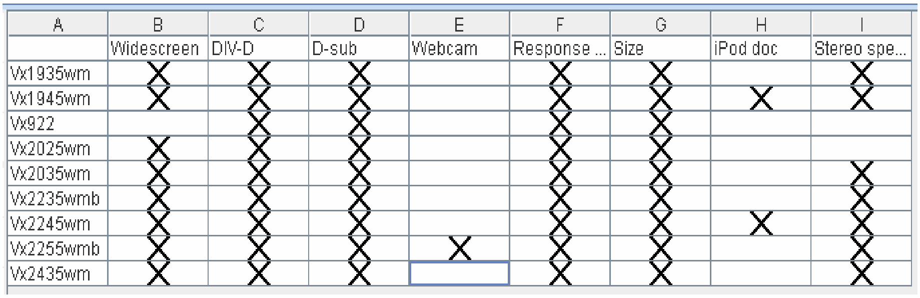
Fig. 2. Algorithm of matching semantic annotations with ontology entities

After the connection between Ontology classes and semantic annotation have been established, data and information inside a textual document can now be put into Ontology as the instance of Ontology class. The role of Ontology in OBKE system is that it provides a way to allow data residing in terms and phrases on each sentences of an input textual document to be extracted when the basic Subject-Verb-Object word order of that natural language sentence is equivalent to a Subject-Relationship-Object essential statement from Ontology. Therefore, by doing so, the correctness of extracted data have been guaranteed. However, mistakes somehow still occur during the process of extracting data due to the complexities of natural languages. To deal with those mistakes, some checking SWRL rules are needed to ensure that the extracted data which is fed to Ontology satisfy the predefined constrained conditions.

3 Implementation

The process of knowledge extraction from document about TFT-LCD monitors is used as example to demonstrate the methodology. A group of TFT-LCD monitors with different the features, as shown in Table 1, was used as input for the FCA to construct the 'draft' ontology, as shown in Figure 3. The consistency of the ontology should be checked by DL reasoners. Figure 4 shows the ontology of TFT-LCD monitor after processed by DL reasoner (RACER).

Table 1. TFT-LCD monitors and their features



A	B	C	D	E	F	G	H	I
	Widescreen	DIV-D	D-sub	Webcam	Response ...	Size	iPod doc	Stereo spe...
Vx1935wm	X	X	X		X	X		X
Vx1945wm	X	X	X		X	X	X	X
Vx922		X	X		X	X		
Vx2025wm	X	X	X		X	X		
Vx2035wm	X	X	X		X	X		X
Vx2235wmb	X	X	X		X	X		X
Vx2245wm	X	X	X		X	X	X	X
Vx2255wmb	X	X	X	X	X	X		X
Vx2435wm	X	X	X		X	X		X

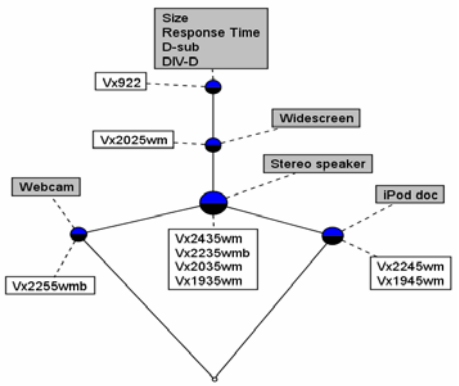


Fig. 3. 'Draft' ontology derived from ConExp

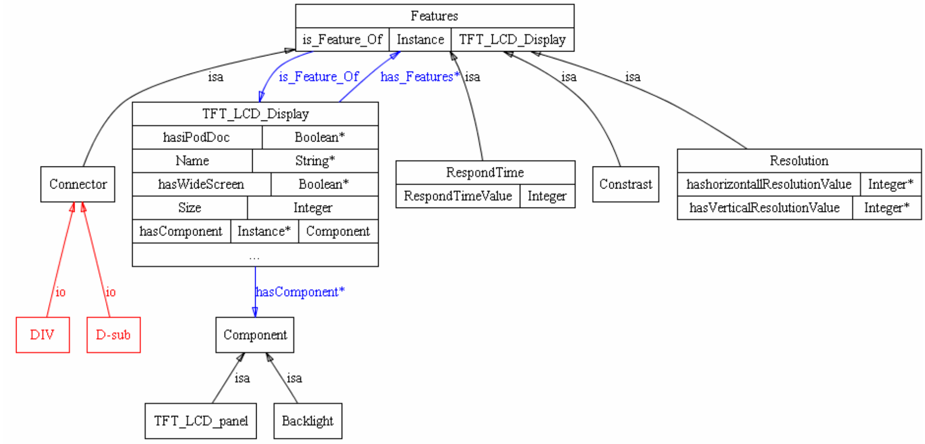


Fig. 4. Ontology for a group of LCD monitors

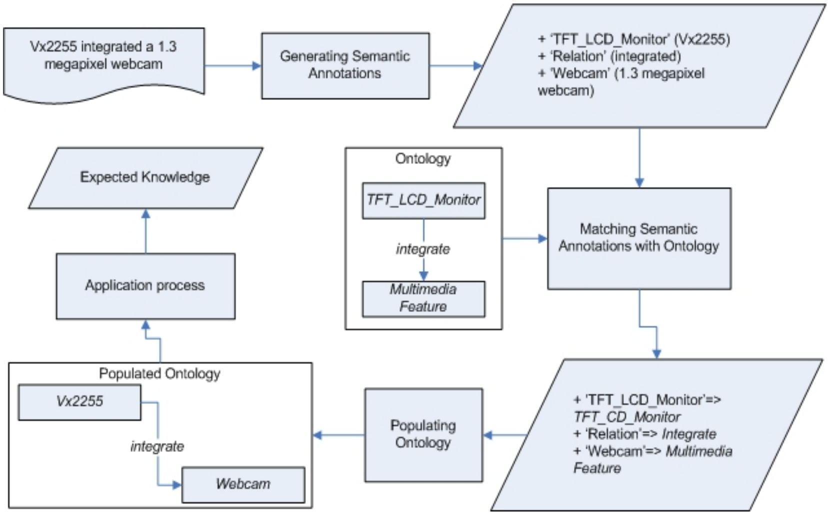
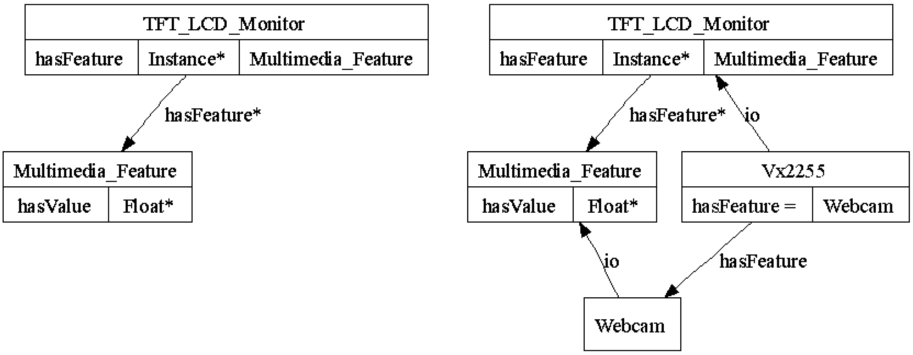

Vx2255 integrated a 1.3 megapixel webcam
Generating Semantic Annotations
+ 'TFT_LCD_Monitor' (Vx2255)
+ 'Relation' (integrated)
+ 'Webcam' (1.3 megapixel webcam)
Expected Knowledge
Ontology
TFT_LCD_Monitor
integrate
Multimedia Feature
Matching Semantic Annotations with Ontology
Application process
Populated Ontology
Vx2255
integrate
Webcam
Populating Ontology
+ 'TFT_LCD_Monitor'=> TFT_CD_Monitor
+ 'Relation'=> Integrate
+ 'Webcam'=> Multimedia Feature

Fig. 5. An example of extracting knowledge from an unstructured document



(a) ontology model before populating

(b) ontology model after populating

Fig. 6. Ontology models before and after populating

As an example, a sentence "Vx2255 integrated a 1.3 megapixel webcam" is used as an input for the system. The extracting process is shown in Figure 5. First, semantic annotations are created based on the vocabulary involved domain knowledge from the input sentence. Here, "Vx2255", "integrated" and "1.3 megapixel webcam" is the name of a TFT LCD Monitor, the potential relation between objects in the sentence and a Webcam respectively. Therefore, semantic annotations ('TFT_LCD_Monitor', 'Relation' and 'Webcam') are generated in order to annotate a words ("Vx2255", "integrated" and "1.3 integrated megapixel webcam"). At the second process, beside semantic annotations, Ontology is another input for system. After being processed, a semantic annotation 'TFT_LCD-Monitor', 'Relation' and 'Webcam' will be matched to "TFT_LCD_Monitor" Ontology class, "integrated" Ontology Object property and

“Multimedia_Feature” Ontology class respectively. Thirdly, the output of OBKE tool is the populated Ontology with extracted data, information found on the input sentence. In this example, “Vx2255” and “webcam” is an object of a “TFT_LCD_Monitor” and “Multimedia_Feature” Ontology class respectively. Finally, the output (i.e. populated Ontology) can be used by other applications such as Knowledge Discovery and Knowledge Sharing to obtained expected knowledge. The ontology models before and after the populating are shown in Figure6.

4 Concluding Remarks

In this paper, a framework for an Automatic Ontology-Based Engineering Knowledge Extraction system has been proposed. The engineering knowledge can be extracted from unstructured engineering documents after the ontology models of interested engineering domain being constructed. The extracted knowledge from the unstructured engineering document stored in Ontology can be accessed and used by other systems or applications such as Knowledge Sharing and Knowledge Discovery. With this framework, a management system to be incorporated to become an Ontology-Based Knowledge Management system or easily integrated to current Ontology-Based Knowledge Management systems.

References

1. McMahon, C., Lowe, A., Culley, S.: Knowledge Management in Engineering Design: Personalization and Codification. Journal of Engineering Design 15(4), 307–325 (2004)
2. Sun, W., Ma, Q.-Y., Gao, T.-Y.: An ontology-based manufacturing design system. Information Technology Journal 8(5), 643–656 (2009)
3. May, P., Ehrlich, H.C., Steinke, T.: ZIB Structure Prediction Pipeline: Composing a Complex Biological Workflow through Web Services. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds.) Euro-Par 2006. LNCS, vol. 4128, pp. 1148–1158. Springer, Heidelberg (2006)
4. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
5. Czajkowski, K., Fitzgerald, S., Foster, I., Kesselman, C.: Grid Information Services for Distributed Resource Sharing. In: 10th IEEE International Symposium on High Performance Distributed Computing, pp. 181–184. IEEE Press, New York (2001)
6. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The Physiology of the Grid: an Open Grid Services Architecture for Distributed Systems Integration. Technical report, Global Grid Forum (2002)
7. Priss, U.: Formal concept analysis in information science. In: Annual Review of Information Science and Technology, vol. 40, pp. 521–543 (2006)
8. Tilley, T.: Tool Support for FCA. In: Eklund, P. (ed.) ICFCA 2004. LNCS (LNAI), vol. 2961, pp. 104–111. Springer, Heidelberg (2004)
9. Haarslev, V., Müller, R.: RACER System Description. In: Goré, R.P., Leitsch, A., Nipkow, T. (eds.) IJCAR 2001. LNCS (LNAI), vol. 2083, pp. 701–705. Springer, Heidelberg (2001)
10. Cunningham, H., Maynard, D., Tablan, V., Ursu, C., Bontcheva, K.: Developing language processing components with GATE (2002), http://www.gate.ac.uk

Exploring EFL Taiwanese University Students' Perceptions of a Collaborative CALL Environment

Yen-Hui Wang[1] and Chia-Nan Wang[2]

[1] Kainan University, Taiwan
[2] National Kaohsiung University of Applied Sciences
ttxyhw@mail.knu.edu.tw

Abstract. This study aimed to investigate the perceptions of EFL Taiwanese university students on a collaborative CALL environment. The participants were 112 intermediate proficiency English as a foreign language Taiwanese third-year university students. The dataset used included the pre-and post-questionnaires on participants' perspectives on the CALL environment. In addition, interview data was collected for more in-depth information on individual participants' perceived views on such a collaborative e-learning program. The results of the study provided encouraging evidence to show that the participants generally perceived that they benefited from the whole process of a collaborative computer assisted language learning program to have positive perspectives on the implemented CALL course itself, to advance their English linguistic knowledge, to construct associated content knowledge, and to foster their affective attitudes towards learning language via a collaborative CALL environment. The findings suggested the need for Taiwanese language teachers, course writers, curriculum planners, policy makers, and educational authorities to consider integrating CALL components into English language lessons to help Taiwanese students improve language competence, gain content knowledge, develop cooperative learning, and strengthen motivation for EFL learning.

Keywords: EFL (English as a Foreign Language), CALL (Computer Assisted Language Learning), perceptions of a collaborative CALL environment.

1 Introduction

With the utilization of computer technology in order to meet the demands of the younger generation of the twenty-first century, education has nowadays changed greatly from purely traditional face-to-face classroom teaching to computer-mediated mode of instruction or blended approaches that combine computer-based applications with face-to-face teaching. In the fields of second language acquisition (SLA) and pedagogy, the adoption of multimedia CALL (Computer Assisted Language Learning) approaches to language teaching has also grown considerably over the past decade (Bonk & Graham, 2006; Zapata & Sagarra, 2007). Technology-based, learner-centered learning is very different from traditional classroom-based, teacher-led learning. CALL instruction offers many advantages that traditional, didactic teaching methods may

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 421–432, 2010.

lack. That is, an e-learning environment enables instant access to resources and effective communication, and allows learners to learn at their own pace as well as to be independent of time and place. In addition, it provides learners not only with repeated production opportunities or unlimited time to complete online assignments/tasks but also with immediate and informative feedback together with interesting reinforcement to promote learners' language acquisition and learning interest (Dunkel, 1987; Murray, 1999; Rivers, 1987). Moreover, it also alleviates teaching burden of instructors to grade a large number of exercises or tests, which enables more time to be spent on lesson planning and effective teaching (Arvan & Musumeci, 2000). Learners' e-learning experience shapes their perceptions of CALL application in language learning, and in turn influences the results of their language acquisition and the success or failure of such e-course instruction. Prior research (e.g., Ginns & Ellis, 2007; Neumeier, 2005) that explores the relations between learners' learning outcomes and their attitudes towards learning with the assistance of computer technology indicates that students' perceptions of the learning environment are highly influential on their learning. Numerous earlier studies dealing specifically with learners' e-learning perceptions with compelling evidence have contributed to developing online teaching/learning materials and constructing CALL environments, but little research has been conducted in a Taiwanese context to investigate the perspectives of Taiwanese EFL learners on the e-learning format. Studies incorporating Taiwanese learners' views on online language teaching and learning in a *social* context are, however, even fewer. As opposed to students' conventional instructor-led mass language learning experiences, collaborative e-learning for language is quite novel to most Taiwanese students. It is therefore important to examine their perceptions of CALL use in an EFL learning context for future CALL applications. Thus, the main research focus of this study is, from Taiwanese university students' perspectives, the effectiveness of a collaborative CALL environment in EFL learning.

2 Literature Review

Hannafin and Land (1997) define a computer-based learning environment as the one involving four main elements: tools, resources, people and design. To be more specific, a computer-based learning environment should be the one where computer technology is used as a tool to provide access to learning resources and as a means to present and distribute information. In addition, computer technology also enables learners or users to read and post messages/comments on any topics for discussions through a synchronous and/or asynchronous online communication tool. Most significantly, a computer-assisted learning environment should be carefully constructed so that all the components are well integrated to effectively support teaching and learning. Learners' perceptions of the exposed learning environment which interrelate with their responsiveness to the particular learning context play a critical role in the learning processes and learning outcomes (den Brok, Brekelmans & Wubbels, 2006; Kern, 1995; Meyer & Muller, 1990; Ramsden, 1991; So & Brush, 2008). Computer assisted language teaching and learning has been applied in many educational settings and has been the subject of numerous L2/FL studies for decades. Up to now, there have been many studies that show the educational value and positive effects of CALL applications on the cultivation of positive language learning attitudes (e.g., Kitao, 1995; Lin, 2009) as well as on the improvement of learners' language competence in terms of vocabulary building (e.g., Pawling, 1999; Sun & Dong, 2004), reading and writing skills (e.g., Lin, 2002), grammar accuracy (e.g., Peterson, 1997), and listening comprehension (e.g., Coniam, 2006). For instance, Ayres (2002) reports in his study that the majority of the subjects perceived the computer-based instruction and online activities promoted their language acquisition. Online environments can offer instant feedback and multiple opportunities for extensive and repeated practice, which facilitates learners to internalize what has been taught for effective learning. As far as affective advantages is concerned, previous research indicates that participating in CALL course lowers L2/FL learners' anxiety about target language learning as well as technology use, and fosters their positive attitudes towards the use of CALL in L2/FL learning and eventually leads to successful language acquisition (Liou, 1997; Ushida, 2005). It is because an e-learning environment allows learners to learn at their own pace without competing with others and feeling embarrassed while making errors, learners will then gradually build self-confidence in both language ability and computer skills. Moreover, as existing studies have shown (e.g., Collentine, 2000; Lee, 2005; Murray, 1999; Stepp-Greany, 2002), language courses with an online component encourage learner autonomy for independent learning, and learners who are exposed to such an environment, rather than acting as passive learners, are more actively engaged in the construction and use of their knowledge and take more responsibilities for their own learning. This is because a computer-based language learning context allows greater learner control over learning process, which, in turn, enhances learners' learning motivation and sense of achievement. In sum, all the above empirical studies show computer assisted language learning application with positive results in support of language courses that incorporate technology.

3 Method

Research setting, participants, and procedures

The research study reported in this paper aimed to explore Taiwanese university students' viewpoints on a particular language teaching and learning context where a technology-based, computer-mediated learning environment was implemented. The course participants did not have prior exposure to learning English as a foreign language within a collaborative e-learning environment, their views on such a collaborative CALL environment were thus worth an examination. The e-learning language course discussed in the present study was offered once a week for three 50-min sessions during the 2010 spring semester, and was implemented through the Digital E-learning Platform constructed by a university located in northern Taiwan. The course entitled Introduction to Language and Culture was an elective course for third-year university students and demanded two years of required English. This course intended to equip course participants with a fundamental understanding of the interrelations of language and culture along with linguistic knowledge associated with language and culture. The participants agreeing to take part in the research were 112 junior students from the School of Humanities and Social Sciences who enrolled in this course, and their English proficiency varied from pre-intermediate to upper intermediate. Participants were informed the purpose of the study, the research procedure, and the confidentiality of the collected data. At the beginning of the course, the participants were given an introduction to the University Digital E-learning Platform and the instructional components of this e-learning course including face-to-face online lectures, online interactive exercises, and collaborative group tasks. They were also offered a technical training session about this e-learning language course. Because this course was taught all through online, a paperless approach was adopted, and all teaching and learning materials distributed and collected were all in electronic form. Each of the four course themes in this e-learning course comprising (1) *verbal communication: the way people speak*, (2) *nonverbal communication: speaking without words*, (3) *cross-cultural conflict and adjustment*, and (4) *education: values and expectations* required two online lectures to be finished, and every finalized theme was followed by self-paced online interactive exercises specific to the course theme presented during the online lectures. With immediate feedback provided, students individually completed a set of web-based online interactive exercises created from the Hot Potatoes suite, the new online exercise application produced by Half-Baked Software, in order to equip themselves with the linguistic knowledge and content knowledge relative to the course theme. Upon the completion of these four course themes, course participants assigned into one of the 8 groups consisting of 14 members were required to log in the online chat room (similar to IRC – Internet Relay Chat using for online synchronous communication) to have online group discussions about the specific task relating to English language education of one of the eight Asian countries. Group members were asked to work together to firstly search useful information from the World Wide Web pertinent to the task topic designated. Subsequently, through the online topic-oriented discussions, students processed the information found online by expressing thoughts, sharing personal experiences and exchanging ideas to gradually construct associated knowledge and develop overall understanding of the key issues. Afterwards, on the basis of the information found from the World Wide Web and the discussions made at the online chat room, every group jointly produced one word-format essay in response to the required task topic, and submitted it as an attachment to an e-mail to the instructor. At last, one designated student from every group made an in-class presentation displayed with PowerPoint format with the aid of a LCD projector. It was such the whole learning process that participants were engaged in a learning environment which started with personal involvement (i.e., online lectures and online interactive exercises) and moved towards meaningful and collaborative social interactions with the assistance of technology (i.e., collaborative group tasks).

Instruments

To probe 112 EFL Taiwanese university students' responsiveness to the implemented collaborative CALL environment, the instruments employed in this study were student perception pre-and post-questionnaires along with a structured interview. In this study, both quantitative and qualitative methods of data gathering and data analysis were used.

Pre-and post-questionnaires

Prior to the e-learning language course, participants were required to fill in the self-completed pre-questionnaire (4 items with a four-point Likert scale) focusing on student reported perceptions of computer assisted language learning, their preferred learning environment as well as their expectations about this online course. At the end of the semester, the post-questionnaire with similar questions to the pre-questionnaire was completed by the participants. The reliability of the pre-and post-questionnaires were (Cronbach's Alpha) 0.81 and 0.867 respectively, and the professional content validity for these questionnaires were established by one TESOL professor who reviewed the questions and the question types to indicate that they were appropriate.

Structured interview

To investigate participants' attitudes towards the technology-driven language course in more depth, a structured interview was conducted after the collaborative CALL program. The face-to-face interview was based on eight guiding questions to ask participants' perspectives on the entire CALL instruction such as:

- whether the e-learning language course helped their learning in a more efficient manner,
- in what ways the e-learning language course helped their learning,
- the advantages of the e-learning language course,
- the disadvantages of the e-learning language course,
- -the part of this e-learning language course (i.e., online lectures, online interactive exercises, collaborative group tasks) they liked the most or disliked the most,
- whether they enjoyed working with group members online,
- what kind of help they received from their group mates,
- whether they would suggest the implementation of a collaborative CALL environment for the future courses.

In terms of the reliability and validity of the interview schedule, the reliabilities of the interview question 3 (5 items with a four-point Likert scale) and interview question 4 (5 items with a four-point Likert scale) were (Cronbach's Alpha) 0.879 and 0.842 individually, and the content validity for this interview schedule was achieved, again, on the basis of the judgments from the TESOL professor reviewing the questions and the question types to verify their appropriateness.

Data analysis methods

Paired-Samples T-Tests were applied to analyze the differences before and after one-semester e-learning language course shown on the pre-and post-questionnaires about participants' self-perceptions of computer assisted language learning as well as their preferred learning environment. Then Pearson's product-moment correlation was conducted to examine the correlation between participants' expectations of this online course and the extent to which their expectations had been fulfilled in terms of linguistic knowledge gain and content knowledge construction. Lastly, the recorded interview protocols of the eight interview questions were transcribed and used to guide more fine-grained analysis of participants' perceived views on the specific features of this collaborative e-learning environment. More specifically, in addition to running simple descriptive statistics based on Microsoft Excel for some interview questions (i.e., Interview Questions 1, 3, 4, 6, and 8), categories concerning participants' responses or comments reported for the other questions (i.e., Interview Questions 2, 5, and 7) were created from the interview qualitative data and descriptively analyzed.

4 Results and Discussions

Student Perception Questionnaire (quantitative measure)

The data from the questionnaire on student perceptions about computer-based language learning indicated that prior to the online course, all course participants involved in this study had experiences in computer and Internet use, but did not have much prior exposure to computer assisted language learning with only 7% of the participants experiencing e-learning format, which implied that traditional mass lecture was the most common instruction mode that participants' previous course instructors adopted in their teaching, and that exposing students to a collaborative CALL environment would confront and challenge their learning experiences as EFL learners. To analyze the data obtained from the student perception pre-and post-questionnaires, Paired-Samples T-Test was calculated to determine if these differences in the participants' perceived computer-assisted EFL learning, through one-semester period of the whole e-learning language course, were significant. As can be seen from the statistical data on Table 1, there were statistically significant differences between the two administrations of the Student Perception Questionnaire at the level of p < .01 regarding student interest in CALL, the effectiveness of CALL, and learning environment preference. This suggested that the e-learning language course produced a beneficial effect on participants' overall satisfaction in the computer-based language learning environment, despite the result that the challenge level of CALL perceived by participants showed no significant difference between the pre-and post-questionnaires. The analysis of the quantitative data on Tables 2 and 3 revealed that students' expectations of this online course had statistically positive relationships with their perceptions of the extent to which their expectations in terms of linguistic knowledge gain and content knowledge construction had been fulfilled. This indicated that the students who made high expectations of the online Language and Culture Course tended to be more satisfied with this course when it came to linguistic knowledge acquisition and content knowledge learning than those who made low expectations of this online course. This was probably because the higher expectations students made for the course, the more attention they devoted to those expected aspects, and the more likely they would accomplish those expectations.

Table 1.

	Paired Differences			
	Mean	Std. Deviation	t	Sig. (2-tailed)
pre-challenge of CALL - post-challenge of CALL	-.071	1.029	-.735	.464
pre-interest in CALL - post-interest in CALL	-.295	1.079	-2.889	.005**
pre-effectiveness of CALL - post-effectiveness of CALL	-.321	1.033	-3.293	.001**
pre-prefer conventional learning environment - post-prefer conventional learning environment	.438	1.432	3.234	.002**
pre-prefer CALL environment - post-prefer CALL environment	-.821	1.344	-6.469	.000**

N=112; *p <.05, ** p <.01

Table 2.

post-expectation fulfillment-linguistic knowledge gain		
pre-expectation-linguistic knowledge gain	Pearson Correlation	.351(**)
	Sig. (2-tailed)	.000

N=112; ** Correlation is significant at the 0.01 level (2-tailed).

Table 3.

	post-expectation fulfillment-content knowledge construction	
pre-expectation-content knowledge construction	Pearson Correlation	.341(**)
	Sig. (2-tailed)	.000

N=112; ** Correlation is significant at the 0.01 level (2-tailed).

Structured Interview (qualitative measure)

This section mainly deals with the qualitative measure related to Taiwanese university students' perceived views on the CALL course in a collaborative learning environment, which was explored through analyzing participants' responses to the structured interview. The responses to Interview Question 1 showed that the agreement on the view that the collaborative e-learning language course had facilitated effective learning was expressed by almost all participants (108 out of a total of 112 participants), demonstrating they all benefited from the CALL application in their language learning. This was corroborated by the participants' feedback to Interview Question 2 which required them to definitely specify the ways in which the collaborative e-learning language course had promoted their language learning. The majority of students reported that this online EFL course motivated their interest in what was taught about language and culture (78%), that the online EFL course provided more up-to-date information than print resources (67%), that the interactive and self-pace online exercises helped them in increasing their English lexical and linguistic knowledge (89%), and that the collaborative group tasks helped them better construct new knowledge about the interrelation between language and culture (76%). For example, Participant S22's comment demonstrated her strong favor to this language course with a CALL component. She said: "*It seems to me that the online interactive exercises are very effective and I benefited much and learned much from those exercises to improve my English vocabulary and grammatical knowledge.*" Moreover, Participants S34 but also S89 appeared to have a similar viewpoint and mentioned that they gained more English linguistic knowledge and content knowledge associated with language and culture during the e-learning process. As for the facilitating effect of the CALL instruction on some participants' learning motivation, Participant S72's response showed her clear reflection about learning interest: "*From my past experience, I get bored easily or fall into sleep while learning English via textbooks. However, learning through the computer arises my interest in what I am learning.*" The results on Interview Question 3 which is in relation to the advantages of the collaborative e-learning language course revealed that slightly more than half (52%) of the participants agreed or strongly agreed that computer assisted EFL teaching enhanced their interest in what they were learning, and almost two-thirds (66%) strongly agreed or agreed that it was more enjoyable to learn through computer assisted language teaching than through conventional mass lectures. The majority (77%) of participants expressed strong to moderate agreement with the statement that 'Online learning is more effective for me than textbook reading', and only a small number (4%) strongly disagreed with this statement. In the words of Participant S10: "*For me, online learning is more effective. It is because I can receive immediate feedback from online exercises to advance my learning from my own errors and reinforce what I've learned in class.*" With regard to the affective advantages of computer-assisted EFL learning, it was found that over two-thirds (69%) of the respondents agreed that they felt less pressured to learn language through an e-learning environment than a conventional classroom learning environment, as referred to in Participant S57's report: "*I have less anxiety and pressure when learning online because I am allowed to work at my own pace and don't need to compare results with others.*" In addition, approximately half (51%) of the participants associated the collaborative e-learning process with gaining a sense of achievement. Participate S101 said that "*During online discussions, my group mates and I worked together well to collaboratively overcome textual comprehension difficulties. It is quite an accomplishment whenever the unknown parts were cleared up.*" However, this was not always the case with some higher level students. The opposite response was found in the case of Participant S6. He noted that "*In the collaborative group tasks, I helped my group members more than they offered assistance for me, and when I was confronted with some linguistic problems I was unable to sort them out, and those problems, on many occasions, were left as incomprehensible.*" As to participants' reported perceptions of the disadvantages of the collaborative e-learning language course (Interview Question 4), the dataset was overall positive. Only the statements No. 5 which is about a learner's computer skills showed a slightly higher proportion of negative responses (21% agreeing with a negative statement), but the weight of opinion did not appear strong. It was encouraging to see that the statement 'Learning online is unorganized or unsystematic' drew disagreement from nearly all participants (108 students or 96%), and more than three-fourths of the students (84%) expressed strong to moderate disagreement with the statement that 'Computer assisted language learning complicates the process of learning'. It seemed that participants showed strong positive regard for computer-assisted EFL learning. This was probably because the well-designed online materials and activities motivated students to learn and led to effective and efficient learning. In addition, it was interesting to see that there was no one who felt embarrassed when working on the collaborative group tasks with their group mates, which conformed to participants' feedback in reply to Interview Question 6 with the relatively high number of students (102 out of a total of 112 participants) who said that they enjoyed working online with their group members. Only a minority (9%) of the students reported that they felt threatened and uncomfortable with learning English via a computer, and less than one-fourth (21%) moderately agreed with the statement that 'Computer assisted language learning requires technical skills'. Piotrowski and Vadonovich (2000) state that a learner's lack of computer competence will bring about fear to work in an online learning context. However, it is generally believed that students nowadays have more chances to use computers and the Internet at home or school, which reduces their anxiety at computer-based learning, and thus learning through a computer will not cause great difficulty for most students in terms of computer operation. Also, as mentioned earlier, students were offered, at the onset of the course, an introduction and a technical training session about this online language course, which facilitated them to be well familiar with the instructional technology and to request less technical support during the whole e-learning procedure. Furthermore, the most frequent responses to Interview Question 5 about what participants liked and disliked the most about this online CALL course were respectively online interactive exercises (63%) and collaborative group tasks (34 students or 30%, compared to 16 students reporting to be uninterested in online lectures, and the others (62 students or 55%) expressing no any course elements that they disliked). The participants offered reasons for their favorite instructional element concerned with instant feedback provided from the online interactive exercises. Take, for instance, Participant S98's reflection, who said in his response to the favorite part of this online course, *"The exercises are interactive and I can receive immediate feedback to know what I got right and what I got wrong on my answers, which is good for my self-learning.*" In fact, in CALL instruction, greater leaner control over the learning process is allotted for learners so that they can work independently to perform exercises in a self-paced manner. Most importantly, multiple opportunities as well as immediate feedback that online exercises or activities provide for extensive practice and improvement promote second/foreign language learning. As to the part of the online course the participants liked the least, time-consuming task completion was some participants' justifications for negative feedback, but it did not appear from these responses that there were sufficient problems to militate against the implementation of a collaborative CALL environment in EFL learning. For example, Participant S78's report was typical of many participants: "*The collaborative group project makes learning easier because you got a chance to share information, discuss problems, and exchange opinions. All members put their heads together so as to maximize comprehension.*" Also, "*We worked together quite well as a team, which not only helped us to create better sense of text but also to enhance our overall knowledge. I found the more I grasped from what had been read, the more familiar I was getting with the content area and the more overall knowledge I gained from it."* was mentioned by Participant S7. It was clear that the emphasis of the above responses was placed on peer cooperation. However, several participants' comments were somewhat different. They did not show favor to the collaborative group tasks. Too much time spent on helping with others' linguistic problems was noted by the participants as the top reason that they did not enjoy collaborative group projects. Take Participant S12's explanation as an example. He reported that his motivation to carry out the collaborative group task with his group members was not so high because sometimes his group mates could not do much to help and often counted on him to paraphrase or explain for them when faced with reading problems in the texts found online. Some students, such as Participants S8, S53, and S89, also produced a similar view to regard time-consuming processing needed in online group discussions as a drawback and did not comment positively. Participant S53 noted that "*During online discussion, it took a lot of time to discuss with my group mates about the reading texts relevant to the task topic because many times I had to help them with their linguistic problems, whereas I can read much faster when I am reading on my own.*" Nevertheless, in spite of 34 students who said that they disliked the collaborative group projects the most, when asked whether grouped participants enjoyed collaborating with group members online while working on the assigned tasks (i.e., Interview Question 6), the majority (102 participants, 91%) admitted this to be the case. In addition, in terms of the kinds of help that individual participants were offered from their group mates while carrying out the collaborative group tasks (i.e., Interview Question 7), one-fourth of the participants (23%) mentioned their group members helped them with technical problems, and approximately two-thirds reported they received help from group mates to deal with language problems (68%), to find information about the topic from the World Wide Web (65%), and to construct textual meaning found online (69%). As demonstrated by Participant S49's response, he stated that "*In addition to providing lexical resources, whenever I had difficulty in understanding certain sentences, my group members with advanced English proficiency would help me with sentence meanings through analyzing syntactical and grammatical structures or elaborating sentence interpretations so that I could comprehend the text better.*" Participants S2 and S66 also had similar viewpoints, as noted by S66 that "*You learn more linguistic knowledge and overall knowledge during cooperative learning. Others may know what you do not know and everybody pools her/his wisdom, we thus know a lot more all together.*" Obviously, based on all the verbally reported responses to these interview questions, we may thus conclude that such collaborative group tasks created an online learner-centered context where group discussion, problem solving, and collaborative learning were encouraged to enable students to become actively engaged in the construction and use of knowledge, which has long be ignored in conventional language instruction in Taiwan. It is therefore that the learning environment constructed in this study apparently satisfied conditions for the formation of a Vygotskian zone of proximal development (Vygotsky, 1978) for foreign language learners. Specifically, the more capable participants offered scaffolding, which in turn consolidated their linguistic knowledge or content knowledge and firmed up their text understanding. The less capable peers received linguistic assistance and learned from their higher counterparts, which was vital in the shaping of their language competence and probably enabled them to achieve more highly than they would have been able to accomplish alone. Finally, in response to Interview Question 8 of whether participants would be willing to try another online course in the future, almost all participants (109 students) concluded that they would take another language course with an online component and recommended a collaborative CALL environment being implemented in the future courses, suggesting that their overall experience with CALL was positive. Reasonably, participants' positive evaluation about the implemented online course resulted in an increase in their motivation to take future courses that incorporate technology and language instruction.

To sum up, the results produced from the questionnaire quantitative data with significant differences found in student self-perceptions of computer-assisted EFL learning were confirmed by the findings brought about from the interview qualitative data concerning the overall positive attitudes of students towards the effectiveness of CALL use and EFL learning. Also, these findings of the research were in agreement with other CALL studies which report positive effects of online language learning (e.g., Blake & Delforge, 2006; Conolelos & Oliva, 1993; Felix, 2001).

5 Conclusion

In conclusion, we see that the findings drawn from both the questionnaire data and the interview protocols were substantially positive to demonstrate that the participants in this study generally supported collaborative computer-assisted EFL learning and perceived its educational effect on their enhanced linguistic knowledge and language/culture content knowledge, motivation and interest in learning English via a computer, as well as cooperative learning. Such valuable views of participants gleaned from this research study can offer Taiwanese teachers and school authorities evidence regarding CALL application in Taiwanese students' English language learning. Also, we can suggest that students' EFL learning performance could be expected to improve when they are exposed to such a collaborative CALL environment. It is because, as the encouraging and convincing evidence provided from the study, that language learners involved in a collaborative e-learning environment which turned language learning into an engaging, interactive, and student-centered process will be likely to learn in a more efficient manner and to improve their linguistic competence and content knowledge accompanied with higher motivation. It is therefore advisable for Taiwanese English language teachers, course writers, curriculum planners, policy makers, and educational authorities to give further consideration to applying an instructional approach which incorporates an CALL component in a collaborative setting to immerse students in online language learning in a social context in order to help students with effective language acquisition and motivated learning attitudes.

References

Arvan, L., Musumeci, D.: Instructor attitudes within the SCALE efficiency Projects. Journal of Asynchronous Learning Network 4(3), 196–215 (2000)

Ayres, R.: Learner attitudes towards the use of CALL. Computer Assisted Language Learning 15, 241–249 (2002)

Blake, R., Delforge, A.M.: Online language learning: The case of Spanish without walls. In: Salaberry, R., Lafford, B.A. (eds.) The art of teaching Spanish: Second language acquisition from research to praxis, pp. 127–148. Georgetown University Press, Washington (2006)

Bonk, C.J., Graham, C.R.: Handbook of blended learning: Global perspectives, local designs. Pfeiffer Publishing, San Francisco (2006)

den Brok, P., Brekelmans, M., Wubbels, T.: Multilevel issues in research using students' perceptions of learning environments: The case of the questionnaire on teacher interaction. Learning Environments Research 9(3), 199–213 (2006)

Collentine, J.: Insights into the construction of grammatical knowledge provided by userbehavior tracking technologies. Language Learning and Technology 3, 44–57 (2000)

Coniam, D.: Evaluating computer-based and paper-based versions of an English-language listening test. ReCALL 18(2), 193–211 (2006)

Cononelos, T., Oliva, M.: Using computer networks to enhance foreign Language/culture Education. Foreign Language Annals 26, 524–534 (1993)

Dunkel, P.: Computer-assisted instruction (CAI) and computer-assisted language learning (CALL): Past dilemmas and future prospects for audible CALL. Modern Language Journal 71, 250–260 (1987)

Felix, U.: The web's potential for language learning: The student's perspective. ReCALL 13, 47–58 (2001)

Ginns, P., Ellis, R.: Quality in blended learning: Exploring the relationships between on-line and face-to-face teaching and learning. The Internet and Higher Education 10(1), 53–64 (2007)

Hannafin, M.J., Land, S.M.: The foundations and assumptions of technology-enhanced student-centered learning environments. Instructional Science 25, 167–202 (1997)

Kern, R.G.: Restructuring classroom interaction with networked computers: Effects on quantity and quality of language production. The Modern Language Journal 79, 457–476 (1995)

Kitao, K.: Students' evaluation of CAI English classes. Doshisha Studies in English 64, 117–160 (1995)

Lee, L.: Using web-based instruction to promote active learning: Learners' Perspectives. CALICO Journal 23(1), 139–156 (2005)

Lin, C.: Constructivism and Second Language Learning: A Web-Based Reading-Writing Activity. Unpublished master thesis. National Taiwan Normal University, Taiwan (2002)

Lin, Y.: The effects of e-books on EFL learners' reading attitudes. Unpublished master thesis. National Taiwan Normal University, Taiwan (2009)

Liou, H.: The impact of WWW texts on EFL learning. Computer Assisted Language Learning 10, 455–478 (1997)

Meyer, J., Muller, M.: Evaluating the quality of student learning: I. An unfolding analysis of the association between perceptions of learning context and approaches to studying at an individual level. Studies in Higher Education 15, 131–154 (1990)

Murray, G.L.: Autonomy and language learning in a simulated environment. System 27, 295–308 (1999)

Neumeier, P.: A closer look at blended learning – parameters for designing a blended learning environment for language teaching and learning. ReCALL 17(2), 163–178 (2005)

Pawling, E.: Modern languages and CD-ROM-based learning. British Journal of Educational Technology 30(2), 163–175 (1999)

Peterson, M.: Software review: The grammar Rom. ONCALL 11(1997), `http://www.cltr.uq.edu.au/oncall/vol11ndx.html` (Retrieved January 3, 2010)

Piotrowski, C., Vadonovich, S.J.: Are the reported barriers to Internet-based instruction warranted?: A synthesis of recent research. Education 121(1), 48–53 (2000)

Ramsden, P.: A performance indicatory of teaching quality in higher education: The course experience questionnaire. Studies in Higher Education 16, 129–150 (1991)

Rivers, W.M.: Interactive language teaching. Cambridge University Press, New York (1987)

So, H.J., Brush, T.A.: Student perceptions of collaborative learning, social presence and satisfaction in a blended learning environment: Relationships and critical factors. Computers & Education 51(1), 318–336 (2008)

Strepp-Greany, J.: Student perceptions on language learning in a technological Environment: Implications for the new millennium. Language Learning and Technology 6, 165–180 (2002)

Sun, Y., Dong, Q.: An experiment on supporting children's English vocabulary learning in multimedia context. Computer Assisted Language Learning 17(2), 131–147 (2004)

Ushida, E.: The role of students' attitudes and motivation in second language learning in online language courses. CALICO Journal 23, 49–78 (2005)

Vygotsky, L.S.: Mind in society: The development of higher psychological processes. Harvard University Press, Cambridge (1978)

Zapata, G.C., Sagarra, N.: CALL on hold: The delayed benefits of an online workbook on L2 vocabulary learning. Computer Assisted Language Learning 20(2), 153–171 (2007)

Performance Evaluation of Telecommunication Industry between China and Taiwan

Yao-Lang Chang[1], Tsung-Ting Shih[1], and Chih-Hong Wang[2]

[1] National Kaohsiung University of Applied Sciences, Taiwan
[2] National Chengchi University, Taiwan
kc88899@gmail.com

Abstract. It becomes more and more difficult for a telecommunication company to retain a successful operational model in the highly competitive environment. In fact, a good performance evaluation method becomes a good management technique for bench marking among industry. This has become an important topic nowadays.

The objective of this research is to provide an effective study to evaluate the telecommunication industry between China and Taiwan, as well as the analysis of efficiency and improvement suggestions for those companies which want to improve the efficiency based on Data Envelopment Analysis (DEA). Realistic data are collected from both China and Taiwan published stock market. Total major 5 companies, which are included around 90% market share, are collected. The results show that 3 companies have very good performances. Moreover, this study provides improvement suggestions for those companies, which performances are not good enough. This method and results are useful for telecommunication industry for management and bench marking between China and Taiwan.

Keywords: Data Envelopment Analysis, Efficiency, performance evaluation.

1 Introduction

Communications in recent years, rapid development, its information delivery of fast, to replace the information delivery, most high-profile focus, one of the most distinctive places, as the borderless network concept, the development of the communications industry, along with the convenience of the flow of information and the same property, competition policy planning and layout, to become important in the development of the telecommunications industry. Therefore, in the telecommunications industry, how to take advantage of the upstream and downstream information sharing between manufacturers, from the message data in finding each other productive factors, so that all participants in the supply chain of mutual benefit, it is important for long. However, in that many of the manufacturers, to how to choose the most appropriate partners, is the most important single issue.

As information technology and the fierce competition of the era, observation of the development of domestic telecommunication industry in recent years, found that the characteristics of the industry and the development track feature service is primarily for

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 433–443, 2010.

the country. Although considered how to go out, the future scope of business in not only in this country. First, the domestic market is limited [1]. Secondly, the overall communications industry needs substantial economies of scale. This business must be globalization.

Major equipment and technology will also come from abroad in the telecom industry. In addition to the base station rentals, almost all machines are bought from overseas. This is huge investment. Therefore, the added value is not high in this industry. In addition, national and commercial market's competitive advantage is limited to the marketing in the fundamental technology or business model innovation. The key technologies could not be breakthrough, and they can only be highly dependent on foreign partners.

In addition, the government opened the folk communications industry into this market opening from the exclusive. The telecom industry is concerned, as the market requirements increase, the government open investment, once had a lot of competitors of the State, however, in the competition and after, to join the communication market survive much, but in the "investment", "the creation of added value small", "market limited" several special features, the existing manufacturers of profitability or improvement.

Also, government policies on management and profit level influences. For example: Telecom, the business of openness and rates are critical. You can even say, this industry firms are easy to earn money under government's policies.

In fact, the communications industry is capital intensive and knowledge intensive and cross counties. However, by following the world standards, the market which is set by the rules of government is very competitive.

Basically the above characteristics, in order to the electronics industry, or in many traditional industry, never appeared, cited this the characteristics of the industry does not say that Taiwan's overall industry generally is limited to only, presented in the communications industry. These industry characteristics mean that the future development of telecom industry in reference before can learn those lessons. More of these problems are valuable issues.

The remainder of this paper is organized as follows. Literature reviews are given in Section 2. Methodology is introduced in Section 3. Experiment design and results analysis are depicted in Section 4. Finally, Section 5 provides some concluding remarks.

2 Literature Reviews

DEA was first proposed by Charnes, Cooper and Rhodes (CCR) [2]. Its original idea comes from the measurement model of production efficiency proposed by Farrell [3]. DEA itself is a non-parametric method for assessing the relative efficiency of decision making units (DMUs) based upon multiple inputs and outputs. The primitive DEA model adopts the concept of production in microeconomics: efficiency = output / input. Banke, Charnes, and Cooper (BCC) [4] developed a new model from the CCR model to understand the problems of pure technical efficiency (PTE) and scale efficiency (SE). Both of the CCR and BCC models are summarized as below.

The CCR model intends to maximize the ratio of weighted outputs against weighted inputs. It reduces multiple outputs to a single "virtual" output, and multiple inputs to a single "virtual" input for each DMU. CCR is good at analyzing the relative efficiency without setting the weights in prior, which makes the CCR model more objective.

Assume that there are n DMU. Each DMU has m inputs and s outputs. Let x_{ij} represent the ith input and y_{rj} represent the rth output of DMU j, respectively. Let u_r and v_i represent the virtual variables of rth output and ith input, respectively. Let h_j represent the relative efficiency of DMU j. Where ε is a relatively small positive number (normally set at 10^{-6}).

The relative efficiency of each DMU can be calculated by solving the following mathematical programming problems:

$$\text{Max } h_j = \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \tag{1}$$

$$\text{subjected to } \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1 \tag{2}$$

$$u_r \geq \varepsilon > 0 \tag{3}$$

$$v_i \geq \varepsilon > 0 \tag{4}$$

r =1,2,3,….,s ; i =1,2,3,…..,m ; j =1,2,3,…..,n

The CCR input model can suggest improvement directions and the values of both outputs and inputs in order to achieve the desired efficiency value of 1, which can be done by calculating the following equations:

$$x_{ij}^{*} = h_j x_{ij} - s_i^{-*} \quad , \text{ i =1,…,m} \tag{5}$$

$$y_{rj}^{*} = y_{rj} + s_r^{+*} \quad , \text{ r =1,…,s} \tag{6}$$

Where s_i^- represents the lacking quantity (slack) of the output and s_r^+ represents the redundant quantity (surplus) of the input. $x^*{}_{ij}$, $y^*{}_{rj}$, s_r^{+*} and s_i^{-*} represent the optimal values of x_{ij} , y_{rj} , s_r^+ and s_i^- , respectively. Note that when the value of objective function $h_j = 1$, s_r^+ =0 and s_i^- =0 (for all r and i). That is, DMU j achieves its optimal efficiency. Note that CCR is the most popular DEA model for evaluation of the total operational efficiency.

The difference between the CCR and BCC models is the use of returns-to-scale. For each DMU, BCC allows variable returns-to-scale, while CCR is characterized by constant returns-to-scale. The BCC model has an additional convexity condition. The input–oriented BCC model can be written as:

$$\text{Max } h_j = \frac{\sum_{r=1}^{s} u_r y_{rj} - u_0}{\sum_{i=1}^{m} v_i x_{ij}} \tag{7}$$

$$\text{subjected to } \frac{\sum_{r=1}^{s} u_r y_{rj} - u_0}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1 \tag{8}$$

$$u_r \geq \varepsilon > 0 \tag{9}$$

$$v_i \geq \varepsilon > 0 \tag{10}$$

r =1,2,3,…,s ; i =1,2,3,…,m ; j =1,2,3,…,n

u_0 is a real number to indicate the intercept of the production frontier. When u_0>0, the production frontier for this DMU is decreasing returns-to-scale (DRS). When u_0=0, the production frontier for this DMU is constant returns-to-scale (CRS). When u_0<0, the production frontier for this DMU is increasing returns–to-scale (IRS). In addition, the suggested improvement directions of the BCC input model are the same as the CCR input model.

Most of DEA applications are focused on performance evaluation. The research of Bhattacharyya [5] *et al.* show that government owned banks possess more operational efficiency than privately owned banks, but less efficiency than foreign banks. Camanho and Dyson [6] enhance cost efficiency measurement methods. The results obtained in the case study show that the DEA models can provide robust estimates of cost efficiency even in situations of price uncertainty. Lee *et al.* [7] discuss a description of a DEA model for analysis of the control performance for a specific context for electronic data interchange (EDI) in the context of finance and trade. Homburg investigates the use of DEA for activity-based management and pros and cons of DEA as applied to benchmark activities [8]. Mota *et al.* [9] provide a quantitative model for activity-based management (ABM). A real case study of a drill factory is used to illustrate the application of the model. Co and Chew use DEA to analyze the performance and R&D expenditures in American and Japanese manufacturing firms [10]. An approach based on DEA is proposed by Chang and Lo [11] for measuring the relative efficiency of an ISO 9000 certified firm's ability to achieve organizational benefits. Luo and Donthu [12] use DEA and Stochastic Frontier (SF) to show that top 100 marketers' advertising spending in print, broadcast, and outdoor media are not efficient and could bring in 20% more sales. Cook and Zhu [13] use DEA for productivity measurement of highway maintenance crews as maximum achievable by reduction in resources without impacting the outputs from the process. Durand and Vargas [14] analyze the ownership, organization, and private firms' efficient by DEA. Forker *et al.* [15] combine nonlinear DEA and linear regression analyses, and then demonstrate that Total Quality Management (TQM) practices are related to performance. However, DEA can not only be used in performance evaluation and can be extended more.

3 Methodology

This research department you want to take advantage of the datagram through complex analysis and interpretation of the case for the telecommunication industry in addition to innovative product development, improve product quality, reduce operating costs, for application integration policy, put forward a viable measure evaluation method management staff more efficient and convenient choice and discretion to enable more effective expansion of telecom markets, acquire new skills, improve operational efficiency, etc.

STEP1: First collection at home and abroad for the communications industry for DEA research data and then look for the Institute from which to explore the theme and use tools, and for the future can affect industry development factors for performance evaluation, through set of targets, to a hypothetical strategic alliances.

STEP2: By collecting the STEP1 related literature, found in control performance and key performance indicators to invest in projects of select elements, and related input of resources and performance data can be routed through a pointer to the data to determine which output/input of variance for a strategic alliance to produce the maximum benefit, then variation the number of items through the correlation coefficient analysis verification, if the values are negative values, you must delete the data to the number of such variation or select, if the value is positive, then shows the correlation coefficient is positive, and DEA model forward dependencies.

STEP3: The STEP2 know correlation coefficient analysis to forward relevant and consistent with the DEA model forward dependencies, to form DEA software for evaluation, this study chose to CCR-I mode to do analysis.

STEP4: By choosing to STEP3 CCR-I pattern analysis, this stage is to start with the two cases between five telecom vendor data to conduct performance appraisals, aims to confirm which manufacturers operating efficiency for the best.

STEP5: From the available communication operators STEP4 manufacturers operating efficiency, and then to complete data, assess enterprise efficiency.

STEP6: The conclusions obtained by STEP5, to produce a recommendation for the enterprise, enabling enterprises to do strategic alliance, can have a better direction, so that we can in this competitive environment, create the best competitive advantage.

4 Experiment Design and Results Analysis

Telecom industry and research: 1996 through telecommunications law, from the telecommunications industry from exclusive to free competition, open to Taiwan, where the domestic, remote telecommunications, Pacific telecom, Weibao Telecom, the number of operators in the telecommunications industry competition, including fixed-line telecommunications, mobile communications, and data communications provides three areas, dedicated circuits, Internet, broadband, intelligent network, enterprise integration services, and all kinds of value-added services, in order to meet the mobile phone and the Internet is widely used by sparking life action and broadband wave, and through alliances, cooperation, expanding the utility, telecommunications and the Internet to develop action-business, network applications, as well as broadband audio and video multimedia and other broadband services [16].

This paper discusses the researches focus on discussion and looking for operators in several industry competitions, for various companies and their respective management policy, according to the input and output of detail, seeking advice on the telecom industry as a whole industry of best phenomenon, as the telecom industry is back in operation in input, and output the reference basis. This study is based on practical considerations for data examined.

According to the DEA's application, use the datagram complex method of the first steps, that is, select the appropriate DMU, this study of industry is the software development industry, so pick the two case studies on target between carriers for. According to the report data, telecom operator, 1998 the this thesis research company listed in the following table 1.

Table 1. The communications industry data [17] [18] [19]

	Capital (100 millions)	Base stations	Employees	Customers	Revenue (million)	EPS (earnings per share)
CHINA MOBILE	2436	360000	160000	314,220	452103	4.51
China Unicom	921.2	107000	200000	276,019	153945	4.66
ChunghwaTelecom	945	4600	24600	279,394	184040	2.83
Taiwan Mobile	380	4000	2400	421,309	57015	4.741
Far Eastone	340	4000	3600	213,014	53740	0.364

Tables on the data provided input on the part of the input data in employment due to the nature of the industry is not a labor intensive high-tech industry, so the number of investment and the actual operational noticeable effect in terms of number of jobs directly and therefore have not been included in the list into the selected is used to output the data in the operating income and EPS after taxes due to nature, operating income of the High Representative and high, therefore EPS data selected only selected the tax data for output after EPS. This study a total of four DMU inputs and outputs total four input variables: capital, base number, number of users, select EPS after taxes will abide by this principle pointer: input and output items may not be negative, selected the following input and output of the project to a variable

1.Input variable 1 (X_1): Capital
2. Input variable 2 (X_2): base stations
3. Output variable 1 (Y_1): customers
4. Output variable 1 (Y_2): EPS

This research used DEA-Solver software to calculate the DEA model. CCR input model was applied for this study. Table 2 shows the description statistics data.

Table 2. Data of description statistics

	Capital (100 millions)	Base stations	Customers	EPS
Max	2436	360000	52200	4.741
Min	340	4000	433	0.364
Average	1004.44	95920	13732	3.421
SD	760.42513	137912.24	20004.579	1.6827342

The table can be learned, statistics from the most is the investment capital and base station with the largest number of China Mobile, the statistical data to minimum is a long-distance telecom, top sheet to: China Mobile related statistical data to the maximum, the reason should be regional GEO-General on the Mainland, the long-distance telecom investment project was put into question a lot, and in the output, China Mobile user number of the largest, the Chinese population inherent advantages, such as statistical data on the difference between SEO analysis, statistics, minimum statistical value of data on behalf of sense cannot factor as far as a result of statistical assessment points, so the data's data interpretation, still need to take into account the relationship and population, statistics on comparison is not controversial.

Table 3. Statistical correlation coefficient must be positive

	Capital	Base stations	Customers	EPS
Capital	1	0.9555006	0.9582122	0.2654588
Base stations	0.9555006	1	0.9998812	0.1341859
Customers	0.9582122	0.9998812	1	0.1483795
EPS	0.2654588	0.1341859	0.1483795	1

This research applied DEA method to select variables to go on correlation analyses to test if each pair of all variables is positively related, which is the basic calculation assumption of DEA. As shown in Table 3, correlations for all variables are positive. Thus the principle is met.

An analysis of the efficiency is calculated by the software of DEA-Solver. The CCR-I mode is used to analysis those companies. The results are show in Table 4.

Table 4. efficiency assessment analysis tables

Rank	DMU	Score
1	ChunghwaTelecom	1
1	Taiwan Mobile	1
1	**CHINA MOBILE**	1
4	China Unicom	0.9493572
5	Far Eastone	0.8408392

From the above table shows that: in efficiency assessment on Chunghwa Telecom. Taiwan's eldest brother. China Mobile's operating efficiency is no problem of long-distance and China but connected business efficiency is not very good, then do business efficiency assessment, finishing get table 5:

Table 5. Input, output analysis table

No.	DMU	Score			
	I/O	Data	Projection	Difference	%
1	ChunghwaTelecom	1			
	Capital	945	945	0	0.00%
	Base stations	4600	4600	0	0.00%
	Customers	817	817	0	0.00%
	EPS	4.51	4.51	0	0.00%
2	Taiwan Mobile	1			
	Capital	380	380	0	0.00%
	Base stations	4000	4000	0	0.00%
	Customers	450	450	0	0.00%
	EPS	4.66	4.66	0	0.00%

Table 5. (*Continued*)

3	Far Eastone	0.8408392			
	Capital	340	285.88534	-54.114664	-15.92%
	Base stations	4000	3363.3569	-636.64311	-15.92%
	Customers	433	433	0	0.00%
	EPS	2.83	2.83	0	0.00%
4	**CHINA MOBILE**	1			
	Capital	2436	2436	0	0.00%
	Base stations	360000	360000	0	0.00%
	Customers	52200	52200	0	0.00%
	EPS	4.741	4.741	0	0.00%
5	China Unicom	0.9493572			
	Capital	921.2	874.54784	-46.652165	-5.06%
	Base stations	107000	101581.22	-5418.7816	-5.06%
	Customers	14760	14760	0	0.00%
	EPS	0.364	2.2491085	1.8851085	517.89%

The previous table may know, China Telecom, China Mobile, Taiwan Mobile, evaluation input-output, with its operating efficiency on its efficiency, its strong competitiveness, but instead return a long-distance telecom, China Unicom, a data analysis view on its assessment of efficiency on a long-distance telecom capital to reduce to $ 285.87 billion, base station number also can reduce the number 3363 bases and operating efficiency, with a group of telecommunications, China Unicom's operating efficiency to be promoted, in accordance with table questions on capital & base number to 5.06%, capital of 874.58 billion NT dollars, base station number is 101581, EPS after tax increased by 517.89%, $ 2.249 per share, so as to let China Unicom operating efficiency is better operating efficiency.

5 Concluding Remarks

This study provides an enterprise to measure the effectiveness of DEA's direction and management to improve the competitiveness of the reference method. This study of selected second input pointer (capital, base station number) and two output pointer (operating income and earnings per share) assessment 5 carrier's operating performance, analysis, a conclusion to Chunghwa Telecom, China Mobile, cell phones, Taiwan companies still maintain efficiency target value is 1, 2 companies transmitting telecommunications, China Unicom, performance data on less than one, 2 companies operating performance improvement and efficiency target value is 1, so to domestic long-distance telecom, reduce capital investment, and reduce the base number, such as domestic companies reduce will let business efficiency into a, but in order to operate on these two input reduce to practical and commercially lower number means base station rocket, feared that will reduce the number of users that will lead, so putting it in the business of reducing machine efficiency increases, but on may lead to address quality issues, thus reducing the number of users, to China Unicom, problem basically and remote telecommunications are similar, but the biggest difference is then to territorial size, map to reduce China Unicom's capital, means to reduce the base number, change the word know sacrifice quality in exchange for helping to improve the operation of machines, a commercial angle, and it is some problems, so China Unicom has a bit of improvement is to increase EPS increased this data, the number of users that will increase, but the user increases the base station to reduce to clients, China Unicom would not be the best choice you, if we can increase the number of customers may need not reduce the amount of base station, increase the number of users will be able to increase its efficiency.

This study only foists industry firms do business performance analysis, for subsequent study and put forward the following suggestions: (1) in the future for carrier performance evaluation, input and output should be in the project as a more extensive collection, where the performance impact of pointer, should be taken into consideration; if a company's internal data will make research results more persuasive; (2) this study only one industry for business efficiency assessment, future follow-up research can join the upstream and downstream third-party discussion of strategic alliances in order to increase the diversity of topics; (3) an in-depth look at the pros and cons of the industry, provides enterprise key management system and information, and how to properly reduced corporate resources to achieve higher performance; these are come with future directions.

References

[1] Google discussion (2008), http://groups.google.com.tw/group/seetooforum/web/%E6%B1%BD%E8%BB%8A%E6%A5%AD%E3%80%81%E9%9B%BB%E4%BF%A1%E6%A5%AD%E8%88%87%E9%87%91%E8%9E%8D%E6%A5%AD

[2] Charnes, A., Cooper, W.W., Rhodes, E.: Measuring the efficiency of decision making units. European Journal of Operational Research 2(6), 429–444 (1978)

[3] Farrell, M.J.: The measurement of productive efficiency. Journal of the Royal Statistical Society, Series A, General 120, 253–281 (1957)

[4] Banker, R.D., Chanes, A., Cooper, W.W.: Some models for estimating technical and scale inefficiencies in data envelopment analysis. Management Science 30, 1078–1092 (1984)

[5] Banker, R.D., Chanes, A., Cooper, W.W.: Some models for estimating technical and scale inefficiencies in data envelopment analysis. Management Science 30, 1078–1092 (1984)

[6] Bhattacharyya, A., Lovell, C.A.K., Sahay, P.: The impact of liberalization on the productive difference of Indian commercial banks. European Journal of Operational Research 98, 332–345 (1997)

[7] Camanho, A.S., Dyson, R.G.: Cost efficiency measurement with price uncertainty: a DEA application to bank branch assessments. European Journal of Operational Research 161, 432–446 (2005)

[8] Chang, D.S., Lo, L.K.: Measuring the relative efficiency of a firm's ability to achieve organizational benefits after ISO certification. Total Quality Management and Business Excellence 16, 57–69 (2005)

[9] Charnes, A., Cooper, W.W., Rhodes, E.: Measuring the efficiency of decision making units. European Journal of Operational Research 2, 429–444 (1978)

[10] Co, H.C., Chew, K.S.: Performance and R&D expenditures in American and Japanese manufacturing firms. International Journal of Production Research 35, 3333–3348 (1997)

[11] Cook, W.D., Zhu, J.: Output deterioration with input reduction in data envelopment analysis. IIE Transactions 35, 309–320 (2003)

[12] Delmas, M., Toka, Y.: Deregulation, governance structures, and efficiency: the U.S. electric utility sector. Strategic Management Journal 26, 441–460 (2005)

[13] Durand, R., Vargas, V.: Ownership, organization, and private firms' efficient use of resources. Strategic Management Journal 24, 667–675 (2003)

[14] Farrell, M.J.: The measurement of productive efficiency. Journal of the Royal Statistical Society 120, 253–281 (1957)

[15] Forker, L.B., Mendez, D., Hershauer, J.C.: Total quality management in the supply chain: what is its impact on performance. International Journal of Production Research 35, 1681–1702 (1997)

[16] Taiwan telecommunication Industry Development Association, http://www.ttida.org.tw/member.php?level1_id=2&level2_id=8

[17] KGI Securities, http://etrade.kgi.com.tw/

[18] National Communication Commission Watch, http://nccwatch.org.tw/node/37570

[19] cnYes.com, http://news.cnyes.com/Content/20100430/KC8RR42QRJLX6.shtml?c=detail

Discuss and Analyze by AHP Techniques the KSF of Managing the NMMBA under the BOT Model

Wen-Ching Lin, Ying-Fung Huang, and Chia-Nan Wang

Department of Industrial Engineering and Management,
National Kaohsiung University of Applied Sciences
Kaohsiung 811, Taiwan, R.O.C.
lin.wching@msa.hinet.net, cn.wang@cc.kuas.edu.tw

Abstract. This paper as a case study will first explore the stance of the National Museum of Marine Biology and Aquarium (NMMBA) as a venue for social education by promoting outsourced management of resources complement between NMMBA and private sectors. Further, this paper is completed through relevant literature, onsite observation and interview, followed by Analytic Hierarchy Progress (AHP) for determining key success factors (KSF) as a reference for policy makers on other public constructions and outsourcing management.

The findings suggested that management strategies yield the greatest impact upon the composite factors for the success of NMMBA's outsourcing management: namely, OT with BOT, public-private partnership (PPP), and willingness of contract renewal. In analyzing the composite factors for NMMBA's outsourced managerial success, PPP demonstrates the most significant weight of 0.424.

Keywords: public-private partnership (PPP), analytic hierarchy progress (AHP), key success factor (KSF), National Museum of Marine Biology & Aquarium (NMMBA).

1 Introduction

On November 18, 1996, the Executive Yuan authorized the organization of National Museum of Marine Biology & Aquarium (NMMBA). In the draft it says, "For those education and research works that involve with great governmental investment but slow in achievements should be continually sustained and invested by the government; for those mainly for-profit and less related to the access of public right of government should be managed by the public-build-private-operate model or enterprises. Thus enable the organization, the human affairs, the finance and so on to have more elasticity and efficiency."Since, the NMMBA started to be planned towards the goal of BOT. The operation model is also innovative in the nation that the first built aquarium adopts OT, and in limited time, BOT the new aquarium.Through the PPC, the Ministry of Education .. so the feasible, in November 26,1999,Internet bulletin. (Huang Caiyun, Wen-Jie Huang, 2003) Hoping by private team and investment, the national social educational institutions can break the old frames, and the willingness of operation could be increased among enterprises. In current diverse society, the public constructions should think about the

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 444–456, 2010.

changing environment and competitors. The traditional operation model rely solely on governmental funds will finally be eliminated. Integrating private resources has become a trend.(Pei-Hsiu wu,2002)

However, the BOT cases in process are often controversial due to lack of policies and experiences. Some are even away from the basic spirit of "government franchise, private sector funds." Recently, there are cases claiming unfairness occurs in selection, politicians collude with businessmen, or doubts of false profits. These incidents postpone the construction and thus cause complaints.

Christina Liu, the professor of finance department in National Taiwan University writes,Why BOT is worked well in South Korea but not in Taiwan? The world does not have a second BOT plan is the same.(Christina Liu,Wen-Yeu Wang,Huang Yulin,2000) It seems the issue is not on the BOT itself but the false interpretation by the government and private sectors. Perhaps too many people do not know how to adopt BOT and its value to national economic development. As a worker serving in the franchisee of the BOT of the NMMBA, I am honored to have participated the OT and BOT of NMMBA since the beginning. This case also becomes a successful example for other national public BOT constructions. In considering that BOT cases are not smoothly applied in Taiwan, I am motivated to research the success key factors of the NMMBA as a reference to other BOT cases which are ongoing or about to happen.

2 Materials and Methods

The key success factors are the best performed works of a company that seeks success. (Daniel, 1961) About to recognize the key success factors way. According to Yung-Ching Ho (1990), popular statistical methods are the as following: (1) regression analysis; (2) factor analysis; (3) Delphi Method; and (4) AHP.This research adopts the AHP as the method of recognizing key success factors.In composition of "review of the operation of the NMMBA" and interviews of both public and private staff members within the BOT plan of NMMBA, the "Frame of success key factors of the BOT of NMMBA" is drawn as Fig. 1.

2.1 Brief of Analysis Method

This research adopts Analytic Hierarchical Process (AHP) to search success key factors of the BOT of NMMBA and their sequence of priority. The AHP includes 3 steps**(Saaty,1970)**: establishing the hierarchy, pair-wisely comparing the establishment and operation of the matrix, and checking the consistency.Since AHP theory has been in use for years, many assisting computing programs have also been in shape. However, the AHP has both strength and weakness in its application. Ossadnik and Lange (1999) mentions in "AHP-based evaluation of AHP-Software," 3 dominant software applying AHP are Auto Man, Expert Choice, and HIPRE. They all base on 12 structures, such as the graphical display of outcomes, the convert of particular AHP, numbers of hierarchical elements, the offering of sensitivity analysis, easy learnability, problem structure adjustment, supplies, understandable manipulate commands, related assistance and explanation of false message, and the initial investment. After analyzing these software through AHP, it is concluded that the Expert Choice is better than the Auto Man, which is better than the HIPRE. Thus this research chooses the Expert Choice as the applied software to analyze recycled questionnaires.

2.2 Frame of Success Key Factors of the BOT of NMMBA

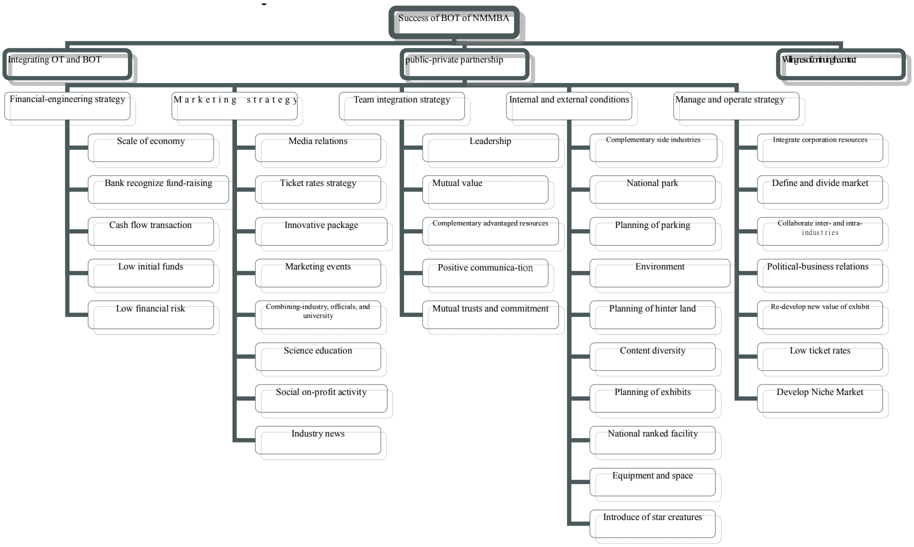

Success of BOT of NMMBA
Integrating OT and BOT
public-private partnership
Willingness of continuing the contract
Financial-engineering strategy
Marketing strategy
Team integration strategy
Internal and external conditions
Manage and operate strategy
Scale of economy
Bank recognize fund-raising
Cash flow transaction
Low initial funds
Low financial risk
Media relations
Ticket rates strategy
Innovative package
Marketing events
Combining-industry, officials, and university
Science education
Social on-profit activity
Industry news
Leadership
Mutual value
Complementary advantaged resources
Positive communica-tion
Mutual trusts and commitment
Complementary side industries
National park
Planning of parking
Environment
Planning of hinter land
Content diversity
Planning of exhibits
National ranked facility
Equipment and space
Introduce of star creatures
Integrate corporation resources
Define and divide market
Collaborate inter- and intra-industries
Political-business relations
Re-develop new value of exhibit
Low ticket rates
Develop Niche Market

Fig. 1. Frame of success key factors of the BOT of NMMBA

3 Review of the Build-Operate-Transfer of the MMBA

3.1 Brief of Plan Content(Wen-Been Chang, 2002)

3.1.1 The Range of Operation and Duration of Permission

A. Range of Operation as Permission

(a) OT(Operate-Transfer): Whale Plaza, Waters of Taiwan (1st exhibit), Coral Kingdom (2nd exhibit), parking lot and related maintenance.

(b) BOT(Build-Operate-Transfer): Waters of the World (3rd exhibit), parking lot C, tourist center, green lands, other facilities.

B. Duration of Operation as Permission: 25 years from July 4th, 2000 to December 31, 2024. The franchised company may propose by before 3 years the contract is due to extend the contract if evaluated to be qualified.

3.1.2 Key Money

A. During the period of building the Waters of the World (from July, 2000 to December, 2005): 50 million of Taiwanese dollars should be collected each year to decrease manufacturer's pressure of funding.

B. After the NMMBA is wholly completed (from January, 2006 to December, 2024): 8.5% of operation income is collected as the base of key money after the Waters of the World starts running (; such a key money is also called proportioned key money with a bottom line).

3.1.3 Deposit

A. Deposit for fulfilling the contract: 200 million TWD.

(a) Guarantee the fulfilling of the BOT contract.
(b) According to the deposit and other guarantee law No.15, do not exceed 15% of the total of purchase.

B. Deposit for construction: 300 million TWD.

(a) Guarantee the complete of building the Waters of the World.
(b) The basis is the accumulated net profit after tax financially calculated upon building the Waters of the World (July, 2000~December, 2005).

3.1.4 Attempt Financial Calculation Basis

A. Self-liquidation ratio: 100% (all funds are raised by the private investor)
B. Stock structure: dept/interests=60：40

C. Permission duration: 25 years starting with the contract is signed.

D. The government does not invest or subsidize, nor lower the amount of private investment.

3.1.5 Promises

As the contract, the franchisee promises that if the exact construct fees of Waters of the World, parking lots, tourist center are lower than 29,780,000 TWD, the fall short part should be unconditionally invested in other NMMBA constructions. The exact construct fee is listed below in Table 1:

Table 1. Fees of side facilities development of franchisee. (NMMBA offers, this research reorganizes).

Facilities Built with Investment	Construct Fee in Contract	Exact Construct Fee	implementation of the case
Waters of the World	2,840,000,000	2,705,970,000	Completed
Parking Lot	64,072,000	65,497,000	Completed
Tourist Center	73,928,000	187,769,388	Completed
Green Plaza		65,846,000	Completed
Total	29,780,000	3,025,082,388	

3.2 Circumstances of the OT of NMMBA

It has been 7 years since the carrying out of the contract at July, 2000. The number of entrance is approximately 6,000~7,000 people per day. The greatest number was over 40,000 people in one day. The public sector estimated to reach 10 million of entrance in 5 years. But in reality, the exact number of entrance has reached 12.2 million, 22% exceeding the estimation. The NMMBA public sector has assisted the franchisee (Hi-Scene World Enterprise CO. LTD.) to earn a good yield of investment. The goal of "profiting the private company" is to bring out the bigger public benefits, which is the essence of "public build-operate-transfer." However, the public-private collaboration, the mutual trust, and accumulated experiences of interactions between the NMMBA (public sector) and the Hi-Scene Enterprise (private sector) come from repetitive steps of "conflicts, argues, compromises, and revision." Instead of "supervise," the public sector has also shifted its management to "integrate." The national BOT policy thus have its new broad way.

3.2.1 The OT of NMMBA

A. Strategy that separate "Operation" and "Research": The public sector has adopted the energy and capacity of private company to operate and market the exhibition of the aquarium. The professional private sectors are responsible for the breeding, raising, and guiding of marine creatures, in addition to advertisement. The public professionals on the other hand execute education and researches, focusing on "exploit new science education" and "e-learning of digital museum" by developing Chinese marine ecology learning website to enhance Taiwan's international science status. Well suited in each own position, the franchised company is in charge of operation and the public sector is in charge of research.

B. Establishing the Specialties of Collaboration: Only by during the early short period of opening (February, 2000~July,2000), the NMMBA was built and operated publicly. The exhibits were later transferred to the private sectors to operate and develop. In the signed contract between the public NMMBA and the private Hi-Scene World Enterprise, the public sector requests to achieve at least 350 days per year (yearly operate days >95%) of operation to best efficiently respond to government's early investment and construction. In the real situation, the private company based on its own profit of investment, achieves the request and even alternates staff members' shifts, making NMMBA the first educational site that runs 365 days per year in Taiwan. Throughout the 6 years of BOT-ing NMMBA, whether the government sectors or the public may both acknowledge that the NMMBA has exactly create its own special marketing strategies and characteristics. The following is the brief of them through organization and analysis. (Lin, 2005)

(a) Expectation of Weeding through the old to bringing forth the new. There have been 12,200,000 visitors since the opening of NMMBA in February 2000 till now (end of 2005). In the beginning of 3 years, there are more than 2 million of visitors on average each year. In the early 5 years, about 1/4 of Taiwanese population ever visited the NMMBA. The 2-day-weekend and domestic travel promotion have stirred up higher visitation in 3 years. The public sector keeps sending a clear message to visitors that what they see today is not the whole of NMMBA, so they may have expectations for more, as if assembling collections. The visitors would also subconsciously know and pass on to others such a message. Almost half of all visitors are aware of the 3rd exhibit, Waters of the World, will be finished and opened in 2006.

(b) Establishing star creatures

(1) The NMMBA contacted Russia to introduce the Beluga Whale in 2002. The public sector and the mass media have successfully given rise to a fever of Beluga Whale in Taiwan. The fever lasts even till now.
(2) The Whale Shark—the biggest fish of the world—came to NMMBA in 2004 made NMMBA the 3rd aquarium in the world raising such a fish. This movement has caused another wave of visitation nation-wide.
(3) In the summer of 2005, groups of penguins including King Penguins arrived from New Zealand. Visitors increased because all these star creatures are rare in Taiwan, but can be seen altogether within one museum.
(4) The 3rd exhibit, Waters of the World officially operated in April, 2006. The virtual tour of sea life and interactions with digital creatures brought in another wave of visitation.

(c) Exposure on the mass media. Introducing the Beluga Whale from Russia was originally a business contract of NMMBA. When the fever of the 2nd exhibit Coral Kingdom was almost over in 2003, the beluga whale became the new selling point. To promote the exposure on the mass media is part of the strategy. (The mass media introduced the 30 more hours of transferring beluga whales in different airports, changing sea water, checking shipping equipments in customs, and staff arrangement, and so on. The wisdom of the public sector and managers is challenged.) Through media exposure, more people in Taiwan care, understand, and learn to respect and love creatures; making the NMMBA the star on news, and also increasing the earnings. For years, the NMMBA continues on its cooperation with electric media to promote and socially educate so that NMMBA has always been the first choice of travel plan among the Taiwanese.

(d) Day or night makes different tours. The biggest difference of NMMBA and common museums is the part of ecological aquarium that introduces living marine creatures. Since each creature behaves differently by day and by night, the NMMBA organized a 2-day-1-nigh tour. This tour is well planned and designed by the professionals to be expected to achieve the best environmental learing effects.

(e) Allying with the hotel and travel services. In order to prolong the management, the NMMBA does not satisfy with the internal profits only. In the aquarium, the information desk helps visitors to consult or register to hotel services in around Heng-Chen Peninsula. All hotels also include NMMBA within their packages. Adding with discounts and convenient services, the companies and visitors assist, profit, and benefit each other to goal in a win-win situation.

(f) Promoting both the front and the back. The aquarium feeds all kinds of rare creatures. The heart makes NMMBA runs well is the most complicated system in the nation. In the 2nd floor of the basement, there are various huge pipes and running machines. When the visitors finish the tour of front marine creatures, the tour of behind the scenes is also available. The visitors may see the system and facilities that sustains the lives of creatures. This is another selling point different to other museums.

(g) Card networking surpassing the Heng-Chuen Peninsula. The NMMBA is located in remote Heng-Chuen Peninsula, but its market is not limited geographically. For example, there is cooperation with Pan Asia telecom to give whale backpack when purchasing a cell phone, or the All Pass card for cash refunds with Cathay United Bank and Jih Sung International Bank bring recognition to card members and returns of visitors, and package tour with EZ-Travel, and the public travel card cooperation with China Trust. Such cooperation also increases the awareness and visitation of NMMBA.

(h) Educating both students and teachers. The NMMBA assist private invested sectors to develop seed teacher camp, educational worksite, marine elementary school, ocean exams, coral camp, and many other camps or trainings along with varying educational tools and materials. Besides training groups of school teachers who are the experts of marine ecology preservation, the NMMBA also makes all schools pay attention to realize marine ecology education. In few years, all schools in Taiwan have included NMMBA in their field trips. This investment in education is the secret weapon bringing more visitors.

(i) Emphasizing the advantages of specialty. NMMBA is special in knowing everything about marine creatures. NMMBA has the marine ecology information learning site which includes the most complete and abundant marine creature information in the database. The information is also printed and published on paper or digitally, such as biological diagrams, technical books, brochures, and disks. The public sector is also entrusted by the Council of Agriculture, Executive Yuan, to be the only shelter for marine creatures in Taiwan. All wild-life protect laws under every municipalities can send the illegal or mistaken caught rare creatures to the NMMBA for cure and release.

(j) Clear vision, innovative future. NMMBA is the only OT integrating BOT, and the only institution that opens to visitor everyday restlessly. The franchise fee is around 100 million since 2006. The manage team strive on developing bio technology and artificial coral breeding system, and actively train related talented people, opening up a new goal and achievement by public funds.

3.2.2 BOT (Building Waters of the World)

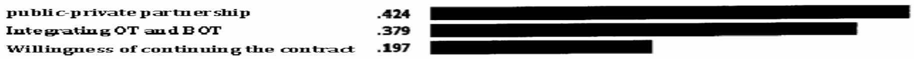
The Waters of the World(The Waters of the World is planned with 4 major themes: Ancient Sea, Kelp Forest, Deep Sea, and Polar Seas; including 4,000 ㎡ of digital exhibition. In addition to demonstrating living creatures, the digital exhibition supplements the ocean where people have never seen, reached, or hardly experienced.) is built with private participation and funds. The franchisee started the construction on February 26, 2004 until the completion on December 16, 2005, and opened to operation on April 28, 2006. The total construction fee is 2,900,000,000 Taiwanese dollars. Under the collaboration of public and private sectors, the Waters of the World is quickly completed in 2 years as the plan, the quality, the budget, and without accidents. Except for satisfying the request on the contract, the NMMBA also opens to a new era.

4 Results and Discussion

This chapter analyze questionnaires by Expert Choice 2000 in 4 parts: the 2nd rank is the analysis of the importance of success factors of the BOT of NMMBA; the 3rd rank is the analysis of impacts of the strategies to success factors of the BOT of NMMBA; the 4th rank is the analysis of the importance of bottom-level and overall effective strategies included in the success factors of the BOT of NMMBA.

4.1 Analysis of Importance of Assemble Factors

In this analysis through AHP, the most important factor is the public-private partnership, with weight as 0.424; followed by the integration of OT and BOT, with weight as 0.379, and finally, the willingness of continuing the contract with weight as 0.197. See Fig. 2.

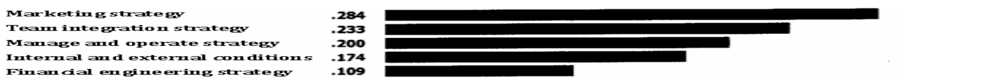


CI = 0.00342

Fig. 2. Analysis of importance of assemble factors

4.2 Analysis of the Impacts of Assemble of Strategies to Success Factors of the BOT of NMMBA

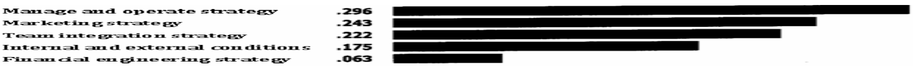
A. Impacts to integrating OT and BOT. In this analysis, the success factor of the BOT of NMMBA is the upper-level criteria, and the lower-level one is the assembly of strategies to success factors of the BOT of NMMBA. The assembly of strategies include: financial engineering strategy, operation strategy, internal and external conditions, team integration strategy, and marketing strategy. The outcomes through AHP analysis show that most remarkable impact is the marketing strategy with 0.284 weight, then the team integration strategy with 0.233 weight, the operation strategy with 0.200 weight, the internal and external conditions with 0.174 weight, and finally the financial engineering strategy with 0.100 weight. See Fig. 3.



CI = 0.04

Fig. 3. Analysis of the impacts to integrating OT and BOT

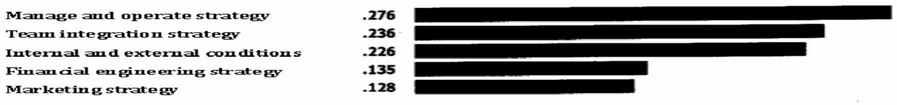
B. Impacts to the public-private partnership. Through AHP analysis, the operation strategy is found most remarkable to impacting the public-private partnership with a 0.296 weight; then the marketing strategy with weight 0.243, the team integration strategy weight 0.222, the internal and external conditions with weight 0.175, and finally the financial engineering strategy with weight 0.063. See Fig. 4.



CI=0.04

Fig. 4. Analysis of impacts to the public-private partnership

C. Impacts to the willingness of continuing the contract. Through AHP analysis, it is known that the most remarkable impact is the operation strategy with a 0.276 weight. Then the team integration strategy with weight 0.236, the internal and external conditions weight 0.226, the financial engineering with weight 0.135, and finally the strategy marketing strategy with weight 0.128. See Fig. 5.

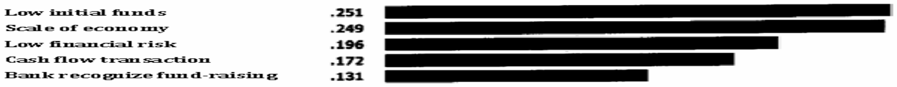


CI=0.03

Fig. 5. Analysis of impacts to the willingness of continuing the contract

4.3 Analysis of the Importance of Bottom-Level and Overall Effective Strategies Included in the Success Factors of the BOT of NMMBA

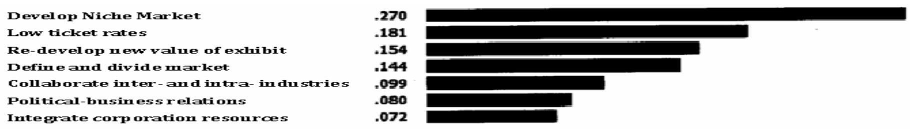
A. Financial engineering strategy. In this analysis, the financial engineering strategy is the upper-level criteria, and be compared with included bottom-level strategies. The bottom-level strategies include scale of economy, recognition of bank fund raising, cash flow transaction, low initial funds, and low financial risk. The results indicate the low initial fund is the most important (0.251), and the least important is the recognition of bank fund raising (0.131). See Fig. 6.



CI=0.04

Fig. 6. Analysis of the importance of the financial engineering strategy and included bottom-level strategy

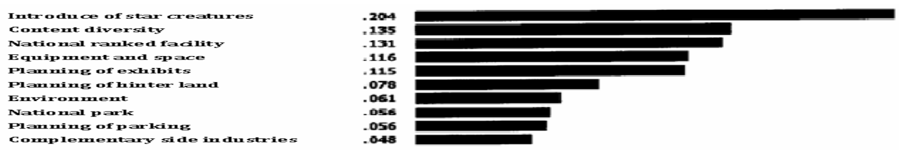
B. Operation strategy. In this analysis, the operation strategy is the upper-level criteria, and be compared with included bottom-level strategies. The bottom-level strategies include, defining and dividing the market, collaborations between inter- and intra-industries, well usage of political and business relations, re-development of new values of exhibition, low ticket rates, and development of Niche Market. The results indicate the development of Niche Market is the most important (0.270), and the least important is the integration of corporation resources (0.072). See Fig. 7.



CI=0.04

Fig. 7. Analysis of the importance of the operation strategy and included bottom-level strategy

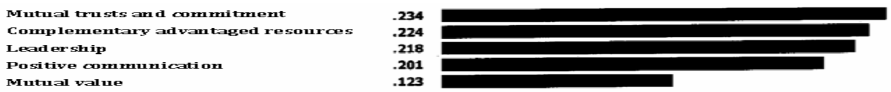
C. Internal and external conditions. In this analysis, the internal and external conditions is the upper-level criteria, and be compared with included bottom-level strategies. The bottom-level strategies include complementary side industries, national park area, planning of parking, environment, planning of hinter land, content diversity, planning of exhibition, nationally ranked facilities, equipment and space, and introduction of star creatures. The results indicate the introduction of star creatures is the most important (0.204), and the least important is the complementary side industries (0.048). See Fig. 8.



CI=0.03

Fig. 8. Analysis of the importance of internal and external conditions and included bottom-level strategy

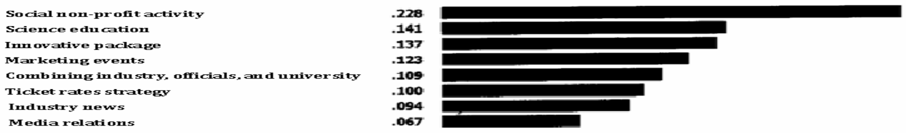
D. Team integration strategy. In this analysis, the team integration strategy is the upper-level criteria, and be compared with included bottom-level strategies. The bottom-level strategies include the leadership, mutual values, complementary advantaged resources, positive communication, mutual trusts and commitments. The results indicate the mutual trusts and commitments is the most important (0.234), and the least important is the mutual values (0.123). See Fig. 9.



CI=0.01

Fig. 9. Analysis of the importance of team integration strategy and included bottom-level strategy

E. Marketing strategy. In this analysis, the marketing strategy is the upper-level criteria, and be compared with included bottom-level strategies. The bottom-level strategies include the media relations, ticket rates strategy, innovative packaging, marketing events, and collaboration of industry-officials-university, science education,



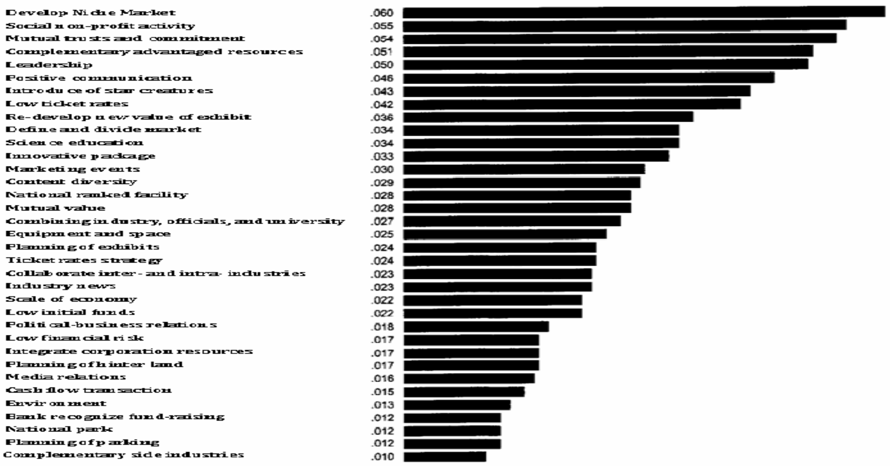
CI = 0.03

Fig. 10. Analysis of the importance of marketing strategy and included bottom-level strategy

social non-profit activities, and industrial news. The results indicate the most important factor is the social non-profit activities (0.228), and the least important is the media relations (0.067). See Fig. 10.

4.4 Overall Importance Comparison of Bottom-Level Strategies

In this nalysis, every priority ratio of bottom-level element is multiplied by the priority ratio of the criteria to receive the overall evaluative weight of factors. Among all factors, the top 10 are: development of Niche Market (0.060), social non-profit activities (0.055), mutual trusts and commitments (0.054), complementary advantaged resources (0.051), leadership (0.050), positive communication (0.046), introducing star creatures (0.043), low ticket rates (0.042), re-development of new value of exhibition (0.036), defining and dividing market (0.034). See Fig. 11.



CI = 0.03

Fig. 11. Analysis of the overall bottom-level strategies

5 Conclusions

This chapter discusses the reasons of the success factors of the BOT of NMMBA in the results and makes conclusions.

5.1 The Collaboration between Public and Private Sectors Is the Most Important Factor

In the analysis of the success factors of BOT NMMBA, we find the most important and highest weight as 0.424 is the public-private partnership. What makes it so? Generally, mutual trust, understanding, and respect come through positive team atmosphere and shared vision. BOT can also be interpreted as "Build on Trust," There is no success without trusts.(LS Sung,2006) The NMMBA realizes the importance of the contract. And only through contract can the mutual trusts be strengthened. The success of BOT contract relies on the public-private partnership to solve tough issues. Otherwise, no matter how excellent the attractions and conditions for investment are, the results may be bad for the government.

5.2 The Operation Strategy Has the Greatest Impacts on the Assembly of Success Factors of BOT, But the Financial Engineering Strategy Is Least Important

In the results according to chapter 4, The operation strategy has affected mostly to factors: integrating OT and BOT, public-private partnership, willingness of continuing contract, followed by the team integration strategy, marketing strategy, internal and external conditions, and finally the financial engineering strategy.

5.3 The Core of Operation Strategy Should Be the Development of Niche Market and Low Ticket Rates

According to the analysis outcomes in chapter 4, the core of operations strategy is the development of Niche Market and low ticket rates, supplemented by defining and dividing the market, collaboration of inter- and intra-industries, application of political and business relations, to upgrade the achievements of operation. However, integrating corporation resources is least effective.

5.4 The Most Important Factor of Team Integration Strategy Is Mutual Trusts and Commitments

According to the analysis outcomes in chapter 4, the most important factor of team integration strategy is mutual trusts and commitments, since the team can only work successfully and smoothly with clear game rules and unconditional dedication of leaders. In addition to countless communications and negotiations to reach consensus and mutual values, team members can offer trust and commitment without any doubts.

5.5 The Most Important Comparison of Internal and External Conditions Is the Introduction of Star Creatures, Followed by the Content Diversity

Within the internal and external conditions of NMMBA, the most important is the introduction of star creatures, followed by the content diversity, which is also an advantage that can be well exercised. However, planning of parking, environment, planning of hinter land, complementary of side industry do not impact much.

5.6 The Most Important Marketing Strategy Is Social Non-Profit Activities, Supplemented by Science Education and Innovative Packaging

Based on the analysis in chapter 4, the social non-profit activity is the most important among the marketing strategies, followed by science education and innovative packaging. However, the industrial news and media relations are less important.

5.7 The Most Important Factor among Financial Engineering Strategies Is the Low Initial Funds, Followed by the Scale of Economy

According to chapter 4, the most important factor in financial engineering strategy is low initial funds, then the scale of economy, low financial risk, and cash flow transactions. The recognition of bank fund raising is less important.

References

1. Caiyun, H., Huang, W.-J.: National public museum the case of Public Private-To four museums for example. Museology Quarterly 17(4), 105–126 (2003)
2. P.-H.: Wu, Public museum services for the privatization of Staff Attitudes Study. Master's thesis. National Sun Yat-sen University (2002)
3. Liu, C., Wang, W.-Y., Yulin, H.: Win-win strategy for BOT. Taipei, Business Ding financial adviser (2000)
4. Daniel, R.D.: Management information crisis. Harvard Business Review 39(5), 111–121 (1961)
5. Ho, Y.-C.: Practical Marketing Management. Hwa-Tai Bookstor. Taipei (1990)
6. Saaty, T.L.: The analytic hierarchy process. McGraw-Hill, New York (1980)
7. Ossadnik, W., Lange, O.: AHP-based evaluation of AHP-Software. European Journal of Operational Research 118, 578–588 (1999)
8. Chang, W.-B.: Gain and loss of BOT model-Case Study of National Museum of Marine Biology. In: Privately run museum in 2002 Policy Theory and Practice Symposium Technology, pp. 81–95 (2002)
9. Chung-Hsiao, L.: Characteristics of Operating the National Museum of Marine Biology and Aquarium. Bi-monthly Journal of Social Education, 45–49 (2005)
10. Sung, L.S.: The study of government encouraging to participate in public construction by OT & BOT mode. Master's thesis. National Taiwan University of Science and Technology (2006)

A DEA Application Model of Production Process for the Chip Resistor Industry

Hsi-Che Teng[1], Shin-Jung Wang[1,*], and Ming-Haur Shih[2]

[1] National Kaohsiung University of Applied Sciences, Taiwan
[2] National Sun Yat-Sen University, Taiwan
cn.wang@cc.kuas.edu.tw

Abstract. Chip resistor is widely applied in various kinds of electronic products nowadays. Almost as long as there are electronic products, chip resistor will be the existence with them. But extensive use of the market created a chip resistor industry expansive development. Because of competitors competing into the ranks of chip resistor manufacturing, leading to oversupply the market and caused strict competition. In order to survive, the same industries sacrifice getting profit and start a price war. And finally show operating losses of the state. To make products reduce costs in a competitive market is to hold competition advantage.

This research as a individual example, based on the data envelopment analysis (Data Envelopment Analysis, referred to DEA), provide a set of systematic model to assess the each process between the input of manpower, working hours, fees and other manufacturing costs. And then to provide the best permutation combination of improvement, the purpose can be reducing manufacturing costs and also upgrade operational efficiency, and improve industry competitiveness. The results demonstrated that the proposed model can be applied to inter-industry analysis of the manufacturing cost and provide effective assessment, improvement and recommendations.

Keywords: Data Envelopment Analysis, Efficiency, Chip Resistor.

1 Introduction

1.1 Industry Introduction

Resistance is divided into fixed, variable type and nonlinear three categories, is one of the three passive components, widely used in electrical, home appliances, information, communication and other kinds of electrical appliances on, its use is mainly to adjust the circuit voltage can also be used to protect components, as snubber or temperature control and sensing purposes.

Chip resistors are fixed resistors, for the past 15 years China began to develop products. As the volume of electronic products meet the characteristics of compact size, SMD type plug-in resistance than the more traditional production advantages, it developed rapidly. Taiwan makers account for about 6 percent of global production, as the world's largest supplier of chip resistors.

[*] Corresponding author.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 457–467, 2010.

Domestic manufacturers in the chip resistor assembly technology costs the country more competitive in Japan, low, but some more for the import of raw materials, including porcelain, ceramic body and ceramic substrates, are all controlled by the hands of foreign companies, domestic firms can provide the substrate in only nine high one, the other Japanese manufacturers are the majority, this year the economy had begun to turn the financial tsunami will be substantial expansion.

Table 1. The main raw material supplier details

	Substrate	Resistive Paste	Package Material
Description	Function material	Screen printing	Paper tape
supply	import	import	import
local Company	Jiu hao	zhong you	Taiwan electron material
Overseas Company	Maruwa、 Nippon Carbide、 Kyocera、Kyoritsu、Meiwa及 Ceramtec	dupont、sumitomo、Ferro、Shoei	Japan printing

1.2 Product Manufacturing Process

Thin film chip resistors are generally divided into thick film chip resistors and chip resistor, the current thin film chip resistors for higher costs than the thick film, the small market at present, but the future of electronic products in response to the trend of slim and light, thin film resistors have a small size, low temperature coefficient high precision, high stability factor advantages, will be the next mainstream product.

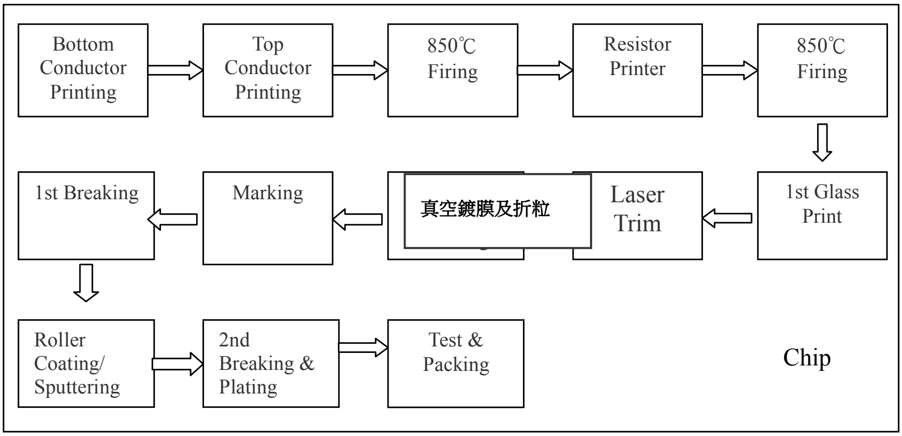


Fig. 1. Thick film chip resistor manufacturing flow chart

Chip resistor product specifications mainly 2512, 2010, 1210, 1206, 0805, 0603, 0402, 0201, at present, the minimum size has grown to 01,005, the substrate size is 60 * 60 * 49.5 size 70 and two substrates. Although large substrates because the available area, yield 30% less substrate.

Chip resistors manufactured by domestic manufacturers, most of the thick film chip resistors is mainly based, also produced by screen printing. The process mainly through the screen printing, the first conductive ink (material for the silver or silver palladium) on alumina ceramic substrates are printed on the back of lead and guide, then print resistance value of resistor ink, and brush on the glass as protection, through the silver coating or vacuum coating printed on the side of the guide way, the final plating method using nickel layer and tin-plated, after testing, packaging paper tape embedded within the storage can be shipped.

2 Concept of the DEA Methodology

DEA was first proposed by Charnes, Cooper and Rhodes (CCR) [19]. Its original idea comes from the measurement model of production efficiency proposed by Farrell [20]. DEA itself is a non-parametric method for assessing the relative efficiency of decision making units (DMUs) based upon multiple inputs and outputs. The primitive DEA model adopts the concept of production in microeconomics: efficiency = output / input. Banke, Charnes, and Cooper (BCC) [21] developed a new model from the CCR model to understand the problems of pure technical efficiency (PTE) and scale efficiency (SE). Both of the CCR and BCC models are summarized as below.

A. CCR Model

The CCR model intends to maximize the ratio of weighted outputs against weighted inputs. It reduces multiple outputs to a single "virtual" output, and multiple inputs to a single "virtual" input for each DMU. CCR is good at analyzing the relative efficiency without setting the weights in prior, which makes the CCR model more objective.

Assume that there are n DMU. Each DMU has m inputs and s outputs. Let x_{ij} represent the ith input and y_{rj} represent the rth output of DMU j, respectively. Let u_r and v_i represent the virtual variables of rth output and ith input, respectively. Let h_j represent the relative efficiency of DMU j. Where ε is a relatively small positive number (normally set at 10^{-6}).

The relative efficiency of each DMU can be calculated by solving the following mathematical programming problems:

$$\text{Max } h_j = \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \tag{1}$$

$$\text{subjected to } \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1 \tag{2}$$

$$u_r \geq \varepsilon > 0 \tag{3}$$

$$v_i \geq \varepsilon > 0 \tag{4}$$

r =1,2,3,….,s ; i =1,2,3,…..,m ; j =1,2,3,…..,n

The CCR input model can suggest improvement directions and the values of both outputs and inputs in order to achieve the desired efficiency value of 1, which can be done by calculating the following equations:

$$x_{ij}^{*} = h_j x_{ij} - s_i^{-*} \quad , \text{ i =1,…,m} \tag{5}$$

$$y_{rj}^{*} = y_{rj} + s_r^{+*} \quad , \text{ r =1,…,s} \tag{6}$$

Where s_i^{-} represents the lacking quantity (slack) of the output and s_r^{+} represents the redundant quantity (surplus) of the input. $x^{*}{}_{ij}$, $y^{*}{}_{rj}$, s_r^{+*} and s_i^{-*} represent the optimal values of x_{ij}, y_{rj}, s_r^{+} and s_i^{-}, respectively. Note that when the value of objective function $h_j = 1$, s_r^{+} =0 and s_i^{-} =0 (for all r and i). That is, DMU j achieves its optimal efficiency. Note that CCR is the most popular DEA model for evaluation of the total operational efficiency.

B. BCC Model

The difference between the CCR and BCC models is the use of returns-to-scale. For each DMU, BCC allows variable returns-to-scale, while CCR is characterized by constant returns-to-scale. The BCC model has an additional convexity condition. The input–oriented BCC model can be written as:

$$\text{Max } h_j = \frac{\sum_{r=1}^{s} u_r y_{rj} - u_0}{\sum_{i=1}^{m} v_i x_{ij}} \tag{7}$$

$$\text{subjected to } \frac{\sum_{r=1}^{s} u_r y_{rj} - u_0}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1 \tag{8}$$

$$u_r \geq \varepsilon > 0 \tag{9}$$

$$v_i \geq \varepsilon > 0 \tag{10}$$

r =1,2,3,…,s ; i =1,2,3,…,m ; j =1,2,3,…,n

u_0 is a real number to indicate the intercept of the production frontier. When u_0>0, the production frontier for this DMU is decreasing returns-to-scale (DRS). When u_0=0, the production frontier for this DMU is constant returns-to-scale (CRS). When u_0<0, the production frontier for this DMU is increasing returns–to-scale (IRS). In addition, the suggested improvement directions of the BCC input model are the same as the CCR input model.

3 Decision Analysis

3.1 Analytical Framework

The performance analysis framework for light Jie company chip resistor manufacturing process for the production of all relevant data and DEA literature, and to be analyzed to provide a basis for managers to improve the decision-making, analytical framework is as follows:

Performance analysis steps:

Step 1: collect process information:

1: The process of collecting data.
2: The process by collecting data to analyze order to find relevant information.
3: Select rating unit (DMU).

Step 2: Select set of input / output variables:

1: Select how to choose a set of input / output variables of the rules.
2: Correlation Analysis of selected suitable input / output variables.
3: According to the input / output variables to describe the understanding of statistical analysis of preliminary data of the picture.

Step 3: A DEA analysis of the ownership process unit (DMU) Performance:

1: CCR-I model first of all make the ranking process.

Step 4: conduct performance RANK DMU classification:

1: RANK RANK mode to conduct performance do sort classification.

Step 5: The overall performance for the center to find the best DMU:

1: find the best performance of DMU

Step 6: The good performance of the DMU is not selected and set to improve the data:

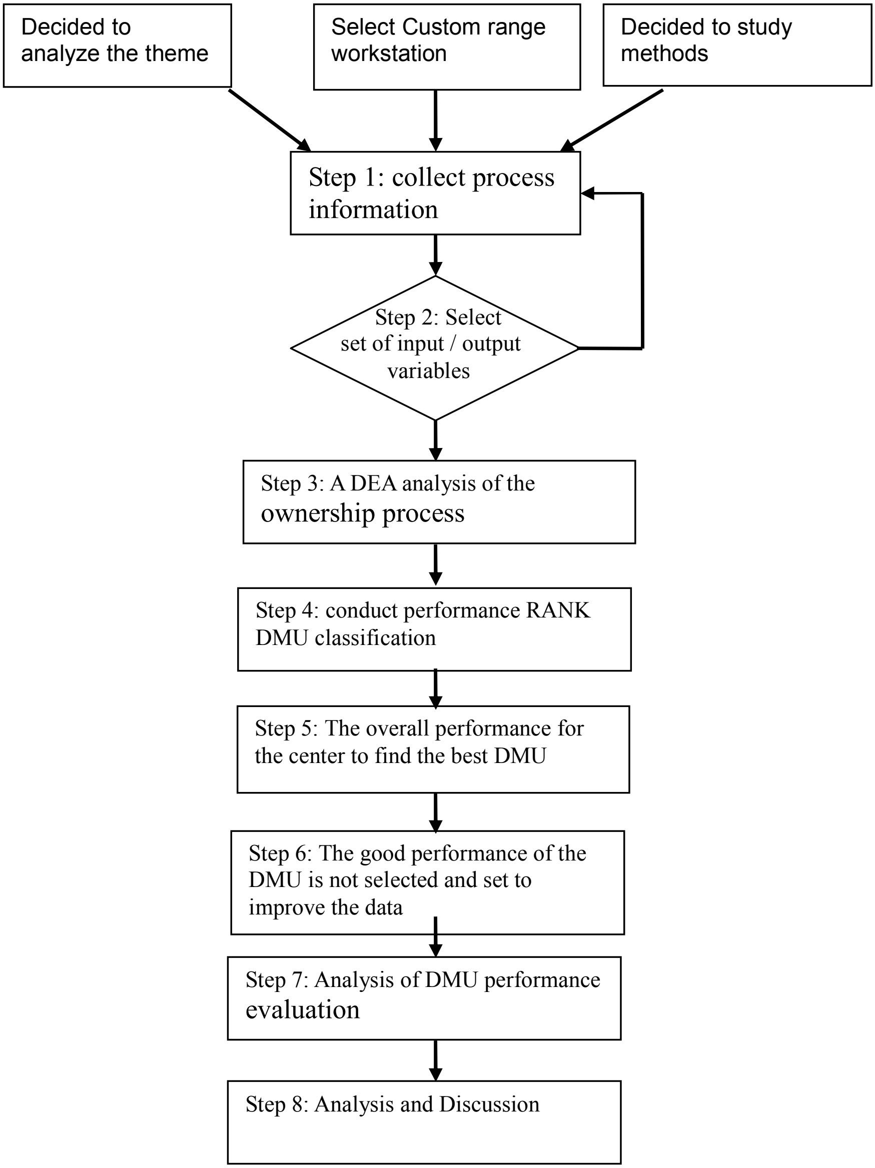

Decided to analyze the theme
Select Custom range workstation
Decided to study methods
Step 1: collect process information
Step 2: Select set of input / output variables
Step 3: A DEA analysis of the ownership process
Step 4: conduct performance RANK DMU classification
Step 5: The overall performance for the center to find the best DMU
Step 6: The good performance of the DMU is not selected and set to improve the data
Step 7: Analysis of DMU performance evaluation
Step 8: Analysis and Discussion

Fig. 2. Performance analysis flowchart

1: Performance is not good the good performance of the DMU DMU and to conduct comparative analysis of input variables

Step 7: Analysis of DMU performance evaluation and improvement of the expected direction:

1: Analysis of the performance expected after the best performance
2: the difference between the variables using analysis of future development to improve the program.

Step 8: Analysis and Discussion:

1: Analysis and discussion of results.
2: suggestions for future research.

3.2 Empirical analysis

3.2.1 The Sample Selection

The performance assessment of the case to the chip resistor factory light Jie Technology Co., Ltd. of the production process as units of analysis, process flow consists of: printing process, conduction process, electroplating process, PQC process, test packet process of five process units (DMU), choose the Performance Evaluation of a critical assessment of factors into inputs and outputs of the relevant data are as follows:

Table 2. The process of basic information

	(I)Person qnty	(I) Overtime	(I) Manufacturing costs	(0) Production quty
PRINTING	26	1828	296882	1008
TERMINATION COATING	16	1763.5	371340	922
PLATING	11	483.5	117860	900
PQC	12	664.5	20456	902
TESTING PACKING	23	1440.5	364341	833

3.2.2 Project Selection and Input-Output Correlation

Variables selected using the principles of DEA to conduct relevant points, so as to verify compliance among the variables are related, that is, the increase of investment shall not be reduced to the reverse output characteristics, can be found among the variables a correlation coefficient values are positively related, consistent with the principles of the use of DEA.

Table 4 for each process variable data for statistical analysis and description of the distribution variables. Variables can be found in the descriptive statistics in the data gaps, this is because the processes are designed to use equipment and materials caused by different factors.

Table 3. Correlation coefficient analysis table

	Person qnty	Overtime	Manufacturing costs	Production quty
Person qnty	1	0.8051	0.7106	0.2997
Overtime	0.8051	1	0.8793	0.3869
Manufacturing costs	0.7106	0.8793	1	0.0184
Production quty	0.2997	0.3869	0.0184	1

Table 4. Variable data analysis table of descriptive statistics

	Person qnty	Overtime	Manufacturing costs	Production quty
Max	26	1828.0	371340	1008
Min	11	483.5	20456	833
Average	17.60	1236.0	234176	913
SD	5.95	559.2	140638	56

3.2.3 Performance Analysis of the Production Process

In this study, DEA-Solver software, using CCR-I model to do assessment work, the first assessment of five production process and the efficiency of using RANK discharged through the entire front of the DMU manager conclusion the following table:

Table 5. The process efficiency rating table

Rank	DMU	Score
1	PLATING	1
1	PQC	1
3	TERMINATION COATING	0.7041
4	PRINTING	0.4738
5	TESTING PACKING	0.4425

Table 6. The difference between the variables of the process

DMU I/O	Score Data	Projection	Difference	%
PRINTING	0.4738			
Person qnty	26	12	-14	-52.62%
Overtime	1828	542	-1286	-70.38%
Manufacturing costs	296882	132001	-164881	-55.54%
Production quty	1008	1008	0	0.00%
TERMINATION COATING	0.7041			
Person qnty	16	11	-5	-29.59%
Overtime	1764	495	-1268	-71.92%
Manufacturing costs	371340	120698	-250641	-67.50%
Production quty	922	922	0	0.00%
PLATING	1.0000			
Person qnty	11	11	0	0.00%
Overtime	484	484	0	0.00%
Manufacturing costs	117860	117860	0	0.00%
Production quty	900	900	0	0.00%
PQC	1			
Person qnty	12	12	0	0.00%
Overtime	665	665	0	0.00%
Manufacturing costs	20456	20456	0	0.00%
Production quty	902	902	0	0.00%
TESTING PACKING	0.4425			
Person qnty	23	10	-13	-55.75%
Overtime	1441	447	-993	-68.95%
Manufacturing costs	364341	109047	-255294	-70.07%
Production quty	833	833	0	0.00%

In the analysis of the relative efficiency of decision-making unit, the inefficient process for the workstation, through the DEA model to understand the difference between the input variables of resource use to identify sources of inefficiency and the corresponding unit of input values should improve the size of the volume level. Analysis of the difference between the variables through the input and output variables in the project view the distribution of resources out whether the phenomenon of excessive or insufficient for the allocation of resources when the variables are not a basis for improvement, this research will analyze the results of process efficiency, to the difference variables analysis, efficiency is not good for manufacturing processes and workstations can increase or decrease the input / output variables convenient to achieve efficiency value of 1, Table 6 for the difference between the variables of the process to provide its participants the overall efficiency improvement test.

4 Conclusion

DEA using the process according to the system input and output data, analyze the performance of each process level, then the difference between the variables based on analysis, can provide a value of 1 less than the efficiency of the DMU, the improvements; Table 6 to break grain processing, for example, the efficiency value of 0.7041, clearly below the efficiency standard value of 1 may be coming to understand the difference between the variables of the efficiency is 1 less than the cause, we can see the need to invest variables to be cut on, to achieve the efficiency value of 1, while the reduction of the Range were: input variables: the number reduced by about 5 to reduce the proportion of 29%, overtime working hours reduced by about 1268 hours, reducing the proportion of 71% manufacturing cost reduction of about 250,641 yuan, 67.5% reduction of the proportion of such information, managers can clearly provide an important basis for decision-making, to achieve maximum effect.

References

[1] Bhattacharyya, A., Lovell, C.A.K., Sahay, P.: The impact of liberalization on the productive difference of Indian commercial banks. European Journal of Operational Research 98, 332–345 (1997)

[2] Camanho, A.S., Dyson, R.G.: Cost efficiency measurement with price uncertainty: a DEA application to bank branch assessments. European Journal of Operational Research 161, 432–446 (2005)

[3] Lee, S., Lee, K., Kang, I.: Efficiency analysis of controls in EDI applications. Information & Management 42(3), 425–439 (2005)

[4] Homburg, C.: Using data envelopment analysis to benchmark activities. International Journal of Production Economics 73(1), 51–58 (2001)

[5] Mota, S., Benzecry, J.H., Qassim, R.Y.: A model for the application of data envelopment analysis (DEA) in activity-based management (ABM). International Journal of Technology Management 17, 861–868 (1999)

[6] Co, H.C., Chew, K.S.: Performance and R&D expenditures in American and Japanese manufacturing firms. International Journal of Production Research 35(12), 3333–3348 (1997)

[7] Chang, D.S., Lo, L.K.: Measuring the relative efficiency of a firm's ability to achieve organizational benefits after ISO certification. Total Quality Management and Business Excellence 16(1), 57–69 (2005)

[8] Luo, X., Donthu, N.: Assessing advertising media spending inefficiencies in generating sales. Journal of Business Research 58, 28–36 (2005)

[9] Cook, W.D., Zhu, J.: Output deterioration with input reduction in data envelopment analysis. IIE Transactions 35(3), 309–320 (2003)

[10] Durand, R., Vargas, V.: Ownership, organization, and private firms' efficient use of resources. Strategic Management Journal 24(7), 667–675 (2003)

[11] Forker, L.B., Mendez, D., Hershauer, J.C.: Total quality management in the supply chain: what is its impact on performance? International Journal of Production Research 35(6), 1681–1702 (1997)

[12] Shaffer, S.: Can Megamergers Improve Bank Efficiency? Journal of Banking and Finance 17, 423–436 (1993)

[13] Worthington, A.C.: Determinants of Merger and Acquisition Activity in Australian Cooperative Deposit-taking Institutions. Journal of Business Research 57(1), 47–57 (2004)

[14] Lubatkin, M., Srinivasan, N.: Merger Strategies and Shareholder Value during Times of Relaxed Antitrust Enforcement: The Case of Large Mergers during the 1980s. Journal of Management 23(1), 59–81 (1997)

[15] Wang, C.-N., Wang, C.-H.: A DEA application model for merger and acquisition in high-tech business. In: Proceedings of International Conference on Engineering Management, pp. 43–47 (2005)

[16] Delmas, M., Toka, Y.: Deregulation, governance structures, and efficiency: the U.S. electric utility sector. Strategic Management Journal 26(5), 441–460 (2005)

[17] Talluri, S., Baker, R.C.: A quantitative framework for designing efficient business process alliances. In: Proceedings of International Conference on Engineering Management, pp. 656–661 (1996)

[18] Shao, B.B.M., Lin, W.T.: Technical efficiency analysis of information technology investments: a two-stage empirical investigation. Information & Management 39(5), 391–401 (2002)

[19] Charnes, A., Cooper, W.W., Rhodes, E.: Measuring the efficiency of decision making units. European Journal of Operational Research 2(6), 429–444 (1978)

[20] Farrell, M.J.: The measurement of productive efficiency. Journal of the Royal Statistical Society, Series A, General 120, 253–281 (1957)

[21] Banker, R.D., Chanes, A., Cooper, W.W.: Some models for estimating technical and scale inefficiencies in data envelopment analysis. Management Science 30, 1078–1092 (1984)

Enhancing Repair Service Quality of Mobile Phones by the TRIZ Method

Yao-Lang Chang, Shu-Chin Lai, and Chia-Nan Wang

Department of Industrial Engineering and Management,
National Kaohsiung University of Applied Sciences,
415 Chien Kung Road, San-Min District,
Kaohsiung 807, Taiwan
kc88899@gmail.com

Abstract. “Fix your Mobile Phone within 30 minutes” is a commitment of mobile phone repair business in Taiwan. Enhancing repair speed and successful repair ratio is the prerequisite of success. However, quantifying repair performance is not easy because customers’ focuses not only on repair speed and repair ratio, but also on the frequency of repeated repair claims. This study applied the Contradiction Matrix and Problem Hierarchy Analysis of the TRIZ (Teoriya Resheniya Izobretatelskikh Zadatch) theory to analyze the issues of mobile phone repair business and propose a strategy for improving repair speed and repair ratio and decrease the frequency of repeated repair claim.

Keywords: TRIZ Theory, Contradiction Matrix, Problem Hierarchy Analysis.

1 Introduction

Owing to popular communication technology, nearly everyone is equipped with a mobile phone, a well-accepted necessity to daily life. Mobile phone popularity also creates a remarkable after service market scale. Versatile brands, multi-functions, and various interface specifications have made mobile phone repair problems increasingly more complicated. Consumer expectation for repair quality and speed has also become higher. Faced with the booming creation of new mobile phone models, repair companies must update service engineer technology soon. The above-mentioned problems highlight repair and administration capability of repair companies.

Two promises provided by Company A; one of leading mobile companies in Taiwan, “Delivery Tonight, Get it the Day After” and “Fix your Mobile Phone within 30 minutes”, make a remarkable impact on the mobile phone service market, forcing other repair companies to pay more attention to maintenance efficiency and accelerate repair speed to provide the same repair performance as Company A. Consumers expect quick and reliable mobile phone repair. If mobile phone problems are not resolved after repeated repairs, consumers will lose confidence in the phone manufacturer and switch to another maker for their next phone purchase. After-service operation is very complicated and involves several aspects, such as original maker regulations, agent cooperation, parts preparation, and repeated maintenance claims etc. Some cases, including requests from remote out of logistic districts, repair claim delay from agents, non-standard procedure

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 468–481, 2010.

claims, customer confirmation is needed if repair fee higher than the high limit of list price, and non-authorized service items as well as pc board damage needing to be forwarded to the original maker, cause maintenance delay. The efficiency of maintenance cases needing to be forwarded to the original maker rely on original maker efficiency. The shortage of parts or part production termination and long testing periods for assuring repair performance definitely decrease efficiency. Dealing with the above obstacles and organizing the standard operation procedure to enhance maintenance efficiency and countermeasures to offer more convenient and faster service are essential to maintaining operation management.

A detailed and systematic analysis of repeated claims for exploring countermeasures also lowers repeated claims. Guaranteed maintenance quality makes maintenance speed and the repair ratio meaningful.

This project focuses on the maintenance procedure, service items, warranted items, and maintenance efficiency of service points of the top four service groups of mobile phones in Taiwan, including companies A, B, C and D. This project applied the Contradiction Matrix and Problem Hierarchy Analysis of the TRIZ theory to analyze the causes and problems of mobile phone repair claims. This work also proposes an improvement strategy for enhancing maintenance speed and repair ratio to decrease repeat repair claims for increasing maintenance quality.

Woo and Fock (1999) suggest that network providers should focus more on transmission quality and network coverage as the core attributes of their service offerings and formulate appropriate pricing policy, rather than competing on customer services and other supplementary services. With the rapid growth in the size of the cellular phone market, performance standards are emerging as the strategic imperative for the service providers in ensuring customer satisfaction. Sharma and Ojha (2004) develop and validate a psychometrically sound multi-item measure of service performance in mobile communications.

The concept is also derived from a Russia inventor, as early as the 1940's; G. Altshuller refused to accept an unreliable unrepeatable psychological personality-dependent approach to creativity. He instead on choosing another way, which is based on an analysis of the results of creativity in technology, that is, inventions. This approach allowed Altshuller to make his conclusion on the basis of information documented the human innovative experience. It is so-called "TRIZ" (Theory of the Solution of Inventive Problems) (Zlotin & Zusman, 1999).

The innovation of interpretation of history and method of TRIZ (The theory of solving inventor's problems) was first proposed by Genrich Altshuller. In the paper, "Suddenly the Inventor Appeared TRIZ, the Theory of Inventive Problem Solving", Genrich Altshuller suggests a very unique point of view to materials and substance-field analysis (Mann, 2002; Fey & Rivin, 2005).

Shulyak, the first scholar to teach children to use TRIZ, suggested three steps for solving a technical contradiction by using a lot of figures and documentations (Altshuller, 1998; Altshuller, 2000). Rantanen proposed the analysis of TRIZ common tools and obtained the ideal final result (IFR) and pointed out nine steps to the Inventive Problem solving (Altshuller, 1999; Mann, 2007; Fey & Rivin, 2005). A detailed explanation to 40 principles of inventions, problem definition and discussion, innovation trend, resource and substance-field analysis was proposed by Mann (Livotov, 2004; Savransky, 2000).

TRIZ is a methodology for solving problems that relies on logic and mass information rather than on hunch, thus accelerating teamwork to innovatively solve problems (Altshuller,1996; Barkan, 2000). The article, "40 Inventive Business Principles with Examples" (Mann & Domb, 1999), presents a good paragon for business management using inventive TRIZ principles. For example, the "Segmentation Principle", is associated with autonomous profit centers or product centers; the "Local Quality Principle" is associated with empowerment, change in salary structure or flexible working hours; the "Asymmetry Principle", is associated with 360 appraisals or avoiding Peter Pyramid; the "Merging Principle", is associated with JIT; the "Universality Principle", is associated with setting up mutual performance to eliminate the need for other parts; the "Equipotentiality Principle" is associated with making horizontal career changes to broaden skills; "The Other Way Round Principle", is associated with home shopping or the park-and-ride scheme in busy cities; the "Spheroidality-Curvature Principle", is associated with mobile commerce; the "Feedback principle", is associated with electronic bulletin boards or CRM; the "Copying Principle", is associated with virtual product service manuals; the "Homogeneity Principle", is associated with co-located project teams or product branding/product families, et cetera.

Judging from the above, this study infers generalizing business behaviors by forty TRIZ principles. By developing and redefining business parameters in TRIZ, this work links a relationship between business strategies and forty Inventive Business Principles to behaviors in companies and industries. Companies and their clients want innovation and impressive business solutions to overcome problems in a severe business environment (Ishida, 2003). This study set up a Matrix of Business Strategy by creating thirty-nine parameters and offering some concrete practice phenomena in the void left by the case study.

This study uses TRIZ as a methodology to establish an originality-evaluated business creativity system, including a knowledge database and a comparable mechanism. The methodology helps the company practice innovation by evaluating the feasibility and value of creativity.

2 The TRIZ Methodology

TRIZ is a romanized acronym for the Russian "Теория решения изобретательских задач" (Teoriya Resheniya Izobretatelskikh Zadatch), meaning "the theory of solving inventor's problems" or "The theory of inventor's problem solving". The Soviet engineer and researcher Genrich Altshuller and his colleagues developed this methodology starting in 1946. Altshuller concluded forty principles of inventions and developed the Contradiction Matrix for solving Technical Contradiction. Contrasted to techniques such as brainstorming (based on random idea generation), TRIZ creates an algorithmic approach to inventing new systems, and refining old systems.

It firstly discovers factors needing improvement (problem), and then characterizes the problem by parameters based on the TRIZ method. Finally, it applies the standard solution to obtain the innovation principle or solving direction to work out the problem rather than using a compromised solution. TRIZ applies four steps to proceed systematic innovation, ie. (1) Define problem → Select tools → Generate Solutions → Evaluate, as shown in Figure 1.

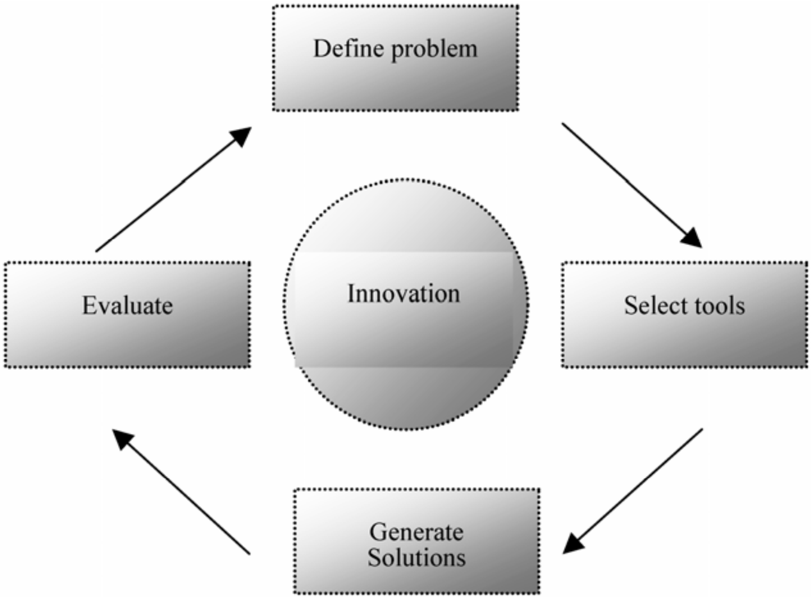


Fig. 1. Steps of systematic innovation

Problem Hierarchy Analysis

Problem Hierarchy Analysis is one of the TRIZ tools, a normal tool used for contraction management and consensual collection. The creative thinking obtained using Problem Hierarchy Analysis pushes all the attendees to collect the best solution. The Problem Hierarchy Analysis result is similar to that of the limited theory, but

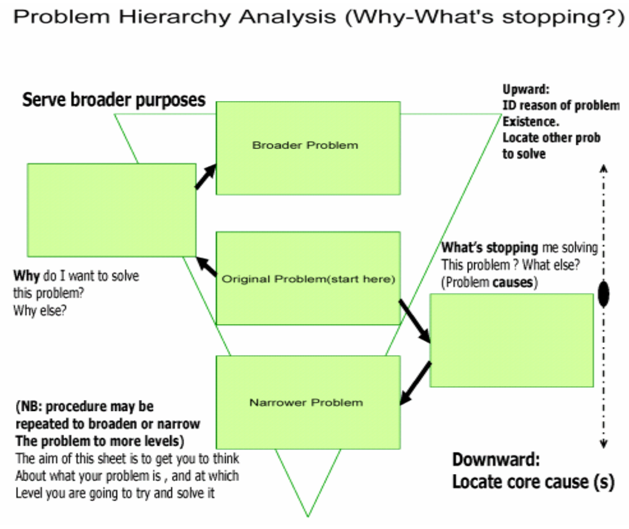

Problem Hierarchy Analysis (Why-What's stopping?)
Serve broader purposes
Broader Problem
Upward:
ID reason of problem
Existence.
Locate other prob
to solve
Why do I want to solve
this problem?
Why else?
Original Problem(start here)
What's stopping me solving
This problem ? What else?
(Problem causes)
Narrower Problem
(NB: procedure may be
repeated to broaden or narrow
The problem to more levels)
The aim of this sheet is to get you to think
About what your problem is , and at which
Level you are going to try and solve it
Downward:
Locate core cause (s)

Fig. 2. Problem hierarchy analysis tool

obtaining the consensual collection is easier and the execution barrier is lower, making it easier to execute innovation (Mann, 2007). Figure 2 shows the main hierarchy.

Contradiction Matrix
Contradiction consists of two categories, including (1) technology contradiction and (2) physics contradiction. Two parameters in this system that contradict represent technology contradiction. Scholars often use the contradiction matrix to solve the technology contradiction. The Contradiction matrix is a two dimensional matrix, where the row represents the characteristic parameters needing improvement, and the column represents the deteriorated characteristic parameters of the system. The corresponding column of two contradicted parameters is the innovation principle proposed by TRIZ.

The current investigation used solution mapping of five stages to analyze the situation, contradiction matrix and innovation principle in this study.

Step 1: Find the parameters needing improvement.

Step 2: Finding out the correspondent improving factors of parameters in the contradiction matrix.

Step 3: List directions for all possible solution.

Step 4: Then, confirm whether the solution direction is associated with contradiction or not.

Step 5: Compare the parameters to find out possible innovation principles.

3 Case Example

The current work analyzed the top four after service groups of mobile phones in Taiwan, companies A, B, C and D. This project applies the Contradiction Matrix and the Problem Hierarchy Analysis of TRIZ theory to analyze the causes and problems of mobile phone repair claims, and proposes an improvement strategy for enhancing maintenance speed and repair ratio and decreasing repeat repair claims.

Major Service Centers Introduction
(1) Company A
Company A, a thirty-year-experienced group of logistics channel and supply chain, offers thirty direct repair centers, 150 maintenance agents, and 9,000 sales agents. The company integrates sales, logistics channels, and repair service. She became a public utility logistics company in 1995 and offers several innovative services such as a ”two year warranty for mobile phones,“ and ” Repair Network” etc. She has built five logistics centers, including centers in Linkou Taiwan, Taichung Taiwan, Melbourne Australia, Bangkok Thailand and Shanghai China. The company provides highly efficient performance, such as placing an order within six minutes, completing an order within fifteen minutes, assembling a CTO computer within six minutes, on-site repair for mobile phones at the repair spot within thirty minutes, and two-day repairs for mobile phones, etc. In 2007, the company processed more than 0.9 million operation orders, delivered 5.6 millions of packed boxes and performed 1.4 million repairs.

(2) Company B
The service centers of Company B include; (1) 200 operation centers associated with a telecom operation counter to establish island-wide distributed agent logistics in Taiwan for providing convenient and instant service, (2) eighty-two direct store shops, and (3) twenty-three repair service centers in Taiwan. The service concepts upheld by the company are “We are everywhere, nothing impossible, in search of excellence”.

(3) Company C.
The company has (1) twelve repair service centers, focusing on professional repair, detailed explanation before sales, instant and firm after-service, original certificated repair technology, self-established professional repair centers, service warranty for reassurance and convenience, and (2) 300 chained stores providing customers the convenience of “a purchase service everywhere”. She has been an experienced mobile communication service for many years. Compared to other logistics channels, she exhibits more professional and plentiful experience and logistics superiority, not only gaining customers’ satisfaction but also helping customers obtain maximum extra added value of mobile communication products. ” We Care, Customers Trust” is not only a slogan but also a belief that pushes her to purchase forever operation devoted to best customer service.

(4) Company C
The company offers six repair centers, 194 communication agents, and six logistics centers. Retailing and acting as an agent for various hot mobile phones, and offers communication and service 24 hours, 365 days a year. The stores of the company directly provide all these services since March 2005. Company C merging with a big communication company is a milestone for integrating a logistics channel and system house..

4 Technology Contradiction Analysis

Contradiction is the core of the TRIZ method. When the contradiction presented in this system causes a harmful or imperfect function, the method requests improving such a system. The work lists the problems encountered in mobile phone repair in the existing systems. By exploring the improved items and the designed keys associated with five-stage solution mapping, the current work analyzes the relationship between the present situation, the contradiction matrix and innovation principles.

The work draws a summary according to repair operation procedures and repair features of companies A, B, C and D which account for the majority of Taiwan’s after service market. Besides, the authors’ actual repair claim experience to these companies inspired the authors about repair service operation, including acceptance and forward procedures. This study summarizes the problems that might be encountered during mobile phone repair as follows:

(1) Incoming unit for repair cannot be registered in time, causing repair delay.
(2) Spare parts shortage.
(3) Repair items not authorized by original maker, taking longer repair time.

(4) Insufficient experiences and technology.
(5) High percentage of repeat repairs.

The improvement items expected are as follows:

(1) Strengthen the collaboration of the repair organization to reduce the wasted time.
(2) Supply the spare parts and tools for repair spots more frequently and efficiently.
(3) Efficient tracking mechanism for repair claims to the original maker.
(4) (4)Establishing a training and knowledge database for experience sharing and accumulation.
(5) Clear description and confirmation of damage cause and providing accessories for testing simultaneously.

The key points of this design focus on the following conditions (1) whether the repair spot delivers within a specified time period, (2) whether the number of spare repairs is sufficient or not, (3) whether the tracking control of repaired mobile phones is sent back to the original makers, and (4) whether service offers the user a tentative phone. Since repeated repair claims take a long time to clarify whether the problem involves accessories or is an accidental case, the repair center should consider a tentative phone service. Training maintenance engineers in repair technology requires a training program and assessment system, a well-documented repair brochure, and a knowledge-sharing database. Table 1 lists the five-stage solution mapping, based on the previous design. Table 2 shows the details.

Table 1. Analysis table for five-stage solution mapping

Problems / stage	No delay to repair claim	Shortage of lack of spare parts	Time consuming for sending back to original maker	Insufficient Repair experience	Repeated claims
Stage1	Time wasted	No. of quantity of materials	Time wasted	Information loss	Problems confirm
Stage2	25	26	26	24	29
Stage3	Increase the degree of adaptability	Ease to repair	Efficient tracking	Capability to judge problems	Problems confirm
Stage4	9	34	9	28	25
Stage5	X	2,32,10,35	X	X	32,26,28,18

Improving one of the parameters or characteristics, deteriorates another parameter or characteristic, then causing contradiction. Once contradiction occurs, a compromised countermeasure should be considered. Since no innovation principle offers an efficient solution, finding the root cause of contradiction reduces the number of innovation principles. The regression method finds common innovation principles. Comparing the relationship of all the above parameters, the high feasibility of innovation principles are:

Table 2. Results of technology contradiction analysis

Method	Content	Improvement
Technology contradiction	39 parameters 40 invention principles Contradiction matrix table	*Decrease the time-consuming once the retailing agent deliver the repair claim within specified period *Accelerate repair speed if parts are sufficient *Decrease the frequency of repeated repair claim if the problems can be confirmed, especially for occasional malfunction and accessory-related problems *Establishment of database, publishing of repair handbook, set up of inheritance and learning platform can enhance the repair speed

(1) Mechanics substitution (28)

Practice: (A) use alternative to mechanical system,

(B) use electromagnetic field in some suitable cases.

(2) Color changes (32)

Practice: (A) change the color of the body or its surroundings,

(B) use color or brightening agent to complete this case

(3) Copying (26)

Practice: (A) use copies to replace the actual one,

(B) use mechanical or photoelectrical skill to complete enlarging | or shrinking a copy.

(4) Parameter changes (35)

Practice: (A) change the physical or chemical status of materials,

(B) use tools to change the characteristics of materials

(5) Taking out (2)

Practice: (A) extract (remove or separate) the expected part or property from the body,

(B) Extract (remove or separate) the unwanted part or property from the body

(6) Preliminary action (10)

Practice: (A) pre-finish all or part of movement,

(B) Pre-set body which is beneficial to the whole movement.

Based on the above discussion, the analysis results of technology contradiction are as follows,

(1) Non-instant repair claim, resulting in time delay, is obtained from the experience of retail stores and agents.
(2) Shortage of spare parts, can be realized from the fact that shortage of parts or spare parts often occurs in the mobile phone brand-new model or out of date model.
(3) Re-transfer to original maker, causes time-delay, can be anticipated from real cases with repeat claims resulting from problem confirmation for avoiding accidental malfunction or the problem related to accessories.
(4) Insufficient experience and technology, can be known from the experience that a higher efficiency and lower repair cost can be obtained in a repair center which has good repair experience, and a well-established database.
(5) Repeated repair claims, can be known from practical experience and internal repair statistic data that repeated repair claims can be lowered by information anarchy, and systematic division of work, which increase repair speed and end repair quality.

Using the TRIZ method, this study works out the above-mentioned items for improving repair speed and the repair ratio, based on innovation principles obtained from technology contradiction analysis and transfer procedure from characterizing problems to a standard problem, and a contradiction matrix.

These concepts, transferred using the thirty-nine characteristic parameters, are still not as specific as those standard solutions provided by innovation principles. Since no specific solution direction can be followed, this study applied the contradiction matrix and problem hierarchy analysis for problem analysis.

5 Problem Hierarchy Analysis of TRIZ Method

The key points for repairing mobile phones are speed and quality. During the repair procedure, repair management problems must be solved. The contradiction matrix and problem hierarchy analysis are useful tools to explore and solve these kinds of problems through problem analysis.

To enhance repair speed and repair ratio, the current investigation analyzed the repair procedure and relationship between agents and repair centers using problem hierarchy analysis (the TRIZ method), shown in Figure 3. Through this analysis, the upstream and downstream hierarchy are described as follows:

The downstream repair center has 4 important items:
(1) consolidate internal management of the repair procedure to establish a standardized repair procedure, (2) provide extra training for new engineers and settle evaluation standards, (3) control spare parts tightly to obtain parts for the new model, and get parts from spare phones, (4) track the retransfer process to the original maker and assign people to track and report progress who instantly report encountered problems. Achieving the above targets could start by consolidating the training of repair engineers, followed by establishing a database for experience accumulation and the evaluation system. Following such an approach the repair company achieves, (1) standardization of the repair

claim procedure, (2) control of repair timing and speed, (3) increase of repair quality to decrease the frequency of repeated claims, (4) establishment of an evaluation system, and (5) setup of information database for experience accumulation.

For an upstream dealer center. Three items are important:
(1) enhance cooperation; (2) deal with the repair claim within the specified time period; (3) keep in touch with the repair centers to get new information.

To maintain excellent repair ratio of every repair center, 90% of repeated claims should be finished within one month. To reach this target, the company should request monitoring people to check repair speed and the repair ratio randomly. This approach results in innovative thinking by the retail agents and repair centers. The top four repair centers for mobile phones in Taiwan approve this approach. Company A case in the next section shows the success of performance evaluation of this approach. Senior repair engineers of mobile phones compile their repair experience into a technical manual and hold a discussion workshop and conduct a technology training course. From the collection of specific problems, self-provided solutions to such problems can be drawn. Repair engineers recognize the solutions obtained through repair engineers with the internal manager in the workshop.

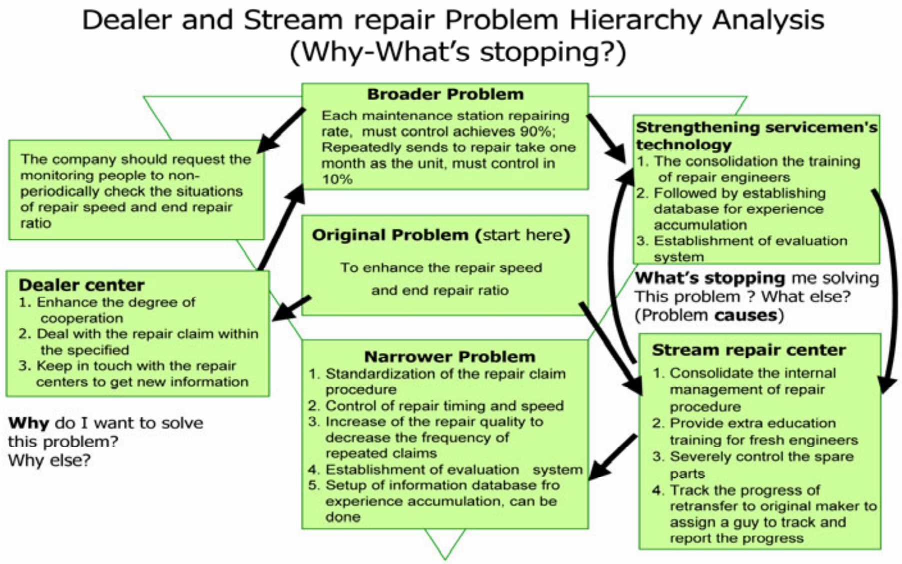

Dealer and Stream repair Problem Hierarchy Analysis
(Why-What's stopping?)
Broader Problem
Each maintenance station repairing rate, must control achieves 90%; Repeatedly sends to repair take one month as the unit, must control in 10%
The company should request the monitoring people to non-periodically check the situations of repair speed and end repair ratio
Strengthening servicemen's technology
1. The consolidation the training of repair engineers
2. Followed by establishing database for experience accumulation
3. Establishment of evaluation system
Original Problem (start here)
To enhance the repair speed and end repair ratio
Dealer center
1. Enhance the degree of cooperation
2. Deal with the repair claim within the specified
3. Keep in touch with the repair centers to get new information
What's stopping me solving This problem ? What else? (Problem causes)
Why do I want to solve this problem? Why else?
Narrower Problem
1. Standardization of the repair claim procedure
2. Control of repair timing and speed
3. Increase of the repair quality to decrease the frequency of repeated claims
4. Establishment of evaluation system
5. Setup of information database fro experience accumulation, can be done
Stream repair center
1. Consolidate the internal management of repair procedure
2. Provide extra education training for fresh engineers
3. Severely control the spare parts
4. Track the progress of retransfer to original maker to assign a guy to track and report the progress

Fig. 3. Repair procedure and relationship between agents and repair centers

6 Performance Evaluation

Company A repair service demonstrates that the approach proposed by the authors is the same as her real practice. This example also proves the innovative research direction of repair operation matching the actual repair situation.

(1) Improvement of repair operation procedure
The repair operation procedure starts from customer delivery of defective mobile phones to the repair of the Company Repair Network. The repair spot requests the customer to explain the defect details. This repair claim is then sent to the repair management system.

(2) Instant Repair Efficiency
Based on the repair conditions, the claim location and the related repair conditions, repair efficiency differs, as follows.

(a) Onsite repair of mobile phone: some mobile phones can be instantly repaired onsite at the direct repair center of Company A. Onsite repair efficiency depends on the defective situation and the service request.
(b) Onsite parts replacement of communication products: some unique communication products or consumables can be instantly repaired onsite in the direct repair center of Company A. Onsite repair efficiency also depends on the defective situation and the service request.

(3) Instant repair service of Company A

Based on the statistics gathered by Company A, nearly 90% of instant repair service claims can be finished within two days, fulfilling the standard of her instant repair. Such high service efficiency is attributed to the highly efficient logistics channel, consisting of three repair headquarters in Taipei, Taichung, and Kaohsiung, Taiwan. Once the damaged equipment is delivered to the direct repair center, it deserves instant repair service. Even though it does not belong to the on-site repair scope.

(4) Cases of non-instant repair efficiency
There are 4 cases and listed below.

(a) Service in a remote area, outside the logistics service area.

(b) Delay associated with the repair not claimed by the acceptance agent to the repair center or claimed by a nonstandard procedure.
(c) The repair fee exceeds the standard price table, needing consumer confirmation.
(d) The damaged item is not authorized by the original maker, and needs to be sent to the original maker. The repair of some pc board damages is transferred to the original maker when the consumer requests an original maker service. Repair efficiency depends on the original maker.
(e) Shortage of spare parts.
(f) The repair needs a longer test time to ensure repair quality.

(5) Standard Repair Fee

The customer is only charged for cases of sabotage, parts consumption, and out of warranty repair according to the damage case. Within warranty, all functional failures are free of charge. The following lists the charged items as follows:

(a) Collection service charge: when the repair shop collects other brand products, the collection service charge is 100 NTD (New Taiwan Dollars, tax included). Such a claim collected by the direct repair center of Company A is free of charge.

(b) Materials cost: if the claim matches the warranty scope, it is free. If it is out of the warranty scope or is not a natural damage and out of the warranty scope, Company A applies a service charge according to the repair conditions.

(c) Examination cost; claim for out of warranty products: when the consumer abandons the repair or the repair cannot be completed due to parts shortage, Lemel charges 300 NT dollars for the examination. The examination charge for personal computer notebooks, and monitors might be higher. The actual charge follows the Company A's charge regulation or the original maker.

(d) Administration charge: non-Company A products delivered by the original maker will be rejected and returned to the delivery unit and the customer will be charged a 300 NT administration fee.

(6) Table 3, 4, and figure 4 show the communication growth rate result for Company A.

(7) Miscellaneous charge

In addition to the repair service charge, Company A will not regulate the service charge for the extra service provided by the retailing agent to the customers, and agreed to by the agent and consumer.

Table 3. 2008 Annual Growth Rate Unit: Hundred millions of New Taiwan Dollars

Communication	2007y	2008y	Year growth rate
	166	221	+33%

Table 4. 2009 January to May Growth Rate Unit : Hundred millions of New Taiwan Dollars

Communication	2008Year January to May	2009 Year January to May	Year growth rate
	83	119	+44%

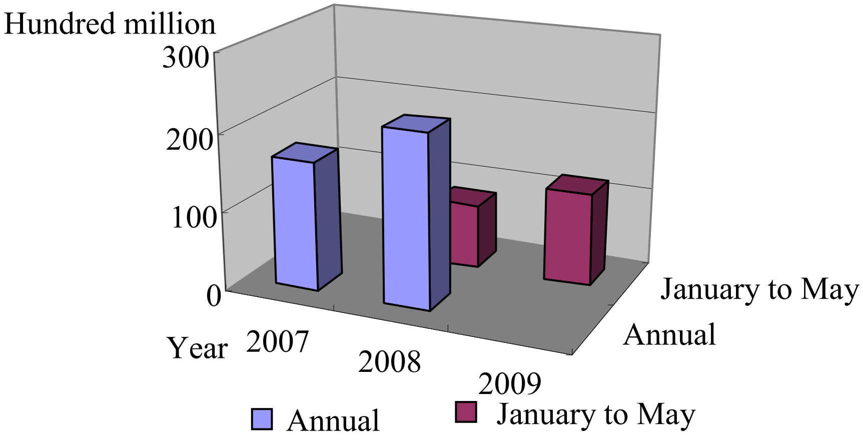


Fig. 4. Growth Rate

According to the repair operation procedure, Company A's present performance is: (a) CTO computer assembly within six minutes, (b) on-site repair of mobile phone within thirty minutes in the repair center, and (c) the mobile phone repair within two days in the repair center. More than 0.9 million orders, 5.6 million delivery cases, and 1.4 million repair cases have been completed by her, proving that the principles proposed by this study effectively enhance repair speed and fixing ratio of the mobile phone. Company A's performance data reveals that controlling repair numbers and repair timing of mobile phones improves repair quality and decreases the number of repeated repair claims.

7 Conclusion

Mobile phone repair operation problems cover a wide range of issues. Some problems are accidental; some relate to users or intrinsic characteristics of mobile phones. Some are associated with repair technology and delay due to the repair delivery procedure and management, such as original maker regulations, the extent of cooperation with the retailing agent, spare parts preparation, repeated repair claims, etc. This study enhances repair speed and the repair ratio, in the mean time maintains repair quality. The current study uses TRIZ tools, such as the Contradiction Matrix and Problem Hierarchy Analysis to identify problems and proposes an improvement method for effectively solving problems.

This paper demonstrated the procedure for using TRIZ methods including thirty-nine Parameters, forty Inventive principles, the Contradiction matrix and Problem Hierarchy Analysis to find out problems and proposes improvement measures for avoiding recurrence of problems. The main contributions are:

(1) Solving the problems for repeated mobile phone repair claims and reduce the problems of repair speed and quality.
(2) Reaching systematic improvement, and establishing a standardized repair operation procedure, including preliminary examination, and evaluation system performance for repair engineers, thus reaching comprehensive improvement and upgrading.
(3) Effectively controlling repair performance and reducing internal repair cost.
(4) Fast learning and development under a systematic and well- structured mode, for saving training fees, shortening the learning curve, and avoiding the occurrence of severe repair mistakes.
(5) Establishing a knowledge database for sharing experiences and effective learning.

Based on the above, the current work suggests directions for further study：

(1) Collecting and reorganizing repair engineer experiences and the solutions to prepare a repair manual. Establish a more comprehensive information inquiring system for repair solutions.
(2) Effectively transforming the repair engineers' experience into a knowledge database through sharing, inheritance, and learning experiences to effectively control internal repair performance and reduce repair expenditure. Significantly reduce training cost and the learning curve for new engineers through a systematic and

well-structured system. Avoid severe repair mistakes of new repair engineers for repair cost-saving.

(3) Finding the proposed improvement parameters by combining TRIZ methods and tools, to establish standard repair operation procedures for mobile phones and consolidating repair quality confirmation. Collecting repair problems and solutions and editing them as a handbook, to intensify repair technology and execute training. Besides, holding regular evaluation meetings and creating a performance index are vital.

References

Abate, F.R.: The Oxford Dictionary and Thesaurus,(Amercian ed.) Oxford University (1997)

Altshuller, G.: 40 Principles: TRIZ Keys to Innovation. Technical Innovation Center, Inc., USA (1996)

Altshuller, G.: And Suddenly the Inventor Appeared: TRIZ. The Theory of Inventive Problem Solving, Technical Innovation Center, USA (1996)

Altshuller, G.: Innovation Algorithm: TRIZ, systematic Innovation and technical creativity Technical Innovation Center, Inc. Ed., (2000)

Altshuller, G.: The innovation algorithm: TRIZ, systematic innovation and technical creativity. Technical Innovation Center, Inc., USA (1999)

Aroca Co. Inc. website, `http://www.arcoa.com.tw`

Aurora Co. website, `http://www.auroracomm.com.tw`

Barkan, M. G.: Situation analysis - a must first step in a problem solving process, (2000), website `http://www.triz-journal.com/archives/2000/04/d/index.htm`

Fey, V., Rivin, E.: Innovation on Demand: New Product Development Using TRIZ. Cambridge University Press, Cambridge (2005)

Ishida, A.: Using TRIZ to Create Innovation Business Model and Products. The TRIZ Journal (December 2003)

Livotov, P.: The undervalued innovation potential. TRIZ-Journal (April 2004)

Mann, D.L.: Hands-on Systematic Innovation. CREAX Press, Belgium (2007)

Mann, D., Domb, E.: 40 Invention Business Principles with Examples. The TRIZ Journal (September 1999)

Savransky, S.D.: Engineering of Creativity: Introduction to TRIZ Methodology of Inventive Problem Solving. CRC Press, Boca Raton (2000)

Senao Co. website, `http://www.senao.com.tw`

Sharma, N., Ojha, S.: Measuring service performance in mobile communications. The Service Industries Journal 24(6), 109–128 (2004)

Synnex Co. website, `http://www.synnex.com.tw`

Woo, K.S., Fock, H.K.Y.: Customer satisfaction in the Hong Kong mobile phone industry. The Service Industries Journal 19(3), 162–174 (1999)

Adaptive Learning Approach of Fuzzy Logic Controller with Evolution for Pursuit–Evasion Games

Hung-Chien Chung and Jing-Sin Liu

Institute of Information Science, Academia Sinica
Nangang, Taipei, Taiwan 11529
onitsukas@iis.sinica.edu.tw, liu@iis.sinica.edu.tw

Abstract. This paper studies a simplified pursuit-evasion problem. We assume that the evader moves with constant speed along a trajectory that is well-defined and known a priori. The objective of steering control of the pursuer modeled as a nonholonomic unicycle-type mobile robot is to intercept the moving evader. An adaptive learning approach of fuzzy logic controller is developed as an inverse kinematics solver of unicycle to enable a mobile robot to use the evader trajectory to adapt its control actions to pursuit-evasion game. In this proposed approach, GA evolves the parameter values of the fuzzy logic control system aiming to approximate the inverse kinematics of pursuer so as to generate a trajectory capturing the evader. Simulation results of pursuit-evasion game illustrate the performance of the proposed approach.

Keywords: Fuzzy logic control, Pursuit-evasion, Inverse kinematics, Genetic algorithm.

1 Introduction

Pursuit-evasion games have become the increasingly important issues in robotics security and surveillance recently. There are many types of pursuit-evasion games, such as rule-based fuzzy system for pursuit-evasion games [10], pursuit-evasion games with searching the nodes of a graph [11], visibility based pursuit-evasion games [13]. [12] presented a time-optimal control strategy of pursuer for the pursuit-evasion game. The authors [14] proposed a new method based on hierarchical reinforcement learning to study multi-agent pursuit-evasion problem, and the experimental result has also showed that based on the option algorithm of hierarchical reinforcement learning, the algorithm efficiency can reduce the complexity of the pursuit-evasion task, avoiding traditional reinforcement learning curse of dimensionality.

Fuzzy system possesses characteristics of linguistic information and logic control, while neural networks have characteristics of parallelism, fault tolerance and association. Therefore, fuzzy systems and neural networks are applied to several control problems [1-2] with satisfactory results and possess the characteristics of universal approximation [3-4]. Traditionally, the fuzzy systems and neural networks are trained by using the gradient descent method. However, such techniques may lead to local optimum. Some researchers have been trying to use evolutionary algorithms, such as genetic algorithms (GAs), to overcome such difficulties [5-7].

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 482–490, 2010.

In this paper, our objective is to develop an adaptive learning approach of fuzzy logic controller with evolution to enable a unicycle-type wheeled mobile robot adapt its actions to pursuit-evasion game, assuming the evader trajectory is well-defined and known. The motion control of pursuer is to intercept the evader. The weighting parameters and membership functions of the fuzzy logic controller are tuned via evolutionary method such as GAs. In addition, experimental results of pursuit-evasion games illustrate the performance of the proposed approach.

This paper is organized as follows. In Sec. 2, statement of the pursuit-evasion game that this paper deals with is formulated as an inverse kinematics problem. A fuzzy logic control algorithm of evader trajectory tracking is proposed in Sec. 3. In Sec. 4, the fuzzy logic controller parameters are tuned by genetic algorithm. In Sec. 5, simulation results of pursuit-evasion are shown to demonstrate the effectiveness of our proposed approach. In Sec. 6, we conclude this paper.

2 Problem Statement

A simple structure of two circular-shape mobile robots in pursuit-evasion game is shown in Figure 1.The kinematic equations of the pursuer robot modeled as a nonholonomic unicycle mobile robot [12] are written as:

$$\begin{aligned} \dot{x}_p &= v\cos\theta \\ \dot{y}_p &= v\sin\theta \\ \dot{\theta} &= \frac{v}{R}\tan u \end{aligned} \tag{1}$$

where (x_p, y_p) denotes the position of the pursuer robot, θ is the orientation, v is the velocity, u is the steering angle and R is the wheel base.

We assume that the evader moves with constant speed along a trajectory (x_e, y_e) that is well-defined and known a priori. We consider a simplified control problem: v is constant and the only control is the steering angle u .Alternatively, for control purpose, define the reference signal to be the visibility line angle $\theta_d = \tan^{-1}(\frac{y_e - y_p}{x_e - x_p})$, and the relative distance between pursuer and evader $d = \sqrt{(x_e - x_p)^2 + (y_e - y_p)^2}$ as the evader information.

The control objective is to design an intelligent controller, specifically an adaptive fuzzy controller for (1) such that by controlling the orientation angle θ of pursuer to guarantee that the pursuer can follow the direction of the evader, i.e. $\theta = \theta_d$. Since the nonholonomic pursuer (1) can't change its orientation θ without changing its position (x_p, y_p) , the capture occurs when the pursuer can touch (or intercept) the evader such that the distance d is less than or equal to the sum of radius of pursuer and evader:

$$d <= r_e + r_p \quad (2)$$

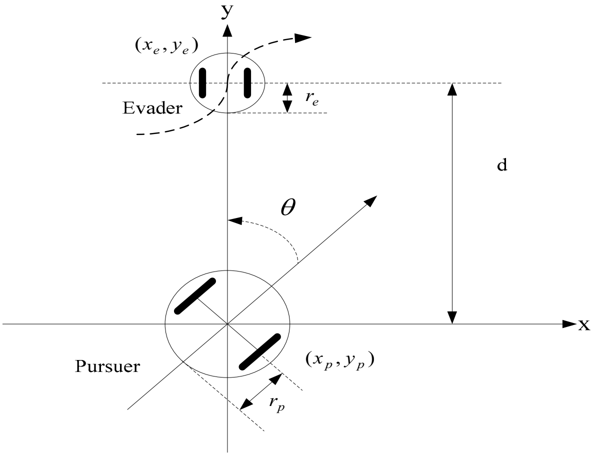


Fig. 1. Two mobile robots in pursuit-evasion

Formally, the pursuit-evasion problem that we deal with in this paper can be stated as follows. Given a well-defined evader trajectory $(x_e(t), y_e(t))$, an initial location (x_0, y_0) of pursuer, find a u that generates a trajectory (x_p, y_p) of pursuer passing through (x_0, y_0), such that the error $e = \theta_d - \theta$ along this trajectory approaches zero and d approaches a value within an accuracy tolerance defined in (2) eventually. The solution to this problem requires that the inverse of the velocity kinematics (1) of pursuer must be solved approximately to find control u that steers the pursuer in the direction aligning to the reference signal and catch the evader. This is achieved by the fuzzy logic control in this paper.

3 Fuzzy Logic Controller for Pursuer Motion

3.1 Description of Fuzzy Logic System

The basic configuration of fuzzy logic systems consists of fuzzy IF-THEN rules and a fuzzy inference engine. The fuzzy inference engine uses the fuzzy IF-THEN rules to perform a mapping from input linguistic variables to output linguistic variables. Given the input data x_q, $q = 1,2,\cdots,n$, and the output data y_p, $p = 1,2,\cdots,m$, the i th fuzzy rule has the following form:

$$\begin{aligned} R^i : IF \quad & x_1 \quad is \quad A_1^i \quad and \cdots x_n \quad is \quad A_n^i \\ THEN \quad & y_1 \quad is \quad w_1^i \quad and \cdots y_m \quad is \quad w_m^i \end{aligned} \tag{3}$$

where i is a rule number, A_q^i 's are the fuzzy sets of the antecedent part, and w_p^i are real numbers of the consequent part. When the inputs $x = [x_1 x_2 \cdots x_n]^T$ are given, the output y_p of the fuzzy inference can be derived from the following equations:

$$y_p(x|w_p) = \frac{\sum_{i=1}^{h} w_p^i (\prod_{q=1}^{n} \mu_{A_q^i}(x_q))}{\sum_{i=1}^{h} (\prod_{q=1}^{n} \mu_{A_q^i}(x_q))} \tag{4}$$

where $\mu_{A_q^i}(x_q)$ is the membership function of A_q^i, h is the number of the fuzzy rules. $w_p = [w_p^1 w_p^2 \cdots w_p^h]^T$ is a weighting vector related to the p th output $y_p(x)$.

3.2 Fuzzy Control as Inverse Kinematics Solver

Now we proceed to fuzzy control design of pursuer motion. Let the output tracking error $e = \theta_d - \theta$. The fuzzy logic controller has two inputs $\mathbf{e} = (e_1, e_2)^T = (e, \dot{e})^T$

$$e_1 = e = \theta_d - \theta \tag{5}$$

$$e_2 = \dot{e} = \frac{1}{1 + (\frac{y_e - y_p}{x_e - x_p})^2} - \frac{v}{R} \tan u \tag{6}$$

By using centroid of area for defuzzifier, the steering angle u is the output, where

$$u_{Fuzzy} = \frac{\sum_{j=1}^{h} w_j \cdot B_j(\mathbf{e})}{\sum_{j=1}^{h} B_j(\mathbf{e})} \tag{7}$$

where h is the number of rules (3) to guarantee a rich enough space of controls, and $B_j(\mathbf{e})$ is a membership function defined as Gaussian basis function

$$B_j(\mathbf{e}) = \prod_{i=1}^{2} \exp(-(\frac{e_i - m_{i+j}}{\sigma_{i+j}})^2), j = 1, \cdots, h \tag{8}$$

4 Tuning of the Fuzzy Logic Controller Parameters via Evolutionary Method

4.1 Genetic algorithm

GA is a stochastic search and optimization technique that imitates natural evolution with Darwinian survival. Traditional GAs performs on the coding of the parameters, therefore, the coding method allows GAs to handle multiparameters or multimodel type of optimization problems easily. The mechanism of a GA, shown in Fig. 2, can be divided into four parts: first, keep a population of solutions coded as artificial chromosomes. Second, choose the better solutions (in terms of fitness) for recombination. Third, perform crossover and mutation on the chromosomes. Fourth, use these offspring to replace original chromosomes and obtain a new generation. Much work has shown that GAs lead to nearly global optimum solutions in many problems, such as control system, image recognition, path planning, and robot learning, etc. For optimal design of fuzzy systems, theoretical and empirical results have demonstrated that GAs are good candidates for selecting and generating fuzzy rules [8-9].

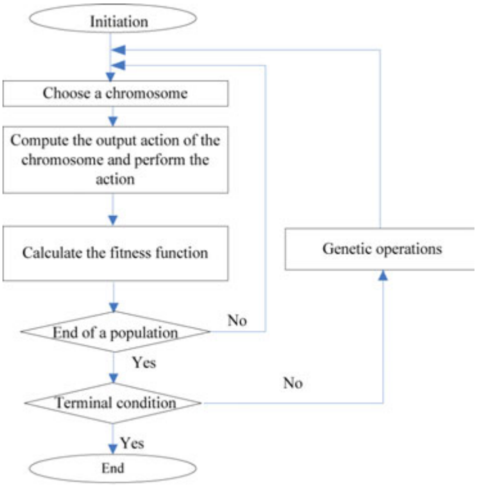


Fig. 2. The process of the adaptive learning approach shows the mechanism of the offline tuning

$m1$	$\sigma1$	W_1	$m2$	$\sigma2$	W_2	$m3$	$\sigma3$	W_3	$m4$	$\sigma4$	W_4	$m5$	$\sigma5$	W_5	$m6$	$\sigma6$	W_6	W_7	W_8	W_9

Fig. 3. The set of parameters of control, encoded as the chromosome, consists of weighting parameters and membership functions of the fuzzy logic system

4.2 Controller Parameters +Tuning via GA

An adaptive learning approach for controlling pursuer motion is to find such a set of parameters of control that steer the pursuer motion (1) in the desired direction. First, the weighting parameters and membership functions of the fuzzy logic controller (7) shown in Fig. 3 are encoded as a chromosome to be evolved by GA. We compute the output action of each chromosome, and perform the action. Next, a fitness function defined by

$$fitness = \left(\sum_{k=1}^{L} d_k \right)^{-1} \tag{9}$$

where $d = \sqrt{(x_e - x_p)^2 + (y_e - y_p)^2}$ and L is the step number. Finally, based on the fitness function, the operations of GA are performed according to Fig. 2, including reproduction, crossover and mutation. This process is repeated until the termination criteria are met. The overall scheme of the proposed controller is shown in Figure 4.

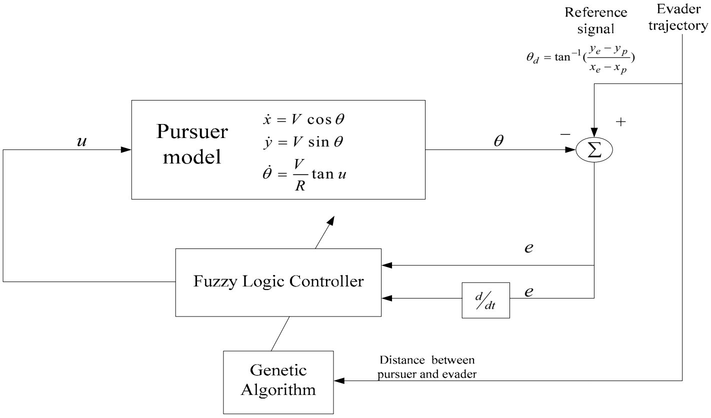


Fig. 4. Block diagram of the adaptive learning control system for pursuer, where the parameters of fuzzy logic controller are tuned via GAs. The evader information is in terms of the range and direction between the pursuer and evader.

5 Simulation Results

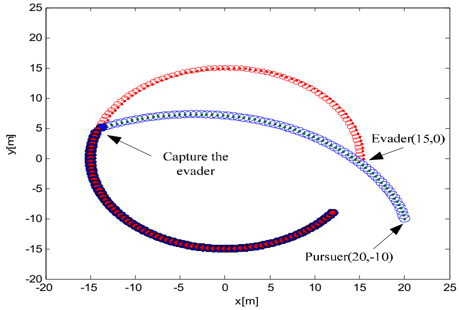
Consider the kinematics equations of the pursuer model in (1). The radius of the pursuer and evader are assumed as $r_p = 0.6m$ and $r_e = 0.4m$, respectively. Simulation are performed with evader trajectory represented by a circle $x_e = 15\cos(t), y_e = 15\sin(t)$, where $t \in [0, 2\pi]$ is a parameter. The initial position of the pursuer and evader are assumed as $(x_p, y_p) = (20,-10)$ and $(x_e, y_e) = (15,0)$, respectively. The initial orientation $\theta = 0\,rad$ and the tricycle wheel base $R = 0.2m$. The speed of the pursuer and evader is equal $V_p = V_e = 0.6m/s$.

For the fuzzy membership functions, we choose three gaussian membership functions. Each membership function has two parameters m and σ, respectively. The fuzzy controller has 2 inputs, so we have 9 rules. Thus, we obtain a chromosome with length 21 genes, as shown in Figure 3. The population consists of 30 chromosomes which are all randomized initially. The crossover and mutation probability are assumed as 65% and 2%, respectively.

The tracking results of the pursuer and evader are shown in Figure 5. One can observes that the proposed controller can steer the pursuer robot to successfully capture the evader. Once the pursuer captured the evader, the pursuer will follow the evader afterwards. Figure 6 shows the control input represents the steering angle of the pursuer, and the capture time is 71 seconds. Figure 7 is the fitness function of the evolution, showing the decrease of distance between pursuer and evader, thus convergence. The capture time of equal speed is tabulated in Table 1. The capture time is monotonically decreased at low speed, but is not necessarily decreased as the speed is higher.

Table 1.

speed ($V_p = V_e$)	**Capture time**
0.1m/s	291 s
0.2m/s	288 s
0.3m/s	179 s
0.4m/s	179 s
0.5m/s	127 s
0.6m/s	71 s
0.7m/s	80 s
0.8m/s	68 s
0.9m/s	53 s
1m/s	80 s



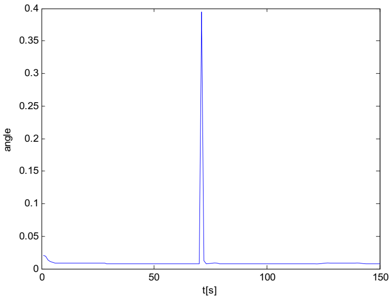


Fig. 5. The tracking results in xy plane shows that the pursuer can capture the evader successfully

Fig. 6. The control input u shows that the capture time is 71 seconds

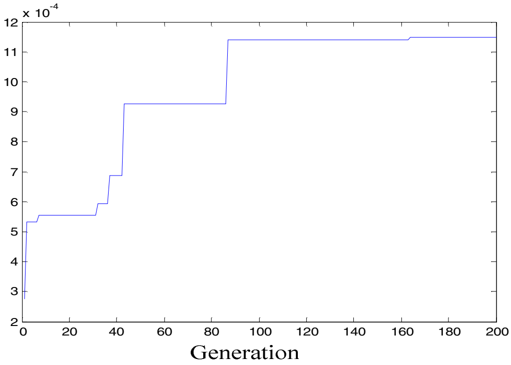


Fig. 7. The fitness function shows the decrease of relative distance between pursuer and evader

6 Conclusion

In this paper, an adaptive learning approach of fuzzy logic controller using evolution is developed as an inverse kinematics solver of unicycle for steering constant speed pursuer motion to capture an evader with well-defined, a priori known trajectory for pursuit-evasion games. GA is employed to tune the set of parameters of control consisting of weighting parameters and membership functions of the fuzzy logic controller using evader trajectory. The applicability and feasibility of the proposed technique is demonstrated by the simulation results.

References

1. Wang, C.H., Wang, W.Y., Lee, T.T., Tseng, P.S.: Fuzzy B-spline membership function (BMF) and its applications in fuzzy-neural control. IEEE Trans. Syst. Man, Cyber. 25, 841–851 (1995)
2. Wang, L.X.: Adaptive fuzzy systems and control: design and stability analysis. Prentice-Hall, Englewood Cliffs (1994)
3. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Networks, 359–366 (1989)
4. Wang, L.X., Mendel, J.M.: Fuzzy basis functions, universal approximation, and orthogonal least squares learning. IEEE Trans. Neural Networks 3, 807–814 (1992)
5. Wang, C.H., Liu, H.L., Lin, C.T.: Dynamic optimal learning rates of a certain class of fuzzy neural networks and its applications with genetic algorithm. IEEE Transactions on Systems, Man and Cybernetics. 31, 467–475 (2001)
6. Wang, W.Y., Lee, T.T., Hsu, C.C., Li, Y.H.: GA-based learning of bmf fuzzy-neural network. In: Proceedings of the 2002 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2002), vol. 2, pp. 1234–1239 (2002)
7. Farag, W.A., Quintana, V.H., Lamberttorres, G.: A Genetic-Based Neuro-Fuzzy approach for modeling and control of dynamical systems. IEEE Trans. on neural networks 9 (1998)
8. Yuan, Y., Zhuang, H.: A genetic algorithm for generating fuzzy classification rules 84, 1–19 (1996)

9. Seng, T.L., Khalid, M.B., Yusof, R.: Tuning of a neuro-fuzzy controller by genetic algorithm. IEEE Trans. Syst. Man, Cyber. Part B 29, 226–236 (1999)
10. Hladek, D., Vascak, J., Sincak, P.: Hierarchical fuzzy inference system for robotic pursuit evasion task. In: 6th International Symposium on Applied Machine Intelligence, SAMI 2008, pp. 273–277 (2008)
11. Kehagias, A., Hollinger, G., Singh, S.: A graph search algorithm for indoor pursuit/evasion. Mathematical and Computer Modelling 50, 1305–1317 (2009)
12. Lim, S.A., Furukawa, T., Dissanayake, G., D-Whyte, H.: A Time-Optimal Control Strategy for Pursuit-Evasion Games Problems. In: Proceedings of the 2004 IEEE International Conference on Robotics and Automation, vol. 4, pp. 3962–3966 (2004)
13. Isler, V., Kannan, S., Khanna, S.: Randomized pursuit-evasion in a polygonal environment. IEEE Transactions on Robotics 21, 875–884 (2005)
14. Liu, J., Liu, S., Wu, H., Zhang, Y.: A pursuit-evasion algorithm based on hierarchical reinforcement learning. In: 2009 International Conference on Measuring Technology, vol. 2, pp. 482–486 (2009)

Comparison of Multispectral Image Processing Techniques to Brain MR Image Classification between Fuzzy C-Mean Method and Geodesic Active Contours of Caselles Level Set Method

Yen-Sheng Chen[1], Shao-Hsien Chen[2], and Chun-Chih Chang[3]

[1] Department of Automation, Lan Yang Institute of Technology, No. 79, Fushin Rd.,
Touchen Town, Yilan County (26141), Taiwan, R.O.C.
yschen1114kimo@yahoo.com.tw

[2] Department of Mechanical Engineering, National Chin-Yi University of Technology, No. 35,
Lane 215, Sec. 1, Chung Shan Rd. Taipei City, Taiping City, Taichung County, 411,
Taiwan, R.O.C.
e6036@ncut.edu.tw

[3] YUJUN Technology Co., Ltd., No.385, Sec. 1, YUAN LIN. TOWN, CHANG HUA,
Taiwan, R.O.C.
jimmy@yu-jun.com.tw

Abstract. Magnetic Resonance Imaging (MRI) has become a widely used modality because it produces multispectral image sequences that provide information of free water, proteinaceous fluid, soft tissue and other tissues with a variety of contrast. The abundance fractions of tissue signatures provided by multispectral images can be very useful for medical diagnosis compared to other modalities. Multiple Sclerosis (MS) is thought to be a disease in which the patient immune system damages the isolating layer of myelin around the nerve fibers. This nerve damage is visible in Magnetic Resonance (MR) scans of the brain. Manual segmentation is extremely time consuming and tedious. Therefore, fully automated MS detection methods are being developed which can classify large amounts of MR data, and do not suffer from inter observer variability. In this paper, we propose two intelligent segmentation methods, fuzzy c-mean and Geodesic Active Contours of Caselles level set method to do the MR image segmentation jobs so as to find the effect they yield. The results show those intelligent methods both do a pretty job than other common image segmentation algorithm.

Keywords: Magnetic Resonance Imaging (MRI), medical image segmentation, multispectral image sequences, fuzzy c-mean (FCM), Caselles level set method.

1 Introduction

Nuclear magnetic resonance (NMR) can be used to measure the nuclear spin density, the interactions of the nuclei with their surrounding molecular environment and those between close nuclei, respectively. It produces a sequence of multiple spectral signals of tissues, which represents a variety of contrast on images, using three magnetic resonance parameters, spin-lattice (T1), spin-spin (T2) and dual echo-echo proton

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 491–498, 2010.

density (PD). Here we utilize those three different spectral image which we call them as patient1_T1, patient1_T2 and patient1_FLAIR. Here, FLAIR means fluid attenuating inversion recovery to do our demonstration.

The anatomical structures that appear in magnetic resonance (MR) or computed tomography (CT) scans are often explicitly extracted or segmented from the image for use in surgical planning, navigation, simulation, diagnosis, and therapy evaluation. By segmentation, we refer to the process of labeling individual voxels in the volumetric scan by tissue type, based on properties of the observed intensities as well as anatomical knowledge about normal subjects. Segmentation is often performed using automated techniques and semi-automated techniques. With CT data, segmentation of some structures can be performed just using intensity thresholding or other low-level image processing.

Brain parenchyma classification and segmentation of normal and pathological tissue is the first step of addressing a wide range of clinical problems. Using the information of volumes, shapes and region distributions of brain tissues, one can find those abnormalities.

Over the past years many computer-assisted methods for MR image classification have been reported in the literature[1]-[6], such as principle component analysis (PCA) in [2], eigenimage analysis in [3], neural networks in [4], fuzzy c-means methods in [5], knowledge-based techniques in [6] etc. For example, eigenimage filter-based approach has shown a promise in segmentation and feature extraction. Hybrid methods combine imaging processing and model-based techniques to improve segmentation. Knowledge-based techniques further allow one to make more intelligent classification and segmentation decisions. As an alternative, neural networks are also proposed to demonstrate their performance in segmentation of brain tissues to classical maximum likelihood methods.

Multiple Sclerosis (MS) is thought to be a disease in which the patient immune system damages the isolating layer of myelin around the nerve fibers. This nerve damage is visible in Magnetic Resonance (MR) scans of the brain. For instance, in MRI FLuid Attenuation Inversion Recovery (FLAIR) damage is visible as small bright spots called lesions or plaques. In both clinical trials and every day diagnostic use, the MR scans are manually read and marked by a human expert. This manual segmentation is extremely time consuming because of the large number of MR slices of each patient. Therefore, fully automated MS detection methods are being developed which can classify large amounts of MR data, and do not suffer from inter observer variability. Therefore, in this paper, we propose two intelligent segmentation method to do the MR image segmentation.

For the first method, we select the Fuzzy C-Mean method have two reasons. One is that it allows to generate background signatures in an unsupervised manner for classification. Another is that it is basically a spatial-based pattern classification technique.

For the second method, we choose Geodesic Active Contours of Caselles Method [7] to do the image segmentation. Geometric active contour is a recent image segmentation method that overcomes previous problems with “snakes” [8]. It is an attractive method for medical image segmentation as it is able to capture the object of interest in one continuous curve. It involves minimizing a so-called energy function which is based on certain properties of the desired object boundary, for example the smoothness of the boundary curve and local gradient of the image.

In recent years, techniques of active contours and curve evolution have been widely investigated and applied to the image segmentation problem. Compared to other methods such as thresholding and edge-based methods, the active contour model has the advantage of being less sensitive to blurred edges and also avoiding broken contour lines. In general, active contour models are based on deforming an initial contour C towards the boundary of the object to be detected, through minimizing a functional designed such that its minimum is obtained at the boundary of the object. Energy minimization involving components controlling the smoothness of the curve and one for pulling the curve closer to the boundary is a common technique.

2 Methodology and Implementation

2.1 Fuzzy Clustering – What Is Data Clustering

Clustering of numerical data forms the basis of many classification and system modeling algorithms. The purpose of clustering is to identify natural groupings of data from a large data set to produce a concise representation of a system's behavior.

For MATLAB, its Fuzzy Logic Toolbox tools allow you to find clusters in input-output training data. You can use the cluster information by several kinds of methods.

2.2 Fuzzy C-Means Clustering Method

Fuzzy c-means (FCM) is a data clustering technique wherein each data point belongs to a cluster to some degree that is specified by a membership grade.

This technique was originally introduced by Jim Bezdek in 1981[9] as an improvement on earlier clustering methods. It provides a method that shows how to group data points that populate some multidimensional space into a specific number of different clusters.

Fuzzy Logic Toolbox command line function fcm starts with an initial guess for the cluster centers, which are intended to mark the mean location of each cluster. The initial guess for these cluster centers is most likely incorrect.

Additionally, “fcm” function assigns every data point a membership grade for each cluster. By iteratively updating the cluster centers and the membership grades for each data point, fcm iteratively moves the cluster centers to the right location within a data set. This iteration is based on minimizing an objective function that represents the distance from any given data point to a cluster center weighted by that data point's membership grade.

The command line function fcm outputs a list of cluster centers and several membership grades for each data point. You can use the information returned by fcm to help you build a fuzzy inference system by creating membership functions to represent the fuzzy qualities of each cluster.

2.3 Steps of Implementation

The steps used in our method are described as follow:

Step 1: Load first image for one sub-band of Patient1_T1 MRI phantom image as row1(:) for every pixel as shown in Fig. 1.

Step 2: Read their image pixel information and adjust its image intensity values or colormap.

Step 3: Load second Patinet1_T2 image file for another sub-band of MRI phantom image as row2(:) for every pixel as shown in Fig. 2.

Step 4: Combine them to a data array.

Step 5: Use following fcm function with 7 clusters.

fcm **Syntax :** [center,U,obj_fcn] = fcm(data,cluster_n)
Description
[center, U, obj_fcn] = fcm(data, cluster_n) applies the fuzzy c-means clustering method to a given data set.

The input arguments of this function are
data: data set to be clustered; each row is a sample data point
cluster_n: number of clusters (greater than one)

The output arguments of this function are described as followed:
center: matrix of final cluster centers where each row provides the center coordinates
U: final fuzzy partition matrix (or membership function matrix)
obj_fcn: values of the objective function during iterations
Step 6: Finding the pixels for each class
Step 7 : Reshapeing the array to a image
Finally, we get a pretty good result as shown in Fig. 3.

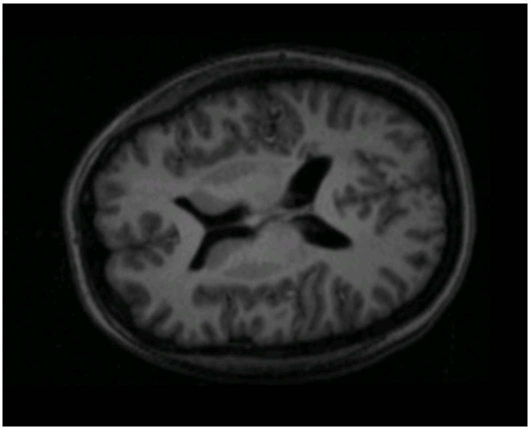


Fig. 1. Load first image for one sub-and of MRI T1 phantom image

2.4 Caselles Method

An important problem in image analysis is object segmentation. It involves the isolation of a single object from the rest of the image that may include other objects and a background.

Here, we focus on boundary detection of one or several objects by a dynamic model known as the "geodesic active contour"

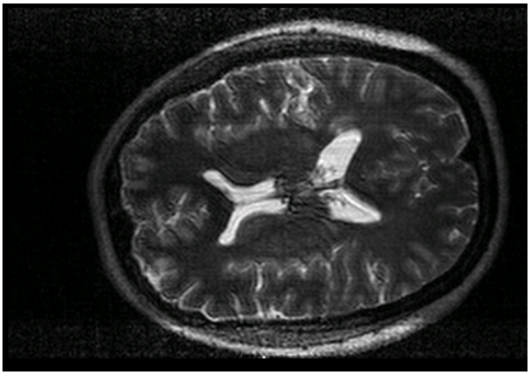


Fig. 2. Load Second image for another sub-band of MRI T2 phantom image

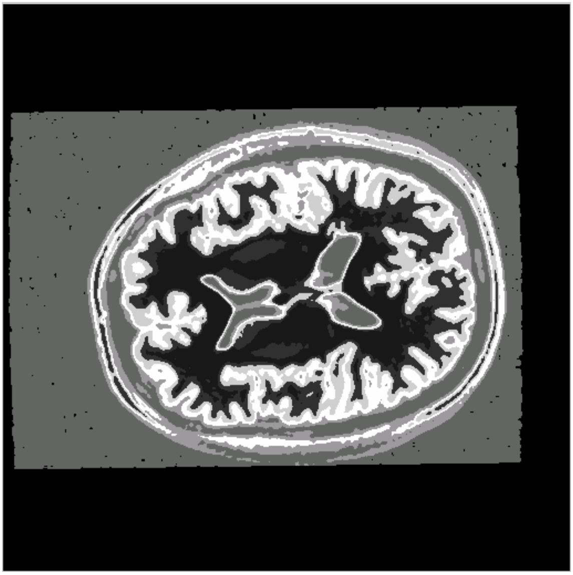


Fig. 3. The Result after fcm with 7 classification

Geodesic active contours were introduced as a geometric alternative for “snakes”, mentioned previously. Snakes are deformable models that are based on minimizing an energy along a curve. The curve, or snake, deforms its shape so as to minimize an “internal” and “external” energies along its boundary. The internal part causes the boundary curve to become smooth, while the external part leads the curve toward the edges of the object in the image.

Actually, a simplified snake model yields the same result as that of a geodesic active contour model, up to an arbitrary constant that depends on the initial parameterization. Unknown constants are an undesirable property in most automated models.

Here, we give some definition and formula first to be a base to implement them.

Energy criterion

$$E(\Gamma) = \int_0^1 g(I(\Gamma(q)))\|\Gamma'(q)\|dq \quad , \tag{1}$$

where

$$g(I) = \frac{1}{1 + \|\nabla(G * I)\|^2} \quad , \tag{2}$$

$I(\cdot)$ corresponds to the image intensity, Γ is the parametric curve and G is a gaussian filter of variance 1.

Evolution equation

$$\frac{\partial \phi}{\partial t}(x) = g(I(x))\|\nabla \phi(x)\|(c + \kappa) + \nabla g(I(x))\nabla \phi(x) \quad . \tag{3}$$

where $\kappa = \left(\frac{\nabla \phi(x)}{\|\nabla \phi(x)\|} \right)$ corresponds to the curvature of the evolving contour and c is a constant that acts as a balloon force.

2.5 Steps of Implementation

The implementation steps are:

Step 1: This algorithm is a contour-based method i.e. the gradient of the image is used to compute the force function. The curve will thus be driven to regions with high gradient.

Step 2: This method does not require any regularization term as it is intrinsic to the method.

Step 3: ϕ is implemented as a signed distance function and is reinitialized at each iteration.

This algorithm has one specific parameter that can be modified by the user, the propagation term c, it is as a constant that acts as a balloon force pushing the contour either inward or outward (default value is set to 1). Therefore, we got a result as shown in Fig. 4 Gray matter segmentation in a MRI brain image. The white contour converges to the outer cortical surface and the black contour converges to the inner cortical surface.

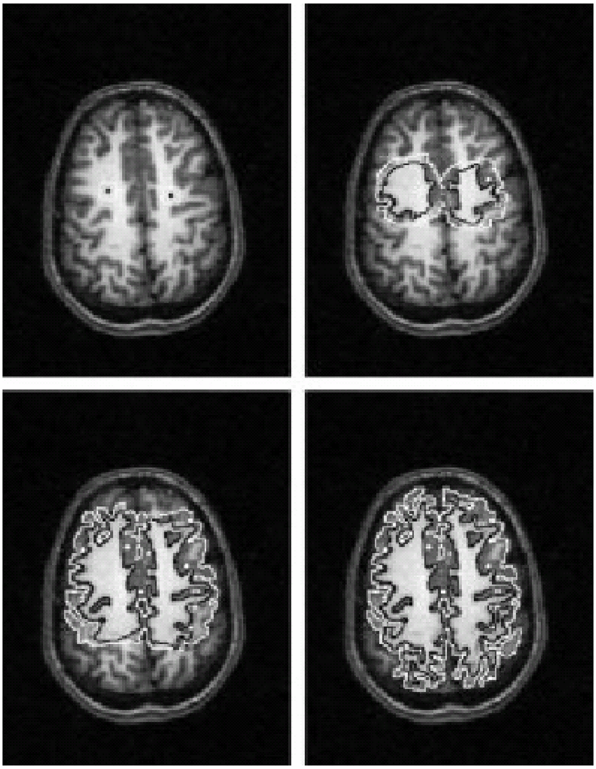


Fig. 4. Gray matter segmentation in a MRI brain image. The white contour converges to the outer cortical surface and the black contour converges to the inner cortical surface.

3 Conclusion

Fuzzy c-means based approach has been successfully applied to remotely sensed images in target detection and classification. It views an MR image sequence as a multispectral image cube and models each pixel vector as a linear mixture of tissue substances resident in the MR pixel vector. Unlike traditional image classification techniques which are carried out on a pure pixel basis, fcm-based classifiers are mixed pixel classification techniques. They use the linear mixture model to generate a fractional image for each object required for classification. The advantages of mixed pixel classification have been demonstrated in the experiment. The experimental results show that the cerebral tissue was classified accurately. As opposed to other approach which only classifies objects of interest, the fcm method classifies all MR image pixel vectors including background pixel vectors into pattern classes. In future, we consider to apply new method to different types of medical images and to compare its effectiveness over other clustering methods. Also, we consider doing segmentation based on the mixture of our method with different methods like active control, Multi scale FCM, Statistical methods and mix the results to have more accurate segmentation in abnormal diagnosis or different important matters in medical images. Also we consider adding useful aspects of other methods to this method.

For Caselles level set method, we just give a short description of the implemented algorithms as mentioned above. For further details about this method, the reader is invited to refer to the cited algorithms. For this method, we give the energy criterion which is minimized, the derived evolution equation and the main properties of the method.

References

1. Johnston, B., Atkins, M.S., Machiewich, B., Anderson, M.: Segmentation of multiple sclerosis lesions in intensity corrected multispectral MRI. IEEE Transactions on Medical Imaging 15(2), 154–169 (1996)
2. Grahn, H., Szeverenyi, N.M., Roggenbuck, N.W., Delaglio, F., Geladi, P.: Data analysis of multivariate magnetic resonance image I. a principal component analysis approach. Chemometrics and Intelligent Lab. Systems 5, 11–322 (1989)
3. Soltantian-Zadeh, H., Saigal, R., Haggar, A.M., Windham, J.P., Yagle, A.E., Hearshen, D.C.: Optimization of MRI protocol and pulse sequence parameters for eigenimage filtering. IEEE Trans. Medical Imaging 13(1), 161–175 (1994)
4. Alirezaie, J., Jernigan, M.E., Nahmias, C.: Automatic segmentation of cerebral MR images using artificial neural networks. IEEE Transactions on Nuclear Science 45(4), 2174–2182 (1998)
5. Stella Atkins, M., Mackiewich, B.T.: Fully automatic segmentation of brain in MRI. IEEE Transactions on Medical Imaging 17(1), 98–107 (1998)
6. Clark, M.C., Hall, L.O., Goldgof, D.B., Velthuizen, R., Murtagh, F.R., Silbiger, S.: Automaic tumor segmentation using knowledge-based techniques. IEEE Transactions on Medical Imaging 17(2), 178–201 (1998)
7. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic active contours. Int. J. of Computer Vision 22, 61–79 (1997)
8. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. IJCV 1, 321–331 (1987)
9. Bezdec, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)

Modeling PVT Properties of Crude Oil Systems Using Type-2 Fuzzy Logic Systems

Sunday Olusanya Olatunji[1], Ali Selamat[1], and Abdul Azeez Abdul Raheem[2]

[1] Intelligent Software Engineering Laboratory, Faculty of Computer Science & Information Systems, University of Technology Malaysia, 81310 UTM Skudai, Johor, Malaysia
oluolatunji.aadam@gmail.com, aselamat@utm.my

[2] Centre for Petroleum and Minerals, the Research Institute, King Fahd University of Petroleum and Minerals (KFUPM), Dhahran 31261, Box: 1105, Kingdom of Saudi Arabia
toazeez@gmail.com

Abstract. This paper presented a prediction model of Pressure-Volume-Temperature (PVT) properties of crude oil systems based on type-2 fuzzy logic systems. PVT properties are very important in the reservoir engineering computations, and its accurate determination is important in the primary and subsequent development of an oil field. Earlier developed models are confronted with several limitations especially in uncertain situations coupled with their characteristics instability during predictions. In this paper, an interval type-2 fuzzy logic based model is proposed and implemented to improve PVT properties predictions. Comparative studies have been carried out and empirical results show that the newly proposed approach outperforms others in general and particularly in the area of stability, consistency and the ability to adequately handle uncertainties. Another unique advantage of the proposed model is its ability to generate prediction intervals without extra computational cost.

Keywords: Type-2 fuzzy logic system, Feedforward neural networks, Empirical correlations, PVT properties, Formation volume factor, Bubble point pressure.

1 Introduction

Characterization of reservoir fluids plays a very crucial role in developing a strategy on how to produce and operate a reservoir. PVT Properties are very crucial for geophysics and petroleum engineers, namely for the utilization in material balance calculations, inflow performance calculations, well log analysis, determining reserve estimates and the amount that can be recovered, the flow rate of oil or gas and numerical reservoir simulations [1, 2].

PVT properties are part of the important properties for oil and gas industry. The phase and volumetric behavior of petroleum reservoir fluids is referred to as PVT (pressure-volume-temperature). PVT properties include Formation volume factor (FVF), Solution gas-oil ratio (GOR), Solution oil-gas ratio (OGR), Liquid specific gravity, API specific gravity, Gas specific gravity, Bubble-Point pressure, Saturation pressure and the likes, [3]. Among those PVT properties, the bubble-point pressure (P_b) and the Oil formation factor (B_{ob}) are the most important, because they are the

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 499–508, 2010.

most essential factors in reservoir and production computations [3]. The more the preciseness of estimating these properties, the better the calculations involved in reservoir simulation, production, and field development. Bubble-point pressure (P_b) is the pressure at which gas first begins to come out of the solution at constant temperature, while Oil Formation Volume Factor (Bob) is defined as the volume of reservoir oil that would be occupied by one stock tank barrel oil plus any dissolved gas at the bubble point pressure and reservoir temperature as stated in [1, 4-6].

To overcome the shortcomings associated with the earlier correlation methods, researchers have utilized artificial intelligence based methods foremost of which is the classical artificial neural network (ANN) and its variants. However, the developed neural networks correlations often do not perform to expectations and are bedeviled with several shortcomings that include; ANN is a black box modeling scheme that is based on the trial-and-error approach with its architectural parameters having to be guessed in advance, such as, number and size of hidden layers, learning rate, training algorithm parameters, initial random weights and most importantly the tendency to be stocked at the local minima, and its inability to handle uncertainties. Among the major problems of ANN is instability and inconsistency, with its non-homogeneous nature such that very different learning curves were obtained for different repeat of the same experiment. Therefore we proposed type-2 FLS to resolve these shortcomings.

The rest of this paper is organized as follows. Section 2 provides a summarized literature review. Section 3 presents the proposed intelligence framework based on type-2 FLS. Section 4 provides data acquisition, statistical quality measures used, and empirical study based on the provided real-world industry data and deep comparative studies. Results and discussions are also presented in this section. The conclusion and future work recommendations are provided in Section 5.

2 Literature Review

The development of correlations for PVT calculations has been the subject of extensive research, resulting in a large volume of publications.

2.1 Common Empirical Models and Evaluation Studies

Standing in [7] presented correlations for bubble point pressure and for oil formation volume factor. Glaso in [8] developed the Glasoempirical correlation for formation volume factor using 45 oil samples from North Sea hydrocarbon mixtures. Al-Marhoun in [9] published his second correlation for oil formation volume factor. For more empirical correlation-related work, discussion, applications, and comparative studies, interested readers can see [10-15].The authors in [16] used neural network with non-iterative approach for predicting the complete PVT behavior of reservoir oils and gas condensates, while in [17], two neural networks are trained separately to estimate the bubble point pressure (P_b) and oil formation volume factor (B_{ob}). The authors in [4] used the feedforward learning scheme with log sigmoid transfer function in order to estimate the formation volume factor at the bubble point pressure, while the authors in [18] developed two new models to predict the bubble point pressure, and the oil formation volume factor at the bubble-point pressure for Saudi crude oils. For further related works, see [1, 6, 19-20].

3 The Proposed Type-2 Fuzzy Based Scheme

In this work, type-2 fuzzy logic systems framework is investigated, developed and then utilized for predicting PVT properties, specifically, bubble point pressure (Pb) and Oil Formation Volume Factor (Bob), being the most important and essential factors in reservoir management and production computations [3]. The proposed type-2 FLS framework that has the ability to take care of all types of uncertainty and imprecision is presented as follows. Thegoal is to completely specify the FLSs using the training data, which is a unique characteristic ofadaptive fuzzy systems. The core structure of the newly proposed framework is shown in fig. 1, and the functioning of each component is briefly described shortly.

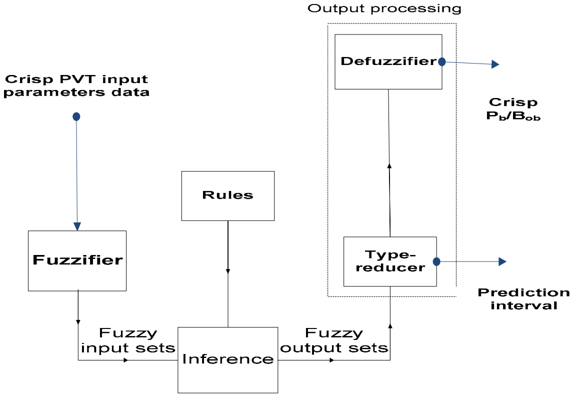


Fig. 1. Schematic diagram of type-2 FLS based framework to build PVT properties models: where the “crisp PVT input parameters data” include the four input parameters of solution gas-oil ratio, reservoir temperature, oil gravity, and gas relative density, while the final output will be B_{ob} / P_b

The fuzzifier takes the four input parameters, namely, solution gas-oil ratio, reservoir temperature, oil gravity, and gas relative density values as inputs. The output of the fuzzifier is the fuzzified measurements which will be the input to the inference engine. The resultant of the inference engine is type-2 fuzzy output sets which can be reduced to type-1 fuzzy set by the type reducer. This type reduced fuzzy set, in this model, is an interval set which gives the predicted external attribute measurement as a possible range of values. The defuzzifier calculates the average of this interval set to produce the predicted crisp external attribute measurement (which could be bubble point pressure (P_b) and Oil Formation Volume Factor (B_{ob}) as the case may be).

3.1 Inferencing in the Proposed Type-2 Fuzzy Logic System

Fuzzy inference engine combines the fired fuzzy rules and maps inputs into type-2 output fuzzy sets. Generally a type-2 FLS is a fuzzy logic system in which at least one of the fuzzy sets used in the antecedent and/or consequent parts and each rule inference output is a type-2 fuzzy set. Consider a type-2 Mamdani FLS having n inputs

$x_1 \in X_1, \ldots\ldots x_n \in X_n$ and one output $y \in Y$. The rule base contains L fuzzy rules expressed in the following form:

R^1: IF x$_1$ is F_1^l and x$_2$ is F_2^l and … ……and x$_n$ is F_p^l THEN y is G^l

Where l=1,2……L, F_i^l and G^l are type-2 fuzzy set.

This rule represents a Type-2 Fuzzy relation between the input space X∈ X$_1$ x X$_2$ x … x X$_n$ and the output space Y of the system. We denote the membership function of this Type-2 relation as:

$$\mu_{F_1^l x \ldots x F_n^l \to G^l}(x, y) \tag{1}$$

where F_1^l x …x F_p^l denotes the Cartesian product of F_1^l, F_2^l,……. F_n^l and x = {x$_1$, x$_2$, …, x$_n$}.

The antecedents in the fuzzy rules are connected by using the meet operation, the firing strength of the input fuzzy sets are combined with output fuzzy sets using the extended sup-star composition, and the multiple rules are combined using the join operation.

However, the computing load involved in deriving the system output from a general type-2 FLS model is high in practice, and the general practice is to use the interval type-2 FLS in which the fuzzy sets F_i^l and G^l are interval fuzzy sets through which the computing of type-2 FLS can be greatly simplified. The membership grades of interval fuzzy sets can be fully characterized by their lower and upper membership grades of the footprint of uncertainty (FOU) separately.

Without the loss of generality, let $\mu_{F_i^l}(x) = [\underline{\mu}_{F_i^l}(x), \overline{\mu}_{F_i^l}(x)]$ and $\mu_{G^l}(y) = [\underline{\mu}_{G^l}(y), \overline{\mu}_{G^l}(y)]$ for each sample (x, y). The firing strength of interval type-2 FLS $\mu_{F^l}(x) = \bigcap_{i=1}^{n} \mu_{F_i^l}(x)$ is an interval i.e., $\mu_{F^l}(X) = [\underline{f}^i(X), \overline{f}^i(X)]$. In the proposed interval type-2 FLS, the meet operation under product t-norm is used, so that the firing strength is an interval type-1 set as shown below:

$$f^i(X) = [\underline{f}^i(X), \overline{f}^i(X)] = [\underline{f}^i, \overline{f}^i] \tag{2}$$

Where $\underline{f}^i(X)$ and $\overline{f}^i(X)$ can be re-written as follows with * representing the t-norm product operation:

$$\underline{f}^i(X) = \underline{\mu}_{\overline{F}_1^j}(x_1) * \ldots\ldots * \underline{\mu}_{\overline{F}_p^j}(x_n) \tag{3}$$

$$\overline{f}^i(X) = \overline{\mu}_{\overline{F}_1^j}(x_1) * \ldots\ldots * \overline{\mu}_{\overline{F}_p^j}(x_n) \tag{4}$$

3.2 Type Reduction

The results from the inference engine are type-2 fuzzy sets. There is then the need to reduce the type-2 fuzzy sets from 2 to type-1fuzzy sets in order to give room for defuzzification in order to generate the final crisp outputs. Centre-of-sets (COS) type-reducer algorithm is made use of in this study because it has reasonable computational complexity compared to others like the expensive centroid type reducer. COS type reducer is made up of two stages, viz: (i) calculating the centroids of type-2 fuzzy rule consequences and (ii) calculating the reduced fuzzy sets.

3.3 Computing the Centroid of Type-2 Fuzzy Rule Consequences

Suppose the output of an interval type-2 FLS is represented by type-2 fuzzy sets G^t, where t=1, ….., T. T is the number of output fuzzy sets. In this first stage, the centroids of all the T output fuzzy sets are calculated and they will be used in calculating the reduced sets in the next stage. The centroid of the ith output fuzzy set y^t is a type-1 interval set which can be expressed in the following formula:

$$y^t = [y_l^t, y_r^t] = \int_{\theta_1 \in J_{y1}} \int_{\theta_z \in J_{yz}} \frac{1}{\sum_{z=1}^{Z} y_z \theta_z / \sum_{z=1}^{Z} y_z \theta_z} \tag{5}$$

Where y_l^t and y_r^t are the leftmost and rightmost point of y^t respectively.

3.4 Computing the Reduced Type-1 Fuzzy Sets

To calculate the type-reduced set, it is sufficient to compute its upper and lower bounds of the reduced set y_l and y_r and can be expressed as follows:

$$y_l = \frac{\sum_{i=1}^{M} f_l^i y_l^i}{\sum_{i=1}^{M} f_l^i} \qquad y_r = \frac{\sum_{i=1}^{M} f_r^i y_r^i}{\sum_{i=1}^{M} f_r^i} \tag{6}$$

where f_l^i and y_l^i are the firing strength and the centroid of the output fuzzy set of ithrule (*i=1,…,M*) associated with y_l respectively. Similarly, f_r^i and y_r^i are the firing strength and the centroid of the output fuzzy set of ith rule (*i=1,…,M*) associated with y_r respectively.

3.5 Defuzzification

We defuzzify the type-reduced set to get a crisp output from the type-2 FLS. The final output of type-2 FLS is thus set to the average of y_l and y_r as shown below:

$$y(x) = \frac{y_l + y_r}{2} \tag{7}$$

3.6 Training the Model with Type-2FLS Learning Procedures Using Steepest Descent Approach

Consider a FLS with Gaussian membership functions, center of sums type-reducer, average defuzzification, ma-product composition, and product implication, it could be expressed by the equation:

$$y(x^{(i)}) = f(x^{(i)}) = \frac{\sum_{l=1}^{M} \bar{y}^{l} \prod_{k=1}^{p} \exp\left[-\frac{\left(x_k^{(i)} - m_{F_k^l}\right)^2}{\left(2\sigma_{F_k^l}^2\right)}\right]}{\sum_{l=1}^{M} \prod_{k=1}^{p} \exp\left[-\frac{\left(x_k^{(i)} - m_{F_k^l}\right)^2}{\left(2\sigma_{F_k^l}^2\right)}\right]} \qquad i = 1,\ldots,N \tag{8}$$

whereM is number of rules, p is number of antecedents and N is number of data points. Given an input-output training pair ($x^{(i)} : y^{(i)}$) also known as data point, we wish to design an FLS so that the error function is minimized. The Steepest Descent approach can be applied to obtain the following recursions to update all the design parameters of this FLS in order to minimize the error function.

$$m_{F_k^l}(i+1) = m_{F_k^l}(i) - \alpha_m [f_s(x^{(i)}) - y^{(i)}][\bar{y}^l(i) - f_s(x^{(i)})] \times \frac{[x_k^{(i)} - m_{F_k^l}(i)]}{\sigma_{F_k^l}^2(i)} \phi_l(x^{(i)}) \tag{9}$$

$$\bar{y}^l(i+1) = \bar{y}^l(i) - \alpha_{\bar{y}} [f_s(x^{(i)}) - y^{(i)}] \phi_l(x^{(i)}) \tag{10}$$

$$\sigma_{F_k^l}(i+1) = \sigma_{F_k^l}(i) - \alpha_\sigma [f_s(x^{(i)}) - y^{(i)}][\bar{y}^l(i) - f_s(x^{(i)})] \times \frac{[x_k^{(i)} - m_{F_k^l}(i)]^2}{\sigma_{F_k^l}^3(i)} \phi_l(x^{(i)}) \tag{11}$$

$$\text{RMSRE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left[\frac{f_s(x^{(i)}) - y^{(i)}}{y^{(i)}}\right]^2} \tag{12}$$

Now, the back propagation algorithm can be applied as follows:

Algorithm 1. Back propagation algorithm for FLS

```
1.  Initialize the parameters of all the membership functions for all the rules, $m_{F_k^l}(0)$,
    $\bar{y}^l(0)$ and $\sigma_{F_k^l}(0)$.
2.  Set an end criterion to achieve convergence.
3.  **Repeat**
        i.   **for all** data points ( $x^{(i)} : y^{(i)}$ ) $i = 1,\ldots,N$
                a)  Propagate the next data point through the FLS.
                b)  Compute error.
                c)  Update the parameters of the membership functions using 9, 10, and 11.
        ii.  **end for***(*end for each input-output pair*)*
        iii. Compute the root mean square relative error (RMSRE) as  in(12).
        iv.  Test the end criterion. If satisfied break.
**Until***(*end for each epoch*)*
```

4 Empirical Study, Discussion, and Comparative Studies

In order to carry out an empirical study, two distinct databases were acquired and necessary preprocessing done on all of them for quality control. To evaluate performance of each modeling scheme, the entire database was divided using the stratified sampling approach. 70% of the data was used for training and building Type-2 fuzzy model (internal validation) and the remaining 30% was reserved for testing/validation (external validation or cross-validation criterion). Generally, once the type-2 fuzzy inference system has been trained, the calibration model becomes ready for testing and evaluation using the cross-validation criterion. For testing and evaluation of the newly developed framework, and to carry out effective comparative studies viz-a-viz the standard neural networks and the three common published empirical correlations: namely, Standing [7]; Glaso[8]; and Al-Mahroun[9] empirical correlations, the most common statistical quality measures that are utilized in both petroleum engineering and data mining journals were employed in this study.

The acquired databases were earlier utilized in distinct published research articles. They include; (a) 160 observations –database, (b) 782-observations. Details of each are as follows:

a. **160-dataset**: This database was drawn from the article of [20].
b. **782-dataset**: This database was obtained from the works of [5, 20].

4.1 Results and Discussions

The results of comparisons using external validation checks have been summarized in Tables 1 and 2. It could be easily observed from these results that the type-2 fuzzy modeling scheme outperforms the neural network and the three most common published empirical correlations reported, especially in terms of stability as indicated by its lowest standard deviation throughout the reported results. The proposed model showed a high accuracy in predicting both B_{ob} and P_b values with a stable and accurate performance, and achieved the lowest standard deviation all through, lowest absolute percent relative error and the highest correlation coefficient in most cases among all the presented correlations for the two distinct data sets used.

Table 1. Testing results for 160- dataset when predicting B_{ob} and $\mathbf{P_b}$. R^2 = Correlation coefficient, SD=Standard deviation, E_a=Average absolute percent relative error (AAPRE)

	B_{ob}			P_b		
Methods	R^2	SD	E_a	R^2	SD	E_a
Standing (1947)	0.974	2.5823	2.724	0.867	25.159	67.73
Glaso (1980)	0.972	2.673	3.374	0.945	25.171	18.52
Al-Marhoun (1992)	0.981	1.2842	2.205	0.906	12.839	20.01
ANN Model	0.993	0.89835	1.209	0.943	4.744	22.68
Type-2 FLS	0.994	0.09166	1.493	0.931	2.461	20.65

Table 2. Testingresultsfor 782- datasetwhenpredicting B_{ob} and $\mathbf{P_b}$. R^2 = Correlation Coefficient, SD=Standard Deviation, E_a=Average absolute percent relative error (AAPRE)

	B_{ob}			P_b		
Methods	R^2	SD	E_a	R^2	SD	E_a
Standing (1947)	0.9953	5.7655	2.7202	0.8657	13.696	24.684
Glaso (1980)	0.9959	5.2274	0.9821	0.9675	25.27	26.551
Al-Marhoun (1992)	0.9935	5.0028	2.0084	0.9701	25.015	8.9416
ANN Model	0.9968	4.7402	1.4592	0.9765	22.73	6.7495
Type-2 FLS	0.9998	0.1625	0.1	0.9894	2.347	0.3432

From the tables presented, it could be easily observed, for instance in estimating P_b based on the 782-dataset used in [5, 25], that type-2 fuzzy based model has the smallest standard deviation, SD =2.347, the highest correlation coefficient, R^2 = 0.9894, and the smallest absolute percent relative error, E_a = 0.3432%, while the neural network is below type-2 FLS scheme with SD = 22.73, R^2 =0.9765, E_a = 6.7495%. The other empirical correlations indicates higher error values and standard deviation with lower correlation coefficients as Standing (1947) has SD = 13.696, R^2 =0.8657, and E_a = 24.684%, Glaso correlation has SD = 25.27, R^2 =0.9675, and E_a = 26.551%, and Al-Marhoun (1992) correlation has SD = 25.015, R^2 =0.9701, and E_a = 8.9416%. Thus from this particular case, type-2 FLS had 97.6% improvement over the best among the other models and 90.7% over the model with the least performance in terms of standard deviation (SD) which is a measure of stability and consistency of the models. While in terms of correlation coefficient (R^2), type-2 FLS had 1.32% improvement over the best among other models and 14.3% over the least model. As for the average absolute percent relative error (E_a), type-2 FLS had 94.9% improvement over the best among other models (i.e. ANN) and 14.3% over the method with the least performance. Similar result patterns are also obtained for other cases.

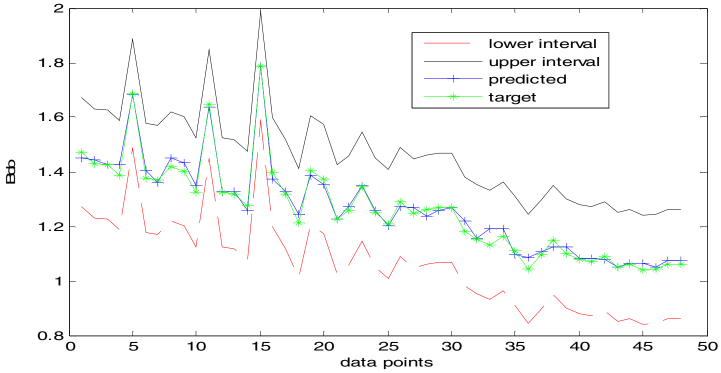


Fig. 2. Prediction intervals for B_{ob} based on type-2 FLS

4.2 Estimation of Prediction-Intervals Using Type-2 FLS

One important additional benefit of type-2 FLS worth mentioning at this point is, compared with existing models, is that the model based on type-2 fuzzy logic will generate not only the PVT properties predictions, but also prediction intervals without incurring extra computational cost. Fig. 2 showsa sample containing the prediction intervals for B_{ob} as generated in this work.

5 Conclusion

In this study, type-2 fuzzy logic system has been investigated, developed and implemented, as predictive solutions that take care of all forms of uncertainties, for predicting both bubble point pressure and oil formation volume factor using the four input variables of solution gas-oil ratio, gas relative density, oil gravity and reservoir temperature. It is an established fact in the petroleum engineering communities that these two predicted properties were the most important PVT properties of crude oil systems[3-5].In-depth comparative studies have been carried out between this new framework and the standard neural networks and the three common published empirical correlations presented in [7], [8] and [9]. Empirical results from simulations show that Type-2 fuzzy logic based model outperformed all the compared models or at least compete favorably with the best, while in terms of performance stability and consistency, type-2 fuzzy based model has no rival, as indicated by its lowest standard deviation, all through, compared to others. Another advantage of the proposed approach is the ability to generate not only the target forecast but also prediction intervals without incurring additional computational cost.

Acknowledgement

The University of Technology Malaysia (UTM) and King Fahd University of Petroleum and Minerals (KFUPM) Saudi Arabia are hereby acknowledged for some of the facilities utilized during the course of this research work.

References

[1] Osman, E.A., Al-Marhoun, M.A.: Artificial Neural Networks Models for Predicting PVT Properties of Oil Field Brines. In: 14th SPE Middle East Oil & Gas Show and Conference, Bahrain (2005)

[2] Omole, O., Falode, O.A., Deng, A.D.: Prediction Of Nigerian Crude Oil Viscosity Using Artificial Neural Network. Petroleum & Coal, International Journal for Petroleum Processing 51(3) (2009)

[3] Standing, M.B.: Oil-System Correlation. In: Frick, T.C. (ed.) Petroleum Production Handbook, 2nd edn., McGraw-Hill Book Co., New York (1962)

[4] Kumoluyi, A.O., Daltaban, T.S.: High-Order Neural Networks in Petroleum Engineering. In: SPE Western Regional Meeting, Longbeach, California, USA (1994)

[5] Goda, H.M., et al.: Prediction of the PVT Data using Neural Network Computing Theory. In: The 27th Annual SPE International Technical Conférence and Exhibition in Abuja, Abuja, Nigeria (2003)

[6] El-Sebakhy, A.E.: Forecasting PVT properties of crude oil systems based on support vector machines modeling scheme. Journal of Petroleum Science and Engineering 64(1-4), 25–34 (2009)

[7] Standing, M.B.: A Pressure-Volume-Temperature Correlation for Mixtures of California Oils and Gases. Drill & Prod. Pract., API, pp. 275–287 (1947)

[8] Glaso, O.: Generalized Pressure-Volume Temperature Correlations. Journal of Petroleum Technology, 785 (May 1980)

[9] Al-Marhoun, M.A.: New Correlation for formation Volume Factor of oil and gas Mixtures. Journal of Canadian Petroleum Technology, 22 (March 1992)

[10] Almehaideb, R.A.: Improved PVT Correlations For UAE Crude Oils. In: The 1997 SPE Middle East Oil Show and Conference, Bahrain (1997)

[11] Elsharkawy, A.M., Elgibaly, A.A., Alikhan, A.A.: Assessment of the PVT Correlations for Predicting the Properties of the Kuwaiti Crude Oils. Journal of Petroleum Science & Engineering 13, 219 (1995)

[12] Mahmood, M.M., Al-Marhoun, M.A.: Evaluation of empirically derived PVT properties for Pakistani crude oils. Journal of Petroleum Science & Engineering 16, 275 (1996)

[13] Hanafy, H.H., et al.: Empirical PVT Correlation Applied to Egyptian Crude Oils Exemplify Significance of Using Regional Correlations. In: SPE Oilfield Chemistry International Symposium, Houston (1997)

[14] Al-Shammasi, A.A.: Bubble Point Pressure and Oil Formation Volume Factor Correlations. In: SPE Middle East Oil Show and Conference, Bahrain (1997)

[15] Al-Shammasi, A.A.: A Review of Bubblepoint Pressure and Oil Formation Volume Factor Correlations. SPE Reservoir Evaluation & Engineering, 146–160 (2001)

[16] Varotsis, N., Gaganis, V., Nighswander, J.: A Novel Non-Iterative Method for the Prediction of the PVT Behavior of Reservoir Fluids. In: SPE Annual Technical Conference and Exhibition, Houston, Texas (1999)

[17] Gharbi, R.B., Elsharkawy, A.M.: Neural-Network Model for Estimating the PVT Properties of Middle East Crude Oils. In: SPE Middle East Oil Show and Conference, Bahrain (1997)

[18] Al-Marhoun, M.A., Osman, E.A.: Using Artificial Neural Networks to Develop New PVT Correlations for Saudi Crude Oils. In: 10th Abu Dhabi International Petroleum Exhibition and Conference (ADIPEC), Abu Dhabi, UAE (2002)

[19] Elsharkawy, A.M.: Modeling the Properties of Crude Oil and Gas Systems Using RBF Network. In: SPE Asia Pacific Oil & Gas Conference, Perth, Australia (1998)

[20] Osman, E.A., Abdel-Aal, R.E.: Abductive Networks: A New Modeling Tool for the Oil and Gas Industry. inAsia Pacific Oil and Gas Conference and Exhibition. Melbourne, Australia (2002)

An Extensional Signed Fuzzy Measure of Signed Rho-Fuzzy Measure

Hsiang-Chuan Liu

Department of Bioinformatics and Medical Informatics, Asia University
Graduate Institute of Acupuncture Science, China Medical University
41354 Taichung, Taiwan, ROC
lhc@asia.edu.tw

Abstract. If some values of fuzzy density function are negative, none of non-negative fuzzy measure can be used, the signed fuzzy measures with real valued fuzzy density function are needed, a univalent signed fuzzy measure satisfying Liu's revised monotonicity, called signed Rho-measure, was proposed by author's previous work. In this paper, for any real valued fuzzy density function, it is proved that the well-known signed additive measure is a signed fuzzy measure satisfying the Liu's revised monotonicity, and a multivalent signed fuzzy measure with infinite many signed fuzzy measure solutions satisfying Liu's revised monotonicity, based on signed Rho-measure, called extensional signed Rho-fuzzy measure, is proposed, this new signed fuzzy measure is an generalization of not only signed Rho-fuzzy measure but also signed addition measure, obviously, it is more useful than above mentioned two signed fuzzy measures, some related properties are also discussed.

Keywords: Fuzzy measure, revised monotonicity, signed fuzzy measure, signed additive measure, signed L-fuzzy measure.

1 Introduction

The most important key issue in the application of Choquet integral [1-3, 6-9] is how to decide an adequate fuzzy measure with the determination of fuzzy density function.

If the given fuzzy density function is a non-negative function, then some useful measures satisfying the monotone condition, called fuzzy measures or monotone measures, may be used, such as the well-known fuzzy measure; Sugeno's λ-measure [3], the smallest fuzzy measure; Zahda's P-measure [4], the largest fuzzy measure; the author's B-measure [7], and the multivalent fuzzy measure with infinite many fuzzy measure solutions including all of above mentioned fuzzy measures; the author's L-measure and completed L-measure [8-9], obviously, the last one is more useful than others. Some times, if the given fuzzy density function is no more a non-negative valued function, then all of above mentioned non-negative fuzzy measures can not be used, and Choquet integral with respect to signed fuzzy measures based on some signed fuzzy density functions are needed [2, 8-12], a univalent signed fuzzy measure satisfying Liu's revised monotonicity, called signed ρ-fuzzy measure, was proposed by author's previous work [8-9].

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 509–519, 2010.

In this paper, for any real valued fuzzy density function, it is proved that the well-known signed additive measure is a signed fuzzy measure satisfying the Liu's revised monotonicity, and a multivalent signed fuzzy measure with infinite many signed fuzzy measure solutions satisfying Liu's revised monotonicity, based on signed ρ -fuzzy measure, called extensional signed ρ -fuzzy measure, is proposed , this new signed fuzzy measure is an generalization of signed ρ -fuzzy measure including not only signed ρ -fuzzy measure but also the signed additive measure, some related properties are also discussed.

This paper is outlined as follows: Signed fuzzy density function, signed fuzzy measures, and revised monotonicity are introduced in section 2. Signed ρ -fuzzy measure is introduced in section 3, The signed additive measures and its property are introduced in section 4. the new signed fuzzy measure; extensional signed ρ -fuzzy measure, and its some theorems are provided in Section 5. Finally, conclusions are summarized in Section 6.

2 Signed Fuzzy Measures

A generalization of fuzzy measures by allowing that the measure function can take negative values, leads to the notion of signed measures for using to non- homogeneous interactions among independent variables, a strict signed measures, called signed fuzzy measure, was introduced by B. Jiao [10], investigated and modified by X. Liu [11], E. Pap emphasized the revised monotonicity of Signed fuzzy measure [12]. In this paper, we emphasize the Normalizing properties.

2.1 Signed Measures

Definition 1. Signed measure **[2, 5]**

A signed measure, μ -measure, on a finite set X, if its measure function $g_\mu : 2^X \to (-\infty, \infty)$ satisfying the following axioms:

$$g_\mu(\phi) = 0 \quad \text{(vanishing at the empty set)} \tag{1}$$

2.2 Signed Fuzzy Density Function

Definition 2. Signed fuzzy density function

A signed fuzzy density function of a signed measure μ on a finite set X is a signed fuzzy density function $d : X \to [-1,1]$ satisfying:

1) $$d(x) = g_\mu(\{x\}), x \in X \tag{2}$$

2) $$|X| = n = n_+ + n_-,\ n, n_+ \in N, n_- \in N \cup \{0\} \tag{3}$$

3) $$d(x)=\begin{cases} d(x_i^+)>0,\ x=x_i^+,\ i=1,2,\ldots,n_+ \\ d(x_j^-)<0,\ x=x_j^-,\ j=1,2,\ldots,n_- \end{cases} \quad (4)$$

$d(x)$ is called the signed fuzzy density of singleton x

Definition 3. Normalized signed fuzzy density function

A normalized signed fuzzy density function of a signed measure μ on a finite set X is a signed fuzzy density function $d: X \to [-1,1]$ satisfying:

$$\sum_{x\in X_+} d(x)=1, -1\le \sum_{x\in X_-} d(x)\le 0 \quad (5)$$

where $X_+=\{x\,|\,d(x)\ge 0, x\in X\}, X_-=\{x\,|\,d(x)<0, x\in X\}$

Note that if $n_-=0$, then

1) The signed fuzzy density function is just a fuzzy density function.
2) The normalized signed fuzzy density function is just a normalized fuzzy density function

2.3 Signed Fuzzy Measure

Definition 4. Signed fuzzy measure **[10-12]**

A signed measure, μ-measure, on a finite set X, is called a signed fuzzy measure, if its measure function $g_\mu: 2^X \to R$ satisfying the following axioms:

1) $$g_\mu(\phi)=0, g_\mu(X_+)=1, -1\le g_\mu(X_+)\le 0 \text{ (Boundary conditions)} \quad (6)$$

where $X_+=\{x\,|\,g_\mu(\{x\})\ge 0, x\in X\}, X_-=\{x\,|\,g_\mu(\{x\})<0, x\in X\}$

2) $\forall A,B\subseteq X, A\cap B=\phi$ (Liu's revised monotonicity)

i) $$g_\mu(A), g_\mu(B)\ge 0,\ g_\mu(A)\vee g_\mu(B)>0 \Rightarrow g_\mu(A\cup B)\ge g_\mu(A)\vee g_\mu(B) \quad (7)$$

ii) $$g_\mu(A), g_\mu(B)\le 0,\ g_\mu(A)\wedge g_\mu(B)<0$$

$$\Rightarrow g_\mu(A\cup B)\le g_\mu(A)\wedge g_\mu(B) \quad (8)$$

iii) $$g_\mu(A)<0<g_\mu(B)\Rightarrow g_\mu(A)\le g_\mu(A\cup B)\le g_\mu(B) \quad (9)$$

Note that signed fuzzy measure is a special case of signed measure.

Theorem 1.

1) A Signed fuzzy measure is a generalization of a fuzzy measure

2) A normalized signed fuzzy measure is a generalization of normalized fuzzy measure

Proof: if n_- =0 then the signed fuzzy measure is just a non-negative fuzzy measure, and the normalized signed fuzzy measure is just a normalized non-negative fuzzy measure.

3 Signed Additive Measure

For any given signed fuzzy density function $d:X\to[-1,1]$ on a finite set

$X=\left\{x_1^+,x_2^+,...x_{n_+}^+,x_1^-,x_2^-,...x_{n_-}^-\right\}$, where $|X|=n=n_++n_-$, $n,n_+\in N, n_-\in N\bigcup\{0\}$

To construct a normalized signed fuzzy measure on X, satisfying Liu's revised monotonicity, we need to specify 2^n-n-1 coefficients satisfying $n_+2^{n_+-1}\left[1+\left(n_-2^{n_--1}\right)\right]$ revised monotone conditions, it is more difficulty than the case of fuzzy measure. The signed additive measure is the well-known one.

Definition 5. Signed additive measure [5]

For any given signed fuzzy density function,

$d:X\to[-1,1]$, The signed additive measure, denoted as S_a-measure, its measure function: $g_{S_a}:2^X\to R$ satisfies

$$g_{S_a}(A)=\sum_{x\in A}d(x)=\sum_{x\in A}g_{S_a}(\{x\}), \forall A\subseteq X \tag{10}$$

Theorem 2. Signed additive measure is a signed fuzzy measure

Proof:

1) To prove $g_{S_a}(\phi)=0$

$$g_{S_a}(\phi)=\sum_{x\in\phi}d(x)=0$$

2) To prove the revised monotonicity

Let $\forall A,B\subseteq X$, $\ni A\bigcap B=\phi$, then we have

$$g_{S_a}(A\bigcup B)=\sum_{x\in A\bigcup B} d(x)=\sum_{x\in A} d(x)+\sum_{x\in B} d(x)=g_{S_a}(A)+g_{S_a}(B) \tag{11}$$

i) if $g_{S_a}(A),g_{S_a}(B)\geq 0, g_{S_a}(A)\vee g_{S_a}(B)>0\Rightarrow$ then

$$g_{S_a}(A\bigcup B)=g_{S_a}(A)+g_{S_a}(B)\geq g_{S_a}(A),g_{S_a}(B)\Rightarrow g_{S_a}(A\bigcup B)\geq\left[g_{S_a}(A)\vee g_{S_a}(B)\right] \tag{12}$$

ii) if $g_{S_a}(A),g_{S_a}(B)\leq 0, g_{S_a}(A)\wedge g_{S_a}(B)<0$ then

$$g_{S_a}(A\bigcup B)=g_{S_a}(A)+g_{S_a}(B)\leq g_{S_a}(A),g_{S_a}(B)\Rightarrow g_{S_a}(A\bigcup B)\leq\left[g_{S_a}(A)\wedge g_{S_a}(B)\right] \tag{13}$$

iii) if $g_{S_a}(A)\leq 0\leq g_{S_a}(B)$ then

$g_{S_a}(A\bigcup B)=g_{S_a}(A)+g_{S_a}(B)\geq g_{S_a}(A)$ and $g_{S_a}(A\bigcup B)=g_{S_a}(A)+g_{S_a}(B)\leq g_{S_a}(B)$

We obtain $g_{S_a}(A)\leq g_{S_a}(A\bigcup B)\leq g_{S_a}(B)$

The proof is completed.

Definition 6. Normalized signed additive measure

For any given normalized signed fuzzy density function, $d:X\rightarrow[-1,1]$, satisfying

$$\sum_{x\in X_+} d(x)=1,\ -1\leq\sum_{x\in X_-} d(x)\leq 0$$

where $X_+=\{x\,|\,d(x)\geq 0, x\in X\}, X_-=\{x\,|\,d(x)<0, x\in X\}$,

Its signed additive measure is called a normalized signed additive measure, denoted as S_{Na} -measure.

Theorem 3. Normalized signed additive measure is a normalized signed fuzzy measure

Proof: It is trivial.

4 Signed ρ -Fuzzy Measure

A signed fuzzy measure satisfying Liu's revised monotonicity rather then signed additive measure, called signed ρ-measure, was proposed by author's previous work [8-9], its formal definition is listed as below;

Definition 7. Signed ρ - fuzzy measure

For a given signed fuzzy density function $d:X\rightarrow[-1,1]$ on a finite set X, the set function $g_\rho:2^X\rightarrow R$ is called a signed ρ-measure, if it satisfies

1) $g_\rho(\phi)=0,\ g_\rho(X_+)=1, -1\le g_\rho(X_-)\le 0$ (Boundary conditions) (14)

where $X_+=\{x\,|\,d(x)\ge 0, x\in X\}, X_-=\{x\,|\,d(x)<0, x\in X\}$

2) $\forall A,B\in 2^X,\ A\cap B=\phi,\ A\cup B\ne X$

$$\Rightarrow g_\rho(A\cup B)=g_\rho(A)+g_\rho(B)-g_\rho(A)g_\rho(B) \tag{15}$$

3) $$\forall A\in 2^X,\ g_\rho(A)=1-\prod_{x\in A}(1-d(x)) \tag{16}$$

Definition 8. Normalized signed ρ - fuzzy measure

For a given normalized signed fuzzy density function $d: X\to[-1,1]$ on a finite set X, its signed ρ - fuzzy measure is called a normalized signed ρ-measure,

Theorem 4. Signedρ-fuzzy measure and normalized signed ρ-fuzzy measure are signed fuzzy measures satisfying Liu's revised monotonicity. [9]

5 Extensional Signed ρ -Fuzzy Measure

In this paper, a generalization of both signed additive measure and signedρ-fuzzy measure, called extensional signedρ-fuzzy measure, is proposed as follows;

5.1 Definition of Extensional Signed ρ -Fuzzy Measure

Definition 9. Extensional signed ρ - fuzzy measure

For a given signed fuzzy density function $d: X\to[-1,1]$ on a finite set X, $0\le\rho\le 1$, the set function $g_\rho: 2^X\to R$ is called an extensional signed ρ-measure, if it satisfies

1) $g_\rho(\phi)=0,\ g_\rho(X_+)=1, -1\le g_\rho(X_-)\le 0$ (Boundary conditions)

where $X_+=\{x\,|\,d(x)\ge 0, x\in X\}, X_-=\{x\,|\,d(x)<0, x\in X\}$

2) $\forall A,B\in 2^X,\ A\cap B=\phi\ \Rightarrow g_\rho(A\cup B)=g_\rho(A)+g_\rho(B)-\rho g_\rho(A)g_\rho(B)$ (17)

3) $\forall A\subset X, 0\le\rho\le 1\Rightarrow$

$$g_\rho(A) = \begin{cases} \sum_{x \in A} d(x) & if \ \rho = 0 \\ \sum_{B \subset A} \left[(-\rho)^{|B|-1} \prod_{x \in B} d(x) \right] & if \ 0 < \rho \leq 1 \end{cases} \tag{18}$$

where $d(x) = g_\rho(\{x\}), x \in X$

ρ is also called the determine parameter of this measure.

5.2 Important Properties of Extensional Signed ρ -Fuzzy Measure

Theorem 5. Extensional signed ρ - fuzzy measure is well defined.

Proof: We need only to prove that Formula (17) is the necessary and sufficient conditions of Formula (18), if $\rho = 0$, it is trivial, now we suppose $0 < \rho \leq 1$

1) First to prove the necessary condition by using the mathematical induction, that is, let $A_m = \{x_1, x_2, ..., x_m\} \in 2^X$, using the condition (17), to prove that

$$g_\rho(A) = \sum_{B \subset A} \left[(-\rho)^{|B|-1} \prod_{x \in B} d(x) \right] \tag{19}$$

i) Let $m = 2$, $A_2 = \{x_1, x_2\} \in 2^X$, then

$$\begin{aligned} g_\rho(A_2) &= g_\rho(\{x_1, x_2\}) = \left[g_\rho(\{x_1\}) + g_\rho(\{x_2\}) \right] - \rho g_\rho(\{x_1\}) g_\rho(\{x_2\}) \\ &= \sum_{B \subset A_2} \left[(-\rho)^{|B|-1} \prod_{x \in B} d(x) \right] \end{aligned} \tag{20}$$

ii) Let $m = k$,

$$A_k = \{x_1, x_2, ..., x_k\} \in 2^X \ satisfies \ g_\rho(A_k) = \sum_{B \subset A_k} \left[(-\rho)^{|B|-1} \prod_{x \in B} d(x) \right], \tag{21}$$

$$A_{k+1} = \{x_1, x_2, ..., x_k, x_{k+1}\} \in 2^X, A_{k+1} = A_k \bigcup \{x_{k+1}\}, A_k \bigcap \{x_{k+1}\} = \phi \tag{22}$$

then $$g_\rho(A_{k+1}) = g_\rho(A_k \bigcup \{x_{k+1}\}) = g_\rho(A_k) + g_\rho(\{x_{k+1}\}) - \rho g_\rho(A_k) g_\rho(\{x_{k+1}\})$$

$$= \sum_{B \subset A_k} \left[(-\rho)^{|B|-1} \prod_{x \in B} d(x) \right] + g_\rho(\{x_{k+1}\}) - \rho g_\rho(\{x_{k+1}\}) \sum_{B \subset A_k} \left[(-\rho)^{|B|-1} \prod_{x \in B} d(x) \right] \tag{23}$$

$$= \sum_{B \subset A_{k+1}} \left[(-\rho)^{|B|-1} \prod_{x \in B} d(x) \right], \tag{24}$$

iii) From i) and ii), it completed the proof

2) Second to prove the sufficient condition by using the mathematical induction, that is, let $A_m = \{x_1, x_2, ..., x_m\} \in 2^X$, using the condition (19), to prove that formula (17) is true.

i) Let $m = 2$, $A_2 = \{x_1, x_2\} \in 2^X$, then

$$\begin{aligned} g_\rho(\{x_1, x_2\}) &= g_\rho(A_2) = \sum_{B \subset A_2} \left[(-\rho)^{|B|-1} \prod_{x \in B} d(x) \right] \\ &= \left[g_\rho(\{x_1\}) + g_\rho(\{x_2\}) \right] - \rho g_\rho(\{x_1\}) g_\rho(\{x_2\}) \\ &= g_\rho(\{x_1\}) + g_\rho(\{x_2\}) - \rho g_\rho(\{x_1\}) g_\rho(\{x_2\}) \end{aligned}$$

ii) Let $m = k$,

$$A_k = \{x_1, x_2, ..., x_k\}, A_{k+1} = \{x_1, x_2, ..., x_k, x_{k+1}\} \in 2^X, \text{ satisfies} \tag{25}$$

$$g_\rho(A_{k+1}) = \sum_{B \subset A_{k+1}} \left[(-\rho)^{|B|-1} \prod_{x \in B} d(x) \right], g_\rho(A_k) = \sum_{B \subset A_k} \left[(-\rho)^{|B|-1} \prod_{x \in B} d(x) \right]$$

Since $g_\rho(A_k \cup (x_{k+1})) = g_\rho(A_{k+1}) = \sum_{B \subset A_{k+1}} \left[(-\rho)^{|B|-1} \prod_{x \in B} d(x) \right]$

$$\begin{aligned} &= \sum_{B \subset A_k} \left[(-\rho)^{|B|-1} \prod_{x \in B} d(x) \right] + g_\rho(\{x_{k+1}\}) - \rho g_\rho(\{x_{k+1}\}) \sum_{B \subset A_k} \left[(-\rho)^{|B|-1} \prod_{x \in B} d(x) \right] \\ &= g_\rho(\{A_k\}) + g_\rho(\{x_{k+1}\}) - \rho g_\rho(\{A_k\}) g_\rho(\{x_{k+1}\}) \end{aligned} \tag{26}$$

iii) Not lose the generality, from i) and ii), it completed the proof.

Theorem 7. For any $\rho \in [0,1]$, extensional signed ρ - fuzzy measure is a signed fuzzy measure

Proof: We need only to prove that it satisfies Liu's revised monotonicity as below;

Let $\forall A, B \in 2^X$, $A \cap B = \phi$, $A \cup B \neq X$

1) If $g_\mu(A), g_\mu(B) \geq 0$, $g_\mu(A) \vee g_\mu(B) > 0$, then form (17), we can get

$$g_\rho(A\bigcup B)=g_\rho(A)+g_\rho(B)-\rho g_\rho(A)\,g_\rho(B)=g_\rho(A)+g_\rho(B)\big[1-\rho g_\rho(A)\big]\geq g_\rho(A) \quad (27)$$

$$g_\rho(A\bigcup B)=g_\rho(A)\big[1-\rho g_\rho(B)\big]+g_\rho(B)\geq g_\rho(B) \quad (28)$$

Then $g_\rho(A\bigcup B)\geq\big[g_\rho(A)\vee g_\rho(B)\big]$

2) If $g_\mu(A), g_\mu(B)\leq 0,\ g_\mu(A)\wedge g_\mu(B)<0$, then form (17), we can get

$g_\rho(A\bigcup B)=g_\rho(A)+g_\rho(B)\big[1-\rho g_\rho(A)\big]\leq g_\rho(A)$ and $g_\rho(A\bigcup B)=g_\rho(A)\big[1-\rho g_\rho(B)\big]+g_\rho(B)\leq g_\rho(B)$

Then $g_\rho(A\bigcup B)\leq\big[g_\rho(A)\wedge g_\rho(B)\big]$

3) If $g_\mu(A)<0<g_\mu(B)$, then

$g_\rho(A\bigcup B)=g_\rho(A)+g_\rho(B)\big[1-\rho g_\rho(A)\big]\geq g_\rho(A)$ and $g_\rho(A\bigcup B)=g_\rho(A)\big[1-\rho g_\rho(B)\big]+g_\rho(B)\leq g_\rho(B)$

Hence $g_\rho(A)\leq g_\rho(A\bigcup B)\leq g_\rho(B)$

The proof is completed.

Theorem 8

1) Extensional signed ρ- fuzzy measure is a multivalent signed fuzzy measure with infinite many signed fuzzy measure solutions.

2) The measure function of extensional signed ρ-fuzzy measure is a deceasing and continuous function of ρ on [0,1].

3) If ρ =0, then extensional signed ρ-fuzzy measure is just a signed additive measure.

4) If ρ =1, then extensional signed ρ-fuzzy measure is just a signed ρ-fuzzy measure.

5) If $X=X_+, X_-=\phi$, then extensional signed ρ-fuzzy measure is just a fuzzy measure.

Proof: Omitted for lack of space.

5.3 Extensional Normalized Signed ρ-Fuzzy Measure

Definition 10. Extensional normalized signed ρ- fuzzy measure

For a given signed fuzzy density function $d : X \to [-1,1]$ on a finite set X, $0 \le \rho \le 1$, an extensional signed ρ- fuzzy measure is called an extensional normalized signed ρ-measure, if it satisfies

$$\sum_{x \in X_+} d(x) = 1, -1 \le \sum_{x \in X_-} d(x) \le 0$$

Where $X_+ = \{x \mid d(x) \ge 0, x \in X\}, X_- = \{x \mid d(x) < 0, x \in X\}, d(x) = g_\rho(\{x\}), x \in X$

Theorem 9.

1) Extensional normalized signed ρ- fuzzy measure is a multivalent signed fuzzy measure with infinite many signed fuzzy measure solutions.
2) The measure function of extensional normalized signed ρ-fuzzy measure is a deceasing and continuous function of ρ on [0,1].
3) If ρ =0, then extensional normalized signed ρ-fuzzy measure is just a normalized signed additive measure.
4) If ρ =1, then extensional signed ρ-fuzzy measure is just a normalized signed ρ -fuzzy measure.
5) If $X = X_+, X_- = \phi$, then extensional signed ρ-fuzzy measure is just a normalized fuzzy measure.

Proof: It is trivial.

6 Conclusion

In this paper, first, the author proves that the well-known signed additive measure is a university signed fuzzy measure satisfying the Liu's revised monotonicity, and points out that the signed Rho-fuzzy measure proposed by author' previous work, is also a university signed fuzzy measure satisfying the Liu's revised monotonicity. Second, a multivalent signed fuzzy measure with infinite many signed fuzzy measure solutions also satisfying Liu's revised monotonicity, called extensional signed Rho-fuzzy measure, is proposed, this new signed fuzzy measure is an generalization of not only signed Rho-fuzzy measure but also signed addition measure, obviously, it is more useful than above mentioned two signed fuzzy measures. Third, some related properties are also discussed. In the future, Chquet integral with respect to this new signed fuzzy measure will be considered to analysis multi-criteria decision making problems.

Acknowledgments. This paper is partially supported by the grant of National Science Council of Taiwan Government (NSC 98-2410-H-468-014).

References

1. Choquet, G.: Theory of capacities. Annales de l'Institut Fourier 5, 131–295 (1953)
2. Wang, Z., Klir, G.J.: Fuzzy Measure Theory. Plenum Press, New York (1992)
3. Sugeno, M.: Theory of fuzzy integrals and its applications, unpublished doctoral dissertation, Tokyo Institute of Technology, Tokyo, Japan (1974)
4. Zadeh, L.A.: Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems 1, 3–28 (1978)
5. Halmos, P.R.: Measure Theory (19 C), 81-82
6. Liu, H.-C., Chen, C.-C., Wu, D.-B., Sheu, T.-W.: Theory and Application of the Composed Fuzzy Measure of L-Measure and Delta-Measures. Wseas Transaction On International Science And Control 4(8), 359–368 (2009)
7. Liu, H.-C.: Extensional Completed L- Measure Based on any Given Fuzzy Measure. In: Proceeding of the 3rd International Conference on Intelligent Information Technology Application (IITA 2009), Nanchang, China, November 21-22, vol. 2, pp. 602–605 (2009) ISBN 978-0-7695-3859-4
8. Liu, H.-C.: A useful theorem for constructing fuzzy measures and fuzzy signed measures. In: The 19th south Taiwan Statistics Conference, Tainan, Taiwan 7.6-7 (2010)
9. Liu, H.-C., Liu, T.-S.: A novel fuzzy measure and its signed fuzzy measure. In: The 10th WSEAS International Conference on Systems Theoryand Scientific Computation (ISTASC 2010), Taipei, Taiwan, August 20-22 (2010)
10. Jiao, B.: Hahn decomposition theory for signed fuzzy measure, Thesis, Hebei University (1992)
11. Liu, X.: Hahn decomposition theory for infinite signede fuzzy measure. Fuzzy Sets and Systems 1, 377–380 (1993)
12. Pap, E.: The Jordan decomposition of the null-additive signed fuzzy measure. Novi. Sad. J. Math. 30(1), 1–7 (2000)

Adaptive Fuzzy Neural Network Control for Automatic Landing System

Jih-Gau Juang* and Li-Hsiang Chien

Department of Communications, Navigation and Control Engineering
National Taiwan Ocean University, Keelung, Taiwan
jgjuang@mail.ntou.edu.tw

Abstract. This paper presents an intelligent automatic landing system that uses adaptive fuzzy neural network controller to improve the performance of conventional automatic landing systems. Functional fuzzy rules are implemented in neural network. In this study, Lyapunov stability theory is utilized to derive adaptive learning rate in the controller design. Stability of the control system is guaranteed. Simulation results show that the fuzzy neural network controller with adaptive learning rate has better performance than PID controller in guiding aircraft to a safe landing in turbulence condition.

1 Introduction

According to the report of NTSB (National Transportation Safety Board) [1], in 2008, there were a total of 1559 aircraft accidents that occurred in the United States. Of these 1559 accidents, 275 were fatal events and fatalities were 495. An accident survey of 1,300 aircraft accidents from 1950 through 2008 categorized the causes [2]. Weather was a contributing factor, the percentage of weather related to total accidents is 12%. The atmospheric disturbances affect not only fly qualities of an airplane but also flight safety. An inadvertent encounter with a low-altitude wind disturbance can be a serious problem even for a skilled pilot. In some cases, pilots will try to abort landing or escape the disturbance [3-4]. For those pilots that are unable to abort landing must face the disturbance and handle the aircrafts manually. It is therefore desirable to develop an intelligent Automatic Landing System (ALS) that expands the operational envelope to include more safe responses under severe wind conditions. The earliest automatic pilots could do no more than maintain an aircraft in straight and level flight by controlling pitch, yaw, and roll movements; and they are still used most often to relieve the pilot during routine cruising. Modern automatic pilots can, however, execute complex maneuvers or flight plans, bring aircraft into approach and landing paths. Conventional automatic landing systems can provide a smooth landing which is essential to the comfort of passengers. However, these systems work only within a specified operational safety envelope. When the conditions are beyond the envelope, such as turbulence or wind shear, they often cannot be used.

Most conventional control laws generated by the ALS are based on the gain scheduling method [5]. Control parameters are preset for different flight conditions

* Corresponding author.

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 520–530, 2010.

within a specified safety envelope which is relatively defined by Federal Aviation Administration (FAA) regulations. According to FAA regulations, environmental conditions considered in the determination of dispersion limits are: headwinds up to 25 knots; tailwinds up to 10 knots; crosswinds up to 15 knots; moderate turbulence, wind shear of 8 knots per 100 feet from 200 feet to touchdown [6]. If the flight conditions are beyond the preset envelope, the ALS is disabled and the pilot takes over. An inexperienced pilot may not be able to guide the aircraft to a safe landing. Recently, many researchers have applied intelligent concepts such as neural networks, fuzzy systems, and genetic algorithms to flight control to adapt to different environments. For ALS, most of the improvements have been on the guidance instruments, such as the GNSS Integrity Beacons, Global Positioning System, Microwave Landing System and Autoland Position Sensor [7-10]. By using improvement calculation methods and high accuracy instruments, these systems provide more accurate flight data to the ALS to make the landing more smooth. However, these researches did not include weather factors such as wind disturbances. There have been some researches on the problem of intelligent landing control [11-17] but most of them do not consider wind disturbances. Here, we present a learning scheme, which uses adaptive fuzzy neural network controller, to guide the aircraft to a safe landing and make the controller more robust and adaptive to the ever-changing environment.

2 Flight Control

To make the ALS more intelligent, reliable wind profiles are necessary. Two wind disturbance models are most common in aircraft flight paths. They are turbulence and wind shear. In this paper, we put focus on turbulence since it is the most encountered condition to deal with during aircraft landing. In this study the Dryden form [9] was used for its demonstration ease. The model is given by

$$u_g = u_{gc} + N(0,1)\sqrt{\frac{1}{\Delta t}}\left(\frac{\sigma_u\sqrt{2a_u}}{s+a_u}\right) \tag{1}$$

$$w_g = N(0,1)\sqrt{\frac{1}{\Delta t}}\left(\frac{\sigma_w\sqrt{3a_w}\,(s+b_w)}{(s+a_w)^2}\right) \tag{2}$$

where $u_{gc} = -u_{510}\left[1+\frac{\ln(h/510)}{\ln(51)}\right]$, $L_w = h$, $a_u = \frac{U_0}{L_0}$, $L_u = 100h^{\frac{1}{3}}$ for h>230, L_u =600 for $h \le$ 230, $a_w = \frac{U_0}{L_w}$, $\sigma_w = 0.2\left|u_{gc}\right|(0.5 + 0.00098h)$ for $0 \le h \le 500$, $\sigma_w = 0.2\left|u_{gc}\right|$ for h>500, $\sigma_u = 0.2\,|\,u_{gc}\,|$, $b_w = \frac{U_0}{L_w\sqrt{3}}$. The parameters are: u_g is the longitudinal wind velocity (ft/sec), w_g is the vertical wind velocity (ft/sec), U_0 is the nominal aircraft speed (ft/sec), u_{510} is the wind speed at 510 ft altitude, L_u and L_w are scale lengths (ft), σ_u and σ_w are RMS values of turbulence velocity (ft/sec), Δt is the simulation time step (sec), $N(0, 1)$ is the Gaussian white noise with zero mean and unity standards deviation, u_{gc} is the constant component of u_g, and h is the aircraft altitude (ft). Figure 1 shows a turbulence profile with a wind speed of 20 ft/sec at 510 ft altitude.

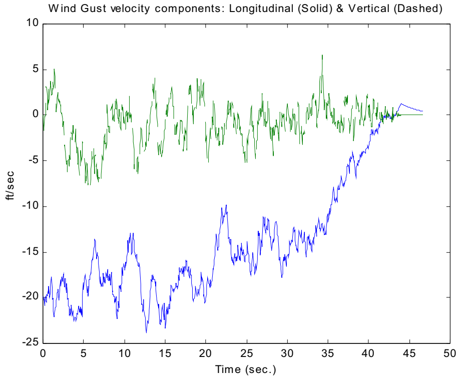


Fig. 1. Turbulence profile

The pilot descends from the cruise altitude to an altitude of approximately 1200 ft above the ground. The pilot then positions the airplane so that the airplane is on a heading towards the runway centerline. When the aircraft approaches the outer airport marker, which is about 4 nautical miles from the runway, the glide path signal is intercepted (as shown in Figure 2). As the airplane descends along the glide path, its pitch, attitude and speed must be controlled. The aircraft maintains a constant speed along the flight path. The descent rate is about 10 ft/sec and the pitch angle is between -5 to 5 degrees. As the airplane descends 20 to 70 feet above the ground, the glide path control system is disengaged and a flare maneuver is executed. The vertical descent rate is decreased to 2 ft/sec so that the landing gear may be able to dissipate the energy of the impact at landing. The pitch angle of the airplane is then adjusted, between 0 to 5 degrees for most aircraft, which allows a soft touchdown on the runway surface.

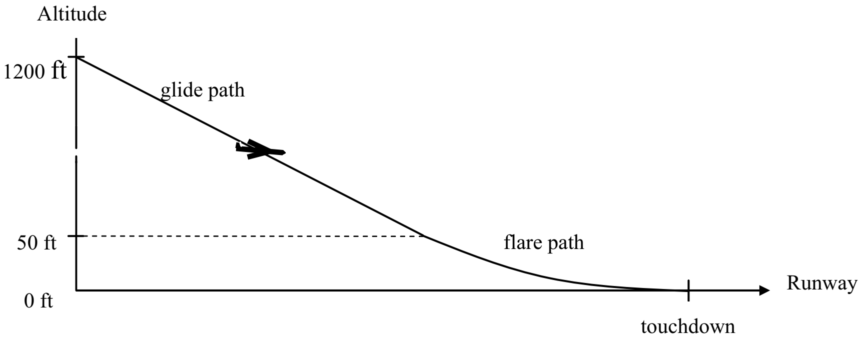


Fig. 2. Glide path and flare path

A simplified model of a commercial aircraft that moves only in the longitudinal and vertical plane was used in the simulations for implementation ease [9]. Since the aircraft maintains a constant speed along the flight path, we assumed that the change in throttle command is zero. The aircraft is thus controlled solely by the pitch command. Detailed descriptions can be found in [9]. A complete landing phase is divided into several stages (intervals). Each stage uses the same fuzzy neural network controller. Wind disturbances are added to each stage in the simulation. The learning process is shown in Figure 3, where FC is the fuzzy neural network controller and AM is the aircraft model. Every learning cycle consists of all stages from S_0 to S_k. The controller is trained by a modified learning-through-time process. The linearized inverse aircraft model (LIAM) calculates the error signals that will be used to back propagate through the controller in each stage [18]. The error $\Delta C_{\mathrm{k}-1}$ is back propagated through the FC to obtain the weight change. The $e_{k\text{-}1}$ is used to further derive $\Delta C_{\mathrm{k}-2}$ and $e_{k\text{-}2}$. The error continues to be back propagated through all k stages with weight changes computed for the controller in each stage. The weight changes from all of the stages obtained from the delta learning rule are added together for the overall update. The FC generates new commands to the AM and the AM responds to new flight conditions in return. This process is repeated until the aircraft is safely landed. The inputs for the fuzzy controller are: altitude, altitude command, altitude rate, and altitude rate command. The output of the controller is the pitch command.

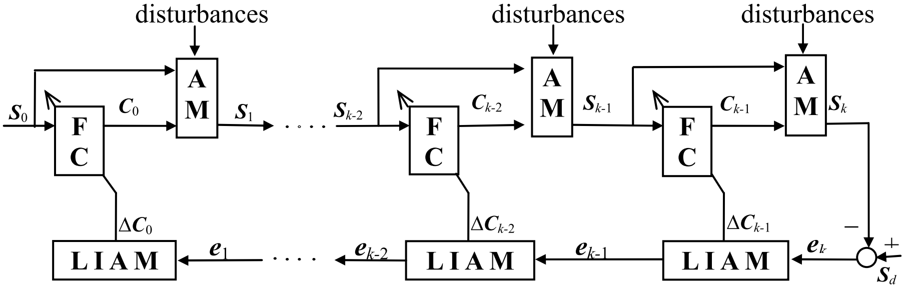


Fig. 3. Learning through time process

3 Adaptive Functional Fuzzy Neural Network Controller

The functional fuzzy neural network is similar to the linguistic fuzzy neural network in network structure. The difference lies in the consequence. The linguistic fuzzy rule has scalar value output, but the functional fuzzy rule is a polynomial function, which makes the resulting defuzzification approach a nonlinear system. The network consists of six layers, and the network structure is shown in Figure 4. The network realizes the following inference method:

$$R^i : \text{IF } x_1 \text{ is } A_{1_{i1}} \text{ and } x_2 \text{ is } A_{2_{i2}} \text{ and} \ldots \text{ and } x_4 \text{ is } A_{4_{i4}} \text{ then } y \text{ is } \varphi_i(x_1, x_2, x_3, x_4) \qquad (3)$$

where $i1$, $i2$, $i3$, and $i4$ is from 1 to 2; i is from 1 to 16. The function in consequence of the fuzzy rule can be any polynomial. In here we use first order linear equation:

$$\varphi_i(x_1, x_2, x_3, x_4) = a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4 \tag{4}$$

where a_i is a scalar and adjusted during network training. The connection weights are modified to identify fuzzy rules and tune the membership functions in the premises using the following rules:

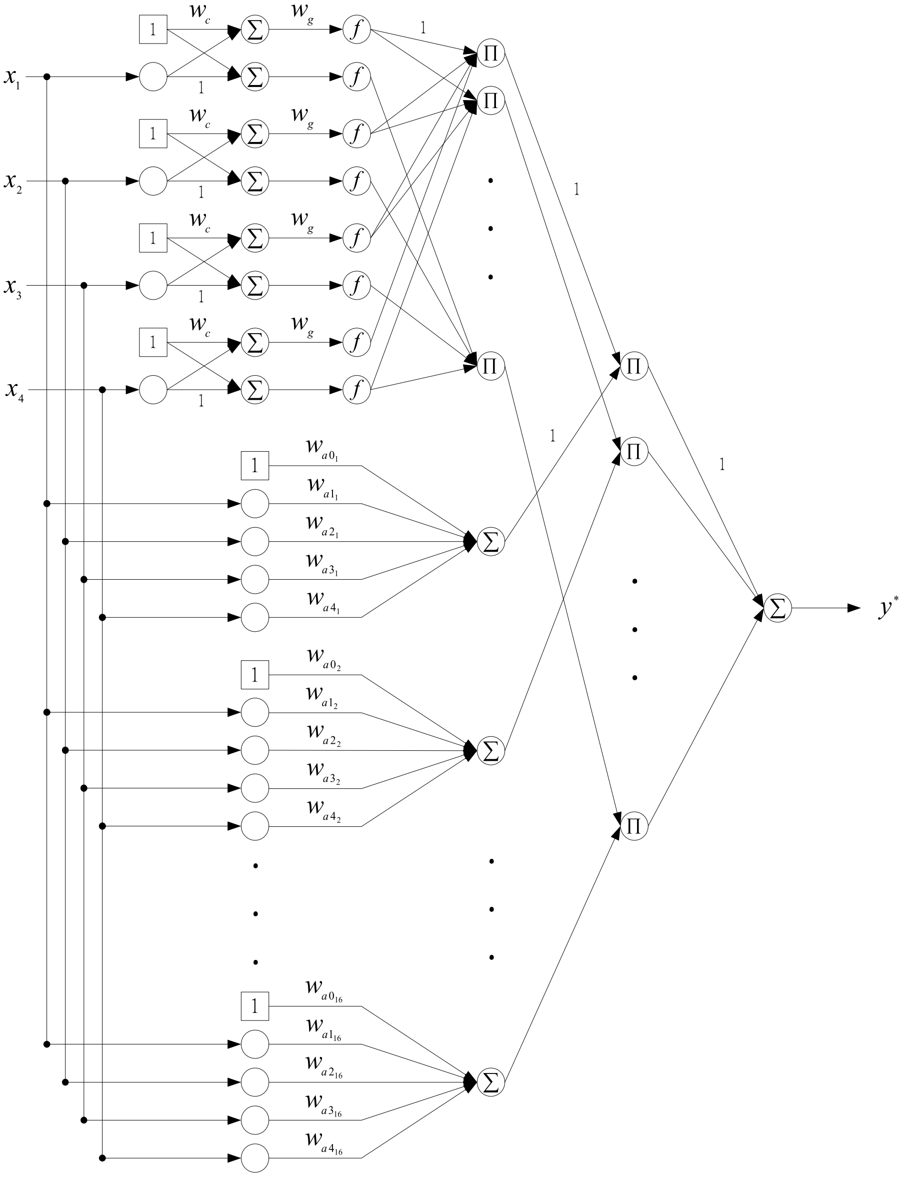


Fig. 4. Fuzzy neural network with functional rule

1. at sixth layer

$$\delta^6 = (d-y)f'(I^6) = d(n) - y(n) \tag{5}$$

2. at fifth layer

$$\delta_i^5 = f'(I_i^5)\sum_k \delta_k^6 \cdot 1 = \delta_k{}^6 , \; i = 1,2...16 , \; k = 1 \tag{6}$$

3. at fourth layer

$$\delta_i^4 = f'(I_i^4)\sum_k \left(\delta_k^5 \cdot \prod_{j\neq i} O_j^4 \right) = \frac{1}{\sum \mu_k} \left(\delta_k^5 \cdot \prod_{j\neq i} O_j^4 \right) \tag{7}$$

$$\delta_i^4 = f'(I_i^4)\sum_k \left(\delta_k^5 \cdot \prod_{j\neq i} O_j^4 \right) = \delta_k^5 \cdot \prod_{j\neq i} O_j^4 \tag{8}$$

$i = 1,2..16$, $j = 1,2...4$

$$w_{a_{ji}}(m+1) = w_{a_{ji}}(m) + \eta\delta^4 x_i \tag{9}$$

$$w_{a_{0i}}(m+1) = w_{a_{0i}}(m) + \eta\delta^4 1 , \tag{10}$$

4. at third layer

$$\delta_i^3 = f'(I_i^3)\sum_k \left(\delta_k^4 \cdot \prod_{j\neq i} O_j^3 \right) = O_3 \cdot (1 - O_3) \cdot \sum_k \left[\delta_k{}^4 \cdot \prod_{j\neq i} O_j^3 \right], \; k = 1,2.....8 \tag{11}$$

$$w_g(m+1) = w_g(m) + \eta\delta^3 O^2 \tag{12}$$

5. at second layer

$$\delta_i^2 = f'(I_i^2)\sum_k \delta_k^3 w_{ki}^3 = \delta_i^3 \cdot w_g , \; i = 1,2.....8 \tag{13}$$

$$w_c(m+1) = w_c(m) + \eta\delta^2 O^1 \tag{14}$$

Let η_i be the learning rate for the *i*th weight or the parameter of the neural network controller and let $P_{L,\max}$ be defined as $P_{L,\max} = \max_q \| P_L(q) \|$, where $P_L(q) = \frac{\partial y_L(q)}{\partial W_L}$, y_L is the output of neural network, $W_L = [w_1 \quad w_2 \quad ... \quad w_n]^T$ represents a change in an arbitrary weighting vector in R^n , and $\| \cdot \|$ is the Euclidean norm in R^n . Then convergence is guaranteed if η_i is chosen as $0 < \eta_i < \frac{2}{P_{L,\max}^2}$ [19].

The varying range of the learning rate η_a of w_a can be obtained by letting

$$P_{Laji}(q) = \frac{\partial y^*(q)}{\partial w_{aji}} = \hat{\mu}_i(q) x_j(q) \tag{15}$$

where i=1,2,…,16 , j=0,1,2,3,4 , w_{0i} is the bias, $\hat{\mu}_i$ is the normalized value of μ_r . Let $\overset{\wedge}{\mu}_{\max}=\max\limits_{q}\left|\hat{\mu}_r(q)\right|$, $\hat{X}=\max\limits_{q}\left|x_j(q)\right|$, we have

$$\left\|P_{La}(q)\right\| < \sqrt{16\times 5\left(\overset{\wedge}{\mu}{}^{2}_{\max}\hat{X}{}^{2}\right)} \tag{16}$$

where R is the normalized numbers of μ_r . Let

$$P_{La,\max} = \sqrt{16\times 5\left(\overset{\wedge}{\mu}{}^{2}_{\max}\hat{X}{}^{2}\right)} \tag{17}$$

Thus

$$0<\eta_a<\frac{2}{16\times 5\left(\overset{\wedge}{\mu}{}^{2}_{\max}\hat{X}{}^{2}\right)} \tag{18}$$

Similarly, the varying range of the learning rates η_g of w_g and η_c of W_c can be obtained as

$$0<\eta_g<\frac{2}{8\left(8\hat{O}{}^{(2)}_{\max}\hat{O}{}^{(3)}_{\max}\overset{\wedge}{\mu}_{\max}\right)^2} \text{ and } 0<\eta_c<\frac{2}{8\left(8\hat{O}{}^{(2)}_{\max}\hat{O}{}^{(3)}_{\max}W_{g,\max}\overset{\wedge}{\mu}_{\max}\right)^2} \tag{19}$$

In the FFNN, we choose a Lyapunov function expressed as:

$$V(q)=\frac{1}{2}e^2(q) \tag{20}$$

thus, the change in Lyapunov function is obtained by

$$\Delta V(q)=V(q+1)-V(q)=\frac{1}{2}\left[e^2(q+1)-e^2(q)\right]=\Delta e(q)\left[e(q)+\frac{1}{2}\Delta e(q)\right] \tag{21}$$

The error difference due to learning can be represented by

$$\Delta e(q)=e(q+1)-e(q)\approx\left[\frac{\partial e(q)}{\partial W_L}\right]^T\Delta W_L(q)=\left[\left(\frac{\partial e(q)}{\partial w_a}\right)\left(\frac{\partial e(q)}{\partial w_g}\right)\left(\frac{\partial e(q)}{\partial w_c}\right)\right]\begin{bmatrix}\Delta w_a(q)\\ \Delta w_g(q)\\ \Delta w_c(q)\end{bmatrix} \tag{22}$$

using the gradient descent method, we have

$$\Delta W_L(q)=-\eta_L\frac{\partial E(q)}{\partial W_L}=\eta_L e(q)\frac{\partial y(q)}{\partial W_L}=e(q)\begin{bmatrix}\eta_a & 0 & 0\\ 0 & \eta_g & 0\\ 0 & 0 & \eta_c\end{bmatrix}\begin{bmatrix}\dfrac{\partial y^*(q)}{\partial w_a}\\ \dfrac{\partial y^*(q)}{\partial w_g}\\ \dfrac{\partial y^*(q)}{\partial w_c}\end{bmatrix} \tag{23}$$

Let

$$P_{L,\max}\equiv\begin{bmatrix}P_{La,\max} & P_{Lg,\max} & P_{Lc,\max}\end{bmatrix}^T=\begin{bmatrix}\max\limits_{q}\left\|\dfrac{\partial y^*(q)}{\partial w_a}\right\| & \max\limits_{q}\left\|\dfrac{\partial y^*(q)}{\partial w_g}\right\| & \max\limits_{q}\left\|\dfrac{\partial y^*(q)}{\partial w_c}\right\|\end{bmatrix}^T$$

Then we know the asymptotic convergence is achieved if η_i are chosen to satisfy

$$0<\eta_i<\frac{2}{\left(P_{Li,\max}\right)^2},\quad i=a,g,c \tag{24}$$

From (18) and (19) we know the learning rates satisfy the above condition. Thus $\Delta V<0$ and ΔV are negative definite. The tracking error $e(q)$ is asymptotically stable, guaranteeing convergence of the learning process.

4 Simulations

In the simulations, initial flight conditions are: $h(0)$=500 ft, $\dot{x}(0)$ =235 ft/sec, $x(0)$=9240 ft, and $\gamma_0=-3$ degrees. With the wind speed of turbulence at 60 ft/sec, the horizontal position at touchdown is 997 ft, horizontal velocity is 234.7 ft/sec, vertical speed is –2.3 ft/sec, and pitch angle is 0.7 degrees, as shown in Figure 5 to Figure 7. Table 1 shows the comparison of using different controllers.

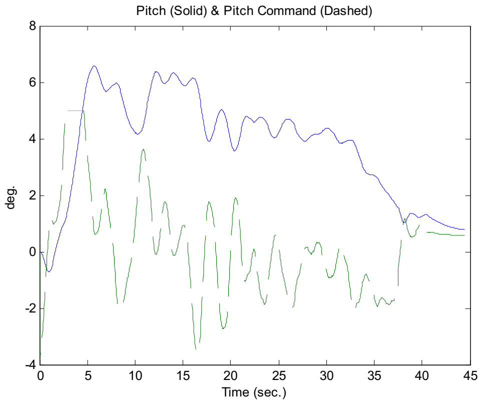


Fig. 5. Aircraft pitch and pitch command

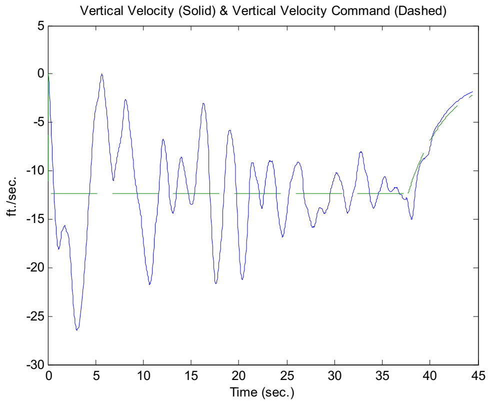


Fig. 6. Aircraft vertical velocity and velocity command

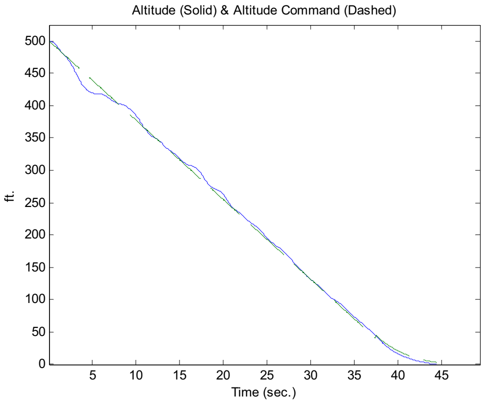


Fig. 7. Aircraft altitude and altitude command

Table 1. Simulation results under different turbulences

Controller	PID [17]	Neural Network [17]	Fuzzy Neural Network [20]	Adaptive Fuzzy Neural Network
Maximal wind speed (ft/sec)	30	40	55	60

5 Conclusion

The purpose of this study is to investigate the use of adaptive fuzzy neural network in automatic landing systems and to make it more intelligent. Current flight control law was adopted in the intelligent controller design. Tracking performance and robustness were demonstrated through software simulations. For the safe landing of an aircraft with a conventional PID controller, the wind speed turbulence limit is 30 ft/sec. From our previous study, a neural network controller can reach 40 ft/sec. The drawback is that the number of required neurons in the hidden layer is uncertain. In general, the number of the hidden neurons must be large enough to form a decision region as complex as required by a given problem. However, if the decision region is too large, then "over-fitting" occurs. In this study, the adaptive fuzzy neural network controller can overcome turbulence to 60 ft/sec without the problem on selecting the number of hidden neurons. The proposed fuzzy controller can successfully expand the controllable environment in severe wind disturbances.

References

1. NASDAC Review of NTSB Weather-Related Accidents, `http://www.ntsb.gov/aviation/Table10.htm`
2. Aircraft accident statistics, `http://www.planecrashinfo.com/cause.htm`
3. Dogan, A., Kabamba, P.T.: Escaping a Microburst with Turbulence. In: Proc. American Control Conference, pp. 1349–1353 (2000)
4. Chen, C.Y.: Optimal Abort Landing Trajectories for a High-Angle-Of-Attack Windshear Encounter. The Chinese Journal of Mechanics 11(1), 75–81 (1995)
5. Buschek, H., Calise, A.J.: Uncertainty Modeling and Fixed-Order Controller Design for a Hypersonic Vehicle Model. Journal of Guidance, Control, and Dynamics 20(1), 42–48 (1997)
6. Federal Aviation Administration, Automatic Landing Systems. AC 20-57A (January 1971)
7. Cohen, C.E., et al.: Automatic Landing of a 737 Using GNSS Integrity Beacons. In: Proc. ISPA (1995)
8. DDC-I: Advanced Auto Landing System from Swiss Federal Aircraft Factory. Real-Time Journal, Sprint (1995)
9. Asai, S., et al.: Development of Flight Control System for Automatic Landing Flight Experiment. Mitsubishi Heavy Industries Technical Review 34(3) (1997)
10. Kaufmann, D.N., McNally, B.D.: Flight Test Evaluation of the Stanford University and United Airlines Differential GPS Category III Automatic Landing System. NASA Technical Memorandum 110354 (June 1995)
11. Malaek, S.M.B., Sadati, N., Izadi, H., Pakmehr, M.: Intelligent Autolanding Controller Design Using Neural Networks and Fuzzy Logic. In: Proc. IEEE 5th Control Conference, vol. 1, pp. 365–373 (2004)
12. Izadi, H., Pakmehr, M., Sadati, N.: Optimal Neuro-Controller in Longitudinal Autolanding of a Commercial Jet Transport. In: Proc. IEEE International Conference on Control Applications, pp. 1–6 (2003)
13. Chaturvedi, D.K., Chauhan, R., Kalra, P.K.: Application of Generalized Neural Network for Aircraft Landing Control System. Soft Computing 6, 118–441 (2002)
14. Iiguni, Y., Akiyoshi, H., Adachi, N.: An Intelligent Landing System Based on Human Skill Model. IEEE Transactions on Aerospace and Electronic Systems 34, 877–882 (1998)

15. Ionita, S., Sofron, E.: The Fuzzy Model for Aircraft Landing Control. In: Proc. AFSS International Conference on Fuzzy Systems, pp. 47–54 (2002)
16. Nho, K., Agarwal, R.K.: Automatic Landing System Design Using Fuzzy Logic. Journal of Guidance, Control, and Dynamics 23, 298–304 (2000)
17. Jorgensen, C.C., Schley, C.: A Neural Network Baseline Problem for Control of Aircraft Flare and Touchdown. Neural Networks for Control, 403–425 (1991)
18. Juang, J.G., Chang, H.H.: Application of Time Delay Neural Network to Automatic Landing Control. In: Proc. of IEEE International Conference on Control Applications, vol. 1, pp. 150–155 (2002)
19. Lee, C.H., Teng, C.C.: Identification and Control of Dynamic Systems Using Recurrent Fuzzy Neural Networks. IEEE Trans. Fuzzy Sets and Systems 8(6) (August 2000)
20. Chin, K.C., Juang, J.G.: Application of Functional Fuzzy Neural Network to Aircraft Automatic Landing Control in Turbulence Condition. In: Proc. of Joint Conference on Fuzzy Systems and Grey Systems, A117 (2003) (in Chinese)

Author Index