Frontiers in Artificial Intelligence and Applications

APPLICATIONS OF DATA MINING IN E-BUSINESS AND FINANCE

Edited by
Carlos Soares
Yonghong Peng
Jun Meng
Takashi Washio
Zhi-Hua Zhou

IOS Press

APPLICATIONS OF DATA MINING IN E-BUSINESS AND FINANCE

Frontiers in Artificial Intelligence and Applications

FAIA covers all aspects of theoretical and applied artificial intelligence research in the form of monographs, doctoral dissertations, textbooks, handbooks and proceedings volumes. The FAIA series contains several sub-series, including “Information Modelling and Knowledge Bases” and “Knowledge-Based Intelligent Engineering Systems”. It also includes the biennial ECAI, the European Conference on Artificial Intelligence, proceedings volumes, and other ECCAI – the European Coordinating Committee on Artificial Intelligence – sponsored publications. An editorial panel of internationally well-known scholars is appointed to provide a high quality selection.

Volume 177

Recently published in this series

ISSN 0922-6389

Applications of Data Mining in E-Business and Finance

Edited by

Carlos Soares

University of Porto, Portugal

Yonghong Peng

University of Bradford, UK

Jun Meng

University of Zhejiang, China

Takashi Washio

Osaka University, Japan

and

Zhi-Hua Zhou

Nanjing University, China

IOS Press

Amsterdam • Berlin • Oxford • Tokyo • Washington, DC

ISBN 978-1-58603-890-8
Library of Congress Control Number: 2008930490

Publisher
IOS Press
Nieuwe Hemweg 6B
1013 BG Amsterdam
Netherlands
fax: +31 20 687 0019
e-mail: order@iospress.nl

Distributor in the UK and Ireland
Gazelle Books Services Ltd.
White Cross Mills
Hightown
Lancaster LA1 4XS
United Kingdom
fax: +44 1524 63232
e-mail: sales@gazellebooks.co.uk

Distributor in the USA and Canada
IOS Press, Inc.
4502 Rachael Manor Drive
Fairfax, VA 22032
USA
fax: +1 703 323 3668
e-mail: iosbooks@iospress.com

PRINTED IN THE NETHERLANDS

Applications of Data Mining in E-Business and Finance
C. Soares et al. (Eds.)
IOS Press, 2008

Preface

We have been watching an explosive growth of application of Data Mining (DM) technologies in an increasing number of different areas of business, government and science. Two of the most important business areas are finance, in particular in banks and insurance companies, and e-business, such as web portals, e-commerce and ad management services.

In spite of the close relationship between research and practice in Data Mining, it is not easy to find information on some of the most important issues involved in real world application of DM technology, from business and data understanding to evaluation and deployment. Papers often describe research that was developed without taking into account constraints imposed by the motivating application. When these issues are taken into account, they are frequently not discussed in detail because the paper must focus on the method. Therefore, knowledge that could be useful for those who would like to apply the same approach on a related problem is not shared.

In 2007, we organized a workshop with the goal of attracting contributions that address some of these issues. The *Data Mining for Business* workshop was held together with the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), in Nanjing, China.[1]

This book contains extended versions of a selection of papers from that workshop. Due to the importance of the two application areas, we have selected papers that are mostly related to finance and e-business. The chapters of this book cover the whole range of issues involved in the development of DM projects, including the ones mentioned earlier, which often are not described. Some of these papers describe applications, including interesting knowledge on how domain-specific knowledge was incorporated in the development of the DM solution and issues involved in the integration of this solution in the business process. Other papers illustrate how the fast development of IT, such as blogs or RSS feeds, opens many interesting opportunities for Data Mining and propose solutions to address them.

These papers are complemented with others that describe applications in other important and related areas, such as intrusion detection, economic analysis and business process mining. The successful development of DM applications depends on methodologies that facilitate the integration of domain-specific knowledge and business goals into the more technical tasks. This issue is also addressed in this book.

This book clearly shows that Data Mining projects must not be regarded as independent efforts but they should rather be integrated into broader projects that are aligned with the company's goals. In most cases, the output of DM projects is a solution that must be integrated into the organization's information system and, therefore, in its (decision-making) processes.

Additionally, the book stresses the need for DM researchers to keep up with the pace of development in IT technologies, identify potential applications and develop suitable

[1] http://www.liaad.up.pt/dmbiz.

solutions. We believe that the flow of new and interesting applications will continue for many years.

Another interesting observation that can be made from this book is the growing maturity of the field of Data Mining in China. In the last few years we have observed spectacular growth in the activity of Chinese researchers both abroad and in China. Some of the contributions in this volume show that this technology is increasingly used by people who do not have a DM background.

To conclude, this book presents a collection of papers that illustrates the importance of maintaining close contact between Data Mining researchers and practitioners. For researchers, it is useful to understand how the application context creates interesting challenges but, simultaneously, enforces constraints which must be taken into account in order for their work to have higher practical impact. For practitioners, it is not only important to be aware of the latest developments in DM technology, but it may also be worthwhile to keep a permanent dialogue with the research community in order to identify new opportunities for the application of existing technologies and also for the development of new technologies.

We believe that this book may be interesting not only for Data Mining researchers and practitioners, but also to students who wish to have an idea of the practical issues involved in Data Mining. We hope that our readers will find it useful.

Porto, Bradford, Hangzhou, Osaka and Nanjing – May 2008
Carlos Soares, Yonghong Peng, Jun Meng, Takashi Washio, Zhi-Hua Zhou

Program Committee

Alípio Jorge	University of Porto	Portugal
André Carvalho	University of São Paulo	Brazil
Arno Knobbe	Kiminkii/Utrecht University	The Netherlands
Bhavani Thuraisingham	Bhavani Consulting	USA
Can Yang	Hong Kong University of Science and Technology	China
Carlos Soares	University of Porto	Portugal
Carolina Monard	University of São Paulo	Brazil
Chid Apte	IBM Research	USA
Dave Watkins	SPSS	USA
Eric Auriol	Kaidara	France
Gerhard Paaß	Fraunhofer	Germany
Gregory Piatetsky-Shapiro	KDNuggets	USA
Jinlong Wang	Zhejiang University	China
Jinyan Li	Institute for Infocomm Research	Singapore
João Mendes Moreira	University of Porto	Portugal
Jörg-Uwe Kietz	Kdlabs AG	Switzerland
Jun Meng	Zhejiang University	China
Katharina Probst	Accenture Technology Labs	USA
Liu Zehua	Yokogawa Engineering	Singapore
Lou Huilan	Zhejiang University	China
Lubos Popelínský	Masaryk University	Czech Republic
Mykola Pechenizkiy	University of Eindhoven	Finland
Paul Bradley	Apollo Data Technologies	USA
Peter van der Putten	Chordiant Software/ Leiden University	The Netherlands
Petr Berka	University of Economics of Prague	Czech Republic
Ping Jiang	University of Bradford	UK
Raul Domingos	SPSS	Belgium
Rayid Ghani	Accenture	USA
Reza Nakhaeizadeh	DaimlerChrysler	Germany
Robert Engels	Cognit	Norway
Rüdiger Wirth	DaimlerChrysler	Germany
Ruy Ramos	University of Porto/ Caixa Econômica do Brasil	Portugal
Sascha Schulz	Humboldt University	Germany
Steve Moyle	Secerno	UK
Tie-Yan Liu	Microsoft Research	China
Tim Kovacs	University of Bristol	UK
Timm Euler	University of Dortmund	Germany
Wolfgang Jank	University of Maryland	USA
Walter Kosters	University of Leiden	The Netherlands
Wong Man-leung	Lingnan University	China
Xiangjun Dong	Shandong Institute of Light Industry	China
YongHong Peng	University of Bradford	UK
Zhao-Yang Dong	University of Queensland	Australia
Zhiyong Li	Zhejiang University	China

This page intentionally left blank

Contents

Applications of Data Mining in E-Business and Finance
C. Soares et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-890-8-1

Applications of Data Mining in E-Business and Finance: Introduction

Carlos SOARES [a,1] and Yonghong PENG [b] and Jun MENG [c] and Takashi WASHIO [d] and Zhi-Hua ZHOU [e]

[a] *LIAAD-INESC Porto L.A./Faculdade de Economia, Universidade do Porto, Portugal*
[b] *School of Informatics, University of Bradford, U.K.*
[c] *College of Electrical Engineering, Zhejiang University, China*
[d] *The Institute of Scientific and Industrial Research, Osaka University, Japan*
[e] *National Key Laboratory for Novel Software Technology, Nanjing University, China*

Abstract. This chapter introduces the volume on Applications of Data Mining in E-Business and Finance. It discusses how application-specific issues can affect the development of a data mining project. An overview of the chapters in the book is then given to guide the reader.

Keywords. Data mining applications, data mining process.

Preamble

It is well known that Data Mining (DM) is an increasingly important component in the life of companies and government. The number and variety of applications has been growing steadily for several years and it is predicted that it will continue to grow. Some of the business areas with an early adoption of DM into their processes are banking, insurance, retail and telecom. More recently it has been adopted in pharmaceutics, health, government and all sorts of e-businesses. The most well-known business applications of DM technology are in marketing, customer relationship management and fraud detection. Other applications include product development, process planning and monitoring, information extraction and risk analysis. Although less publicized, DM is becoming equally important in Science and Engineering.[2]

Data Mining is a field where research and applications have traditionally been strongly related. On the one hand, applications are driving research (e.g., the Netflix prize[3]and DM competitions such as the KDD CUP[4]) and, on the other hand, research results often find applicability in real world applications (Support Vector Machines in Computational Biology[5]). Data Mining conferences, such as KDD, ICDM, SDM, PKDD

[1]Corresponding Author: LIAAD-INESC Porto L.A./Universidade do Porto, Rua de Ceuta 118 6º andar; E-mail: csoares@fep.up.pt.

[2]An overview of scientific and engineering applications is given in [1].

[3]http://www.netflixprize.com

[4]http://www.sigkdd.org/kddcup/index.php

[5]http://www.support-vector.net/bioinformatics.html

and PAKDD, play an important role in the interaction between researchers and practitioners. These conferences are usually sponsored by large DM and software companies and many participants are also from industry.

In spite of this closeness between research and application and the amount of available information (e.g., books, papers and webpages) about DM, it is still quite hard to find information about some of the most important issues involved in real world application of DM technology. These issues include data preparation (e.g., cleaning and transformation), adaptation of existing methods to the specificities of an application, combination of different types of methods (e.g., clustering and classification) and testing and integration of the DM solution with the Information System (IS) of the company. Not only do these issues account for a large proportion of the time of a DM project but they often determine its success or failure [2].

A series of workshops have been organized to enable the presentation of work that addresses some of these concerns.[6] These workshops were organized together with some of the most important DM conferences. One of these workshops was held in 2007 together with the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD). The *Data Mining for Business* Workshop took place in beautiful and historical Nanjing (China). This book contains extended versions of a selection of papers from that workshop.

In Section 1 we discuss some of the issues of the application of DM that were identified earlier. An overview of the chapters of the book is given in Section 2. Finally, we present some concluding remarks (Section 3).

1. Application Issues in Data Mining

Methodologies, such as CRISP-DM [3], typically organize DM projects into the following six steps (Figure 1): business understanding, data understanding, data preparation, modeling, evaluation and deployment. Application-specific issues affect all these steps. In some of them (e.g., business understanding), this is more evident than in others (e.g., modeling). Here we discuss some issues in which the application affects the DM process, illustrating with examples from the applications described in this book.

1.1. Business and Data Understanding

In the business understanding step, the goal is to clarify the business objectives for the project. The second step, data understanding, consists of collecting and becoming familiar with the data available for the project.

It is not difficult to see that these steps are highly affected by application-specific issues. Domain knowledge is required to understand the context of a DM project, determine suitable objectives, decide which data should be used and understand their meaning. Some of the chapters in this volume illustrate this issue quite well. Ni *et al.* discuss the properties that systems designed to support trading activities should possess to satisfy their users [4]. Also as part of a financial application, Sousa and Costa present a set of constraints that shape a system for supporting a specific credit problem in the retail banking industry [5]. As a final example, Wu *et al.* present a study of economic indicators in a region of China that requires a thorough understanding of its context [6].

[6] http://www.liaad.up.pt/dmbiz

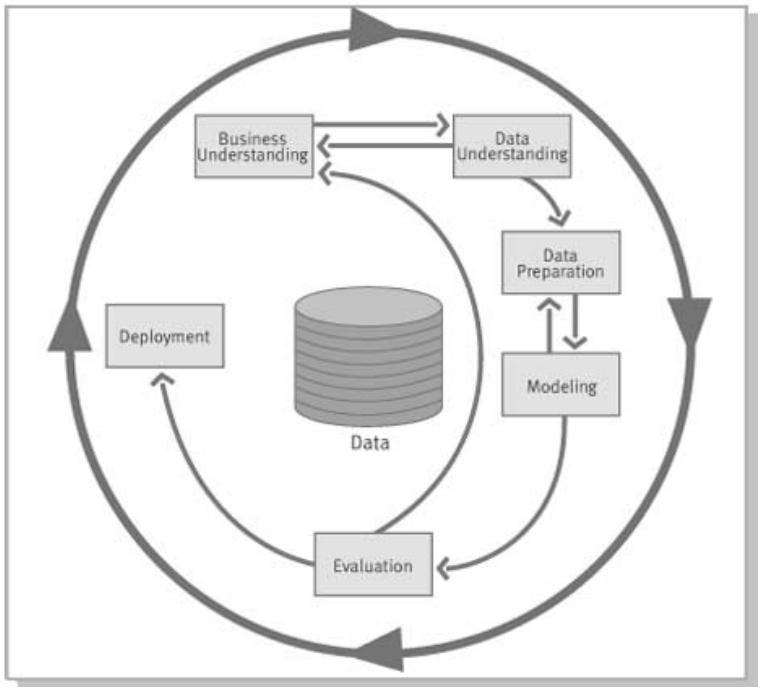

Figure 1. The Data Mining Process, according to the CRISP-DM methodology (image obtained from `http://www.crisp-dm.org`)

1.2. Data Preparation

Data preparation consists of a diverse set of operations to clean and transform the data in order to make it ready for modeling.

Many of those operations are independent of the application operations (e.g., missing value imputation or discretization of numerical variables), and much literature can be found on them. However, many application papers do not describe their usage in a way that is useful in ther applications.

On the other hand, much of the data preparation step consists of application-specific operations, such as feature engineering (e.g., combining some of the original attributes into a more informative one). In this book, Tsai *et al.* describe how they obtain their data from corporate blogs and transform them as part of the development of their blog search system [7]. A more complex process is described by Yuan *et al.* to generate an ontology representing RSS feeds [8].

1.3. Modeling

In the modeling step, the data resulting from the application of the previous steps is analyzed to extract the required knowledge.

In some applications, domain-dependent knowledge is integrated in the DM process in all steps except this one, in which off-the-shelf methods/tools are applied. Dong-Peng *et al.* described one such application where the implementations of decision trees and

association rules in WEKA [9] are applied in a risk analysis problem in banking, for which the data was suitably prepared [10]. Another example in this volume is the paper by Giuffrida *et al.*, in which the Apriori algorithm for association rule mining is used on an online advertising personalization problem [11].

A different modeling approach consists of developing/adapting specific methods for a problem. Some applications involve novel tasks that require the development of new methods. An example included in this book is the work of Datta *et al.*, who address the problem of predicting resource demand in project planning with a new sequence mining method based on hidden semi-Markov models [12]. Other applications are not as novel but have specific characteristics that require adaptation of existing methods. For instance, the approach of Ni *et al.* to the problem of generating trading rules uses an adapted evolutionary computation algorithm [4]. In some applications, the results obtained with a single method are not satisfactory and, thus, better solutions can be obtained with a combination of two or more different methods. Kaya *et al.* propose a method for risk analysis which consists of a combination of support vector machines and logistic regression [13]. In a different chapter of this book, Toro-Negro *et al.* describe an approach which combines different types of methods, an optimization method (evolutionary computation) with a learning method (k-nearest neighbors) [14].

A data analyst must also be prepared to use methods for different tasks and originating from different fields, as they may be necessary in different applications, sometimes in combination as described above. The applications described in this book illustrate this quite well. The applications cover tasks such as clustering (e.g., [15]), classification (e.g., [13,14]), regression (e.g., [6]), information retrieval (e.g., [8]) and extraction (e.g., [7]), association mining (e.g., [10,11]) and sequence mining (e.g., [12,16]). Many research fields are also covered, including neural networks (e.g., [5]), machine learning (e.g., SVM [13]), data mining (e.g., association rules [10,11]), statistics (e.g., logistic [13] and linear regression [6]) and evolutionary computation (e.g., [4,14]) The wider the range of tools that is mastered by a data analyst, the better the results he/she may obtain.

1.4. Evaluation

The goal of the evaluation step is to assess the adequacy of the knowledge in terms of the project objectives.

The influence of the application on this step is also quite clear. The criteria selected to evaluate the knowledge obtained in the modeling phase must be aligned with the business goals. For instance, the results obtained on the online advertising application described by Giuffrida *et al.* are evaluated in terms of clickthrough and also of revenue [11]. Finding adequate evaluation measures is, however, a complex problem. A methodology to support the development of a complete set of evaluation measures that assess quality not only in technical but also in business terms is proposed by Luo *et al.* [16].

1.5. Deployment

Deployment is the step in which the knowledge validated in the previous step is integrated in the (decision-making) processes of the organization.

It, thus, depends heavily on the application context. Despite being critical for the success of a DM project, this step is often not given sufficient importance, in contrast to other steps such as business understanding and data preparation. This attitude is illustrated quite well in the CRISP-DM guide [3]:

> In many cases it is the customer, not the data analyst, who carries out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the customer to understand up front what actions need to be carried out in order to actually make use of the created models.

This graceful handing over of responsibilities of the deployment step by the data analyst can be the cause for the failure of a DM project which, up to this step, has obtained promising results.

In some cases, the model obtained is the core of the business process. Deployment, thus, requires the development of the software system (e.g, program or website) that will serve as the wrapper to the model. An example is the blog search system developed by Tsai *et al.* [7].

Despite the complexities of the development of new systems, it is often simpler than integrating the model with an existing Information System (IS), as illustrated in Figure 2. In this case, there are two scenarios. In the first one, the model is integrated in an existing component of the IS, replacing the procedure which was previously used for the same purpose. For instance, the model developed by Giuffrida *et al.* for personalization of ads replaces the random selection procedure used by a web ad-server [11]. Another example is the work of Sousa and Costa, in which a DM application generates a new model for deciding whether or not to pay debit transactions that are not covered by the current balance of an account [5]. In the second scenario, integration consists of the development of a new component which must then be adequately integrated with the other components of the IS. In this volume, Datta *et al.* describe how the sequence mining algorithm they propose can be integrated into the resource demand forecasting process of an organization [12].

In either case, integration will typically imply communication with one or more databases and with other modules. It may also be necessary to implement communication with external entities, such as users or hardware. Finally, because it cannot be guaranteed that a model developed with existing data will function correctly forever, monitoring and maintenance mechanisms must be implemented. Monitoring results should be fed back to the data analyst, who decides what should be done (e.g., another iteration in the DM process). In some cases it is possible to implement an automatic maintenance mechanism to update the model (e.g., relearning the model using new data). For instance, the model for personalization of ads used by Giuffrida *et al.* is updated daily with new data that is collected from the activity on the ad-server [11].

Additionally, development of typical DM projects uses real data but it is usually independent of the decision process which it aims to improve. Very often, the conditions are not exactly the some in the development and the deployment contexts. Thus, it may be necessary in some cases to carry out a gradual integration with suitable live testing. The development of mechanisms to support this kind of integration and testing implies changes to the IS of the organization, with associated costs. Again, Giuffrida *et al.* describe how a live evaluation of their system is carried out, by testing in parallel the effect of the ad personalization model and a random selection method [11].

2. Overview

The chapters in this book are organized into three groups: finance, e-business and miscellaneous applications. In the following sections we give an overview of their content.

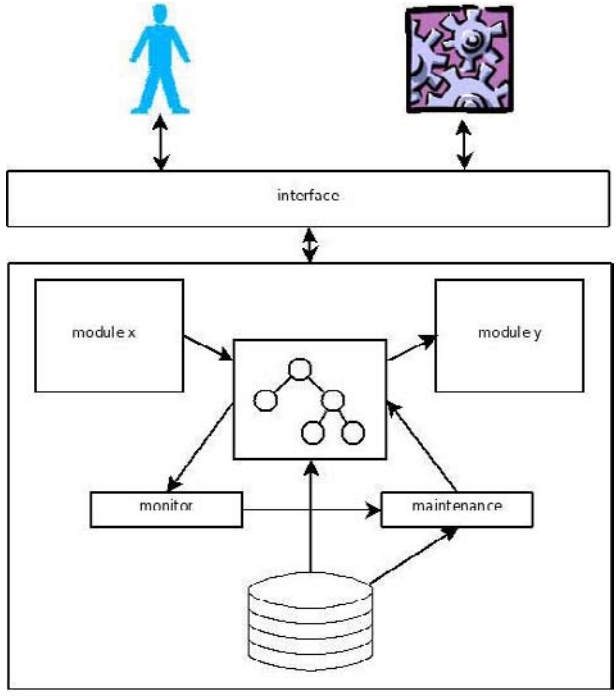

Figure 2. Integration of the results of Data Mining into the Information System.

2.1. Finance

The chapter by Ni *et al.* describes a method to generate a complete set of trading strategies that take into account application constraints, such as timing, current position and pricing [4]. The authors highlight the importance of developing a suitable backtesting environment that enables the gathering of sufficient evidence to convince the end users that the system can be used in practise. They use an evolutionary computation approach that favors trading models with higher stability, which is essential for success in this application domain.

The next two chapters present credit risk modeling applications. In the first chapter, Kaya *et al.* try three different approaches, by transforming both the methods and the problem [13]. They start by tackling the problem as a supervised classification task and empirically comparing SVM and logistic regression. Then, they propose a new approach that combines the two methods to obtain more accurate decisions. Finally, they transform the problem into one of estimating the probability of defaulting on a loan.

The second of these chapters, by Peng *et al.*, describes an assessment of client credibility in Chinese banks using off-the-shelf tools [10]. Although the chapter describes a simple application from a technical point of view, it is quite interesting to note that it is carried out not by computer scientists but rather by members of a Management and Economics school. This indicates that this technology is starting to be used in China by people who do not have a DM background.

The last chapter in this group describes an application in a Portuguese bank made by Sousa and Costa [5]. The problem is related to the case when the balance of an account is

insufficient to cover an online payment made by one of its clients. In this case, the bank must decide whether to cover the amount on behalf of the client or refuse payment. The authors compare a few off-the-shelf methods, incorporating application-specific information about the costs associated with different decisions. Additionaly, they develop a decision process that customizes the usage of the models to the application, significantly improving the quality of the results.

2.2. E-Business

The first chapter in this group, by Giuffrida *et al.*, describes a successful application of personalization of online advertisement [11]. They use a standard association rules method and focus on the important issues of actionability, integration with the existing IS and live testing.

Tsai *et al.* describe a novel DM application [7]. It is well known that blogs are increasingly regarded as a business tool by companies. The authors propose a method to search and analyze blogs. A specific search engine is developed to incorporate the models developed.

The next chapter proposes a method to measure the semantic similarity between RSS feeds and subscribers [8]. Yuan *et al.* show how it can be used to support RSS reader tools. The knowledge is represented using ontologies, which are increasingly important to incorporate domain-specific knowledge in DM solutions.

The last paper in this group, by Mashayekhy *et al.*, addresses the problem of identifying the opponent's strategy in a automated negotiation process [15]. This problem is particularly relevant in e-business, where many opportunities exist for (semi-)autonomous negotiation. The method developed uses a clustering method on information about previous negotiation sessions.

2.3. Other Applications

The first chapter in this group describes a government application, made by Luo *et al.* [16]. The problem is related to the management of the risk associated with social security clients in Australia. The problem is addressed as a sequence mining task. The actionability of the model obtained is a central concern of the authors. They focus on the complicated issue of performing an evaluation taking both technical and business interestingness into account.

The chapter by Toro-Negro *et al.* addresses the problem of network security [14]. This is an increasingly important concern as companies use networks not only internally but also to interact with customers and suppliers. The authors propose a combination of an optimization algorithm (an evolutionary computation method) and a learning algorithm (k-nearest neighbors) to address this problem.

The next paper illustrates the use of Statistical and DM tools to carry out a thorough study of an economic issue in China [6]. As in the chapter by Peng [10], the authors, Wu *et al.*, come from an Economics and Management school and do not have a DM background.

The last chapter in this volume describes work by Datta *et al.* concerned with project management [12]. In service-oriented organizations where work is organized into projects, careful management of the workforce is required. The authors propose a se-

quence mining method that is used for resource demand forecasting. They describe an architecture that enables the integration of the model with the resource demand forecasting process of an organization.

3. Conclusions

In spite of the close relationship between research and practise in Data Mining, finding information on some of the most important issues involved in real world application of DM technology is not easy. Papers often describe research that was developed without taking into account constraints imposed by the motivating application. When these issues are taken into account, they are frequently not discussed in detail because the paper must focus on the method and therefore knowledge that could be useful for those who would like to apply the method to their problem is not shared. Some of those issues are discussed in the chapters of this book, from business and data understanding to evaluation and deployment.

This book also clearly shows that DM projects must not be regarded as independent efforts but they should rather be integrated into broader projects that are aligned with the company's goals. In most cases, the output of the DM project is a solution that must be integrated into the organization's information system and, therefore, in its (decision-making) processes.

Some of the chapters also illustrate how the fast development of IT, such as blogs or RSS feeds, opens many interesting opportunities for data mining. It is up to researchers to keep up with the pace of development, identify potential applications and develop suitable solutions.

Another interesting observation that can be made from this book is the growing maturity of the field of data mining in China. In the last few years we have observed spectacular growth in the activity of Chinese researchers both abroad and in China. Some of the contributions in this volume show that this technology is increasingly used by people who do not have a DM background.

Acknowledgments

We wish to thank the organizing team of PAKDD for their support and everybody who helped us to publicize the workshop, in particular Gregory Piatetsky-Shapiro (www.kdnuggets.com), Guo-Zheng Li (MLChina Mailing List in China) and KMining (www.kmining.com).

We are also thankful to the members of the Program Committee for their timely and thorough reviews, despite receiving more papers than promised, and for their comments, which we believe were very useful to the authors.

Last, but not least, we would like to thank the valuable help of a group of people from LIAAD-INESC Porto LA/Universidade do Porto and Zhejiang University: Pedro Almeida and Marcos Domingues (preparation of the proceedings) Xiangyin Liu (preparation of the working notes), Zhiyong Li and Jinlong Wang (Chinese version of the webpages), Huilan Luo and Zhiyong Li (support of the review process) and Rodolfo Matos (tech support). We are also thankful to the people from Phala for their support in the process of reviewing the papers.

The first author wishes to thank the financial support of the Fundação Oriente, the POCTI/TRA/61001/2004/Triana Project (Fundação Ciência e Tecnologia) co-financed by FEDER and the Faculdade de Economia do Porto.

References

[1] Robert L. Grossman, Chandrika Kamath, Philip Kegelmeyer, Vipin Kumar, and Raju R. Namburu. *Data Mining for Scientific and Engineering Applications*. Kluwer Academic Publishers, Norwell, MA, USA, 2001.

[2] R. Kohavi and F. Provost. Applications of data mining to electronic commerce. *Data Mining and Knowledge Discovery*, 6:5–10, 2001.

[3] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. *CRISP-DM 1.0: Step-by-Step Data Mining Guide*. SPSS, 2000.

[4] J. Ni, L. Cao, and C. Zhang. Evolutionary optimization of trading strategies. In C. Soares, Y. Peng, J. Meng, Z.-H. Zhou, and T. Washio, editors, *Applications of Data Mining in E-Business and Finance*, pages 13–26. IOS Press, 2008.

[5] M. R. Sousa and J. P. Costa. A tripartite scorecard for the pay/no pay decision-making in the retail banking industry. In C. Soares, Y. Peng, J. Meng, Z.-H. Zhou, and T. Washio, editors, *Applications of Data Mining in E-Business and Finance*, pages 47–52. IOS Press, 2008.

[6] G. Wu, Z. Li, and X. Jiang. Analysis of foreign direct investment and economic development in the Yangtze delta and its squeezing-in and out effect. In C. Soares, Y. Peng, J. Meng, Z.-H. Zhou, and T. Washio, editors, *Applications of Data Mining in E-Business and Finance*, pages 123–137. IOS Press, 2008.

[7] F. S. Tsai, Y. Chen, and K. L. Chan. Probabilistic latent semantic analysis for search and mining of corporate blogs. In C. Soares, Y. Peng, J. Meng, Z.-H. Zhou, and T. Washio, editors, *Applications of Data Mining in E-Business and Finance*, pages 65–75. IOS Press, 2008.

[8] M. Yuan, P. Jiang, and J. Wu. A quantitative method for RSS based applications. In C. Soares, Y. Peng, J. Meng, Z.-H. Zhou, and T. Washio, editors, *Applications of Data Mining in E-Business and Finance*, pages 77–87. IOS Press, 2008.

[9] I. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 2000.

[10] Y. Dong-Peng, L. Jin-Lin, R. Lun, and Z. Chao. Applications of data mining methods in the evaluation of client credibility. In C. Soares, Y. Peng, J. Meng, Z.-H. Zhou, and T. Washio, editors, *Applications of Data Mining in E-Business and Finance*, pages 37–35. IOS Press, 2008.

[11] G. Giuffrida, V. Cantone, and G. Tribulato. An apriori based approach to improve on-line advertising performance. In C. Soares, Y. Peng, J. Meng, Z.-H. Zhou, and T. Washio, editors, *Applications of Data Mining in E-Business and Finance*, pages 53–63. IOS Press, 2008.

[12] R. Datta, J. Hu, and B. Ray. Sequence mining for business analytics: Building project taxonomies for resource demand forecasting. In C. Soares, Y. Peng, J. Meng, Z.-H. Zhou, and T. Washio, editors, *Applications of Data Mining in E-Business and Finance*, pages 139–148. IOS Press, 2008.

[13] M.E. Kaya, F. Gurgen, and N. Okay. An analysis of support vector machines for credit risk modeling. In C. Soares, Y. Peng, J. Meng, Z.-H. Zhou, and T. Washio, editors, *Applications of Data Mining in E-Business and Finance*, pages 27–35. IOS Press, 2008.

[14] F. de Toro-Negro, P. Garcìa-Teodoro, J.E. Diaáz-Verdejo, and G. Maciá-Fernandez. A deterministic crowding evolutionary algorithm for optimization of a KNN-based anomaly intrusion detection system. In C. Soares, Y. Peng, J. Meng, Z.-H. Zhou, and T. Washio, editors, *Applications of Data Mining in E-Business and Finance*, pages 113–122. IOS Press, 2008.

[15] L. Mashayekhy, M. A. Nematbakhsh, and B. T. Ladani. Comparing negotiation strategies based on offers. In C. Soares, Y. Peng, J. Meng, Z.-H. Zhou, and T. Washio, editors, *Applications of Data Mining in E-Business and Finance*, pages 89–100. IOS Press, 2008.

[16] D. Luo, L. Cao, C. Luo, C. Zhang, and W. Wang. Towards business interestingness in actionable knowledge discovery. In C. Soares, Y. Peng, J. Meng, Z.-H. Zhou, and T. Washio, editors, *Applications of Data Mining in E-Business and Finance*, pages 101–111. IOS Press, 2008.

This page intentionally left blank

Applications of Data Mining in E-Business and Finance
C. Soares et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-890-8-11

Evolutionary Optimization of Trading Strategies

Jiarui NI [1], Longbing CAO and Chengqi ZHANG
Faculty of Information Technology, University of Technology, Sydney, Australia

Abstract. It is a non-trivial task to effectively and efficiently optimize trading strategies, not to mention the optimization in real-world situations. This paper presents a general definition of this optimization problem, and discusses the application of evolutionary technologies (genetic algorithm in particular) to the optimization of trading strategies. Experimental results show that this approach is promising.

Keywords. evolutionary optimization, genetic algorithm, trading strategy optimization

Introduction

In financial literatures and trading houses, there are many technical trading strategies [1]. A trading strategy is a predefined set of rules to apply. In the stock market, it is critical for stock traders to find or tune trading strategies to maximize the profit and/or to minimize the risk. One of the means is to backtest and optimize trading strategies before they are deployed into the real market. The backtesting and optimization of trading strategies is assumed to be rational with respect to repeatable market dynamics, and profitable in terms of searching and tuning an 'optimal' combination of parameters indicating higher likelihood of making good benefits. Consequently, the backtesting and optimization of trading strategies has emerged as an interesting research and experimental problem in both finance [2,3] and information technology (IT) [4,5,6,7,8] fields.

It is a non-trivial task to effectively and efficiently optimize trading strategies, not to mention the optimization in real-world situations. Challenges in trading strategy optimization come from varying aspects, for instance, the dynamic market environment, comprehensive constraints, huge quantities of data, multiple attributes in a trading strategy, possibly multiple objectives to be achieved, etc. In practice, trading strategy optimization tackling the above issues essentially is a problem of multi-attribute and multi-objective optimization in a constrained environment. The process of solving this problem inevitably involves high dimension searching, high frequency data stream, and constraints. In addition, there are some implementation issues surrounding trading strategy optimization in market condition, for instance, sensitive and inconstant strategy performance subject to dynamic market, and complicated computational settings and develop-

[1]Corresponding Author: Jiarui Ni, CB10.04.404, Faculty of Information Technology, University of Technology, Sydney, GPO Box 123, Broadway, NSW 2007, Australia. E-mail: jiarui@it.uts.edu.au.

ment in data storage, access, preparation and system construction. The above issues in trading strategy optimization are challenging, practical and very time consuming.

This paper tries to solve this optimization problem with the help of evolutionary technologies. Evolutionary computing is used because it is good at high dimension reduction, and generating global optimal or near-optimal solutions in a very efficient manner. In literature, a few data mining approaches [7,9], in particular, genetic algorithms [10,11,12,8] based evolutionary computing has been explored to optimize trading strategies. However, the existing research has mainly focused on extracting interesting trading patterns of statistical significance [5], demonstrating and pushing the use of specific data mining algorithms [7,9]. Unfortunately, real-world market organizational factors and constraints [13], which form inseparable constituents of trading strategy optimization, have not been paid sufficient attention to. As a result, many interesting trading strategies are found, while few of them are applicable in the market. The gap between the academic findings and business expectations [14] comes from a few reasons, such as the over-simplification of optimization environment and evaluation fitness. In a word, actionable optimization of trading strategies should be conducted in market environment and satisfy trader's expectations. This paper tries to present some practical solutions and results, rather than some unrealistic algorithms.

The rest of the paper is organized as follows. First of all, Section 1 presents the problem definition in terms of considering not only attributes enclosed in the optimization algorithms and trading strategies, but also constraints in the target market where the strategy to is be identified and later used. Next, Section 2 explains in detail how genetic algorithm can be applied to the optimization of trading strategies, and presents techniques that can improve technical performance. Afterwards, Section 3 shows experimental results with discussions and refinements. Finally, Section 4 concludes this work.

1. Problem Definition

In market investment, traders always pursuit a 'best' or 'appropriate' combination of purchase timing, position, pricing, sizing and objects to be traded under certain business situations and driving forces. Data mining in finance may identify not only such trading signals, but also patterns indicating either iterative or repeatable occurrences. The mined findings present trading strategies to support investment decisions in the market.

Definition A trading strategy actually represents a set of individual instances, the trading strategy set Ω is a tuple defined as follows.

$$\begin{aligned}\Omega &= \{r_1, r_2, \ldots, r_m\} \\ &= \{(t, b, p, v, i) | t \in T, b \in B, p \in P, v \in V, i \in I)\}\end{aligned} \tag{1}$$

where r_1 to r_m are instantiated individual trading strategy, each of them is represented by instantiated parameters of t, b, p, v and an instrument i to be traded; $T = \{t_1, t_2, \ldots, t_m\}$ is a set of appropriate time trading signals to be triggered; $B = \{\text{buy, sell, hold}\}$ is the set of possible behavior (i.e., trading actions) executed by trading participants; $P = \{p_1, p_2, \ldots, p_m\}$ and $V = \{v_1, v_2, \ldots, v_m\}$ are the sets of trading price and volume matching with corresponding trading time; and $I = \{i_1, i_2, \ldots, i_m\}$ is a set of target instruments to be traded.

With the consideration of environment complexities and trader's favorite, the optimization of trading strategies is to search an 'appropriate' combination set Ω' in the whole trading strategy candidate set Ω, in order to achieve both user-preferred technical ($tech\_int()$) and business ($biz\_int()$) interestingness in an 'optimal' or 'near-optimal' manner. Here 'optimal' refers to the maximal/minimal (in some cases, smaller is better) values of technical and business interestingness metrics under certain market conditions and user preferences. In some situations, it is impossible or too costly to obtain 'optimal' results. For such cases, a certain level of 'near-optimal' results are also acceptable. Therefore, the sub-set Ω' indicates 'appropriate' parameters of trading strategies that can support a trading participant a ($a \in A$, A is market participant set) to take actions to his/her advantages. As a result, in some sense, trading strategy optimization is to extract actionable strategies with multiple attributes towards multi-objective optimization [15] in a constrained market environment.

Definition An optimal and actionable trading strategy set Ω' is to achieve the following objectives:

$$\begin{aligned} tech\_int() &\rightarrow max\{tech\_int()\} \\ biz\_int() &\rightarrow max\{biz\_int()\} \end{aligned} \tag{2}$$

while satisfying the following conditions:

$$\begin{aligned} \Omega' = \{e_1, e_2, \ldots, e_n\} \\ \Omega' \subset \Omega \\ m > n \end{aligned} \tag{3}$$

where $tech\_int()$ and $biz\_int()$ are general technical and business interestingness metrics, respectively. As the main optimization objectives of identifying 'appropriate' trading strategies, the performance of trading strategies and their actionable capability are encouraged to satisfy expected technical interestingness and business expectations under multi-attribute constraints. The ideal aim of actionable trading strategy discovery is to identify trading patterns and signals, in terms of certain background market microstructure and dynamics, so that they can assist traders in taking the right actions at the right time with the right price and volume on the right instruments. As a result of trading decisions directed by the identified evidence, benefits are maximized while costs are minimized.

1.1. Constrained Optimization Environment

Typically, actionable trading strategy optimization must be based on a good understanding of organizational factors hidden in the mined market and data. Otherwise it is not possible to accurately evaluate the dependability of the identified trading strategies. The actionability of optimized trading strategies is highly dependent on the mining environment where the trading strategy is extracted and applied. In real-world actionable trading strategy extraction, the underlying environment is more or less constrained. Constraints may be broadly embodied in terms of data, domain, interestingness and deployment aspects. Here we attempt to explain domain and deployment constraints surrounding actionable trading strategy discovery.

Market organization factors [13] relevant to trading strategy discovery consist of the following fundamental entities: $M = \{I, A, O, T, R, E\}$. Table 1 briefly explains these entities and their impact on trading strategy actionability. In particular, the entity $O = \{(t, b, p, v)|t \in T, b \in B, p \in P, v \in V\}$ is further represented by attributes T, B, P and V, which are attributes of trading strategy set Ω. The elements in M form the constrained market environment of trading strategy optimization. In the strategy and system design of trading strategy optimization, we need to give proper consideration of these factors.

Table 1. Market organizational factors and their impact to trading strategy actionability

Organizational factors	Impact on actionability
Traded instruments I, such as a stock or derivative, I={stock, option, feature, ... }	Varying instruments determine different data, analytical methods and objectives
Market participants A, A = {broker, market maker, mutual funds, ... }	Trading agents have the final right to evaluate and deploy discovered trading strategies to their advantage
Order book forms O, O = {limit, market, quote, block, stop}	Order type determines the data set to be mined, e.g., order book, quote history or price series, etc.
Trading session, whether it includes call market session or continuous session, it is indicated by time frame T	Setting up the focusing session can greatly prune order transactions
Market rules R, e.g., restrictions on order execution defined by exchange	They determine pattern validity of discovered trading strategies when deployed
Execution system E, e.g., a trading engine is order-driven or quote-driven	It limits pattern type and deployment manner after migrated to real trading system

In practice, any particular actionable trading strategy needs to be identified in an instantiated market niche m ($m \in M$) enclosing the above organization factors. This market niche specifies particular constraints, which are embodied through the elements in Ω and M, on trading strategy definition, representation, parameterization, searching, evaluation and deployment. The consideration of specific market niche in trading strategy extraction can narrow search space and strategy space in trading strategy optimization. In addition, there are other constraints such as data constraints D that are not addressed here for limited space. Comprehensive constraints greatly impact the development and performance of extracting trading strategies.

Constraints surrounding the development and performance of actionable trading strategy set Ω' in a particular market data set form a constraint set:

$$\Sigma = \{\delta_i^k | c_i \in C, 1 \leq k \leq N_i\}$$

where δ_i^k stands for the k-th constraint attribute of a constraint type c_i, $C = \{M, D\}$ is a constraint type set covering all types of constraints in market microstructure M and data D in the searching niche, and N_i is the number of constraint attributes for a specific type c_i.

Correspondingly, an actionable trading strategy set Ω' is a conditional function of Σ, which is described as

$$\Omega' = \{(\omega, \delta)|\omega \in \Omega, \delta \in \{(\delta_i^k, a)|\delta_i^k \in \Sigma, a \in A\}\}$$

where ω is an 'optimal' trading pattern instance, and δ indicates specific constraints on the discovered pattern that is recommended to a trading agent a.

2. Optimization with GA

In this section, we first describe in detail one type of classical strategies based on moving average in order to illustrate how technical trading strategies work. Afterwards, we will discuss how to use genetic algorithms (GA) to optimize trading strategies.

2.1. Moving Average Strategies

A moving average (MA) is simply an average of current and past prices over a specified period of time. An MA of length l at time t is calculated as

$$M_t(l) = \frac{1}{l}\sum_{i=0}^{l-1} P_{t-i} \tag{4}$$

where P_{t-i} is the price at time $t-i$.

Various trading strategies can be formulated based on MA. For example, a double MA strategy (denoted by MA(l_1, l_2)) compares two MAs with different lengths l_1 and l_2 where $l_1 < l_2$. If $M_t(l_1)$ rises above $M_t(l_2)$, the security is bought and held until $M_t(l_1)$ falls below $M_t(l_2)$. The signal S_t is given by

$$S_t = \begin{cases} 1 \quad \text{if } M_t(l_1) > M_t(l_2) \\ \qquad \text{and } M_{t-k}(l_1) < M_{t-k}(l_2) \\ \qquad \text{and } M_{t-i}(l_1) = M_{t-i}(l_2), \\ \qquad \forall\, i \in \{1, \cdots, k-1\} \\ -1 \text{ if } M_t(l_1) < M_t(l_2) \\ \qquad \text{and } M_{t-n}(l_1) > M_{t-n}(l_2) \\ \qquad \text{and } M_{t-i}(l_1) = M_{t-i}(l_2), \\ \qquad \forall\, i \in \{1, \cdots, n-1\} \\ 0 \quad \text{otherwise} \end{cases} \tag{5}$$

where k and n are arbitrary positive integers, 1 means 'buy', -1 means 'sell', and 0 means 'hold' or 'no action'.

Figure 1 shows an example of double MA strategy, where $l_1 = 10$, $l_2 = 30$, and the upward arrows indicate buy signals, and downward arrows indicate sell signals.

A filtered double MA strategy is more conservative than the original double MA strategy in that it only takes action when $M_t(l_1)$ rises above (or falls below) $M_t(l_2)$ by more than a certain percentage θ. The next subsection will use such a filtered double MA strategy as example for the illustration of optimization with genetic algorithm.

It should be noted that the values of l, l_1, l_2 and θ in the above equations are not fixed. They are usually selected by experience or experiments.

MA strategies give one 'sell' signal after one buy signal and vice versa. There are no consecutive buy signals nor consecutive sell signals. However, other trading strategies, such as those explained in the next sub sections, may give consecutive buy or sell signals.

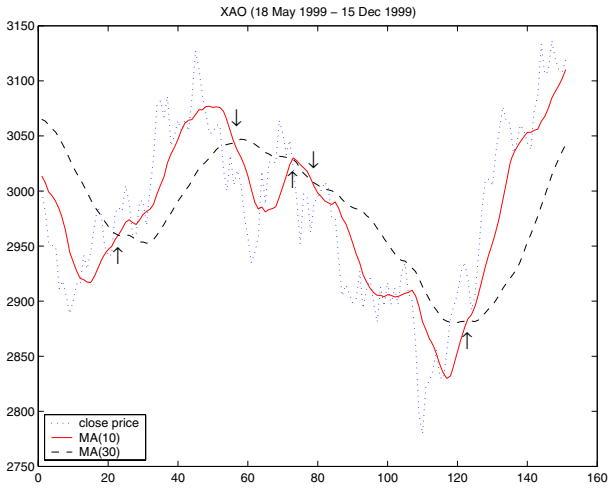

Figure 1. An example of double MA strategy.

2.2. Optimization with Genetic Algorithm

GAs have been widely investigated and applied in many areas since it was developed by John Holland in 1975 [16].

The GA procedure can be outlined as follows:

1. Create an initial population of candidates.
2. Evaluate the performance of each candidate.
3. Select the candidates for recombination.
4. Perform crossover and mutation.
5. Evaluate the performance of the new candidates.
6. Stop if a termination criterion is met, otherwise go back to step 3.

To implement a GA, one has to decide several main issues: fitness function, encoding, population size, selection, crossover and mutation operators, etc. In this subsection, we discuss how they are properly set for the optimization of the MA trading strategy.

2.2.1. Fitness Function

For the optimization of trading strategies, the fitness function is defined by the trader based on his business objectives. Return on investment is a good fitness function for aggressive traders, while the combination of a reasonable return and a low risk may be a better fitness function for conservative investors.

2.2.2. Encoding and Search Space

Encoding of chromosomes is the first question to ask when starting to solve a problem with GA. It depends on the problem heavily. In this problem, it is natural to define the chromosome as the tuple consisting of all the parameters of a certain trading strategy. The parameters can not take arbitrary values. Instead, they are limited by various constraints. Firstly, the type and value range are constrained by their domain specific meaning and relationship. Secondly, for practical reasons, we limit the value ranges to define

a reasonable search space. Further, we also limit the precision of real values since overly high precision is meaningless for this problem. Table 2 lists the parameters and their constraints for the MA trading strategy we test, where 'I' means an integer and 'R' means a real number.

Table 2. Parameters and their constraints

Parameters	Type	Constraints	Precision
short MA length (l_1)	I	$1 \leq l_1 \leq 20$	10^0
long MA length (l_2)	I	$4 \leq l_2 \leq 50$; $l_2 > l_1 + 3$	10^0
filter (θ)	R	$0 \leq \theta \leq 0.04$	10^{-4}

2.2.3. Population Size

The population size defines how many chromosomes are in the population (in one generation). Theoretically, if there are too few chromosomes, GA have few possibilities to perform crossover and only a small part of search space is explored. On the other hand, if there are too many chromosomes, GA slows down. Research shows that after some limit (which depends mainly on encoding and the problem) it is not useful to use very large populations because it does not solve the problem faster than moderate sized populations.

Pilot tests shows that the execution time increases constantly in proportion to the population size, while there is little or no improvement to the fitness after the population size goes above 40. As a result, the population size is set to 40 in the experiments that were carried out to the determine the crossover rate and mutation rate.

2.2.4. Crossover

Each trading strategy may have different number of parameters, therefore the chromosome lengths are not fixed. To make our crossover method useful for different kinds of trading strategies, we choose a crossover method which is independent of the chromosome length. It works as follows:

1. Let the two parents be $M = \{m_1, m_2, \ldots, m_n\}$, $D = \{d_1, d_2, \ldots, d_n\}$;
2. Randomly select one parameter x from $1 \ldots n$, and a random number β from $(0, 1)$;
3. Calculate $b_x = m_x - \beta \times (m_x - d_x)$, $c_x = d_x + \beta \times (m_x - d_x)$;
4. Calculate 2 children: $C_1 = \{m_1, m_2, \ldots, b_x, \ldots, m_n\}$ and $C_2 = \{d_1, d_2, \ldots, c_x, \ldots, d_n\}$;
5. Check if the children satisfy all constraints. If not, go back to Step 2; repeat until two valid children are obtained.

Pilot tests show that the business performance (in terms of the final wealth) of this GA does not rely much on the exact crossover rate value. A larger crossover rate usually gives better business performance but the improvement is not significant, especially when the execution time is taken into consideration. For the simple trading strategies tested in the work, the overhead execution time resulted from the crossover operation is quite prominent. Consequently, when the execution time is not a problem (e.g., when computation resources are abundant), a large crossover rate is preferred for better busi-

ness performance, but if the execution time is stringent, a small crossover rate may be a good choice as it would not heavily degrade the business performance. In the next section, a crossover rate of 0.8 is used in the test of mutation rate.

2.2.5. Mutation

The GA parameter mutation rate defines how often parts of chromosome will be mutated. If there is no mutation, offspring are generated immediately after crossover (or directly copied) without any change. If mutation is performed, one or more parts of a chromosome are changed. If mutation rate is 100%, whole chromosome is changed, if it is 0%, nothing is changed.

Mutation generally prevents the GA from falling into local optima. Mutation should not occur very often, because then GA will in fact change to random search.

In our experiment, the disturbance mutation method is used. That is to say, one parameter is randomly selected and replaced with a random value (subject to its constraint). Again, this method is independent of the number of parameters and thus can be applied to various trading strategies.

Pilot tests show that mutation rate is almost irrelevant in this problem. Neither the business performance nor the execution time is evidently affected by mutation rate.

2.2.6. Evaluation History vs. Evaluation Time

The most time consuming task in a GA is the evaluation of fitness function for each chromosome in every generation. During the evolution of a GA, there may be identical chromosomes in different generations. There are two methods to deal with these repeatedly appeared chromosomes. One is to evaluate every chromosome regardless of whether it has been evaluated in previous generations, the other is to keep a history of fitness values of all the chromosomes that have been evaluated since the GA starts and re-use the fitness value when a chromosome re-appears in a new generation. The latter method requires more memory to store the evaluation history and extra time to search the history but may save the total execution time by reducing the number of evaluations, especially when the trading strategy is complicated and the evaluation time is long.

Pilot tests show that the use of evaluation history generally saves the total execution time by about 25 — 30 percent in our test.

2.3. Performance Boost

2.3.1. Data Storage

Stock transaction data have become very detailed and enormous with the introduction of electronic trading systems. This makes it a problem to store and to access the data in later analyses such as mining useful patterns and backtesting trading strategies.

Various storage methods have been used to store stock transaction data. One simple and widely used method is a formatted plain text file, such as a comma separated values (CSV) file. A second storage method is to use a relational database. There are also other customized storage methods, e.g, FAV format used by SMARTS[2]. All these methods have their own strengths and limitations and are not always suitable for the optimization

[2]http://www.smarts.com.au

of trading strategies. Therefore a new dynamic storage method has been proposed in [17], which is briefly introduced below.

In this storage schema, compressed text files are used to store stock transaction data. There is one folder for each stock. In this folder, the data will be further split into several files. The size of each file (before compression) is controlled to be around S, which is to be tuned according to performance tests. Each file may contain data for a single day or multiple days, depending on the number or transaction records. An index file is used to map the date to data files. With the help of the index file, it is possible to quickly identify the data file containing the data for any specific date. And because the size of each data file is restricted to somewhere around S, the maximum data to be read and discarded in order to read the data for any specific date is also limited. Further, since plain text files are used here, any standard compression utilities can be used to compress the data file to save storage space. Besides, when new data are available, it is also easy to merge new data with existing data under this storage method. Only the index file and the last data file (if necessary) have to be rebuilt.

This storage method provides a flexible mechanism to balance between storage space and access efficiency. With this method, it is easy to trade storage space for access efficiency and vice versa.

2.3.2. Parallel GA

When the trading strategy is complicated and the search space is huge, it is very time-consuming to run GA for optimization. A straightforward way to speed up the computation is to parallelize its execution. Once a parameter set is given, a processing element (PE) can execute the whole trading strategy by itself. Therefore it is easy to parallelize the GA with master-slave model. Slave PEs simply execute the trading strategy with a given parameter set, while master PE takes charge of the whole optimization process, as shown in Figure 2.

1. Create an initial population of n parameter sets;
2. Partition the population into N equal groups;
3. Send the parameter sets in each group to one slave process;
4. Receive the fitness value for each triple from the slaves;
5. Stop if a termination criterion is met;
6. Select the parameter sets for recombination;
7. Perform crossover and mutation;
8. Go back to step 2;

Figure 2. Algorithm – Master process.

3. Application and Refinements

This section applies the techniques discussed in Section 2 to the real market to test the business performance of the optimized trading strategies. Further refinements are also introduced to boost the business performance.

3.1. Empirical Studies in the Asx Market

As a first step to study the business performance of the optimized trading strategies, we carry out a series of experiments with historical data. The experiment settings and results are discussed below.

We carry out our experiments over the historical daily data of 32 securities and 4 indices traded on the Australian Securities eXchange (ASX). The securities are selected according to their price, liquidity and data availability. The data are downloaded from Commonwealth Securities Ltd. (www.comsec.com.au) for free as CSV files and contain open price, close price, highest price, lowest price and volume. The time range is from June 1996 to August 2006 and comprises about 2500 trading days.

Five technical trading strategies are applied in a series of tests described below.

MA Filtered double moving average
BB Bollinger band with stop loss
KD Slow stochastic oscillator cross/rising and stop loss
RSI Relative strength indicator with retrace and stop loss
CBO Filtered channel break out

The filtered double MA strategy has been explained in detail in Section 2.1. The details of the other four trading strategies are not discussed here. Interested readers can refer to [18,19] for further information. The typical settings of various trading strategies are also obtained from [18]. Besides, the simple buy-and-hold strategy (BH) is also tested for comparison purpose.

During the experiments, the trading strategies are always applied to a security/index with an initial capital of AU$10,000 for a security or AU$1,000,000 for an index. A transaction cost of 0.2% of the traded value is charged for each trade (buy or sale).

Four tests have been designed to evaluate the effectiveness of GA. The data are divided into in-sample and out-of-sample data. The period of in-sample data comprises 2 years (1997 and 1998), while the period of out-of-sample data comprises of 7 years (from 1999 to 2005). To compare the various trading strategies, we are only concerned about the profit over the out-of-sample period.

Test 1 applies the 5 trading strategies with typical settings over the out-of-sample data. It shows the performance of the trading strategies with a typical fixed setting. These settings are usually given by professionals or set default by some trading tools. In our work, the typical settings of various trading strategies are obtained from [18].

Test 2 applies GA over the out-of-sample data to see the maximal profits that various trading strategies might achieve. These are only theoretical values since the optimal settings can only be found out when the historical data are available. No one knows them beforehand and therefore can not trade with these optimal settings in a real market.

Test 3 applies GA over the in-sample data to find the best parameters for the in-sample period and then apply these parameters to the out-of-sample data. This kind of usage reflects exactly how backtesting is used in reality, that is, finding the strategies or parameters that worked well in the past data and applying them to the future market, with the hope that they will keep working well.

Test 4 is a moving window version of Test 3. Here, the in-sample data and out-of-sample data are no longer fixed as mentioned at the beginning of this sub section. Instead, we repeatedly use 2 years' data as the in-sample data and the following year's data as

the out-of-sample data. Again, we find the best parameters for the in-sample periods and then apply them to the corresponding out-of-sample data.

The other 5 trading strategies, namely, MA, CBO, BB, RSI and KD, all go through the 4 tests described above. For Test 1, each trading strategy is executed once for each security/index. For Tests 2, 3 and 4, every pair of trading strategy and security/index is tested for 30 times and the average result of these 30 tests is used to measure the business performance of the optimized trading strategy applied to the security/index.

Table 3 is a comparison table of the performance of various trading strategies in the above-mentioned tests, where $P_i > P_{BH}, i \in \{1, 2, 3, 4\}$ means the performance of the relevant strategy in Test i is better than the performance of the BH strategy, $P_i > P_j, i, j \in \{1, 2, 3, 4\}$ means the performance of the relevant strategy in Test i is better than that in Test j, and the numbers in the table mean for how many securities/indices $P_i > P_{BH}$ or $P_i > P_j$ holds for the relevant strategy. From this comparison table, we can draw several conclusions.

Table 3. Comparison of test results

	MA	CBO	BB	RSI	KD
$P_1 > P_{BH}$	4	3	3	0	4
$P_2 > P_{BH}$	28	27	23	33	35
$P_3 > P_{BH}$	8	3	2	1	2
$P_4 > P_{BH}$	4	4	1	1	3
$P_2 > P_1$	36	36	36	36	36
$P_3 > P_1$	25	27	25	35	29
$P_4 > P_1$	22	26	21	36	36
$P_4 > P_3$	14	19	16	24	17

First of all, the performance of the typical settings (Test 1) are rather poor. For very few securities, the typical settings can beat the BH strategy. It suggests nothing else than that we can not rely on the typical settings given by the professional traders or the trading software.

Secondly, for some securities, even when the trading strategies are optimized (Test 2), they still can not beat the BH strategy. This usually happens for securities whose prices rise stably. Given that GA executes pretty fast, it can help rule out these fruitless trading strategies for such securities quickly.

Thirdly, although Test 2 shows that in most cases the optimized trading strategies are better than the BH strategy, the results of Test 3 and Test 4 show that it is practically not achievable. The optimal parameter settings can only be calculated after the trading data are available, which means there is no chance to trade with such optimized trading strategies. Theoretically, there is an optimal setting, while practically it does not exist beforehand.

Fourthly, for a large portion of the tested securities, the optimized trading strategies (Test 3 and Test 4) work better than those with typical settings (Test 1). This means that our optimization is able to provide some performance improvement.

Lastly, there is no apparent difference between the result of Test 3 and that of Test 4. The change of the lengths of in-sample and out-of-sample periods show no effect in this experiment.

3.2. Business Performance Stabilization

One big problem that emerges from the experiment results of Section 3.1 is that the trading strategies optimized during the in-sample period do not work well for the out-of-sample period. This results in that the trading strategies performs worse than the BH strategy for most of the securities in Test 3 and Test 4.

This problem comes from the fact that we try to find the parameter settings that give the best business performance for the in-sample period without considering any other factors such as stability. As a result, the best parameter setting found by the optimization algorithm may get over optimized and is so fragile that any small disturbance to the trading data may result in a big performance degradation. When it is applied to the out-of-sample period, as the market conditions evolve as time goes on, it is very hard for this ex-optimal parameter setting to continue its performance. Obviously, this is not desirable. The question is whether it is possible to find a parameter setting that works well for the in-sample period and also works well for the out-of-sample period.

In this work, we try to answer this question with a stabilized GA which finds a stable parameter setting instead of the fragile optimal parameter setting. We achieve this by adjusting the fitness function of the GA. Besides the final total wealth, we also take it into consideration whether the final total wealth is resilient to the disturbance of parameters. The new fitness function is calculated as follows:

1. Let the parameter setting to be evaluated be $P = (p_1, p_2, \ldots, p_n)$;
2. Calculate $2n$ new parameter settings which are disturbed versions of P:

$$P_{i1} = (p_1, p_2, \ldots, p_i \times (1 + \delta), \ldots, p_n)$$

$$P_{i2} = (p_1, p_2, \ldots, p_i \times (1 - \delta), \ldots, p_n)$$

   where $i = 1, 2, \ldots, n$, and δ is a disturbance factor;
3. Calculate the final total wealth for P, P_{i1}, P_{i2}, denoted by W, W_{i1}, W_{i2} ($i = 1, 2, \ldots, n$), respectively;
4. Calculate the maximum difference between W_{i1}, W_{i2} ($i = 1, 2, \ldots, n$) and W, denoted by D_{max}:

$$D_{max} = max\{|W_{ij} - W||i = 1, 2, \ldots, n; j = 1, 2\}$$

5. Let the initial wealth be C;
6. Calculate fitness value

$$F = (1 - \frac{D_{max}}{C}) + \frac{W}{C}$$

This new fitness function depends on two factors: the absolute business performance and its resilience to parameter disturbance. The expectation is that such a fitness function will guide GA to find parameter settings with stable high business performance. To check the effect of this fitness function, we again carry out Test 3 and Test 4 over the 36 securities/indices as described in Section 3.1. These two tests with new fitness functions are denoted by Test 3* and Test 4*, respectively.

Table 4 shows the comparison between Test i* and Test i (i = 3, 4). Although the effect on Test 3 is chaotic, the effect on Test 4 is promising with the business performance of Test 4* equal to or better than that of Test 4 for all five trading strategies. Further calculation shows that in average the business performance of Test 4* is 1.1% better than that

of Test 4 for each pair of trading strategy and security, while the business performance of Test 3^* is 0.46% worse than that of Test 3.

Table 4. Comparison of test results II

	MA	CBO	BB	RSI	KD
$P_{3^*} > P_3$	17	18	14	19	13
$P_{4^*} > P_4$	19	19	24	19	18

The different effects of the new fitness function on Test 3 and Test 4 may result from the different in-sample/out-of-sample lengths. In Test 3, the length of the out-of-sample period is 7 year, which is too long compared with the 2 years' in-sample length, and no optimal parameter settings can keep working well for such a long period. However, in Test 4, the out-of-sample length is only 1 year for each 2 year's in-sample period, and the stability of the parameter setting can be kept much better for the out-of-sample period.

4. Conclusions

This paper presents the trading strategy optimization problem in detail and discusses how evolutionary algorithms (genetic algorithms in particular) can be effectively and efficiently applied to this optimization problem. Experimental results show that this approach is promising. Future work will focus on the exploration of solutions with more stable performance during the out-of-sample period.

Acknowledgement

This work was supported in part by the Australian Research Council (ARC) Discovery Projects (DP0449535 and DP0667060), National Science Foundation of China (NSFC) (60496327) and Overseas Outstanding Talent Research Program of Chinese Academy of Sciences (06S3011S01).

References

[1] E. Acar and S. Satchell, editors. *Advanced Trading Rules*. Butterworth-Heinemann, Oxford, 2 edition, 2002.

[2] C.-H. Park and S. H. Irwin. What do we know about the profitability of technical analysis? *Journal of Economic Surveys*, 2006.

[3] R. Sullivan, A. Timmermann, and H. White. Data-snooping, technical trading rule performance, and the bootstrap. *Journal of Econometrics*, 54:1647–1691, 1999.

[4] B. Kovalerchuk, E. Vityaev, and E. Vityaev. *Data Mining in Finance: Advances in Relational and Hybrid Methods*. Kluwer Academic Publishers, 2000.

[5] M. Gavrilov, D. Anguelov, P. Indyk, and R. Motwani. Mining the stock market: Which measure is best? In *Proc. of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 487–496, 2000.

[6] K. V. Nesbitt and S. Barrass. Finding trading patterns in stock market data. *IEEE Computer Graphics and Applications*, 24(5):45–55, 2004.

[7] D. Zhang and L. Zhou. Discovering golden nuggets: data mining in financial application. *IEEE Transactions on Systems, Man, and Cybernetics: Part C*, 34(4):513–522, 2004.

[8] L. Lin and L. Cao. Mining in-depth patterns in stock market. *Int. J. Intelligent System Technologies and Applications*, 2006.

[9] G. J. Deboeck, editor. *Trading on the Edge: Neural, Genetic, and Fuzzy Systems for Chaotic Financial Markets*. John Wiley & Sons Inc., 1994.

[10] S.-H. Chen, editor. *Genetic Algorithms and Genetic Programming in Computational Finance*. Kluwer Academic Publishers, Dordrecht, 2002.

[11] L. Davis, editor. *Handbook of Genetic Algorithms*. Van Nostrand Reinhold, New York, 1991.

[12] D. E. Goldberg. *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley Professional, 1989.

[13] A. Madhavan. Market microstructure: A survey. *Journal of Financial Markets*, 3(3):205–258, Aug. 2000.

[14] F. O. G. Ali and W. A. Wallace. Bridging the gap between business objectives and parameters of data mining algorithms. *Decision Support Systems*, 21(1):3–15, sep 1997.

[15] A. A. Freitas. A critical review of multi-objective optimization in data mining: a position paper. *ACM SIGKDD Explorations Newsletter*, 6(2):77–86, dec 2004.

[16] J. H. Holland. *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, Ann Arbor, USA, 1975.

[17] J. Ni and C. Zhang. A dynamic storage method for stock transaction data. In *Proc. of the IASTED International Conference on Computational Intelligence (CI'2006)*, pages 338–342. ACTA Press, 2006.

[18] Ashcon. http://www.ashkon.com/predictor/strategies.html, 2006.

[19] Stockcharts. http://stockcharts.com/education/IndicatorAnalysis/, 2006.

Applications of Data Mining in E-Business and Finance
C. Soares et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-890-8-25

An Analysis of Support Vector Machines for Credit Risk Modeling

Murat Emre KAYA [a,1], Fikret GURGEN [b] and Nesrin OKAY [c]
[a] *Risk Analytics Unit, Mashreqbank, 1250, Dubai, UAE*
[b] *Department of Computer Engineering, Bogazici University, 34342, Istanbul, Turkey*
[c] *Department of Management, Bogazici University, 34342, Istanbul, Turkey*

Abstract. In this study, we analyze the ability of support vector machines (SVM) for credit risk modeling from two different aspects: credit classification and estimation of probability of default values. Firstly, we compare the credit classification performance of SVM with the widely used technique of logistic regression. Then we propose a cascaded model based on SVM in order to obtain a better credit classification accuracy. Finally, we propose a methodology for SVM to estimate the probability of default values for borrowers. We furthermore discuss the advantages and disadvantages of SVM for credit risk modeling.

Introduction

Banks use credit risk modeling in order to measure the amount of credit risk which they are exposed to. The most commonly used technique for this purpose is logistic regression. In this paper, we compare the credit risk modeling ability of support vector machines (SVM) with logistic regression for two different types of applications. The aim of the first application is to classify the borrowers as "good" or "bad" so that the borrowers which are classified as "bad" are not granted any credit. The number of risk classes can be more than 2 as well e.g. 1 to k, "class 1" having the lowest risk and "class k" having the highest risk. By analyzing the distribution of the borrowers into the risk classes, management can take several decisions such as determining the margins or reducing the credit limits for the risky borrowers.

Another application of credit risk modeling is the estimation of the probability of default (PD) values. This application became more popular especially after the Basel 2 Accord. The Basel Committee on Banking Supervision released a consultative paper called New Basel Capital Accord in 1999 with subsequent revisions in 2001 and 2003 and new international standards for computing the adequacy of banks' capital were defined (see [1,2,3]). The new Basel Accord introduces the three major components of a bank's risk as: market risk, operational risk and credit risk. Among these components, banks are exposed to substantial amount of credit risk. One of the parameters which is required to calculate credit risk capital is the probability of default (PD) and therefore

[1]Corresponding Author: Manager Analytics, Risk Analytics Unit, Mashreqbank, 1250, Dubai, UAE; E-mail:me.kaya@gmail.com

most of the banks started to build PD models to estimate the probability of defaulting of their borrowers.

In the literature, several statistical and machine learning techniques were developed for credit risk modeling. One of the first statistical methods was linear discriminant analysis (LDA) (see [4,5]). The appropriateness of LDA for credit risk modeling has been questioned because of the categorical nature of the credit data and the fact that the covariance matrices of the good and bad credit classes are not likely to be equal. Credit data are usually not normally distributed, although Reichert reports this may not be a critical limitation [6]. Logistic regression (see [7]) model was studied to improve credit risk modeling performance. Also, the non-parametric k-NN (see [8]) model was tested on the problem of modeling credit risk. Other researchers have investigated classification trees [9] and various neural networks [10,11]. Classification trees have been shown to demonstrate the effect of individual features on the credit decision.

This paper includes the following sections: in the second section, we compare SVM with logistic regression for credit classification on German credit data set. In the third section, we propose a cascaded model based on SVM to obtain a more accurate model than the stand-alone SVM and logistic regression models. In the fourth section, we propose a methodology in order to estimate the PD values by using SVM model. Finally we discuss the advantages and disadvantages of SVM for credit risk modeling.

1. Comparison of SVM and Logistic Regression

In this section, we compare the performances of SVM and logistic regression for credit classification. The SVM algorithm has found various applications in the literature (see [12,13,14,15]) and is a global, constraint, optimized learning algorithm based on Lagrange multipliers method. SVM tries to separate two classes (for binary classification) by mapping the input vectors to a higher dimensional space and then constructing a maximal separating hyperplane which achieves maximum separation (margin) between the two classes. Solution of this problem in the high dimensional feature space is costly, therefore SVM uses a "kernel trick" instead of applying the ϕ function to project the data. Finally SVM tries to solve the following quadratic optimization problem in order to find the parameters w and b which define the maximal separating hyperplane (see [13]):

$$Minimize\ \frac{||w||^2}{2} + C\sum_i \xi_i \tag{1}$$

$$subject\ to\ \ y_i(w^T\phi(x_i)\ +\ b)\ \geq\ 1-\xi_i \tag{2}$$

$$\xi_i \geq 0. \tag{3}$$

Logistic regression [7] is a statistical technique where the dependent variable is a Bernoulli variable. It models the logarithm of the odds as a linear function of the independent variables, X_i. The model takes the following form:

$$logit(p_i) = ln(p_i/(1-p_i)) = \alpha + \beta_1 x_{1,i} + ... + \beta_k x_{k,i} \tag{4}$$

$$i = 1, ..., n \tag{5}$$

Logistic regression uses the maximum likelihood method to estimate the coefficients of the independent variables. Once the coefficients are estimated, the probability p_i (probability of default in our case) can be directly calculated:

$$p_i = Pr(Y_i = 1|X) = \frac{e^{\alpha+\beta_1 x_{1,i}+...+\beta_k x_{k,i}}}{1 + e^{\alpha+\beta_1 x_{1,i}+...+\beta_k x_{k,i}}} \tag{6}$$

In order to build a classification model with logistic regression, the following rule can be used:

$$c = \begin{cases} 1, & \text{if } p_i \leq 0.5 \\ -1, & \text{if } p_i > 0.5 \end{cases}$$

1.1. Experiments and Results

For the experiments, we used the German credit data set which was available in the UCI Repository [2]. The data set consists of 20 attributes (7 numerical and 13 categorical) and there are totally 1000 instances (300 bad and 700 good cases). It was produced by Strathclyde University and is also associated with several academic work [3]. Our aim in this paper was not to come up with the best model which out-performs all previously used models. We rather aimed to compare logistic regression and SVM in terms of their classification performance and tried propose a methodology for SVM to build PD models. The reason of using German credit data set was since it is one of the few freely available credit data sets.

The RBF kernel was used for the SVM model (see equation 7) which has the parameter γ. The optimum values of kernel parameter γ and penalty parameter C were found by using Grid search. Grid search tries different (C, γ) values within a specified range in order to find the optimum values. For this search, we used a Python script called "grid.py" which is available in the LIBSVM web site[4]. Finally the value of $C = 32.768$ and γ value of $\gamma = 0.000120703125$ were used to train the model.

$$K(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2),\ \gamma > 0. \tag{7}$$

The models were compared based on their accuracy on the German credit data set by using 10-fold cross validation. We divided the data set into ten partitions. Then, we iteratively took one of the ten partitions as the test set and the combination of the other nine were used to form a training set. The accuracy of a hold-out partition was defined as the number of correct classifications over the total number of instances in the partition. Accuracy of the 10-fold cross validation procedure was calculated by dividing the sum of the accuracies of all hold-out partitions by ten.

As seen in table 1, SVM has a slightly better accuracy than logistic regression. It should be noted that, we used only one data set and this is not enough to draw a general conclusion that SVM is a better credit classification technique than logistic regression. However, we can conclude that SVM gave a slightly better result than logistic regression for credit classification on German credit data set. From a model user's point of view,

[2] http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/

[3] http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)

[4] www.csie.ntu.edu.tw/˜cjlin/libsvm/

logistic regression always has the advantage of transparency over the black-box SVM and is still preferable in case of slight improvements. In our experiments, SVM did not give significantly higher accuracy than logistic regression, therefore logistic regression is still a better option. By using logistic regression, the model user can know which variables are used and how much important in the model.

Table 1. Accuracy comparison of SVM and logistic regression

Model	Accuracy
SVM	75.8%
LR	75.6%

2. A Two Layer Cascade Model based on SVM

2.1. Idea and Analysis

In this section, we aim to obtain a more accurate classification model by using SVM. The idea started with the following question: "How accurate is the classification of the instances which are close to the separating hyperplane?". In order to answer this question, we divided the feature space into two regions called "critical region" and "non-critical region", see figure 1.

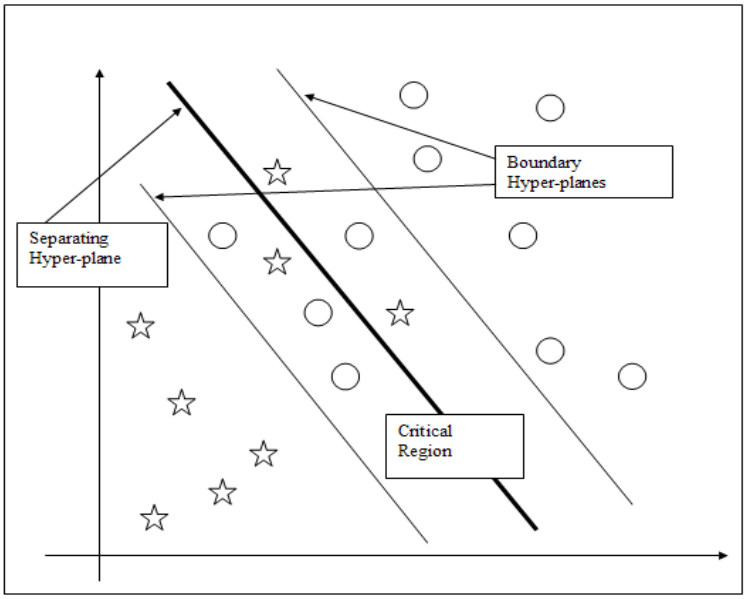

Figure 1. Boundary Hyperplanes.

We defined the critical region by using two hyperplanes which are parallel and close to the separating hyperplane $wx + b = 0$:

$$w.\phi(x) + b = \epsilon \quad \{boundary\ hyperplane\ 1\} \tag{8}$$

$$w.\phi(x) + b = -\epsilon \ \{boundary\ hyperplane\ 2\} \tag{9}$$

We then performed some experiments in order to answer the question above. We divided the German credit data into training and validation partitions which contained 490 good borrowers, 210 bad borrowers and 210 good borrowers, 90 bad borrowers, respectively. We then built the SVM model on the training partition and determined ϵ as $\epsilon = 0.4$ by trial and error. Table 2 shows the accuracies of SVM model for "critical" and "non-critical" regions on the validation partition. As shown in the table, nearly half of the incorrectly predicted data instances (35 of 72) lie in the critical region which is defined as the area between boundary hyper-planes. Accuracy in the critical region (43.5%) is also very low, however accuracy in the non-critical region is good with a value of 84.5%. That is, most of the predictions in the critical region are erroneous (56.5 %), so it can be concluded that, it is risky to trust on these predictions. By rejecting to classify 20.7 % percent of data instances (62 of 300 instances which lie in critical region), 84.5 % of accuracy can be obtained on the classified instances in the non-critical region.

Table 2. Region Comparison: Critical and Non-critical

Region	#(False)	#(True)	Total	Accuracy	Error Rate
Critical Region	35	27	62	43.5 %	56.5 %
Non-critical Region	37	201	238	84.5 %	15.5 %
Overall	72	228	300	76.0 %	24.0 %

2.2. The Cascade SVM-LR Model and Experiments

According to the analysis in the last section, SVM has a good performance for the classification of the instances which are not in the critical region. Coming from this idea, we propose a cascaded model which has a first layer model as SVM which will not classify the instances in the critical region. Rejected instances will be forwarded to the second layer model logistic regression. So the proposed cascaded model consists of two stages: a SVM stage and a logistic regression (LR) stage as in figure 2.

The classification rule of the first layer SVM model is modified as below:

$$c = \begin{cases} 1, & \text{if } w.\phi(x) + b \geq \epsilon \\ -1, & \text{if } w.\phi(x) + b \leq -\epsilon \\ Reject\ and\ forward\ to\ second\ layer, & \text{else} \end{cases}$$

There are several reasons to use logistic regression as a second layer. Firstly, it is very commonly used for credit risk modeling and gives good results. Secondly, if the stages of a cascaded model come from different theories and algorithms, it is probable that they will complement each other as stated in [15] which makes sense for SVM and logistic regression.

For the experiments, we divided the German credit data into training and validation partitions which contained 490 good borrowers, 210 bad borrowers and 210 good borrowers, 90 bad borrowers, respectively. We then built the SVM and logistic regression models on the training partition. We used $\epsilon = 0.4$ and according to the results, the cas-

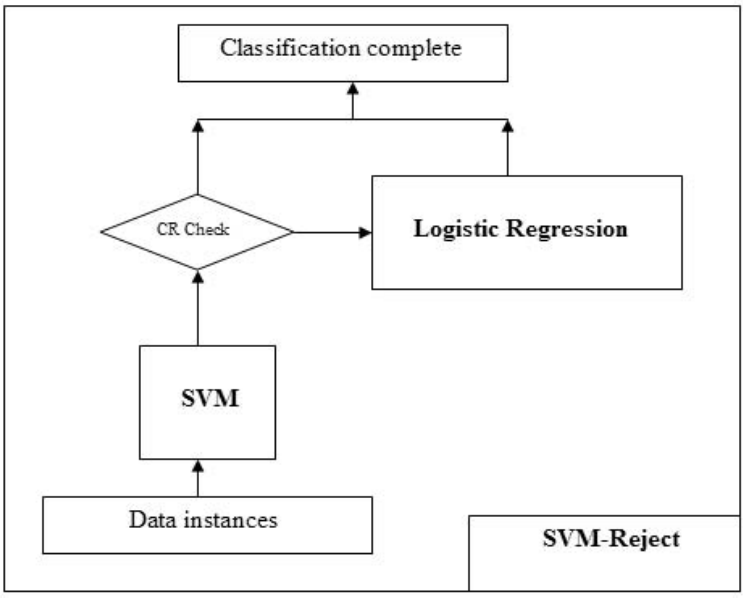

Figure 2. Cascade SVM and Logistic regression.

caded SVM-LR model has a better accuracy than the standalone SVM and LR models. It should be noted that, the accuracy difference between cascaded model and logistic regression is not very big but also is not insignificant as well. Therefore, the cascaded SVM-LR model can be good option for a model user since it brings a better prediction ability.

Table 3. Accuracy analysis of the cascaded model

Model	Accuracy
SVM-LR	% 80.3
SVM	% 76.0
LR	% 78.7

3. Probability of Default Modeling with SVM

In this section, we propose a methodology to estimate probability of default values by using SVM model. Here, instead of classifying the borrowers as "good" and "bad", the aim is to assign probability values ranging from 0 to 1 according to the following logic: If a borrower is "good", he/she should be assigned a low probability of default and if a borrower is "bad", he/she should be assigned a high probability of default.

Logistic regression is the most common technique for building PD models due to several reasons. Firstly, it is a transparent model (as we mentioned in the previous sections). Secondly and maybe most importantly, PD is directly available from the logit

score. Logistic regression models the log-odds (in our case, the logarithm of the ratio of bad borrowers to good borrowers) as a linear function of the explanatory variables as in the following formula:

$$logit(p_i) = ln(\frac{p_i}{1-p_i}) = \alpha + \beta_1 x_{1,i} + ... + \beta_k x_{k,i} \tag{10}$$

Here p_i represents the probability of the borrower to default and it is straightforward to calculate $p_i = P(Y_i = 1)$ from the logit (output of logistic regression function):

$$p_i = P(Y_i = 1) = \frac{e^{logit(p_i)}}{1 + e^{logit(p_i)}} \tag{11}$$

PD is not directly available from the output of SVM like logistic regression. SVM algorithm basically finds a separating hyper-plane in the high dimensional credit space to separate the two classes for binary classification:

$$w\phi(x) + b = 0 \tag{12}$$

Each credit data instance is then classified according to the following rule (class 1 corresponds to a "good" borrower and class -1 corresponds to a "bad" one):

$$c = \begin{cases} 1, & \text{if } w.\phi(x) + b \geq 0 \\ -1, & \text{if } w.\phi(x) + b \leq 0 \end{cases}$$

In order to estimate the PD values, we propose to use the output $w\phi(x) + b$. Instead of applying a classification rule on the output, we need to find a function which will transform each output value into a PD value. Our methodology makes the following assumption: "As output value increases, PD decreases exponentially" and includes the following steps:

1. Sort the instances of the training data according to the output values in ascending order.
2. Assign the N instances to k buckets in such a way that each bucket contains N/k instances. Due to the assumption above, as k increases the ratio of the bad borrowers in the bucket R_Bad_k should decrease. Therefore bucket 1 should have a very low R_Bad_k value and bucket k should have a very high R_Bad_k value.
3. Calculate the average default rate (DR) of each bucket B_i:

$$DR_{B_i} = (\sum_{x_j \epsilon B_i} y_j)/N_i \tag{13}$$

   where N is the number of instances in the bucket and y_j is the real label of the data instance x_j (0 if the borrower is "good" and 1 if the borrower is "bad").
4. Calculate the average output (O) of each bucket B_i:

$$O_{B_i} = (\sum_{x_j \epsilon B_i} w.\phi(x_j) + b)/N_i \tag{14}$$

   where N_i is the number of instances in the bucket and x_j is the input vector.

5. Run a linear regression between O and $ln(DR)$ (logarithm is applied due to the assumption of PDs decreasing exponentially with increasing output values) by using the values $(O_{B_i}, ln(DR_{B_i}))$ related to k buckets.

Linear regression determines the parameters a and b for the function $ln(PD) = a.o + b$ where o is the output value of a data instance. By this way, each output value can be converted into a PD by using the following function:

$$PD = e^{a.o+b} \tag{15}$$

4. Conclusion

In this study, we analyzed the ability of support vector machines for credit risk modeling from two different aspects: credit classification and estimation of probability of default values.

Firstly, we compared the credit classification performance of SVM with the commonly used technique logistic regression. SVM gave slightly better results than logistic regression on the German credit data set. It should be noted that, we used only one data set and this is not enough to draw a general conclusion that SVM is a better credit classification technique than logistic regression. It is also important to remember that logistic regression always has the advantage of transparency for a model user and still preferable even in case of small improvements.

Then, we proposed a cascaded model based on SVM in order to obtain a better credit classification accuracy. We observed a weak classification accuracy of SVM near the separating hyperplane (crtitical region) and strong classification ability in the non-critical region. Originating from this idea, we proposed a cascaded model based on SVM and achieved to obtain a better classification accuracy than the standalone SVM and logistic regression models.

In the final section, we proposed a methodology for SVM to estimate the probability of default (PD) values for borrowers. In the SVM model, PD is not directly obtainable like logistic regression. By using our methodology, the output of the SVM model can be transformed into a PD value.

References

[1] Basel Committee on Banking Supervision, 1999, *A new capital adequacy framework*, Bank for International Settlements, Basel, Switzerland.

[2] Basel Committee on Banking Supervision, 2001, *The new Basel Capital Accord, second consultative paper*, Bank for International Settlements, Basel, Switzerland.

[3] Basel Committee on Banking Supervision, 2003, *The new Basel Capital Accord, third consultative paper*, Bank for International Settlements, Basel, Switzerland.

[4] Henley, W.: Statistical aspects of credit scoring. Ph.D. thesis, The Open University, Milton Keynes, UK (1995).

[5] Altman, E.I.: Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. Journal of Finance 23 (1968) 589-609.

[6] Reichert, A.K., Cho, C.C., Wagner, G.M.: An examination of the conceptual issues involved in developing credit-scoring models. Journal of Business and Economic Statistics 1 (1983) 101–114.

[7] Galindo, J., Tamayo, P.: Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications. Computational Economics 15 (2000) 107–143.

[8] Henley, W.: A k-nearest neighbor classifier for assessing consumer credit risk. Statistician 44 (1996) 77-95.
[9] Tam, K.Y., Kiang, M.Y.: Managerial applications of neural networks: the case of bank failure predictions. Management Science 38 (1992) 926–947.
[10] Altman, E.I.: Corporate distress diagnosis: comparisons using linear discriminant analysis and neural networks (the Italian experience). Journal of Finance 18 (1994) 505–529.
[11] West, D.: Neural network credit scoring models. Computers & Operations Research 27 (2000) 1131-1152.
[12] Vapnik V.: The Nature of Statistical Learning Theory. Springer-Verlag, New York, (1995).
[13] Burges C.J.C.: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2 (1998) 121–167.
[14] Shawe-Taylor J., Cristianini N.: Kernel Analysis for Pattern Analysis, Cambridge University Press, (2004).
[15] Alpaydin E.: Introduction to Machine Learning, The MIT Press, London, England (2004).

This page intentionally left blank

Applications of Data Mining in E-Business and Finance
C. Soares et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-890-8-35

Applications of Data Mining Methods in the Evaluation of Client Credibility

Yang DONG-PENG, Li JIN-LIN, Ran LUN, Zhou CHAO

School of Management and Economics,
Beijing Institute of Technology, Beijing 100081, China
E-mail:ydp@bit.edu.cn

Abstract. Client credibility plays an important role in the financial and banking industry. This paper combines C4.5 and Apriori algorithms in the data mining process and discusses the relationship between these two algorithms. It then tests them using the WEKA software to acquire information from the analysis of the historical data of credit card clients. Finally, it offers decision-making support for the evaluation of client credibility.

Keywords. Data Mining, Decision Tree, Association Rules

Introduction

Data mining technology is a new and developing cross-disciplinary subject. With the development of computer technology, the application of data mining technology to analyze large amounts of realistic data has become a most advanced orientation in the field of data mining in the world. Data mining technology is often used together with many theories and technologies such as data bases, artificial intelligence, machine learning, statistics etc, and it is also applied in the industries of finance, insurance, telecommunications, and retail, which have accumulated a great quantity of data [1].

With the popularization and development of the credit card business in China, the data which has been accumulated over a long period has formed an information data base. The analysis of these data is not made simply because of the need for research, but mostly to offer real, valuable information to the decision-making of the policy-making body of banks.

The amount of a credit card means the upper limit of the money that can be used, which depends on the credibility of the customer. The amount of credit is evaluated according to the materials of application and documents provided by the proponent. A bank would calculate several factors such as data, job, savings and the condition of housing. When estimating the credit of a proponent, each item should be examined according to some criteria. Once the bank finishes assessing these risks and factors, the credit can be confirmed [2].

This article tries to find useful information in the enormous data by means of data mining to help banks understand the rules for evaluating clients' credit by analyzing the historical information of several primary commercial banks. It also tries to provide

support for decision-making so that banks can evaluate customers' credit applications correctly.

1. Classification Rules and Decision Tree Methods

The concept of supervised classification in data mining [3] is to learn a classification function or construct a classification model based on the known data, which is also called a classifier. This function or model maps the data in the data base to the target attribute, and can, therefore, be used to forecast the class of new data. The decision tree algorithm is one of the classification algorithms in data mining and it is based on principles from information theory. A decision tree algorithm builds models that can be deployed automatically, be easily interpreted by the user, and deal with unknown and noisy data. We can use this algorithm to learn only if the training instance can be expressed by the mode of "attribute-value".

A decision tree consists of inside nodes, branches and leaf nodes, which represent the structure of decision trees. The top node of the tree is called the root node; inside nodes represent tests that are carried out on the values of attributes, branches represent different results on the tests; and leaf nodes represent the classification of the examples that fall in the node.

2. C4.5 Algorithm in Decision Tree Methods

2.1. Application of Entropy

Entropy is the measurement of uncertainty or confusion of a system, and the information quantity of the system is the measurement of its degree of systematization [4]. When the instance is clearly known, the entropy is zero; when the possible state or outcome is variable, the entropy will increase.

The formula of entropy is:

$$S = -\sum_{I}(p_i * \log(p_i)) \tag{1}$$

The information gain in this paper means the decrement of entropy, according to which we can confirm what kind of attribute should be tested on a certain level.

2.2. Principle of C4.5 Algorithm

C4.5 algorithm is an influential and widely used algorithm in machine learning that contains some improvements to the earlier ID3 algorithm, including a different pruning method. We enumerate its merits as follows [5]:

- The C4.5 algorithm calculates the gain ratio of each attribute and the attribute that has the highest information gain ratio will be selected for the node. Using the gain ratio to choose the test attribute addresses a deficiency of ID3, which uses the information gain;

- The pruning may be carried out during the process of the growing the tree or after it has finished;
- It can deal with continuous attributes in discrete ways;
- It can deal with missing data;
- It can generate rules based on the tree.

We suppose that there are two classes, P and N, and that x and y in set S mean the number of records of the class P and N, respectively. So, the entropy of S is :

$$Info(S) = Info(S_p, S_n) = -(\frac{x}{x+y} * \log\frac{x}{x+y} + \frac{y}{x+y} * \log\frac{y}{x+y}) \tag{2}$$

We take variable D as the root node of the decision tree, and partite S to child node $\{S_1, S_{2,} ... S_k\}$, and each S_i (i = 1,2, ... k) includes x_i which means the number belongs to class P and y_i belongs to class N. So, the entropy of child node is:

$$Info(D,S) = \sum_{i=1}^{k} \frac{x_i + y_i}{x + y} * Info(Sip, Sin) \tag{3}$$

The information gain is:

$$Gain(D) = Info(S) - Info(A,S) \tag{4}$$

Then we can conclude the definition of information gain function:

$$Gain(D,S) = Info(S) - Info(D,S) \tag{5}$$

$$Info(S) = I(P) = I(P_1, P_2, \cdots, P_k) = I\left(\frac{|C_1|}{|S|}, \frac{|C_2|}{|S|}, \cdots, \frac{|C_k|}{|S|}\right)$$
$$= -(p_1 * \log p_1 + p_2 * \log p_2 + \cdots + p_k * \log p_k) \tag{6}$$

$$Info(D,S) = \sum_{i=1}^{n} (|S_i| / |S|) * Info(S_i) \tag{7}$$

We can see from the algorithm principle that the decision tree formed by the C4.5 algorithm is not only a model that estimates the state of the data, but what is most valuable is the meaning that the structure of the decision tree represents: the extent to which each factor of the decision tree influences the target attribute. Generally speaking, if a certain attribute is totally correlated to the target attribute, then the diversification of target attribute can be speculated by it.

According to the principles of the C4.5 algorithm, we know that the split on each node of the decision tree is carried out on the attribute which has the largest information gain ratio, that is to say, the attribute of each node represents the factor that influences the target attribute the most. The closer a certain attribute is to the root node, the more it can influence the target attribute.

3. Association Rules

3.1. Application of Association Rules

The purpose of association analysis is to figure out hidden associations and some useful rules contained in the data base, and can use these rules to speculate and judge unknown data from the already known information [6].

We suppose I=$\{i_1, i_{2,} ... i_m\}$ is a collection, and its elements are called items, D is a collection of transactions T, and T can be seen as the collection of items, and T⊆I. There is certain sign of each transaction, TID is its ID number, X is a collection, and if T⊆I, then we say T includes X.

An association rule is a formula as X⇒Y, X⊂I, Y⊂I, and X∩Y=Φ. The support of a rule X⇒Y is the proportion of transactions that include X and Y, i.e.

$$support(X \Rightarrow Y) = P(X \cup Y) \tag{8}$$

The confidence of a rule X⇒Y is the proportion of transactions that contain X which also contain Y, i.e.

$$confidence(X \Rightarrow Y) = P(X / Y) \tag{9}$$

Given a certain set of transactions D, the goal of association analysis is to figure out the rules whose support and confidence are bigger then certain minimum values (minsupport and minconfidence).

3.2. Apriori Algorithm

Apriori is a widely used algorithm in the mining of association rules. An important characteristic that Apriori takes advantage of is the following: if the support of a rule I is less than the minimum support, i.e., $P(I) < s$, then adding a new item A to I, then the support of the new rule with $(I \cup A)$ is also less than the minimum support, i.e., $P(I \cup A) < s$ [7].

This characteristic is used by Apriori during the search process, avoiding the exploration of rules which are guaranteed not to have minimum support and, thus, improving its efficiency.

The algorithm starts with a set with a single item I, search for another item L_1 and add L_1 to produce C_1, which is also a set. Then filter all the items of C_2 to produce L_2, and go on till L_k becomes an empty set.

4. Relationship Between the Two Algorithms

According to the representation of these two kinds of algorithms, we can find that they are similar to some extent. Their models can be represented by: $x \to y$ .The result of the classification rules is a fixed attribute, which is single. The result of the association rules is the combination of one or many items, such as: $x \to y \wedge z$, and it is uncertain. So, if we assign a fixed attribute as the target attribute, the association rules can become classification rules [8]. That is to say, if the result of association rules is a fixed item, like y, then we only need to find the rules associated with y, then the process turns to be the induction of classification rules. We can also say that classification rules are special instances of association rules.

5. Application Analysis

5.1. Data Preparation

We take some historical client information as the data source. After the collection, cleaning, integration and some other preprocessing work, we get a data set as the one in Table 1.

Table 1. Interface of data source, after the preprocessing work on the noisy, disorderly data sets that we collected. The table shows the orderly and integrated data sets.

Sex	Age	Marital status	Education	…
Male	>50	Single	High	…
Female	<30	Married	High	…
Male	30-50	Single	High	…
Male	30-50	Single	High	…
Male	>50	Single	High	…
Male	30-50	Single	High	…
Male	>50	Single	High	…
Male	30-50	Single	High	…
Male	>50	Married	High	…
Male	<30	Married	Low	…
…	…	…	…	…

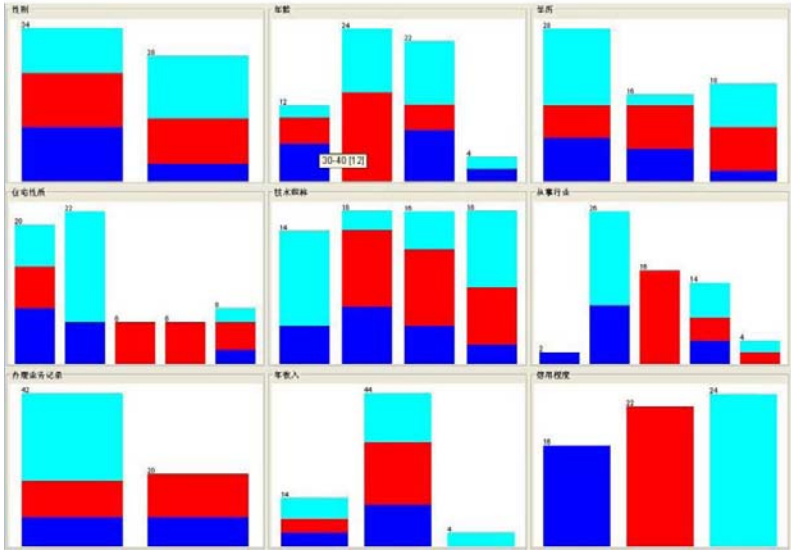

Figure 1. Here we visualize all the attributes of data sets after the preprocessing work.

5.2. Application of WEKA

Here we use WEKA software to do the mining process [9].

5.2.1. C4.5 Algorithm Training

We use the C4.5 algorithm to deal with the data and build the classification decision tree (Figure 2). We use the cross-validation method to estimate the error of the tree. In other words, we split the data randomly into 10 folds, and iteratively make each set to be the test data, and use the remaining as training data. Finally we compare the results, and select the results that have the highest accuracy as the resulting decision tree (Figure 3).

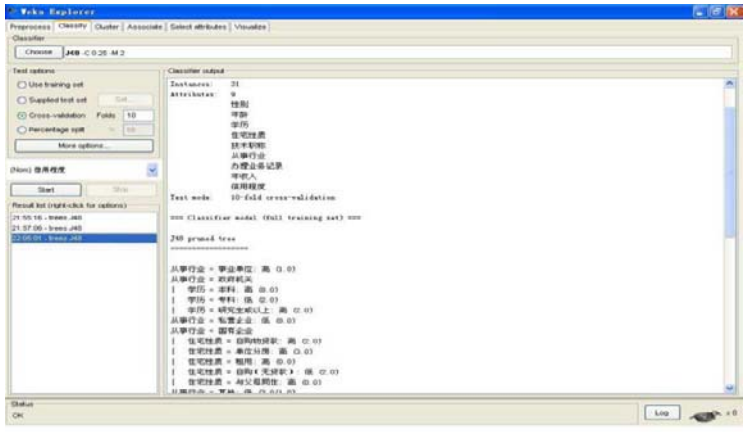

Figure 2. User interface of C4.5 algorithm.

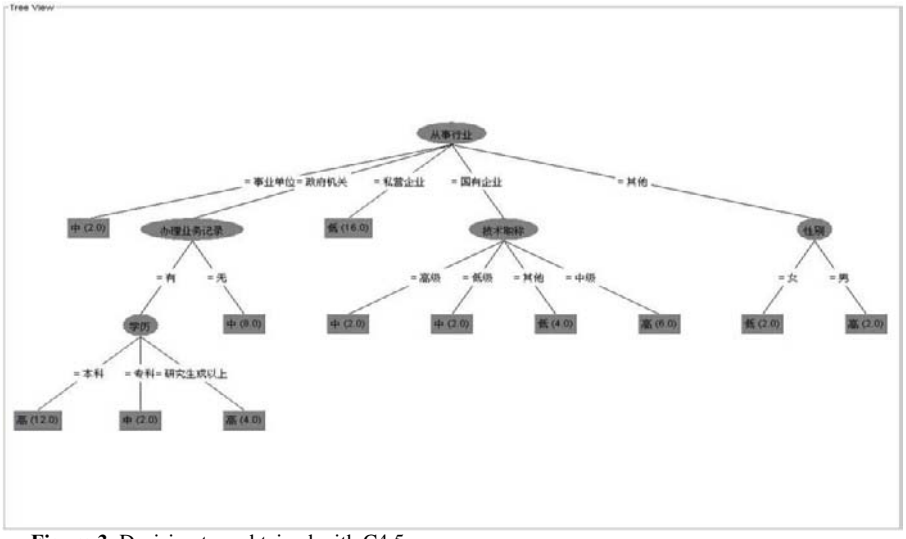

Figure 3. Decision tree obtained with C4.5.

5.2.2. Apriori Algorithm

We also use WEKA implementation of the Apriori algorithm (Figure 4).

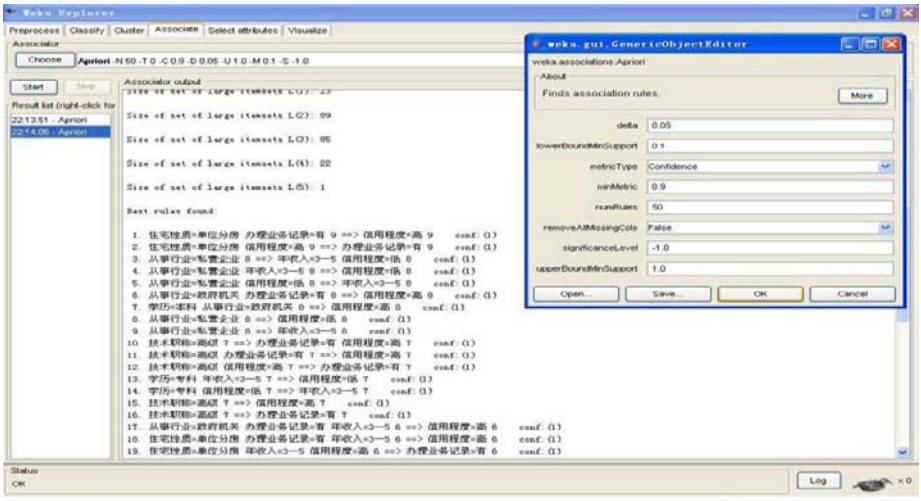

Figure 4. User interface of the Apriori algorithm.

5.3. Results

Based on the results obtained with the two algorithms, a number of evaluation rules of credibility degree can be concluded, which are listed as follows:

- People who have relatively more stable professions tend to acquire a higher degree of credibility, such as public officials of the national government, and employees of the enterprises with good benefits. In another instance, if the

monthly income of the applicant is comparatively low, but the nature of his profession is related to government and housing is provided, then he may also receive higher degree of credibility.

- The nature of personal housing can indicate the economical situation of the individual. A higher degree of credibility can be obtained if the individual's working unit supplies personal housing.
- The professional post is a testimony of the client's working ability. Comparatively, clients who have professional posts such as engineers with varied rankings, economists, chartered accountants, outstanding teachers tend to be favoured by the banks. Consequently, their degree of credibility would be promoted.
- There is not much difference between a junior college degree of qualification and a bachelor's degree when it comes to evaluating the degree of credibility. However, this rule would not apply when a higher degree of qualification is involved.
- People who have a detailed economic income certification and a stable income, as well as a long-term prospect for increased income would receive relatively high degree of credibility.
- If the client has accounts or credit cards in the bank, and there is regular cash flow, then those who used to support the develop of the client would generally receive higher credibility degree.

6. Conclusion

Through the application of data mining methods, this paper has collected and analyzed the historical information of clients of several major commercial banks, and has found a number of evaluation regulations which are shared by those banks when they evaluate the clients' degree of credibility. Moreover, the evaluation regulations obtained can give effective policy-making support to banks in evaluating their future clients' degree of credibility.

At present, there is a great deal of research which has combined the decision trees method with other methods such as association rules, Bayesian, Neural Network and Support Vector Machine [10].

However, further research is called for on how to apply those methods in the financial fields outside the banks.

References

[1] Shan Siqing, Chen Yin, Cheng Yan: Data Mining- Concept, Model, Method, and Algorithm[M]. Tsinghua University Publishing Company, Beijing (2003)
[2] Guo Ai, Liang Shiben. Evaluation and Analysis of Customer's Credibility[j]. System Engineering (2001)
[3] Tang Huasong, Yao Yaowen. The Discussion of Decision Tree Method in Data Mining[J]. Computer Application (2001)
[4] Margaret H.Dunham. Data Mining[M]. Tsinghua University Publishing Company,Beijing (2005)
[5] Wang Xiaoguo. Construction of Classification Decision Tree by Applying of C4.5 Algorithm[J]. Computer Engineering (2003)
[6] Agrawal R, Imielinski T, Swami A. Mining Association rules between sets of items in large databases[A]. Proceedings of the ACM SICMOD Conference on Management of Data[C].

[7] Bi Jianxin, Zhang Qishan. Algorithms of Association Rules[M]. China Engineering Science, Beijing (2005)
[8] Jiawei Han, Micheline Kamber. Concept and Technology of Data Mining[M].Mechanical and Industry Publishing Company, Beijing (2001)
[9] Ian H. Witten, Eibe Frank. Data Mining [M]. Mechanic and Industry Publishing Company, Beijing (2003)
[10] Luo Minxia. Technology and Application of Data Mining and Knowledge Discovery, Journal of Yuncheng University (2005)

This page intentionally left blank

Applications of Data Mining in E-Business and Finance
C. Soares et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-890-8-45

A Tripartite Scorecard for the Pay/No pay Decision-Making in the Retail Banking Industry

Maria Rocha SOUSA [a] and Joaquim Pinto da COSTA [a]
[a] *Faculdade Ciências Universidade Porto, Porto, Portugal*
e-mail: maria.rochasousa@portugalmail.pt
e-mail: jpcosta@fc.up.pt

Abstract. Traditionally retail banks have supported the credit decision-making on scorecards developed for predicting default in a six-month period or more. However, the underlying pay/no pay cycles justify a decision in a 30-day period. In this work several classification models are built on this assumption. We start by assessing binary scorecards, assigning credit applicants to good or bad risk classes according to their record of defaulting. The detection of a critical region between good and bad risk classes, together with the opportunity of manually classifying some of the credit applicants, led us to develop a tripartite scorecard, with a third output class, the review class, in-between the good and bad classes. With this model 87% decisions are automated, which compares favourably with the 79% automation rate of the actual scorecards.

Keywords. Pay/no pay decision, mass-market, tripartite scorecard

Introduction

The ubiquity of digital communications has led to the generalization of online payments in individuals' Demand Deposit Accounts (DDAs). Retail banks have to assure a prompt answer for those payment requests, which can be in the order of millions a day. When the DDA has insufficient balance the bank has to decide whether or not to pay that debit transaction (a pay/no pay decision-making) in a process named Non Sufficient Funds (NSF). This pay/no pay decision must be performed at the latest by the end of the day, to fit the Financial Net Settlement System service level's requirements. Optimizing this decision-making entails the decision to be uniform, objective and fast, with the minimum of mistakes and losses.

Currently at a retail bank, most of the decisions (79%) are automatically managed, while critical decisions are left for manual assessment. However, the automatic behavioural scoring models in use were developed to predict default in a six-month period; furthermore, to keep the implementation straightforward, they do not entirely emulate human reasoning. Therefore, some distinctive features of the problem are not materialized in them. Both customers' earnings and NSF process cycles take one month to be completed. Hence, if a "pay" decision is made, it is expected that the DDA cures within

30 days (the DDA is cured when it does not exceed its balance and overdraft limits). This led us to consider the development of a specific model to classify short-term credit risk for mass-market customers of the retail bank.

1. A Credit Model for the Pay/no pay Decision

The research summarized here was conducted by using a large real-life dataset (comprising 187733 records) from the credit data of the leading Portuguese retail bank. We choose the SAS Enterprise Miner package to perform all computations. Each customer's DDA in the sample was labelled as *good* if it cured within 30 days and as *bad*, otherwise. Rejected transactions were also included in the sample. They were assigned to *good* or *bad* classes according to their balance in the following month. The adopted class definition is based on Portuguese economic practices, as well as specific market segments, and therefore to the pattern of NSF and customers' earnings cycles.

In pay/no pay decision-making, human evaluation is usually supported in the existing information of the three-month period preceding the decision day. This human procedure was incorporated in the models using a three-month observation window. The sample comprises DDAs with pay/no pay decisions of an entire month, the decision period. For those DDAs, data were collected for the previous three-month period – the observation window. The performance was evaluated for each customer's DDA according to his behaviour in the 30-day period after having had a pay/no pay decision – the performance window [1]. The driving idea was to look to pay/no pay decisions in the decision period and evaluate whether the DDA cured in the following 30-day period. If so, the DDA was labelled as non-defaulter; if not, as defaulter.

The information gathered for each costumer's DDA comprises 47 characteristics related with the DDA transactional pattern (e.g. the structure and volume of monthly debits and credits and balance cycles) and customers' behaviour in their relation with credit. The current or past flawed experiences with financial institutions were included in the sample as well, such as missing payments and bankruptcy status.

The original sample dataset was randomly divided into three subsets: 70% for the training set, 20% and 10% to be the validation and test groups, respectively. The proportion of each target class in the actual population, 18% defaulter and 82% non-defaulter, was kept in the sample dataset.

The classifiers were trained both with an equal loss matrix and a loss matrix that integrates the cost of misclassification, empirically estimated using a sample of historical decisions.

1.1. Estimation of the loss matrix

In this application models are required to minimize the total loss associated with the decisions, rather than the number of errors. One of the most efficient approaches to build models that are sensitive to non-uniform costs of errors is to make the decision based on both the estimated probabilities of the unseen instances and a measure of business performance (profit, loss, volume of acquisitions, etc) [2]. We adopted the expected loss value for each possible decision.

For each approval in pay/no pay decision-making, the bank charges an amount M that equals the maximum between a fixed fee and the interests. When the transaction is a cheque, the bank charges an additional fee: f_+ if the cheque is paid, f_- otherwise. The estimation of the loss matrix was based on the following principles:[1]

- The error of classifying an actual defaulter as non-defaulter generates a loss that is equal to the value of the transaction;[2] since the mean value of cheques is higher, the costs of misclassifications was differentiated by group of transactions. Therefore, the expected cost of a bad decision in cheques, l_c, and the expected cost of a bad decision in other cases, l_o, was weighted by the expected proportions of cheques and other transactions in the true population, p_c and p_o. The expected loss is therefore $p_c\, l_c + (1 - p_c) l_o$.
- The error of classifying an actual non-defaulter as defaulter produces a loss corresponding to the fees that the Bank dos not charge/collect and the revenue from charging the fee f_-, in the case of cheque refusal. Weighting those fees by the corresponding proportion, the loss is given by $p_c\,(f_+ - f_-) + M$.

Although fees and interests are pre-defined, some scenarios can correspond to exclusions, decreasing the amount to be charged. Hence, rather than using the standard predefined fees, which would lead to unrealistic and inflated profits, matrix parameters were estimated empirically using a sample of historical decisions. Mean charged fees and the expected costs were then calculated for each of the two groups, cheques and others. Loss matrix parameters (normalized values) were estimated as 0.28, 0.22, 0.44, and 0.34 for p_c, M, f_+ and f_-, respectively. The normalized cost of misclassification is 18 if the transaction is a cheque, and is 10 for other type of transactions.

These principles allow a practical evaluation of the expected loss of a single decision in the pay/no pay decision-making, and can be summarized in a loss matrix that puts more weight on costumers wrongly predicted as non-defaulters with the proportion 1:49.

2. Binary Scorecard

Several standard binary classification models, based on logistic regression, classification trees [3] and neural networks [4,5], were designed from the same input dataset. More than just discriminating between the two classes, the models yielded a scored dataset as a result of their training.[3] Two different strategies were gauged: training the models to estimate only the probabilities of each class of the target variable, without incorporating any business objectives for which the predictor will be used. This strategy corresponds in adopting the equal-loss matrix, with which both types of errors are equally weighted. In a second strategy the training of the models incorporates the estimated business costs,

[1] Our approach for evaluating the loss of a pay/no pay decision does not incorporate indirect profits such as commercial benefits from keeping relation with good customers active, neither the costs of preserving bad customers. Although quantifying them would lead to valuable results it would also require considering some non-trivial business assumptions. As that would take us beyond the objectives of the current work, they were not considered.

[2] Although in practice the loss of misclassifying a defaulter is less than the value of the transaction, we considered the worst scenario in which the credit is totally lost.

[3] A scored dataset consists of a set of posterior probabilities for each level of the target variable, corresponding to the probability of defaulting and not defaulting.

focusing not in the minimization of the misclassification rate but in the optimization of the profit or loss. The selection of the cutoff for each case is easily determined. If the probability of defaulting p_d of a given costumer is known, the best cutoff for a general loss matrix $\begin{bmatrix} l_1 & l_2 \\ l_3 & l_4 \end{bmatrix}$ is determined by comparing the expected loss of predicting as defaulter, $l_1\ p_d + l_3\ (1 - p_d)$, with the expected loss of predicting as non-defaulter, $l_2\ p_d + l_4\ (1 - p_d)$. The resulting cutoff is $\left(1 + \frac{l_2 - l_1}{l_3 - l_4}\right)^{-1}$. For the equal-loss matrix case, the threshold is 0.5; for the estimated loss matrix, the threshold is 0.02. The best results are summarized in Table 1. For each model, the minimum loss, the sensitivity

Table 1. Results for the best binary models.

(a) Minimization of business rules.

Model	Loss	Specificity	Sensitivity	Error rate
Logistic Regression	0.724	27.2%	98.6%	59.9%
Decision Tree	0.697	28.4%	98.8%	58.9%
Neural Network	0.697	34.3%	98.2%	54.1%
Naïve 1	0.820	0.0%	100.0%	82.0%
Naïve 2	8.820	100.0%	0.0%	18.0%

(b) Minimization of the error rate.

Model	Loss	Specificity	Sensitivity	Error rate
Logistic Regression	0.094	98.4%	55.2%	9.4%
Decision Tree	0.083	97.8%	64.5%	8.3%
Neural Network	0.089	98.1%	59.6%	8.9%
Naive 1	0.820	0.0%	100.0%	82.0%
Naive 2	0.180	100.0%	0.0%	18.0%

(percentage of actual defaulters predicted as defaulters), the specificity (percentage of actual non-defaulters predicted as non-defaulters), and the error rate are provided. As a reference performance, the results for two baseline classifiers are also presented. The Naive 1 model refuses all examples, while Naïve 2 model classifies all as non-defaulters.

Models tuned with the matrix incorporating the business rules have high sensitivity (above 98%), while their specificity is low (below 35%). This strategy led to models with high error rates. When the models were developed to minimize the error rate the results were essentially reversed. The error rate of these models is mostly due to misclassified defaulters in the set.

2.1. Tripartite Scorecard

The results attained with the binary classifiers show that none could discriminate the defaulter from the non-defaulter in a satisfactory way. We also observed a certain overlap between the distribution of the defaulter and of the non-defaulter, when analysed over the predicted probability of defaulting, meaning that the models were not effective in distinguishing them. When varying the cutoff value we are just trading off between the two types of possible errors. Pushing the cutoff near the values obtained for the estimated

matrix, almost all defaulters are correctly predicted, while most of the non-defaulters are incorrectly predicted as defaulters. Relaxing the cutoff to values around the value obtained for the default matrix, the errors are reversed. When deploying a system of this kind in a retail bank, there is the opportunity of defining a third type of decision, the review class: an example predicted as review will be evaluated manually by human experts, possibly making use of some additional information. Therefore, we investigated the possibility of designing models with three output classes [6]: defaulter, review and non-defaulter. Bipartite and tripartite scorecards have been used in the industry before, but only in an ad hoc way, with no effort being made to find the optimal division [7].

Table 2. Confusion matrix for a three-output model.

	Predicted Defaulter	Review	Predicted Non-Defaulter
True Defaulter	p_1	p_2	p_3
True Non-Defaulter	p_4	p_5	p_6

Considering the generic confusion matrix for a three-output model (Table 2) the training of the models was driven to find two cutoffs simultaneously that provide low error rates p_3 and p_4 (assuming that all manual decisions are correct) and high automation rate. The lack of standard formulations and implementations to solve a problem of this kind, led us to start with a simple approach. Starting on the models previously designed, the two cutoffs were determined as follows:

- A cutoff was initialized as 0.0. Next, it was iteratively raised until a predefined probability p_3 (= 0.025) was obtained.
- A cutoff was initialized as 1.0. Next, it was iteratively lowered until a predefined probability of error p_4 (= 0.050) was obtained.

Finally, the percentage of automatic correct decisions ($p_1 + p_6$), the percentage of defaulters in the approved set (p_3/p_6), the error rate ($p_3 + p_4$) and the automation ($p_1 + p_3 + p_4 + p_6$) were measured. Three models, presented in Table 3, were chosen from all under evaluation.

Table 3. Tripartite Scorecard Results.

Model	Cutoff Low	Cutoff High	Specificity	Sensitivity	Approved defaulters	Error rate	Automation
Logistic Regression	10.5%	30.5%	92.6%	84.8%	3.5%	7.2%	82.1%
Decision Tree	7.0%	23.7%	93.2%	80.2%	4.9%	8.0%	86.6%
Neural Network	11.0%	27.3%	92.8%	84.2%	3.7%	7.4%	85.0%

The three-class output models have more balanced measures of sensitivity and specificity, with a better prediction of the true classes. About 15% of the decisions, corresponding to the overlapping region, are left for human assessment.

Assuming that the percentage of actual defaulters approved automatically does not consider the effects of the recovery actions that can be performed, we accepted a value up to 5%. The Tree based model was considered the most adequate for the pay/no pay decision-making, providing 87% of automatic decisions. Furthermore, a Decision Tree model is suitable for deployment and explanation of the decisions.

3. Discussion

This study focuses on the development of a scorecard for supporting the evaluation of default risk in the pay/no pay decision-making of a retail bank.

Binary classification models were developed based on well-known classification techniques. Although an extensive study was conducted, the attained discrimination between the two classes (default and non-default) was not satisfactory. When the weights of the two types of errors are heavily asymmetric, the boundary between the two classes is pushed near values where the highest cost error seldom happens. For equal misclassification losses, the boundary is biased to predict accurately the dominant class. Therefore, the research continued with the development of tripartite scorecards, with a third output class, the review class, in-between the good and bad classes. The final model enables 87% of automatic decisions, comparing favourably with the actual scorecards.

More principled approaches for optimally determining the boundaries between the good, review and bad classes are currently being investigated. A complementary model is also being developed for managing the resulting credit in arrears.

References

[1] Lyn C. Thomas, David B. Edelman, and Jonathan N. Crook. *Credit Scoring and its applications*. SIAM, 2002.

[2] Roger M. Stein. The relationship between default prediction and lending profits: Integrating roc analysis and loan pricing. *Journal of Banking & Finance*, 29:1213–1236, 2005.

[3] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[4] B.D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

[5] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[6] K. B. Schebesch and R. Stecking. Support vector machines for classifying and describing credit applicants: detecting typical and critical regions. *Journal of the Operational Research Society*, 56:1082–1088, 2005.

[7] David J. Hand, So Young Sohn, and Yoonseong Kim. Optimal bipartite scorecards. *Expert Systems with Applications*, 29:684–690, 2005.

Applications of Data Mining in E-Business and Finance
C. Soares et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-890-8-51

An Apriori Based Approach to Improve On-line Advertising Performance

Giovanni GIUFFRIDA[a], Vincenzo CANTONE[b] and Giuseppe TRIBULATO[c]

[a] *Università di Catania-Italy, Dipartimento di Matematica e Informatica,*
giovanni.giuffrida@dmi.unict.it
[b] *Proteo, Catania-Italy, Research & Development,*
vincenzocantone@supereva.it
[c] *Università di Catania-Italy, Dipartimento di Matematica e Informatica,*
tribulato@dmi.unict.it

Abstract. On-line advertising is booming. Compared to traditional media, such as Press and TV, Web advertising is cheap and offers interesting returns. Thus, it is attracting more and more consideration by the industry. In particular, it is now a consistent part of the marketing mix, that is, the set of different approaches to advertise a product. Data mining based optimization on Web advertising can take place at many different levels. From a data miner perspective, Internet advertising is a very interesting domain as it offers a very large amount of data produced at fast pace with a rich and precise amount of details. It also offers the valuable possibility of live hypothesis testing. Here we discuss an Apriori based optimization experiment we performed on live data. We show how effective such optimization is.

Keywords. Internet advertising, Targeting, Apriori, Association rules, one-to-one targeting.

Introduction

Today online advertising is booming, year after year it is experiencing a double digit growth. While, in absolute terms, investments on internet advertising still represent a small percentage on the overall advertising industry expenditure – especially compared to Press and TV – its share is expected to reach a considerable 20% in the next few years [1]. Reasons for such terrific growth are manifold. For instance it was found that the immediate recall of an online static banner ad was 40% compared with the 41% for a 30-second television commercial [2]. Considering the enormous cost difference between the two types of ad, the reason for today's large interest on Web advertising seems quite obvious. Compared to other media, online advertising exhibits the following unique interesting features.

Precise measurability: Performance of Internet campaigns can be measured very precisely compared to other media. In its simplest form, it is enough to count the number of clicks received in relation to the number of exposures in each campaign. Data are usually stored on detailed text log files, which can be easily pre-processed to be analyzed with different models.

Dynamic banner exposure: The banners shown on a web site can change at any time in a very dynamic way [12]. The choice may be function of many parameters such as time of the day, day of the week, daily limits, etc. Perhaps, this represents the most dramatic change compared to other media, such as Press, where the advertising is chosen before publishing. Internet editors need only design the placeholders for banner ads; those spots are later filled dynamically during users' browsing.

Contextuality: This is another very powerful feature of online advertising. Banners can be tailored in a dynamic way depending on the context surrounding it [1]. For instance, if one is reading an article on newborns the banner next to it can show powder milk.

User tracking: User activity can be easily tracked over time in different ways. A typical approach is through usage of cookies stored on customers' computers. This allows sophisticated analysis to be carried out using the accumulated data.

One-to-one targeting: This is the counterpart of the previous point. In particular an online campaign can be tailored for each specific user based on his/her interests and needs. For instance, if we find out through our data analysis that a user is looking to buy a new car we can target the ads he/she sees while browsing the web site. Of course, an intelligent model has to take into consideration that once she buys the car, her interests will rapidly change. Thus, the targeting system should detect as quickly as possible this loss of interest and change the type of ads she is exposed to.

Data volume and availability: Internet logs produced by medium to large web sites grow at a really fast pace and can produce gigabytes of data in a matter of few days. For data analysts (statisticians as well as data miners) this is in general good news, as complex models require large data size to be properly trained. In general, these logs are immediately available, thus, real-time (or semi real-time) models can be developed.

Dynamic model testing: This is a very interesting feature for data miners and statisticians. On the Internet it is straightforward to set up a model testing environment in order to optimize the model under development. For instance, in order to compare two different models (or to test a variation to an existing one), we can split the Internet users visiting our site in two panels. The different models treat users in each panel during the same time period. Hence, performance of the two approaches can be compared accurately. Models can even change many times within the same day. In some cases reactions of users to different models can be measured in real time. This can trigger continuous tuning and refinement of the model under development.

In Table 1 we compare the applicability of the above seen features to Internet compared to other two traditional media: TV and Press.

Table 1. Feature comparison by different media

Feature	TV	Press	Internet
Dynamic content placement	X		X
Direct measurability			X
Data precision			X
Contextualization	X	X	X
One-to-one targeting			X
Dynamic control	X		X
Easy/Dynamic model testing			X
Very large dataset size for modeling			X

Basically, intelligent online advertising helps everyone in the advertising chain, from the advertisers to the users. Today people continuously experience an overload of commercial information. The advertising industry has mostly been following a sort of broadcast approach: let us hit many; out of these many some (typically a very small percentage) will be interested in our product. After all, this approach works and it is justified by the characteristics of the old media. Today, new media such as the Internet, mobile phones, and interactive TV allow a more targeted approach to advertising. The intelligent application of available technologies and analysis to target better online advertising is beneficial for both the advertisers (the big spenders) as they can spend less and for the people as they may see fewer, but more interesting, ads.

In the following we discuss a live experiment we conducted on real data on a real Web site. We will be showing that our approach provides great improvement for both the click-through-rate (CTR)[1] of a specific banner ad as well as the CTR of the entire system.

1. Related Work

While data mining techniques have been used quite extensively on E-Commerce systems there is a lack of scientific data mining work on Internet advertising.

Sharif et al. [3] use Apriori as recommendation engine in a E-commerce system. Based on each visitor's purchase history the system recommends related, potentially interesting, products. It is also used as basis for a CRM system as it allows the company itself to follow-up on customer's purchases and to recommend other products by e-mail. Ale and Rossi [11] propose a temporal extension to the Apriori rule extraction to take into consideration the product market entering time. That is, a newly introduced product that sells very well may go below support in comparison with longer life products — even though they do not sell as well in more recent times. Moreover, a high-support product may have some other temporal restrictions (i.e., it may go out of the market), thus, it may be necessary to dismiss association rules associated with it. They introduce the concept of temporal support.

Other temporal extensions to Apriori have also been proposed in the past. The seasonality problem was well stated by Ozden et al. [4] (they call it "Cyclic"). Sales may be characterized by seasonality trends. For instance, certain products (e.g., ice tea) sell better at summertime. This concept may get more granular with certain correlations among products increasing on the first days of each month or on certain hours of the day. They propose a method to discover cyclic association rules. The same problem was also considered by others, Verma and Vyas [5] propose an efficient algorithm to discover calendar-based association rules: the rule "egg → coffee" has a strong support in the early morning but much smaller support during the rest of the day. Zimbrao et al. [6] extend the seasonality concept just mentioned to include also the concept of product lifespan during rule generation and application.

1 The click-through-rate is the ratio between the number of clicks a banner ad receives over the number of times such banner was seen by any user. It is a basic measure of banner performance.

2. Problem Statement and Our Approach

In our experiments we model the market basket concept [7] for banner ads running on a real web site. That is, for each user on the site, we collect all banners the user has clicked on. Thus, in our approach, a basket is defined by the set of clicked banners made by each individual user, while an item is the banner itself. Basically, we assume that if the same user clicks on two (or more) different banners, these are somehow correlated. As in the Apriori approach, the more users click on the same combination of banners the stronger the correlation among these banners grows [8].

The experiments described here have been fully integrated with a real ad-serving system running on real sites and thus tested on a real setting. An ad-server is a complex software application to manage all advertising needs on a set of web sites. An online advertising campaign is composed by different banner ads that can appear on different sites on different pages. Each campaign can be restricted in various way such as: start/end date, time of the day, day of the week, number of banner impressions[2] per day, overall number of impressions, number of unique users per day, etc. Upon a request for a banner, an ad-server first filters out all banners that do not satisfy at least one of the stated restrictions. Then, among all remaining banners, it picks one based on same selection algorithm and shows it to the user. For the sake of our discussion we can assume that such selection is performed in a random way. This is the point where we used our Apriori based selection algorithm.

For each user visiting any of the sites managed by the ad-server provided, the set of banners each user clicks on is stored into the browser cookie. For each clicked banner we also store info on the time when the click took place. Moreover a log of each user activity is stored into a log file that is passed to the back-end at nighttime. We run Apriori once a day on the set of log files provided by the different servers. The generated rules are then pushed back to the front-end servers. Such front-ends integrate the produced Apriori rules into their run-time serving logic.

2.1. Apriori Rule Extraction

We use Apriori to extract a set of rules of the form A1&A2&…&An → B, that state a certain probability of clicking on banner B once A1,A2,…,An have been already clicked.

We first preprocess all log files produced by each front-end to prepare the data in a format that could be processed by the Apriori algorithm — we used the Apriori implementation described in [9]. The data preprocessor collects all banners (items) clicked by a single user within a specified time-frame and puts them together (basket) on a single line of text. Then we run Apriori on this file.

An example of some Apriori rules we extracted from a relatively small data sample is the following:

54807 ← 54808 (0.1/157, 32.5)
76188 ← 76299 (0.0/40, 22.5)
61211 ← 75381 (0.0/46, 28.3)
75702 ← 75907 (0.0/86, 22.1)

2 By banner impression we mean a banner exposure, in other words, every time a user sees a banner on a web site we refer to it as a banner impression.

55577 ← 75860 (0.0/96, 22.9)
54807 ← 65301 52976 (0.0/25, 36.0)
64931 ← 70257 70713 (0.0/26, 30.8)
52976 ← 54808 54807 (0.0/51, 37.3)
54807 ← 54808 52976 (0.0/46, 41.3)

Consider the first rule: 54807 and 54808 are banner ids; 0.1 states the rules support in percentage over the total training data size; 157 is the actual number of baskets used to generate the rule; 32.5% is the rule confidence.

Once rules are generated they are post-processed into a specific proprietary format and pushed back to the front-end servers.

2.2. Rule Application

As discussed above, rule application takes place at run-time right after the ad-server filters out all non-applicable banners. At this point we are left with a certain numbers of banners, which are all eligible to be shown to the user.

Thus, upon a user's request for a banner impression, we use the extracted association rules according to the following algorithm:

```
For each eligible banner B
    If there is a rule R:"A1,...,An -> B" then
            If the user has clicked A1,...,An in the past and did not click yet on B
            then
                    Pick B with probability: conf(R)*decay(B)
            Endif
    Endif
Endfor
```

Basically, every banner on any rule right-hand-side is selected with a probability proportional to the rule confidence itself. We also use a decay function as multiplier to model a wear out effect. That is, the more often a banner is shown to the same user (without being clicked) the more we reduce the probability that such banner is selected for showing (to that user). Such decay has been computed empirically by observing historical data; its trend is shown in Figure 1.

In order to model such decay trend for each single user we also store in each user's cookie the information on the number of times each banner has been seen by that user.

2.3. Performance Measurements

We measured two types of performance: banner-based and overall. In the first case we are more interested in maximizing the improvement of a specific banner due to our approach. In the second case we measure the overall performance of the entire ad-server system. Depending upon specific requests one type of optimization could be of more interest than the other one. We now discuss these two types of measures in more detail.

2.3.1. Banner-based Performance Measure

In this case we are interested in measuring the CTR improvement of a specific banner due to the application of a discovered association rule. Let us assume to have discovered the rule: A → B. In order to measure the improvement, we split the customer base into two sets: C_A and $C_{\neg A}$. In particular, C_A includes all users who already clicked on banner A in the past and did not click on B yet; while $C_{\neg A}$ comprises all other users. We performed our experiment by showing banner B to all users in C_A according to the algorithm discussed above. At the same time all users in $C_{\neg A}$ were shown banner according to random selection. After a long enough period of time we measure the CTR of banner B in both sets. We define the lift in this case as the ratio between the CTR of banner B measured on C_A over the one measured on $C_{\neg A}$.

A banner specific optimization can be of great interests for specific advertisers who could pay a premium price for optimized performance of the ad-server for their specific ads.

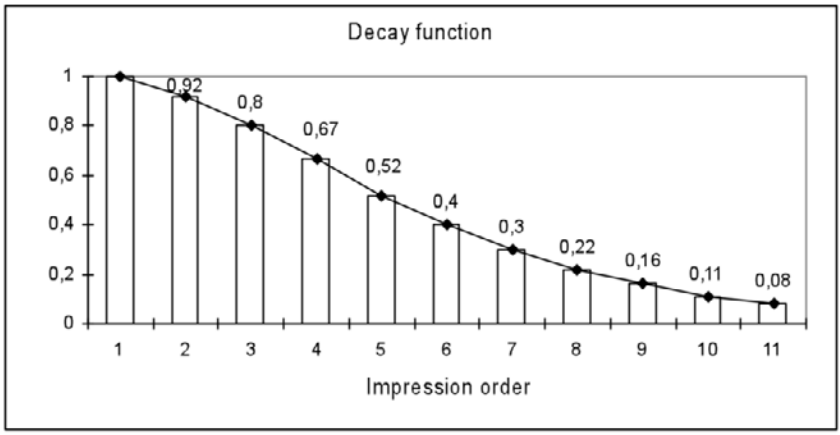

Figure 1. Impression order decay function to simulate wear-out effect. We multiply this value for the estimated CTR of a banner to a certain user. The more the user sees the same banner the more we reduce the estimated CTR.

2.3.2. System Overall Performance Measure

In this case we measure the system overall performance improvement. Hence we do not focus on any particular banner improvement per se but on the overall system improvement. In other words, by shuffling banner impressions among users based on the Apriori rules we measure the increment of the total number of clicks the system generates. This type of measurement is particularly important for the ad-serving service provider as it can provide a more efficient service for all advertisers on its network. Basically, they get a better return on their investment as with the same amount invested they receive more visits on their site.

In order to have an accurate measure of our performance we split randomly the customer base into two panels: test and execution. Users in the test panel are shown banners according to a random selection algorithm without the Apriori based optimization. While users in the execution panel are subject to the Apriori optimization. In this case, our lift is the ratio: ctr(execution panel) / ctr(test panel).

By splitting the user base into two groups and comparing the ctr of the two panels running in parallel during the same time period we get unbiased performance estimation. That is, we remove from our estimation any error due to time based banner CTR change. Banner CTR may change over time (from day to day and even from hour to hour); it may also depend upon environmental conditions.

Notice also that in this specific case we only optimize on banner CTR. Even though banner revenue information was available we only use such information for final reporting on performance of our model. Including revenue information, although possible, would complicate the model as certain business constraints need to be taken into consideration. For the purpose of our research such additional complication was deemed not to be relevant. However, our model could easily be modified to consider revenue in the objective function.

3. The Experiments

The basic aim of our experiment is to exploit Apriori rules [9] derived by banner click logs in order to increase the CTR of some banners. In this section we discuss first the data we used and then the results of the experiment itself.

3.1. Data Collection

As already mentioned, data on Internet are not a scarce resource. In our testing environment we collected more than 30 Million log lines every single day. Each line describes one of the two possible actions we are interested in: impression and click. For each of those actions we collect a substantial set of information such as user-id, time, position on the site, banner-id, etc. All data collected in our logs are anonymous, that is, no personal and/or sensitive data of any form are necessary in our system. Notice also that we do not need to know any specific graphical feature of the banners, that is, we just rely on their unique id without considering things like size, shape, colors, category, etc. For our experiment we took a sample of data composed of about 60 Million log lines collected over a month period. This sample includes only click logs as impression data are not relevant for deriving association rules. These logs are produced run-time by the system numerous front-ends. Such logs are divided into a set of files of equal number of lines. They are passed at nighttime to the backend system. On the backend we run a custom application that produce all files ready to be processed by Apriori.

3.2. Results

By running A-Priori [9] on our data set we collect a set of association rules where all items in the rules represent banners clicked by users [10]. We will refer to the click-through rate of banner A as CTR(A). We will use the notation CTR(A | B) to denote the click-through rate of banner A among all users who have previously clicked on banner B.

We now show the results of applying our system to the discovered rule 31 → 189, where 31 and 189 are two different banner ids. In our experiment we considered only highly performing banners in order to have a wider data set and make our analysis more robust. The click-through rate of banner 189 computed during the considered

time period among all users is 3.47%. This value was pretty much constant on every day of the month long period we considered in our experiment. Table 2 shows the day-by-day results of our experiment.

Table 2. Day-by-day lift detail by applying a specific association rule

Day	Impr(31)	Impr(189|31)	Clicks(189|31)	CTR(189|31)	Lift
1	610785	2	0	0,0%	0.00
2	457023	1	0	0,0%	0.00
3	359995	1	0	0,0%	0.00
4	81270	0	0	0,0%	0.00
5	89166	1	0	0,0%	0.00
6	101865	0	0	0,0%	0.00
7	519119	4	0	0,0%	0.00
8	542609	659	82	12,4%	3.58
9	369882	1142	129	11,3%	3.25
10	419101	90	11	12,2%	3.52
11	313450	833	102	12,2%	3.53
12	137743	1113	128	11,5%	3.31
13	130238	1027	88	8,6%	2.47
14	225958	1154	107	9,3%	2.67
15	256673	1077	112	10,4%	2.99
16	331203	1022	84	8,2%	2.37
17	358906	1013	92	9,1%	2.62
18	273730	904	73	8,1%	2.33
19	153346	850	71	8,4%	2.41
20	171469	912	97	10,6%	3.06
21	343501	1450	120	8,3%	2.38
22	224614	1367	125	9,1%	2.66
23	246541	1266	112	8,8%	2.55
24	163615	1124	95	8,5%	2.43
25	117600	922	82	8,9%	2.56
26	47568	718	71	9,9%	2.85
27	51261	825	65	7,9%	2.27
28	117272	924	100	10,8%	3.12
29	119402	817	60	7,3%	2.12
30	106071	736	67	9,1%	2.62
31	7280	32	1	3,1%	0.90

On the 1st column we have the day of the month. The 2nd column reports the number of impressions of banner 31 on each day. The 3rd and 4th column contain, respectively, the number of impressions and the number of clicks of banner 189 among users who have been previously clicked on banner 31. In the 5th column we find the click-through rate of banner 189 computed only among users who previously clicked on banner 31. The last column reports the lift computed as the ratio between the CTR on the 5th column, that is, computed only among users who clicked on both banner 31 and 189 over the CTR of 189 computed among all other users.

On the first seven days of the month we intentionally did not force the application of the rules, thus, the small values reported on the 3rd column are only by chance. That is by a random probability that banner 189 is shown to a user who previously clicked on 31.

For this specific rule, the CTR banner improvement for the entire period was 2.76, which means that by applying that rule we increase the probability of somebody clicking on banner 189 by 2.76 times. This is a quite impressive results considering how simple the model is.

We now analyze the performance of five banners optimized through our approach on another site. In Table 3 we report the lift performance for the entire period (fifteen days) of application of our discovered rules.

Table 3. Lift detail for five banners optimized by means of Apriori

Banner id	Optimized CTR	Non-optimized CTR	Lift
52976	0.151%	0.073%	2.06
75436	1.195%	0.869%	1.38
64931	8.025%	1.634%	4.91
53586	2.759%	1.618%	1.71
75926	4.498%	0.938%	4.80

The 1st column is the banner id. The 2nd and 3rd report the banner CTR computed, respectively, with and without our optimization. The 4th column reports the lift. Notice that there is an improvement in all cases. Moreover, such improvement gets really interesting in some cases going up to 491% (third row) and 480% (last row). In the worst case we get a CTR improvement of 38% (second row), which is still a very respectable performance.

We now analyze the overall improvement we were able to achieve using the proposed system. In Figure 2 we show the overall click improvement by different minimum support thresholds. In all cases we obtained a lift compared to the non-optimized algorithm (column "real" in the figure). It is interesting to notice how decreasing the Apriori minimum support yields a better overall performance. The improvement flattens out after the value of 50 for the minimum support. In this case we improved the overall system performance, measured by overall number of clicks produced, by 45%. This was considered an outstanding result.

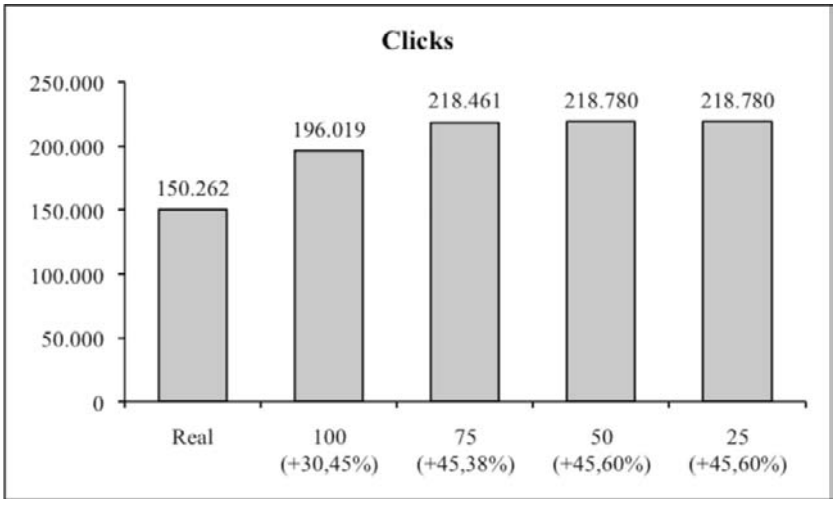

Figure 2. Overall improvement in the number of clicks obtained by varying minimum support.

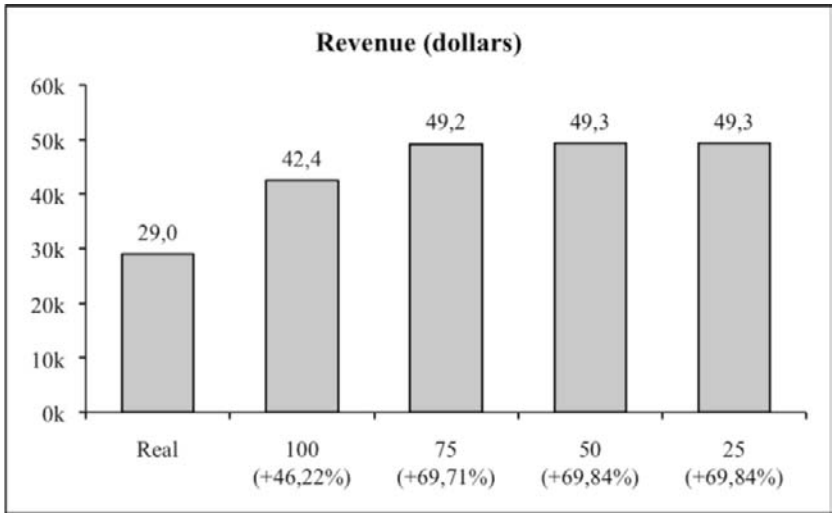

Figure 3. Overall improvement in revenue obtained by varying minimum support.

Similarly, in Figure 3 we present the overall improvement in the revenue obtained by varying the Apriori minimum support. This chart was obtained by multiplying the revenue per banner by the number of additional clicks generated. This has a very similar trend to the chart of Figure 2 as clicks and revenue are correlated.

4. Conclusions and Further Work

Above we discussed the experimental results of an Apriori based optimization performed on a real setting. We showed how a simple application of association rules to banner ads can significantly improve the overall number of clicks generated by the ads shown on some websites. This has very interesting economic implications as it means that advertisers may get a bigger return on their investment for the same money. Moreover, visitors on the website receive a more tailored advertising, which, in general, yields to a greater satisfaction and loyalty.

We tested our approach on some real websites as we developed a software application that was inserted into a real ad-server running on real clients. The achieved optimization was deemed to be impressive by domain experts that also stated how difficult, if not impossible, could have been to achieve such an improvement with any other possible approach.

We planned various improvements to our first model. In particular:

- Taking more variables into consideration such as time of the day and day of the week.
- Applying a combination of models. Given the richness of data available, we believe there is not one single model better than the others, but a mix of them can yield to the best overall performance.
- Testing our system under different circumstances and on different sites.
- Combining our model with the text surrounding the banner at the time it is shown.

- Extend our model with temporal Apriori as we strongly believe that in online advertising, as largely proven in Ecommerce, there are seasonality trends that are not captured by standard Apriori.

The results so far have been so encouraging that we believe this domain of application will attract lots of attention from both the industry and the research community. In our opinion, the scientific data mining community should be extremely interested by the richness and precision of the data to analyze. While the advertising industry should be sensitive to the great return on its investment it could achieve thank to the data mining contribution.

Acknowledgments

We really owe a special thank to Neodata Group s.r.l. (www.neodatagroup.com) as they allowed us to test our approach by integrating our software into their live system. This allowed us to feel more confident about the validity of our approach and about the non-intrusive nature of it. In particular we would like to thank Andrea Fiore and Antonio Butera for the effort they put to support our experiment.

References

[1] Adland's Test Tube, The Economist, December 13th 2006.
[2] M. Bayles. Just How 'Blind' Are we to Advertising Banners on the Web? Department of Psychology, Wichita State University. Usability News, 2000.
[3] M.N.A. Sharif, M. Bahari, A. Bakri, N.H. Zakaria. Using A Priori for supporting E-Commerce System. Department of Information Systems, Faculty of Computer Science & Information Systems, 81310, UTM, Skudai, Johor, at ICOQSIA, 2005.
[4] B. Ozden, S. Ramaswamy, A. Silberschatz. Cyclic association rules. In Proc. 1998 Int. Conf. Data Engineering (ICDE'98), pages 412-421, Orlando, FL. Feb. 1998.
[5] K. Verma, O. P. Vyas. Efficient calendar based temporal association rule. SIGMOD Record 34(3): 63-70, 2005.
[6] G. Zimbrão, J. Moreira de Souza, V. Teixeira de Almeida, W. Araújo da Silva. An Algorithm to Discover Calendar-based Temporal Association Rules with Item's Lifespan Restriction, The Second Workshop on Temporal Data Mining, July 23, 2002 - Edmonton, Alberta, Canada.
[7] P. J. Hoen, S. Bohte, E. Gerding, H. La Poutre'. Implementation of a Competitive Market-Based Allocation of Consumer Attention Space. Proceedings of the 4th Workshop on Agent Mediated Electronic Commerce at AAMAS, 2002.
[8] R. Agrawal, T. Imielinski and A. Swami. Mining Association Rules between Sets of Items in Large Databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data May 1993.
[9] C. Borgelt. Implementation of Apriori algorithm. University of Magdeburg, February 2000.
[10] R. J. Bayardo Jr., R. Agrawal. Mining the Most Interesting Rules. Conference on Knowledge Discovery in Data. Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. San Diego, California, United States, 1999.
[11] J..M. Ale, G. H. Rossi. An approach to discovering temporal association rules, Proceedings of the ACM symposium on applied computing, March 2000.
[12] P. Baudisch, D. Leopold. User configurable advertising profiles applied to Web page banners. Institute for Integrated Information and Publication Systems IPSI, German National Research Center for Information Technology (GMD), 1997.

This page intentionally left blank

Applications of Data Mining in E-Business and Finance
C. Soares et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-890-8-63

Probabilistic Latent Semantic Analysis for Search and Mining of Corporate Blogs

Flora S. TSAI [a,1], Yun CHEN [a] and Kap Luk CHAN [a]
[a] *School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore, 639798*

Abstract. Blogs, or weblogs, have rapidly gained in popularity over the past decade. Because of the huge volume of existing blog posts, information in the blogosphere is difficult to access and retrieve. Existing studies have focused on analyzing personal blogs, but few have looked at corporate blogs, the numbers of which are dramatically rising. In this paper, we use probabilistic latent semantic analysis to detect keywords from corporate blogs with respect to certain topics. We then demonstrate how this method can represent the blogosphere in terms of topics with measurable keywords, hence tracking popular conversations and topics in the blogosphere. By applying a probabilistic approach, we can improve information retrieval in blog search and keywords detection, and provide an analytical foundation for the future of corporate blog search and mining.

Keywords. Weblog search, blog mining, probabilistic latent semantic analysis, corporate blog, business blog, web mining

Introduction

A blog, or weblog, is a type of website where entries are made in a reverse chronological order. Blogs often provide commentary or news on a particular subject, but many function more as personal online diaries. Blogosphere is the collective term encompassing all blogs as a community or social network. Because of the huge volume of existing blog posts and their free format nature, the information in the blogosphere is rather random and chaotic. As a result, effective access and retrieval techniques are needed to improve the quality of the search results. Currently, blog search engines are still in their infancy [1], and many blog-specific search engines index only XML (Extensible Markup Language) feeds, which usually consist of the summary or the first few sentences of the blog entries [2]. Moreover, a study on blog search [1] concluded that blog searches have different interests than normal web searches, suggesting that blog searches tend to track references to known entities and focuses on specific themes or topics. Therefore, to increase the quality and accuracy of the search results, more complex models and information retrieval techniques are required for mining the blogs.

With the amazing growth of blogs on the web, the blogosphere affects much of the media. For example, the rumors of the acquisition of YouTube by Google first surfaced

[1] Corresponding Author: Flora S. Tsai; fst1@columbia.edu.

on a blog site, Michael Arrington's TechCrunch[2], and was later reported by the New York Times and Wall Street Journal, citing the original blog site as the source of the story [3]. In 2002, Microsoft was caught using a fake advert that claimed people were switching from Macs to Windows PCs. The advert on Microsoft's website supposedly recounted the story of a former Apple Mac user who had converted to using Windows, but net users revealed that the supposed "switcher" actually worked for a marketing company employed by the Seattle giant [4].

Studies on the blogosphere include measuring the influence of the blogosphere [5], analyzing blog threads for discovering the most important bloggers [6], determining the spatiotemporal theme pattern of blogs [7], learning contextualized topics in blogs [8], detecting growth trends of blogs [9], tracking the propagation of discussion topics in the blogosphere [10], and detecting cyber security threats in blogs [11].

Many studies have focused on analyzing personal blogs, but few have looked at corporate blogs, which are published and used by organizations. Corporate blogs differ from personal blogs in that information is more focused and targeted, and the language is generally more formal. Although existing techniques used to analyze personal blogs may also be used to analyze corporate blogs, the methods may need to be modified to deal with the corporate ontologies which define semantics share the same meaning for a collection of topics. Existing studies on corporate blogs have focused on blogging strategies, case studies, and conformance checking [12,13,14], and not on blog search and mining. In this paper, we define corporate blogs as a combination of business blogs, which are blogs providing commentary or analysis of companies, and external company blogs, which can be an important link to customers and potential clients. Internal company blogs, which are blogs used within an organization to improve collaboration and internal business intelligence, are not included in our analysis.

Although not as common as personal blogs, corporate blogs are not new. More than 8% of the Fortune 500 companies blog [15] externally, and market research shows that 35% of large companies planned to institute corporate blogs in 2006 [16]. According to the research, nearly 70% of all corporate website operators were expected to implement corporate blogs by the end of 2006 [16].

In our work, we first built a blog search engine which differs from existing blog-specific search engines in that it searches the full text of the blog entry and ranks the results based on similarity measures. Moreover, to broaden the usefulness of the blog search engine, an additional function was created to detect the keywords of various topics of the blog entries, hence tracking the trends and topics of conversations in the blogosphere. Probabilistic Latent Semantic Analysis (PLSA) was used to detect the keywords from various corporate blog entries with respect to certain topics. By using PLSA, we can present the blogosphere in terms of topics represented by a probability distribution of keywords.

The paper is organized as follows. Section 1 reviews the related work on blog search and mining. Section 2 describes an overview of the Probabilistic Latent Semantic Analysis model for mining of blog-related topics. Section 3 presents experimental results, and Section 4 concludes the paper.

[2] http://www.techcrunch.com

1. Review of Related Work

This section reviews related work in developing blog-specific search engines and extraction of useful information from blogs.

1.1. Blog-specific Search Engines

Search engines have been widely used on the Internet for decades, with Google and Yahoo currently the most popular. As blogging becomes a hit in the air, blog-specific search engines are created suit the demand. Past studies [5,2] have examined the characteristics, effectiveness and distinctiveness of blog-specific search engines.

As of October 2006, Technorati[3] has indexed over 56 millions blogs. It is believed to be tracking the largest number of links in real-time. RSS (Really Simple Syndication) or XML feeds automatically send notification from blogs to Technorati quickly, and the thousands of updates per hour that occur in the blogosphere are tracked by Technorati. Technically, Technorati supports open microformat standards, and indexes and searches posts tagged with `rel-tag`, using the full boolean search technique on the RSS/XML files. This full boolean search functionality provides `AND`, `OR` and `NOT` functionality, which narrows down the searching results, and has made tagging more valuable. In order to provide the users with more options, search results of synonymous queries are listed. For example, when “car" is the search query, there is an option of refining the results on “auto", “automotive", “automobile", “autos", “life", “vehicles", “work".

Similarly, many other blog-specific search engines, such as Bloglines[4], Feedster[5], and BlogPulse[6], index and search the RSS/XML feeds of blogs, using boolean search. However, the ranking of the results are not in the order of relevance, but rather in the time of posting, the “popularity" (number of links to the blog), or the update frequency. More importantly, the existing blog search engines do not necessarily use complex models to achieving better results, due to the constraints of the real-time nature of blogs and its extremely rapid speed of update in the blogosphere. Thus, there is room for improvement of the quality of blog search results.

1.2. Extraction of Useful Information from Blogs

Current blog text analysis focuses on extracting useful information from blog entry collections, and determining certain trends in the blogosphere. NLP (Natural Language Processing) algorithms have been used to determine the most important keywords and proper names within a certain time period from thousands of active blogs, which can automatically discover trends across blogs, as well as detect key persons, phrases and paragraphs [9]. A study on the propagation of discussion topics through the social network in the blogosphere were carried out using algorithms developed to detect the long-term and short-term topics and keywords, which were then validated on real blog entry collections [10]. The suitability of methods for ranking term significance on an evolving RSS feed corpus was evaluated using three statistical feature selection methods were implemented: χ^2,

[3] `http://www.technorati.com`
[4] `http://www.bloglines.com`
[5] `http://www.feedster.com`
[6] `http://www.blogpulse.com`

Mutual Information (*MI*) and Information Gain (*I*). The conclusion was that χ^2 method seems to be the best among all, but a full human classification exercise would be required to further evaluate such a method [17]. A probabilistic approach based on PLSA was proposed in [7] to extract common themes from blogs, and also generate the theme life cycle for each given blog and the theme snapshots for each given time period. In addition, PLSA was also used in [11] to detect cyber security threats in weblogs. These studies illustrate that the PLSA-based approach using probabilistic mixture models can be effectively used for viewing spatiotemporal life cycle patterns as well as security threats of blogs.

Our work differs from existing studies in two respects: (1) We focus on corporate blog entries which may contain less extraneous and more useful information than personal blogs, potentially resulting in higher quality search results and (2) we have combined a blog search engine with topic and keyword extraction, and use probabilistic models to extract popular keywords for each topic.

2. Probabilistic Latent Semantic Analysis Model for Blog Mining

Probabilistic Latent Semantic Analysis (PLSA) [18] is based on a generative probabilistic model that stems from a statistical approach to LSA (Latent Semantic Analysis) [19]. PLSA is able to capture the polysemy and synonymy in text for applications in the information retrieval domain. Similar to LSA, PLSA uses a term-document matrix which describes patterns of term (word) distribution across a set of documents (blog entries). By implementing PLSA, topics are generated from the blog entries, where each topic produces a probability distribution of words for that topic, using a maximum likelihood estimation method, the expectation maximization (EM) algorithm.

The starting point for PLSA is the *aspect model* [18]. The aspect model is a latent variable model for co-occurrence data associating an unobserved class variable $z_k \in \{z_1, \ldots, z_k\}$ with each observation, an observation being the occurrence of a keyword in a particular blog entry. There are three probabilities used in PLSA:

1. $P(b_i)$ denotes the probability that a keyword occurrence will be observed in a particular blog entry b_i,
2. $P(w_j|z_k)$ denotes the class-conditional probability of a specific keyword conditioned on the unobserved class variable z_k,
3. $P(z_k|d_i)$ denotes a blog-specific probability distribution over the latent variable space.

In the collection, the probability of each blog and the probability of each keyword are known, while the probability of an aspect given a blog and the probability of a keyword given an aspect are unknown. By using the above three probabilities and conditions, three fundamental schemes are implemented:

1. select a blog entry b_i with probability $P(b_i)$,
2. pick a latent class z_k with probability $P(z_k|b_i)$,
3. generate a keyword w_j with probability $P(w_j|z_k)$.

As a result, a joint probability model is obtained in asymmetric formulation, where, for each blog b, a latent class is chosen conditionally to the document according to

$P(z|b)$, and a keyword is then generated from that class according to $P(w|z)$. After the aspect model is generated, the model is fitted using the EM algorithm. The EM algorithm involves two steps, namely the expectation (E) step and the maximization (M) step. The E-step computes the posterior probability for the latent variable, by applying Bayes' formula. The M-step is to update the parameters based on the expected complete data log-likelihood depending on the posterior probability resulted from the E-step.

The EM algorithm is iterated to increase the likelihood function until specific conditions are met and the program is terminated. These conditions can be a convergence condition, or a cut-off stopping, which enables achieving a local maximum solution, rather than a global maximum.

In short, the PLSA model selects the model parameter values that maximize the probability of the observed data, and returns the relevant probability distributions which quantify relationships between blogs and topics. Based on the pre-processed term-document matrix, the blogs are classified onto different topics. For each topic, the keyword occurrence, such as the probable words in the class-conditional distribution $P(w_j|z_k)$, is determined. Empirical results have shown that there are advantages of PLSA in reducing perplexity, and improving performance in terms of precision and recall in information retrieval applications [18]. Furthermore, LSA can be used to better initialize the parameters of a corresponding PLSA model, with the result combining the advantages of both techniques [20].

3. Experiments and Results

We have created a corporate blog data set, built a blog search system using the latent semantic analysis model, and applied the probabilistic model for blog mining on our data set of corporate blogs. The blog search system provides a ranking of corporate blog entries by similarity measures, and allows for full-text searching in different categories. We extract the most relevant categories and show the topics extracted for each category. Experiments show that the probabilistic model can reveal interesting patterns in the underlying topics for our data set of corporate blogs.

3.1. Data Set

For our experiments, we created a corporate blog data corpus that focuses on blogs created by companies or about companies. During the period from April to September 2006, we extracted a set of corporate blogs through the following methods:

1. Search corporate blog entries from various CEOs' blog sites[7].
2. Search corporate blog entries from various companies' blog sites [15].
3. Search particular corporate blog entries using the existing blog search engines, such as Bloglines, Technorati, and Google Advanced Blog Search [8].

Meaningful blog entries from these blog sites were extracted and stored into our database. There are a total of 86 companies represented in the blog entries and Table 1

[7] `http://blogwrite.blogs.com` Note: only a few CEOs' blogs write about purely business matters.
[8] `http://blogsearch.google.com/blogsearch/advanced_blog_search` RSS news subscription feeds were eliminated from the search results.

summarizes the top companies in the corporate blog data corpus, and Table 2 lists the fields in the database.

Table 1. List of Top Companies.

Company
Microsoft
eBay
Samsung
Dell
Amazon
Sony
Google
Apple
Palm
Yahoo

Table 2. Database schema for blog entry.

Field	Type
ID	int
Title	text
Author	text
Publish_Date	date
URL	text
Content	text
Category_ID	int
Type_ID	int
Company_ID1	int
Company_ID2	int

We then categorize the 1269 blog entries into four categories based on the contents or the main description of the blog: Company, Finance, Marketing, and Product. The Company category deals with news or other information specific to corporations, organizations, or businesses. The Finance category relates to financing, loans and credit information. The Marketing category deals with marketing, sales, and advertising strategies for companies. Finally, the Product category describes the blog entries on specific company products, such as reviews, descriptions, and other product-related news. Table 3 summarizes the distribution of blog entries in each category.

Table 3. Categories for the Corporate Blog Data Corpus.

Categories	Distribution
Company	20.9%
Finance	27.4%
Marketing	21.2%
Product	30.5%

Each blog entry is saved as a text file in its corresponding category, for further text preprocessing. For the preprocessing of the blog data, we performed lexical analysis by removing stopwords and stemming using the Porter stemmer [21]. The text files are then used as the input for the Text to Matrix Generator (TMG) [22] to generate the term-document matrix for input to the blog search and mining system.

3.2. Blog Search System

We implemented a blog search system for our corporate blog data. We used the LSA model [19] for constructing the search system, as LSA is able to consider blog entries with similar words which are semantically close, and calculate a similarity measure based on documents that are semantically similar to the query terms. The similarity measure employed is the cosine similarity measure:

$$\cos\theta_j = \frac{d_j^T q}{\|d_j\|_2 \|q\|_2} \tag{1}$$

where θ_j is the angle between the query vector and the document vector, d_j is the j^{th} document vector, and $q = (q_1, q_2, \ldots, q_m)^T$ is the query vector.

The value of the cosine ranges from -1 to 1, hence the cosine similarity usually has a threshold value. The threshold should be positive, where the higher the threshold value, the smaller number of documents retrieved, as fewer document vectors are closer to the query vector.

In this system, a user can select the type and the category, and enter a word or a phrase as the query to search for blog entries matching the query. The results are ranked in the order of similarity. The title of each searched blog entry is then displayed in order of similarity. Clicking on the title shows the full text of the blog entry, as well as the original hyperlink. Figure 1 shows the overview of the blog search system, and Figure 2 shows a screenshot of the system. Figure 3 shows the top ten blog entry results for a search on the company Dell.

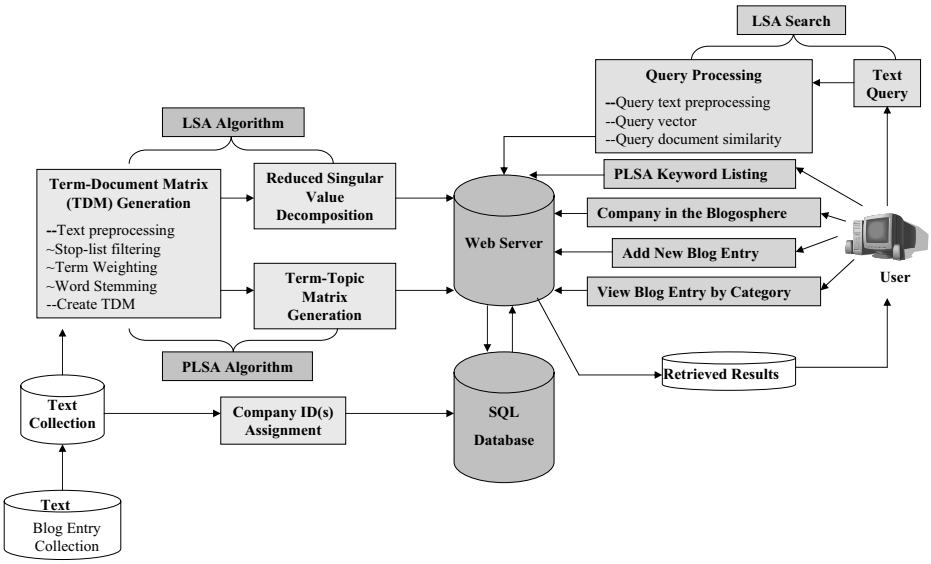

Figure 1. Overview of blog search system

Figure 2. Screenshot of blog search system

1. Jeff Clarke on Dell's OptiPlex 745 and More (Similarity = 0.724232)
2. Dell's Smallest Notebook Dell's Smallest Notebook (Similarity = 0.705834)
3. Green Recycling Options - (Similarity = 0.696217)
4. Dell @ LinuxWorld (Similarity = 0.679831)
5. Dell OpenManage Strategy (Similarity = 0.607306)
6. Real People are Here and We're Listening (Similarity = 0.598994)
7. XPS 700 Order Cancelled? - (Similarity = 0.586906)
8. Dell Store Opens its Doors in Dallas (Similarity = 0.575105)
9. Are We Having Fun Yet? Heck Yeah! (Similarity = 0.570931)
10. Un-concreting the Cow Path (Similarity = 0.565244)

Figure 3. Blog search results for query on "Dell", ranked in order of similarity.

3.3. Results for Blog Mining of Topics

We conducted some experiments using PLSA for the blog entries. Figure 4 shows a screenshot of the blog mining of topics, and Tables 4-7 summarize the keywords found for each of the four topics (Company, Finance, Marketing, and Product).

These keyword listings identify the popular topics in the blogosphere. It is worthy to mention the interesting findings from the PLSA keyword listing, when searching for four topics in all categories. The four topics from this search actually matches the four categories in the database, as shown in Table 3. This reflects that the categorization of the blog entry collection is reasonable, and this categorization can be applied in corporate blogs at large. By looking at the various topics listed, we are able to see that the probabilistic approach is able to list important keywords of each topic in a quantitative fashion. The keywords listed can relate back to the original topics. For example, the keywords detected in the Product topic features items such as mobile products, batteries, phones, and umpc (ultra mobile PC). In this way, it is possible to list popular keywords and track the hot topics in the blogosphere.

The power of PLSA in corporate blog applications include the ability to automatically detect terms and keywords related to business and corporate blog trends in product,

Figure 4. Screenshot of blog mining of topics

Table 4. List of keywords for Topic 1 (Company).

Keyword	Probability
blog	0.010993
ebay	0.009240
amazon	0.007343
google	0.005461
web	0.005218
develop	0.005128
api	0.005019
site	0.004786
search	0.004493
product	0.004252

Table 5. List of keywords for Topic 2 (Finance).

Keyword	Probability
save	0.015760
money	0.014106
debt	0.007432
year	0.006805
financ	0.006256
financi	0.006219
credit	0.005940
card	0.005919
college	0.005915
invest	0.005702

Table 6. List of keywords for Topic 3 (Marketing).

Keyword	Probability
market	0.011830
company	0.007158
custom	0.006505
busi	0.005788
firm	0.004737
advertise	0.003846
brand	0.003701
product	0.003584
corpor	0.003282
client	0.003247

Table 7. List of keywords for Topic 4 (Product).

Keyword	Probability
mobile	0.013204
battery	0.008454
device	0.008264
phone	0.008256
window	0.006740
tablet	0.006506
umpc	0.006034
samsung	0.005558
keyboard	0.004759
apple	0.004746

marketing, and finance matters. By presenting blogs with measurable keywords, we can improve our understanding of business issues in terms of distribution and trends of current conversations and events. This has implications for companies wishing to monitor

real-time business trends present in weblogs or other related documents.

4. Conclusions

This paper presents results using probabilistic and latent semantic models for search and analysis of corporate blogs. We have created a corporate blog data corpus for this study, and categorized the data set into four classes. We have also developed a corporate blog search system that is based on latent semantic analysis, which is able to rank the results in terms of blog document similarity to the query. Our experiments on our data set of corporate blogs demonstrate how our probabilistic blog model can present the blogosphere in terms of topics with measurable keywords, hence tracking popular conversations and topics in the blogosphere. Our approach is unique because it specifically targets the corporate blog knowledge repositories, which can benefit businesses and companies.

Possible advantages for end-users include automatically monitoring and identifying trends in corporate blogs. This can have some significance for organizations and companies wishing to monitor real-time trends in business marketing, finance, and product information present in weblog conversations and the blogosphere. We hope that this work will contribute to the growing need and importance for search and mining of corporate blogs.

References

[1] Mishne, G., de Rijke, M.: A Study of Blog Search. ECIR '06 (2006)
[2] Pikas, C.K.: Blog Searching for Competitive Intelligence, Brand Image, and Reputation Management. Online. **29(4)** (2005) 16–21
[3] Weil, D.: The New York Times and Wall Street Journal source TechCrunch to break the news about Google's possible acquisition of YouTube. Available at: `www.blogwriteforceos.com/blogwrite/2006/10/is_this_an_infl.html` (2006)
[4] BBC: Web users turn tables on Microsoft. Available at: `news.bbc.co.uk/1/hi/technology/2329519.stm` (2002)
[5] Gill, K.E.: How Can We Measure the Influence of the Blogosphere? WWW '04 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics (2004)
[6] Nakajima, S., Tatemura, J., Hino,Y., Hara,Y., Tanaka, K.: Discovering Important Bloggers based on Analyzing Blog Threads. WWW '05 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics (2005)
[7] Mei, Q., Liu, C., Su, H., Zhai, C.: A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs. WWW '06 (2006)
[8] Avesani, P., Cova, M., Hayes, C., Massa, P. : Learning Contextualised Weblog Topics. WWW '05 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics (2005)
[9] Glance, N.S., Hurst, M., Tomokiyo, T: BlogPulse: Automated Trend Discovery for Weblogs. WWW '04 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics (2004)
[10] Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A.: Information Diffusion Through Blogspace. WWW '04 (2004)
[11] Tsai, F.S., Chan, K.L.: Detecting Cyber Security Threats in Weblogs Using Probabilistic Models. In C.C. Yang et al (Eds.): Lecture Notes in Computer Science (LNCS) **4430** (2007) 46–57
[12] Cass, J., Munroe, K., Turcotter, S.: Corporate blogging: is it worth the hype? Available at: `www.backbonemedia.com/blogsurvey/blogsurvey2005.pdf` (2005)
[13] Lee, S. , Hwang, T., Lee, H-H.: Corporate blogging strategies of the Fortune 500 companies. Management Decision. **44(3)** (2006)
[14] Song, D., Bruza, P., McArthur, R., Mansfield, T.: Enabling Management Oversight in Corporate Blog Space. AAAI'06 (2006)

[15] Anderson C., Mayfield R.: Fortune 500 Business Blogging Wiki. Available at: socialtext.net/bizblogs (2006)
[16] Dowling, W.G., Daniels, D.: Corporate Weblogs: Deployment, Promotion, and Measurement. The JupiterResearch Concept Report (2006)
[17] Prabowo, R., Thelwall, M.: A Comparison of Feature Selection Methods for an Evolving RSS Feed Corpus. Information Processing and Management. **42** (2006) 1491–1512
[18] Hofmann, T.: Probabilistic Latent Semantic Indexing. SIGIR'99 (1999)
[19] Deerwester, S., Dumais, S., Landauer, T., Furnas, G., Harshman, R.: Indexing by latent semantic analysis. In Journal of the American Society of Information Science, 41(6) (1990) 391–407
[20] Farahat, A., Chen,F.: Improving Probabilistic Latent Semantic Analysis with Principal Component Analysis. EACL '06 (2006)
[21] Porter,M.F.: An algorithm for suffix stripping. Program. **14(3)** (1980) 130–137
[22] Zeimpekis, D., Gallopoulos, E.: TMG: A MATLAB Toolbox for generating term-document matrices from text collections. Grouping Multidimensional Data: Recent Advances in Clustering. Springer (2005) 187–210

This page intentionally left blank

Applications of Data Mining in E-Business and Finance
C. Soares et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-890-8-75

A Quantitative Method for RSS Based Applications

Mingwei YUAN[a], Ping JIANG[a,b,1] and Jian WU[a]
[a] *Department of Information and Control, Tongji University, China;*
[b] *Department of Computing, University of Bradford, Bradford, BD7 1DP, UK.*
Email: yuan.mingwei@hotmail.com, p.jiang@bradford.ac.uk, tongjiwujian@hotmail.com

Abstract. The RSS technique provides a fast and effective way to publish up-to-date information or renew outdated content for information subscribers. So far, RSS information is mostly managed by content publishers and Internet users have less initiative to choose what they really need. More attention needs to be paid on techniques for user-initiated information discovery from RSS feeds. In this paper, a quantitative semantic matchmaking method for RSS based applications is proposed. The semantic information of an RSS feed can be described by numerical vectors and semantic matching can then be conducted in a quantitative form. The ontology is applied to provide a common-agreed matching basis for the quantitative comparison. In order to avoid semantic ambiguity of literal statements from distributed RSS publishers, fuzzy inference is used to transform an individual-dependent vector into an individual-independent vector and semantic similarities can then be revealed as the result.

Keywords. RSS feeds, Feature vector, Matchmaking, Semantics.

Introduction

Semantic web technologies include a family of tools and languages for expressing information in a machine interpretable form [1] that enable machines to process web-contents and helps humans to perceive changes in the web automatically. RSS (RDF Site Summary/Really Simple Syndication) [2,3,4] could be regarded as an application of semantic web techniques and provides Internet users with an automatic mode of information acquisition. Users are able to explicitly specify which RSS feeds to monitor by subscribing them and no longer need to regularly check websites and find new updates manually.

In general, RSS is a metadata language for describing content changes. Nowadays RSS is adopted by almost all mainstream websites. But research shows that awareness of RSS is quite low among Internet users: 12% of users are aware of RSS, and 4% have knowingly used RSS [5]. So how to extend RSS based applications and make them more convenient and accessible to ordinary Internet users is an interesting research point.

It is worth noting that RSS feeds include many content entries but not all information in an RSS feed is relevant to a subscriber's needs. In this paper a new quantitative method for RSS based applications is proposed to help users select the

[1] Corresponding author.

information that they are looking for. The main idea of the quantitative method is that the information in RSS feeds is transformed into numerical vectors based on an ontology. As a result, semantic matching of information can be conducted by correlation computation. Integrating this method into RSS readers could increase the precision of information acquisition.

1. Related Research

Nowadays the research on RSS rests mostly on how to aggregate/syndicate content effectively [6] and on how to extend its function modules [7,8,9]. Although RSS based applications have been growing significantly, most studies adopt classical textual document methods [10,11,12,13] to handle RSS documents [14,15].

It is obvious that such classical textual analysis methods could be applied to RSS document processing since RSS documents are formatted as textual documents. But the semantic information in the RSS document is often neglected.

The idea to use an ontology as a semantic bridge [16] between news items and dispersed users is adopted in this paper. The method proposed here is not limited to personalization of news contents but also applicable to any content based system such as news publishing, business content distribution services, enterprise internal information publishing, etc. It is a general quantitative method to measure the semantic similarity between RSS feeds and user interests. The method proposed in this paper is for general formats of RSS feeds, and hence RSS 1.0, RS 2.0 and any other RSS-like format (e.g. Atom) can be used.

2. Methodology

RSS based applications need to extract useful information from RSS feeds. An RSS feed is composed of a channel and a series of items. The RSS <item> is where the updates are stored, which defines a summary of an article or a story. It provides an information prototype for RSS based applications. In order to simplify the description, only the headline of information, i.e. the title of an item, is considered in this paper. It summarizes the published information and is the foremost factor influencing information selection. However, the method is not limited to headlines and the content in the description tag of an RSS item can also be included for richer information discovery.

2.1. Ontology Support And Semantic Distance

We suppose that the RSS publishers and subscribers share the same knowledge background which is represented by a domain ontology. Here the ontology is defined by domain experts manually to denote the common knowledge.

The ontology could be defined as $\Omega_C := < E, R >$. $E = \{e_i \mid 1 \le i \le N\}$ represents the set of entities, e_i $(i = 1,..., N)$ represents an entity (concept, terminology, property, attribute) used in a community. R represents the set of relations between entities, $R = \{r_{ij} \mid e_i \times e_j \rightarrow r_{ij}, 1 \le i \le N, 1 \le j \le N\}$. A concept could be represented by a class or an instance in the ontology described in web ontology language-OWL

(http://www.w3.org/TR/2004/REC-owl-ref-20040210/). "Class" is more abstract and general than "instance" and the relations between classes are inherited by their instances. Hence "class" is used to represent a domain concept. Concepts defined in the ontology are semantically relevant. There are three kind of relations between two classes: "IS-A", "PART-OF" and others. Only "IS-A" is of strict format, it is expressed by "*rdfs:subClassOf* ". The other two kind of relations can be found in the expression of "*owl:ObjectProperty*". Regard a concept as a node and a relation as an edge, and then an ontology could be expressed by a semantic graph.

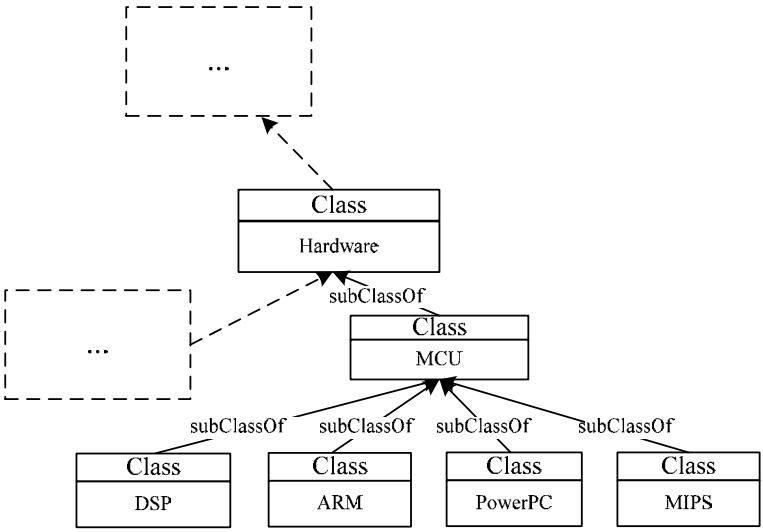

Figure 1. DVR (Digital Video Recorder) R&D ontology illustration.

Figure 1 shows a part of DVR (Digital Video Recorder) R&D ontology. A Digital Video Recorder is a video/audio product that can record and play back video using compression standards, which has been widely used in security surveillance. Developers use the knowledge of MCU (Micro-Controller Unit) to design and develop the hardware of a DVR. The dashed boxes denote other parts of the ontology. The ontology is edited using Protégé editor (http://protege.stanford.edu/plugins/owls/index.html) and is expressed in OWL.

To avoid computational mess and complexity, the concepts in the ontology are parsed into a concept list and the concepts are arranged by following a "Breadth-First" scan of a hierarchical ontology from the root concepts; the higher level concepts will be allocated to more frontal positions in the list. If a concept has multiple parent concepts, i.e. the ontology is not a hierarchical tree, following the "Breadth-First" scan, the first appearance of the parent concept will determine the index of the concept.

Concepts defined in the ontology are semantically relevant. Hence semantic distance is introduced as a measure of semantic difference between any two concepts, which has been applied widely in semantic web matchmaking [17] and data mining [18]. After flattening an ontology graph into a concept list, a semantic distance matrix can be obtained by algorithm 1 to depict the semantic difference between any two concepts in the ontology definition ΩC. An ontology exhibits a natural hierarchical structure, so the semantic distance in this paper is defined using the shortest path length between two concepts.

Algorithm 1: Calculate Semantic Distance Matrix

Step 1 Obtain an *initialMatrix*={ *e*(*i*,*j*) }:*e*(*i*,*j*) denotes the distance value between the i^{th} element and the j^{th} element which have a direct link.

The initial relationship matrix (*initialMatrix*) denotes the semantic distance between two concepts, e_i and e_j, which have a direct link in the ontology graph. The element in the *initialMatrix* $e_{i,j}$ is defined as follows.

$$e_{i,j} = \begin{cases} 0 & if\ i = j \\ 1 & if\ there\ exists\ \text{rdfs : subClassOf}\ \ between\ concept\ i\ and\ concept\ j, and\ i \neq j \\ 2 & if\ there\ exists\ \text{owl : ObjectProp ert}y\ \ between\ concept\ i\ and\ concept\ j, and\ i \neq j \\ X & otherwise \end{cases}$$

and $e_{i,j} = e_{j,i}$.

In this initial relationship matrix, only the direct links are initiated and *X* represents indirect links that will be computed in step 2 according to the initial matrix. The resulting semantic distance matrix satisfies the properties of the general semantic distance mentioned before.

Step 2 Obtain a semantic distance Matrix, *DisMatrix* = {*d(i,j)*}: *d(i,j)* denotes the distance value in the i^{th} element and the j^{th} element.

A semantic distance matrix takes into account indirect relationships between any two concepts based upon the *initialMatrix*, and the distance can be computed recursively.

$\forall$ Concept A, B and B is not the descendant of A in the ontology graph.

$Distance(A,B) = \text{Min}\{ Distance(A, Parent(A)_i) + Distance(Parent(A)_i, B) \}$

*Parent(A)*denotes the set of parents of concept node A in an ontology graph and $Parent(A)_i$ denotes the i-th element in the set.

Thus the generated semantic distance matrix can be represented as:

$$\text{DisMatrix} = \begin{bmatrix} 0 & d(1,2) & \cdots & d(1,N) \\ d(2,1) & 0 & \cdots & d(2,N) \\ & & \cdots & \\ d(N,1) & d(N,2) & \cdots & 0 \end{bmatrix}. \quad (1)$$

d(i, j) is a semantic distance between concept e_i and concept e_j. The dimension of the semantic distance matrix is *N*, which equals to the number of concepts in the ontology.

2.2. Parse RSS Feeds Into Ontology Instances

Since Internet users are geographically distributed, RSS feeds, i.e. ontology instances, are written with concepts defined in the ontology. Hence matchmaking of RSS feeds is transformed to comparison of ontology instances but, first, RSS feeds need to be parsed into ontology instances.

Firstly, some preprocessing is needed: parse RSS feeds to obtain the title and the link in each item, and then extract domain-related concepts appearing in the title.

Secondly, arrange the concepts extracted from each title of an item into a hierarchical concept graph, i.e. an ontology instance. That is to set the URI (link) of an

item as a root and then construct a hierarchical concept graph according to the given ontology, as explained next.

All the concepts in the title are descendants of the root. The relative relationships in the ontology should remain. If two concepts in the title have an ancestor-descendant relation in the ontology, the relationship should be kept as a hierarchy. So should those concepts which have sibling relationship.

For example, a processed headline "Software, MPEG, Video, Compression_Standard" can be transformed into a concept graph as.

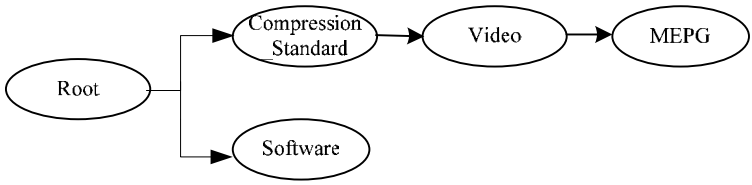

Figure 2. The ontology instance graph of a headline

A content that is expected to be relevant for a subscriber of an RSS feed could also be parsed into an ontology instance and described in an OWL document.

In the following algorithm, an instance (concept graph) is transformed into a feature vector that is a numerical expression of an RSS item by taking into account semantics implied by an ontology (e.g. the one defined in Figure 1).

Algorithm 2: A feature vector for an instance can be represented as:

$$V(i)=[s_1,s_2, \ldots, s_N]^{\mathrm{T}}. \tag{2}$$

The element s_i in V(i) has a one-to-one correspondence to the concept e_i defined in the ontology concept vector. The $s_i \in [0,1]$ indicates the semantic closeness between a concept, e_i, and the root of an ontology instance.

$$s_i = \begin{cases} e^{-\alpha Dis(e_i,root)} & \text{if } e_i \text{ appears in the instance} \\ 0 & \text{if } e_i \text{ does not appear in the instance} \end{cases}. \tag{3}$$

The $Dis(e_i,root)$ is a semantic distance between the entity e_i and the root presented in the last section. The α is a steepness measure [19] for fuzzy modelling, which is often selected to be -7/MAX(Dis) because $e^{-7} \approx 0$ when $Dis(e_i,root)$ reaches its maximum. For example the numerical representation of the instance in Figure 2 based on the ontology of Figure 1 is [0, 0, 0, $e^{-\alpha*1}$, $e^{-\alpha*1}$, 0, …, $e^{-\alpha*2}$, 0, …, $e^{-\alpha*3}$, 0, …, 0].

Now the information from an RSS feed and a user have been represented formally in accordance with the ontology definition. Semantics based matchmaking is a process to compare similarity between concept graphs. In the next section a quantitative method is proposed to measure the semantic similarity between ontology instances.

2.3. Semantic Matchmaking

Both titles extracted from an RSS feed and a user favourite could be transformed into two feature vectors, *Vp* and *Vs* respectively, measuring the similarity between published

information from websites and user favourites is a process of matchmaking. Due to the distributed nature of the Internet applications, it is impossible to force all participants to use strictly consistent terminologies and sentences in their web documents even though they share common domain knowledge. The above-mentioned feature vectors are therefore individual-specific.

In fact the concepts in an ontology instance used by participants are fuzzy; any concept implies some aspects of others due to the semantic correlations that can be defined by a grade of membership in fuzzy set theory [20]. Suppose the entities $\{e_1, e_2, ..., e_N\}$ in the ontology form a universe of discourse in the community. Any announcement *i*, such as a title from an RSS feed containing "design driver program using C", is a linguistic variable. Then the corresponding feature vector $V(i)=[s_1, s_2, ..., s_N]^T$ in (2) is a fuzzy representation of *i* from an individual's point of view, where s_i , $i=1...N$, is a grade of membership corresponding to the i^{th} entity in the universe.

An individual-dependent *V(i)* could be transformed into a fuzzy variable *VI(i)* that becomes individual-independent by taking account of semantic relations among concepts ($e_1...e_N$).

Algorithm 3: Obtain an individual-independent feature vector

$$\begin{aligned} VI(i) &= V(i) \vee . \wedge r(\Omega) \\ &= \begin{bmatrix} s_1 & s_2 & \cdots & s_N \end{bmatrix} \vee . \wedge \begin{bmatrix} 1 & r(1,2) & \cdots & r(1,N) \\ r(2,1) & 1 & \cdots & r(2,N) \\ & & \cdots & \\ r(N,1) & r(N,2) & \cdots & 1 \end{bmatrix} \cdot \\ &= \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix} \end{aligned} \tag{4}$$

$r(\Omega_c)$ is a fuzzy relation matrix and each element of r_{ij} reflects correlation or similarity between entity e_i and entity e_j based on the ontology Ω_c . In this case, similar or close entities can be taken into account even though they are not explicitly cited in their RSS files. The fuzzy relation $r(\Omega_c)$ can be obtained from the ontology definition Ω_C, e.g., of Fig.1, which tells us the closeness between two concepts in ontology Ω_C. It can be calculated as an inverse of the distance matrix accordingly. $r(i,j)=e^{-\alpha d(i,j)}$ and α is a steepness measure. $d(i,j)$ can be obtained from the semantic distance matrix by algorithm 1. $\vee.\wedge$ denotes an inner product for fuzzy inference, such as max-min composition [19]. So

$$x_i = Max(\min(s_1, r(1,i)), \min(s_2, r(2,i)), \cdots, \min(s_N, r(N,i))) \tag{5}$$

After the fuzzy inference in (4) the individual-independent feature vector *VI(i)* is a fuzzy variable considering the semantic correlations of entities in the ontology.

Now the information from a RSS feed and a user could be extracted and represented as a set of individual-independent vectors. Selecting the interested information from RSS publishers is converted into a correlation check between feature vectors. Assume *VIs* and *VIp* denote the favorite vector and the RSS feed vectors respectively. *Ui* is the resulted utility. The following formula is used to filter RSS information.

$$U_i = \frac{\left|VI_s \wedge VI_p\right|}{\left|VI_s\right|} \geq \rho \quad . \tag{6}$$

$\rho \in [0,1]$ is a vigilance parameter set by a RSS subscriber. $VIs \wedge VIp$ is a vector whose i^{th} component is equal to the minimum of VIs_i and VIp_i and $|VIs|$ is the norm of VIs which is defined to be the sum of its components. If every element in VIp is equal to or greater than that in VIs, then $\rho = 1$ and VIp is regarded as a perfect match of VIs. Otherwise VIp_i which makes U_i over ρ will be considered as a sufficient match.

3. An RSS Filter Agent for Job Hunting

We design an RSS filter agent for the job publishing-finding case to illustrate the proposed approach. Suppose a job hunter wants to find a job via RSS feeds from the website http://hotjobs.yahoo.com/jobs/. The job hunter is only interested in the jobs in a specific knowledge domain, for example represented by a DVR developing ontology (Figure 1). The job hunter provides the favorite profile of jobs. The objective of a RSS filter agent is to detect the relevant information and prompt the user.

Several software packages were used:

- Protégé (http://protege.stanford.edu/plugins/owl/index.html) is used to edit a DVR-related technology ontology and export an OWL document.
- Jena RSS package (http://jena.sourceforge.net/) is adopted to parse an RSS feed.
- Jena Ontology API (http://jena.sourceforge.net/) is used to create or parse an OWL document.

The resulted concept list according to the "Breadth-First" scan method is:

{DVR_Related_Technology, Standard, Hardware, Software, Compression_Standard, Digital_Signal_Processing, Network, Hardware_Description_Languages, Circuitry, MCU, Operation_System, Programming_Language, Driver_Development, Database, Video, Audio, FFT, filter, RTP, TCP, IP, RTCP, VHDL, Verilog, PLD, CPLD, FPGA, ASIC, PCB, ARM, MIPS, PowerPC, DSP, Embedded_Operation_System, Macintosh_Operation_System, Windows, DOS, High_Level_Language, Assembly_Language, RS-232, USB, RS-485, IDE, PCI, DB2, MS_SQL, Oracle, MySQL, MPEG, H.26x, AAC, MP3, WMA, WinCE, ucLinux, VxWorks, Linux, Java, C++, C, VB, C#, MPEG-4, MPEG-4_AVC, MPEG-2, MPEG-1, H.263, H.264, H.261, H.262}

The corresponding initial relationship matrix is obtained and the distance matrix can then be computed according to algorithm 1, which is a 70*70 matrix (Figure 3). The greyscale indicates the semantic distance between a concept in the x axis and a concept in the y axis.

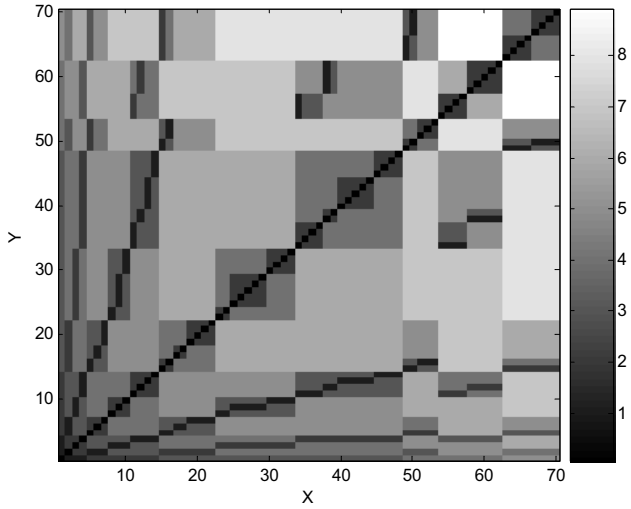

Figure 3. The distance matrix illustration.

The following job titles were received from http://hotjobs.yahoo.com/jobs/:

- J0: “Software Engineer, MPEG, Video, Compression”
- J1: “Senior Firmware Engineer W/ MPEG and ARM”
- J2: “Software Engineer - Programmer - Developer - C++ - Java”
- J3: “C/C++/Linux/Oracle Developers”
- J4: “Embedded Software Engineer –embedded OS, C, Assembly, DSP, Video”
- J5: “Application Engineer,Audio/Video,Hardware,ASIC, PCB”
- J6: “MPEG ASIC/Hardware Engineer”
- J7: “Video Systems, H.264, MPEG, Decoder, FPGA, HDTV”

There are three users who want to find jobs and announce themselves as:

- user0: “A hardware engineer experienced in using CPLD/FPGA with Verilog for design entry”
- user1: “Software Engineer, experienced in C language,Video Compression standard such as H.264 , MPEG ”
- user2: “H.264, MPEG, Video, Assembly, FPGA, DSP”

From these statements, it is easy to observe that they are individual-dependent and do not follow a strict format. The extracted concept graphs can be expressed according to algorithm 2. α is set to 1 in this example. According to algorithm 3 and formula 7, the resulting utility corresponding to every job Ji is:

- user0: [0.0, 0.0, 0.0, 0.0, 0.0, 0.475, 0.475, 0.175]
- user1: [0.833, 0.045, 0.333, 0.122, 0.577, 0.122, 0.045, 0.212]
- user2: [0.135, 0.098, 0.0, 0.0, 0.634, 0.268, 0.098, 0.465]

The relationship between published jobs and announcements of users is illustrated in Figure 4.

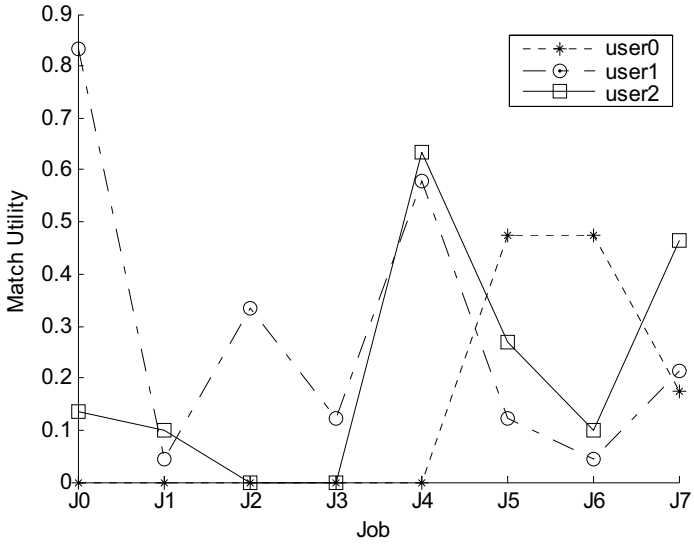

Figure 4. The relationship between published jobs and user profiles.

In Figure 4, for user0, the most related job information is J5 and J6. The method in this paper could be applied to content publisher or content subscriber. For example, on the business website, this method could be applied to find the potential customers and, on the user side, this method could be used to filter favourite information according to a certain level by adjusting the vigilance value of ρ .

The value of parameter ρ determines the information retrieval precision. When ρ is bigger, the precision is higher and more information is filtered. The relation between parameter ρ and the numbers of chosen jobs for each user is shown as follows.

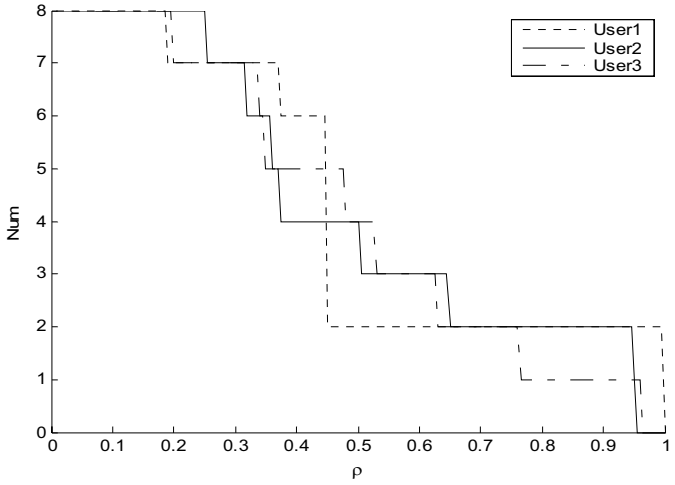

Figure 5. The information filtering for users.

4. Conclusion

In this paper, a method for client-oriented active choice of information from RSS feeds was discussed. A quantitative method for matchmaking between the information the subscriber is looking for and RSS items was proposed, which converts semantic information of headlines in an RSS feed into a numerical vectors and semantic closeness could be measured consequently. The ontology acted as the bridge to link dispersed publishers and subscribers under the assumption of sharing a common knowledge background. Concept graphs are firstly extracted from the title of each RSS item and transformed into an individual-dependent feature vector with the aid of an ontology. In order to eliminate ambiguity inherited from the specific expression of individuals, a fuzzy inference is applied to obtain the grade of membership in terms of ontology, which is an individual-independent feature vector. A job seeking case was used to illuminate the method and the results showed the validity of the method.

Further research needs to be done. Firstly the concept graphs and concept vector are based on an ontology, so the size of the ontology will affect the computation complex. The algorithms need to be improved to handle complex knowledge descriptions represented with large-scale ontologies. Secondly the performance of this method needs to be evaluated in practical settings.

References

[1] Lee, T.B.: Semantic Web Road map (1998) http://www.w3.org/DesignIssues/Semantic.html
[2] Hammersley, B.: Content Syndication with RSS, O'Reilly, ISBN: 0-596-00383-8, (2003)
[3] RSS1.0 specification: http://web.resource.org/rss/1.0/spec
[4] RSS2.0 specification: http://blogs.law.harvard.edu/tech/rss
[5] Grossnickle, J., Board, T., Pickens, B., Bellmont, M.: RSS-crossing into the mainstream (2005) http://publisher.yahoo.com/rss/RSS_whitePaper1004.pdf
[6] Sandler, D., Mislove, A., Post, A., Druschel, P.: FeedTree: Sharing Web micronews with peer-to-peer event notification. In Proceedings of the 4th International Workshop on Peer-to-Peer Systems, Ithaca, NY , USA (2005) 141-151
[7] Glance, N.S., Hurst, M., Tomokiyo, T.: BlogPulse: automated trend discovery for weblogs. In Proceedings of the 13th International WWW Conference: Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, New York, USA (2004) 1-8
[8] Jung, E.: UniRSS: A new RSS framework supporting dynamic plug-In of RSS extension modules. In Proceedings of the 1st Aisan Semantic Web Conference, Beijing, China (2006) 169-178
[9] IBM developerworks: http://www-128.ibm.com/developerworks/rss/
[10] Cancedda, N., Gaussier, E., Goutte, C., Renders, J.: Word-sequence kernels. Journal of machine learning research, 3(2003) 1059-1082
[11] Lodhi, H., Cristianini, N., Shave-Taylor, J., Watkins, C.: Text classification using string kernel. Advances in Neural Information Processing System, 13(2001) 563-569
[12] Szczepaniak, P.S., Niewiadomski, A.: Clustering of documents on the basis of text fuzzy similarity. Abramowicz W. (Eds.): Knowledge-based Information Retrieval and Filtering from the Web (2003) 219–230
[13] Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In Proceedings of the Fourteenth International Conference on Machine Learning, San Francisco, USA (1997) 412–420
[14] Prabowo, R., and Thelwall, M.: A comparison of feature selection methods for an evolving RSS feed corpus. Information Processing and Management, 42(2006) 1491-1512
[15] Wegrzyn-Wolska, K., Szczepaniak, P.S.: Classification of RSS-formatted documents using full text similarity measures. In Proceedings of the 5th International Conference on Web Engineering, Sydney, Australia (2005) 400-405
[16] Conlan, O., O'Keeffe, I., Tallon, S.: Combining adaptive hypermedia techniques and ontology reasoning to produce dynamic personalized news services. In Proceedings of the Fourth International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, Dublin, Ireland (2006) 81-90

[17] Sycara, K.P., Klusch, M., Widoff, S., Lu, J.: Dynamic service matchmaking among agents in open information environments, ACM Special Interests Group on Management of Data Record, 28(1999) 47-53
[18] Akoka, J., Comyn-Wattiau, I.: Entity-relationship and object-oriented model automatic clustering. Data and Knowledge Engineering, 20 (1996) 87-117
[19] Williams, J., Steele, N.: Difference, distance and similarity as a basis for fuzzy decision support based on prototypical decision classes. Fuzzy Sets and Systems, 131(2002) 35-46
[20] Zadeh, L. A.: Fuzzy sets. Information and Control, 8(1965) 338-353

This page intentionally left blank

Applications of Data Mining in E-Business and Finance
C. Soares et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-890-8-87

Comparing Negotiation Strategies Based on Offers

Lena MASHAYEKHY [a], Mohammad Ali NEMATBAKHSH [b] and Behrouz Tork LADANI [b]
[a] *Young Researchers Club of Arak, Arak, Iran*
[b] *Computer Eng. Department, University of Isfahan, Isfahan, Iran*
Email: {lmashayekhy, nematbakhsh, ladani}@eng.ui.ac.ir

Abstract. Negotiation is a process between self-interested agents trying to reach an agreement on one or multiple issues in an ecommerce domain. The knowledge of an agent about the opponents' strategies improves the negotiation outcome. However, an agent negotiates with incomplete information about its opponent. Given this, to detect the opponent's strategy, we can use the similarity between opponents' strategies. In this paper we present a method for measuring the similarity between negotiators' strategies. Offers are generated by the agent's strategy therefore our similarity measure is based on the history of offers in negotiation sessions. We extended the Levenshtein distance technique to detect similarity between strategies. We implement this measure and experimentally show that the result of using the measure improves the recognition of the opponent's strategy.

Keywords. Automated Negotiation, Measure, Similarity, Strategy.

Introduction

Automated negotiation is a key form of interaction in complex systems composed of autonomous agents. Negotiation is a process of making offers and counteroffers with the aim of finding an acceptable agreement [1]. The agents (negotiators) decide for themselves what actions they should perform, at what time, and under what terms and conditions [1,2]. The outcome of negotiation depends on several parameters such as the agents' strategies and the knowledge which one agent has about the opponents [2,3,4,5]. In recent years, the problem of modeling and predicting a negotiator behavior has become increasingly important since this can be used to improve negotiation outcome and increase satisfaction of result [2,3,4,6].

Similarity is a fundamental notion that has to be defined before applying various statistical, machine learning, or data mining methods [5]. Previous works have attempted to exploit the information gathered from opponent's offers during the negotiation process to infer similarity between offers of the opponent to predict future offers. Bayesian classification [7] and similarity criteria [2,3] are examples of such efforts. In this work we want to detect similarity between strategies of negotiators not only during negotiation but also after negotiations to discover knowledge from past

experience. When an agent has knowledge about the opponent's strategy, it can use this knowledge to negotiate better deals for itself [1,6]. However, an agent negotiates with incomplete information about an opponent therefore the use of data mining techniques can assist the negotiator to discover valuable knowledge [6]. Some of these techniques, such as clustering, need a distance function or similarity measure. The main problem is that there are no measures for calculating similarity between negotiators' strategies.

A sequence of offers is a common form of data in negotiation that an agent can use to discover valuable knowledge in order to achieve its goal [2]. A session is defined as an ordered sequence of offers that an agent creates during a negotiation based on its strategy [3]. To detect the similarity between negotiators' strategies, we use data of sessions. Like sequences, one method is to reduce sessions to points in a multi-dimensional space and use the Euclidean distance in this space to measure similarity, but in negotiation, different sessions do not have the same length. One solution discussed in [8] for sequences, is to select n offers of each session to calculate the Euclidean distance. The problem with this approach is: which n offers in each session best represent the strategy of the negotiator. Another method is to represent sessions in k-dimensional space using k features for each session [8]. Using the feature vector representation not only needs the definition of features to model the strategy of the negotiator, but also the problem of sessions' similarity is transformed into the problem of finding similar features in k-dimensional space.

In this paper we consider the problem of defining similarity (or distance) between strategies. We start with the idea that similarity between negotiators should somehow reflect the amount of work that has to be done to convert one negotiation session to another. We formalize this notion as a Levenshtein or edit distance [8,9] between negotiations. We apply a dynamic programming approach for computing the extension of Levenshtein distances and show this similarity measure is efficient in practice.

In detail, the paper is organized as follows. In Section 1 we present the statement of the problem in automated negotiation. In Section 2 we make a review on the negotiation protocol. Some types of negotiation strategies are discussed in Section 3. The definition of similarity between negotiation strategies is given in Section 4. In Section 5 we evaluate the effectiveness of using this measure for detecting similar negotiators strategies and we proceed with the analysis of the results obtained. Section 6 contains conclusions and remarks about future directions.

1. Problem Statement

To model a negotiation, we consider a given set $S= (O_1, ..., O_m)$ of m offers that a negotiator exchanges during the negotiation session S. An offer O consists of one or multiple issues.

The basic problem we consider in this paper is how one should define a concept of similarity or distance between negotiation strategies. Such a notion is needed in any knowledge discovery application on negotiation. The offers exchanged during a negotiation show the negotiator's session strategy [1,2,3,4,10]. To find negotiators with similar strategies, if one can not say when two negotiation sessions are close to each other, the possibility for making contrast is quite limited. For example, consider three buyers that negotiate with a seller who wants to compare the behavior of these buyers. The seller's view of these sessions (received offers) is shown in Figure 1. Each one of the buyers has its initial offer, deadline and strategy to generate offers.

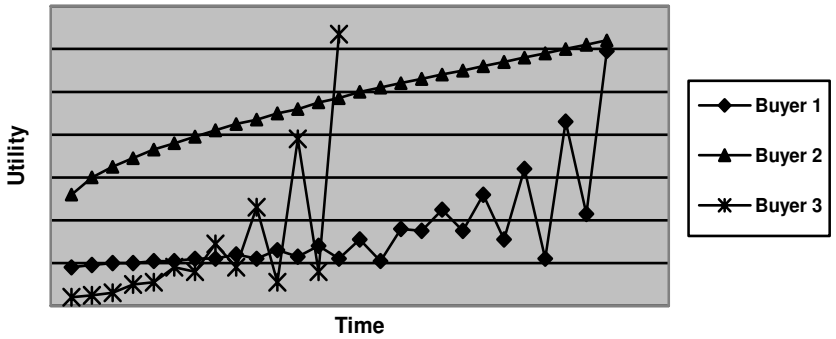

Fig. 1. Buyers' offers

Consider the problem of clustering these three buyers. When comparing two buyers to find if they are similar, we need a similarity measure. The meaning of similarity may vary depending on the domain and the purpose of using similarity. For example someone could group buyer 1 and 2 together, with buyer 3 as the other group because of the number of exchanged offers. But in this paper we want to measure the similarity between negotiators based on their strategies. When a seller observes that the changes of offers received from different buyers are similar during their sessions, then this seller finds that these buyers have similar strategies. In Section 4 we discuss about this similarity measure.

2. Negotiation Protocol

In order to understand the notation used in our model, we firstly describe its basics. Automated negotiation consists of a set of agents equipped with a common protocol. In this negotiation model, we use bilateral negotiation in which a seller and a buyer negotiate on one or multiple issues. We adopt an alternating offers protocol; that is both of them can send and receive offers and decide whether to accept or reject an offer received until they reach their own deadline [5,11]. Each of them has incomplete information about its opponent.

Let $a \in \{buyer, seller\}$ represent the negotiating agents, a' denotes the opponent of agent a. Let $j \in \{1,...,n\}$ be the issues under negotiation therefore an offer O is described as $(o_1, \ldots, o_n)$ where o_j^a shows the value of issue j for agent a. Negotiations can range over quantitative (e.g., price, delivery time) or qualitative (e.g., quality of service) issues. Quantitative issues are defined over a real domain that is acceptable to agent a (i.e., $o_j^a \in D_j^a = [min_j^a, max_j^a]$) and qualitative issues are defined over a partially ordered set (i.e., $o_j^a \in D_j^a = \{q_1, q_2, ..., q_p\}$).

Each agent has a utility function $U_j^a : D_j^a \rightarrow [0,1]$ that gives the score it assigns to a value of issue j in the range of its acceptable values. For convenience, scores are kept in the interval [0, 1].

The relative importance that an agent assigns to each issue under negotiation is modeled as a weight, w_j^a, that gives the importance of issue *j* for agent *a*. We assume the weights of both agents are normalized, i.e., $\sum_{1 \le j \le n} w_j^a = 1$, for all $a \in \{buyer, seller\}$. An agent's utility function for an offer, $O = (o_1, \ldots, o_n)$ in the multi-dimensional space defined by the issues' value ranges is then defined as:

$$U^a(O) = \sum_{1 \le j \le n} w_j^a \cdot U_j^a(o_j) . \quad (1)$$

The agents alternately propose offers at times in $T = \{0, 1, \ldots\}$. Each agent has its deadline. T^a denotes the deadline of agent *a*, when the agent must complete the negotiation.

Let $O^t_{a \to a'}$ denote the offer proposed by agent *a* at time *t*. The agent who makes the first offer is chosen randomly. When an agent receives an offer from its opponent at time *t*, it rates the offer using its utility function U^a and generates a response that is defined as [4]:

$$Action^a(t, O^t_{a' \to a}) = \begin{cases} Quit & If\ t > T^a \\ Accept & If\ U^a(O^t_{a' \to a}) \ge U^a(O^{t+1}_{a \to a'}) \\ Offer\ O^{t+1}_{a \to a'} & Otherwise \end{cases} \quad (2)$$

Offers are generated by the agent's strategy which is discussed in Section 3.

If the agent's deadline passes, the agent withdraws from the negotiation and the negotiation outcome is *"reject"*.

An agent accepts an offer when the value of the offer received is higher than the offer which the agent is ready to send at that moment in time. Therefore the outcome of the session is *"accept"* and the session is closed.

A negotiation session between *a* and *a'* at time *t* ($0 < t < T^a$) is a finite sequence of offers from one agent to the other ordered over time:

$S^t_{a \to a'} = (O^1_{a \to a'}, O^2_{a' \to a}, \ldots)$

The last element of the sequence is from the set {*accept, reject*} based on the outcome of the negotiation session.

3. Negotiation Strategies

Offers are generated by negotiation strategy [4]. A strategy generates a value for each of the issues in negotiation. Two types of strategies which we used in our work are Time dependent and Behavior dependent.

3.1. Time Dependent

This strategy is parameterized and hence it covers a large number of distinct strategies. As time passes, the agent will concede more rapidly trying to achieve an

agreement before coming to the deadline. The offer to be uttered by agent *a* for an issue *j* at time *t* ($0 < t < T^a$) is computed as follows [1]:

$$o_j^t = \begin{cases} min_j^a + \varphi^a(t)(max_j^a - min_j^a) \\ \qquad \text{if } U_j^a \text{ is a decreasing function} \\ min_j^a + (1-\varphi^a(t))(max_j^a - min_j^a) \\ \qquad \text{if } U_j^a \text{ is an increasing function} \end{cases} \quad (3)$$

where $\varphi^a(t)$ is a function depending on time ($0 \le \varphi^a(t) \le 1$) and is parameterized by a value β.

$$\varphi^a(t) = \left(\frac{t}{T^a}\right)^{\frac{1}{\beta}} . \quad (4)$$

A wide range of time-dependent strategies can be defined by varying the way in which $\varphi^a(t)$ is computed [3]. However, depending on the value of β, three qualitatively different patterns of behavior can be identified:

- Boulware if $\beta < 1$
- Linear if $\beta = 1$
- Conceder if $\beta > 1$.

3.2. Behavior Dependent

The key feature of this type of strategy is that it proposes offers based on the opponent's behavior [4].

$$o_j^{t+1} = \begin{cases} min_j^a & If \;\; P \le min_j^a \\ max_j^a & If \;\; P > max_j^a \\ P & Otherwise \end{cases} \quad (5)$$

The parameter *P* determines the type of imitation to be performed. We can find the following families:

Relative Tit-For-Tat. The agent reproduces, in percentage terms, the behavior that its opponent showed $\delta > 1$ steps ago.

$$P = \frac{o_j^{t-2\delta}}{o_j^{t-2\delta+2}} o_j^{t-1} . \quad (6)$$

Absolute Tit-For-Tat. The same as before, but in absolute terms.

$$P = o_j^{t-1} + o_j^{t-2\delta} - o_j^{t-2\delta+2} \ . \quad (7)$$

Averaged Tit-For-Tat. The agent applies the average of percentages of changes in a window of size $\lambda \geq 1$ of its opponent's history.

$$P = \frac{o_j^{t-2\lambda}}{o_j^t} o_j^{t-1} \ . \quad (8)$$

We compute the values for the issues under negotiation according to each strategy.

4. Similarity Measure

We propose a new session similarity measure and use this measure for calculating the similarity between strategies of negotiators. In this section, we define two key concepts: first distance between two sessions and second distance between two offers.

4.1. Distance Between Sessions

A session is defined as an ordered sequence of offers which an agent creates during a negotiation based on its strategy [3]. To detect similarity between negotiators' strategies, we use data obtained from sessions. Therefore we just select offers proposing by an agent.

The idea behind our definition of similarity, or distance, between negotiation sessions is that it should somehow reflect the amount of work needed to transform one negotiation session to another [8,9]. The definition of similarity is formalized as an extended Levenshtein distance *eld(S, R)* for two sessions *S* and *R*.

Operations. To calculate the Levenshtein distance we need to define a set of transformation operations. We have chosen to use three operations:

- *ins(O)*: inserts an offer of the type *O* to the negotiation session.
- *del(O)*: deletes an offer of the type *O* from the negotiation session.
- *update(O, O')*: change an existing offer from *O* to *O'* in the negotiation session.

Cost of Operations. Instead of checking the equality between two offers O_S and O_R from two sessions *S* and *R* respectively, for each operation we associate a cost *cost(op)* based on the distance between offers. The cost of an insertion operation is defined by Eq. (9) where *O'* is a previous offer of *O* in the negotiation session.

$$cost(ins(O)) = distance(O', O) \ . \quad (9)$$

With this definition the cost of adding an outlying offer into the negotiation session is higher than the cost of adding in a neighboring offer. In Section 4.2 we discuss about the distance between two offers.

The cost of a deletion operation is defined to be the same as the cost of an insert-operation. It is proved that if the cost of insertion is equal to the cost of deletion then for all negotiation sessions S and R we have [9]:

$$eld(S, R)=eld(R, S) . \quad (10)$$

The cost of an update-operation is defined as Eq. (11) where V is a constant value.

$$cost(update(O, O'))=V.distance(O, O') . \quad (11)$$

With this definition a low distance has a lower cost that a higher distance.

Definition of Distance. If the cost of an operation op_i is $cost(op_i)$, and k is the number of operations in the sequence Op_j, Eq. (12) calculates the cost of operation sequence $Op_j = op_1, op_2, \ldots, op_k$.

$$cost(Op_j) = \sum_{i=1}^{k} cost(op_i) . \quad (12)$$

The distance $eld(S, R)$ is defined as the sum of costs of the cheapest sequence of operations transforming S into R as shown in Eq. (13).

$$eld(S, R) = min\{cost(Op_j) \mid Op_j \text{ is an operation sequence transforming a session } S \text{ into a session } R \} . \quad (13)$$

That is $eld(S, R)$ is the minimum sum of costs of operations transforming S into R.

We use a dynamic programming approach to find the extended Levenshtein distance of two sessions S and R ($eld(S, R)$). This is dynamic programming approach to calculate the distance between two sessions.

```
CostType eld(OfferType R[], OfferType S[]) {
  CostType cost[n+1][m+1];
  cost[0][0]=0;

  //for session S
  for (int i=1;i<=m;i++)  //session S has m offers
    cost[0][i] = cost[0][i-1]+ins(S[i]);
  //for session R
  for (int i=1;i<=n;i++)  //session R has n offers
    cost[i][0] = cost[i-1][0]+del(R[i]);

  for (int i=1; i<=n;i++)
    for (int j=1;j<=m;j++)
      cost[i][j]=min{cost[i-1][j]+del(R[i]),
                     cost[i][j-1]+ins(S[j]),
                     cost[i-1][j-1]+update(R[i], S[j])};

  return cost[n][m];
}
```

4.2. Distance Between Offers

The distance between two offers in insert, delete and update operation can be defined in a different way for each type of negotiation. Let *O* and *O'* be two offers.

In a single issue negotiation, distances do not need normalization. If each offer has a quantitative value such as price, we define *distance(O, O')* as Eq. (14).

$$distance(O, O') = | O - O' | . \quad (14)$$

For a qualitative issue distance can be calculated based on equality. In that case, the distance between any two offers is defined to be 0 if they are equal; and a positive value if they are not equal. This value can be chosen based on fuzzy logic or based on the utility function.

In multiple issue negotiation, for each issue o_j the distance is calculated based on its quantitative or qualitative value. Then the Euclidean distance is used for calculating the distance between offers. If the issues have different importance, their importance have influence on the distance value. Eq. (15) shows how to calculate the distance between two offers.

$$distance(O, O') = \sqrt{\sum_{1 \le j \le n} w_j . distance(o_j, o'_j)^2} . \quad (15)$$

where w_j is an appropriate weight representing the importance of the issue *j* and $\sum_{1 \le j \le n} w_j^a = 1$.

In multiple issue negotiation, the distance between issues needs to be normalized because it influences the distance between offers. For this purpose, for each issue o_j and o'_j the distance is calculated as:

$$distance(o_j, o'_j) = \left| U_j(o_j) - U_j(o'_j) \right| . \quad (16)$$

5. Experimental Results

In this section we describe how we have evaluated the effectiveness of using this measure for detecting similar negotiator strategies under different negotiation situations. In this experiment we generate 2500 negotiation sessions. In each session a buyer and a seller negotiate for price as a single issue. They choose one of the implemented strategies that were discussed above (Conceder, Linear, Boulware, Relative TFT, Absolute TFT, and Average TFT). An agent's utility function is defined as:

$$U_j^a(o_j^t) = \begin{cases} \dfrac{max_j^a - o_j^t}{max_j^a - min_j^a} & if\ a = buyer \\ \dfrac{o_j^t - min_j^a}{max_j^a - min_j^a} & if\ a = seller \end{cases} \quad (17)$$

Buyers and sellers save information about their strategies, outcome and all exchanged offers during the process of negotiation. Information about buyers and sellers' strategies is shown in table 1 and 2.

Table 1. Percent of Buyers' Strategies

Strategy	Percent
Relative TFT	15.4
Random Absolute TFT	19.6
Average TFT	17.6
Boulware	15.8
Linear	15.4
Conceder	16.2
Total	100.0

Table 2. Percent of Sellers' Strategies

Strategy	Percent
Relative TFT	17.2
Random Absolute TFT	12.8
Average TFT	18.2
Boulware	16.8
Linear	16.4
Conceder	18.6
Total	100.0

After gathering data from all sessions, we choose sessions with “accepted” outcome. In each session we choose buyer offers to detect similarity of buyers strategies. We use our method for calculating the distance between these sessions to determine the distance between buyers' strategies.

After calculating all distances we use the *k*-medoids algorithm [12] to cluster the sessions based on these distances, in order to evaluate our measure. This algorithm is helpful because the center of each cluster is one of the points existing in the data belonging to that cluster. Therefore, the cluster centers are negotiation sessions. This characteristic is important because in this work, we have distances between sessions and do not need to know the offers made during the sessions; therefore, to find a cluster center we just need a session which has minimum distance with other sessions in the cluster. As a result, the comparison between sessions and the cluster center is simple. Furthermore to cluster a new buyer we can compare it with cluster centers if we have the offers of the cluster center session to calculate distance. If a cluster center is not one

of the existing sessions, we do not have real offers of the cluster center to compute the distance between the cluster center and the offers of a new buyer.

Since a buyer saves information about the strategy used in his session, we use this information to analyse our method. To demonstrate that our method is practical for clustering and that the clusters are created based on the similarity between strategies, we check the following: if two buyers use similar strategies and these are located in the same cluster by the clustering, and if two buyers use dissimilar strategies and are located in different clusters, our method to measure strategies similarity is efficient. In fact all the buyers that use the same strategies in their negotiation sessions should form one cluster.

As we know the number of strategies of buyers, we choose k=6 for k-medoids.

After clustering, we check each cluster and find the most common strategy which buyers in that cluster use in his sessions. Table 3 shows the most common strategy in the sessions of each cluster.

Table 3. Percentage of the most common strategy in each cluster

Number of cluster	Strategy	Percent
1	Relative TFT	98%
2	Random Absolute TFT	100%
3	Average TFT	90%
4	Boulware	88%
5	Linear	89%
6	Conceder	100%

These results show that our method is useful for calculating the similarity between the strategies of buyers because each cluster contains buyers with similar strategies. For example in the cluster number 1, 98% of buyers use the Relative TFT strategy in their negotiation sessions.

But in some clusters such as 5, not all the strategies are the same; this is because one buyer uses a strategy which is, nevertheless very close to the strategies of the other buyers in the cluster. The data in this cluster show that some of the other buyers' strategies are Boulware with $\beta \cong 1$ which is similar to a Linear strategy. Therefore, the results show that buyers in each cluster have similar behavior.

Fig. 3 shows changing offers of some sessions in cluster number 2. In Fig. 4 some sessions of cluster number 5 are shown. This cluster contains some Boulware and Conceder strategies which are close to the Linear strategy.

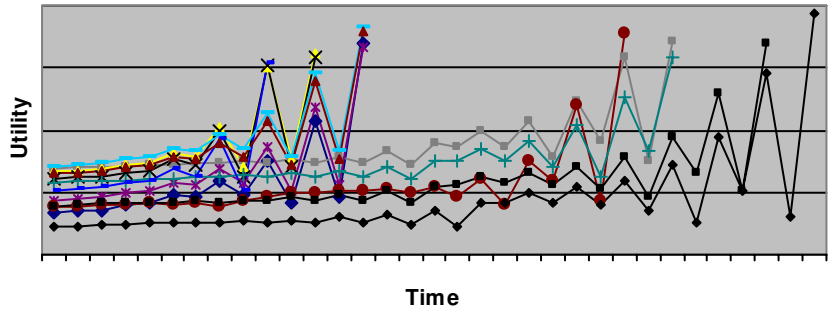

Fig. 3. Sessions in the first cluster

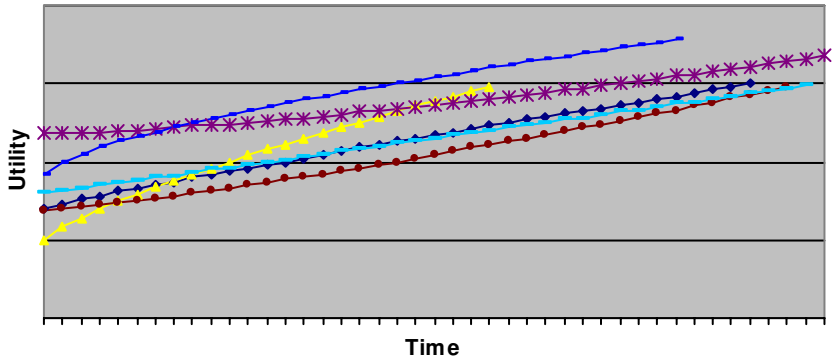

Fig. 4. Sessions in the second cluster

The experiments are repeated with different numbers of clusters and with different negotiation strategies. All experiments show each cluster has buyers which use similar strategies.

As we mentioned above our experiment was based on data of buyers with an outcome of "accepted", but for other data one can do similar experiments.

In this paper we mainly consider a simplified model of negotiation, where each offer has only one issue. As we discussed in Section 4 the presented method can be extended for multiple issue negotiation.

6. Conclusion

The outcome of negotiations depends on several parameters such as the strategies of agents and the knowledge that one agent has about the others. The problem of modeling and predicting a negotiator's behavior is important since this can be used to improve the outcome of negotiations and increase satisfaction with the results. Finding similar behavior is one way to solve this problem. We have described a simple method for defining the similarity between negotiation strategies. This method is based on the sequence of offers during a negotiation. This characteristic gives the method significant

practical value in negotiation because a negotiator has incomplete information about his opponents. Results can be used in knowledge discovery.

This method is implemented using dynamic programming and it is tested with a simple model of negotiation. Results of comparing strategies using our measure to find similar strategies are illustrated. The results show that this measure is efficient and can be used in clustering and any other techniques which need a similarity measure.

For the future, there are two ways in which this research can be extended. Firstly, we would like to consider the performance of our method against additional strategies. Secondly, in this work we only consider single issue negotiation model, our method could be applied to other negotiation models.

We plan to experimentally use this method for predicting opponent's strategy during negotiation.

References

[1] Braun, P., Brzostowski, J., Kersten, G., kim, J.B., Kowalczyk, R., Strecker, S., and Vahidov, R.: e-Negotiation Systems and Software Agents: Methods, Models and Applications. In J. Gupta; G. Forgionne; M. Mora (Hrsg.): Intelligent Decision-Making Support System: Foundation, Applications, and Challenges, Springer, Heidelberg ea., Decision Engineering Series, (503 p, 105 illus, Hardcover. ISBN: 1-84628-228-4) (2006).

[2] Coehoorn, R.M., Jennings, N.R.: Learning an opponent's preferences to make effective multi-issue negotiation tradeoffs. In: 6th International Conference on Electronic Commerce (ICEC2004), pp 113–120, Delft, The Netherlands (2004).

[3] Faratin, P., Sierra, C., and Jennings, N.R.: Using Similarity Criteria to Make Issue Trade-Offs in Automated Negotiations. In Artificial Intelligence 142, pp 205-237, (2002).

[4] Hou, C.: Modelling Agents Behaviour in Automated Negotiation. Technical Report KMI-TR-144. Knowledge Media Institute, The Open University, Milton Keynes, UK (2004).

[5] Lai, H., Doong, H., Kao, C., and Kersten, G.E.: Understanding Behavior and Perception of Negotiators from Their Strategies. Hawaii International Conference on System Science (2006).

[6] Mashayekhy, L., Nematbakhsh, M.A., and Ladani, B.T.: E-Negotiation Model based on Data Mining. In Proceedings of the IADIS e-Commerce 2006 International Conference, pp 369-373, ISBN: 972-8924-23-2, Barcelona, Spain (2006).

[7] Tesauro, G.: Efficient Search Techniques for Multi-Attribute Bilateral Negotiation Strategies. In Proceedings of the 3rd International Symposium on Electronic Commerce. Los Alamitos, CA, IEEE Computer Society, pp 30-36 (2002).

[8] Hetland, M.L.: A Survey of Recent Methods for Efficient Retrieval of Similar Time Sequences. First NTNU CSGSC (2001).

[9] Jagadish, H.V., Mendelzon, A.O., and Milo, T.: Similarity-based queries. pp 36–45 (1995).

[10] Fatima, S.S., Wooldridge, M., and Jennings, N.R.: An agenda-based framework for multi-issue negotiation". Artificial Intelligence, 152(1), pp 1-45 (2004).

[11] Li, C., Giampapa, J., and Sycara, K.: A Review of Research Literature on Bilateral Negotiations. In Tech. report CMU-RI-TR-03-41, Robotics Institute, Carnegie Mellon University (2003).

[12] Han, J., Kamber, W.: Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers (2000).

Applications of Data Mining in E-Business and Finance
C. Soares et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-890-8-99

Towards Business Interestingness in Actionable Knowledge Discovery[1]

Dan LUO [a], Longbing CAO [a], Chao LUO [a], Chengqi ZHANG [a] and Weiyuan WANG [b]
[a] *Faculty of Information Technology, University of Technology, Sydney, Australia*
e-mail: {dluo, lbcao, chaoluo, chengqi}@it.uts.edu.au
[b] *A2 Consulting Pty Limited, Sydney, Australia*
e-mail: {weiyuan.wang}@gmail.com

Abstract. From the evolution of developing a pattern interestingness perspective, data mining has experienced two phases, which are Phase 1: technical objective interestingness focused research, and Phase 2: technical objective and subjective interestingness focused studies. As a result of these efforts, patterns mined are of significant interest to technical concern. However, technically interesting patterns are not necessarily of interest to business. In fact, real-world experience shows that many mined patterns, which are interesting from the perspective of the data mining method used, are out of business expectations when they are delivered to the final user. This scenario actually involves a grand challenge in next-generation KDD (Knowledge Discovery in Databases) studies, defined as actionable knowledge discovery. To discover knowledge that can be used for taking actions to business advantages, this paper addresses a framework that extends the evolution process of knowledge evaluation to Phase 3 and Phase 4. In Phase 3, concerns with objective interestingness from a business perspective are added on top of Phase 2, while in Phase 4 both technical and business interestingness should be satisfied in terms of objective and subjective perspectives. The introduction of Phase 4 provides a comprehensive knowledge actionability framework for actionable knowledge discovery. We illustrate applications in governmental data mining showing that the considerations and adoption of the framework described in Phase 4 has potential to enhance both sides of interestingness and expectation. As a result, knowledge discovered has better chances to support action-taking in the business world.

Keywords. business interestingness, actionable knowledge discovery

Introduction

Patterns that are obtained with data mining tools are often non-actionable to real user needs in the business world [1,2,3,4]. There could be many reasons associated with this scenario. Most importantly, we believe, it is because *business interestingness* is rarely considered in existing data mining methodology and framework. For instance, in stock data mining [5], mined trading patterns are normally evaluated in terms of technical interestingness measures such as the correlation coefficient. On the other hand, traders

[1]This work is sponsored by Australian Research Council Discovery and Linkage Grants (DP0773412, LP0775041, DP0667060), and UTS internal grants.

who use these discovered patterns usually only check business expectations like profit and return. However, due to no checking of such business interestingness during pattern discovery, the identified trading patterns in most cases are useless for real-life trading support.

Such situations are increasingly recognized in current data mining research, especially actionable knowledge discovery. Initial research has been done on developing subjective and business related interestingness, which mainly aims at developing standard and general measures [6,7]. However, due to domain-specific characteristics and constraints [8,2,3,9,10,11] heavily affecting real-life data mining, it is difficult or even hardly possible to capture and satisfy particular business expectations in a general manner. As a result, the gap between technical interestingness and business expectations has not been filled or reduced as expected by business users. Therefore, it is essential to involve business expectations into the process of knowledge discovery. To this end, a practical and effective manner is to study business interestingness in terms of specific domains while developing a domain driven data mining methodology [8,2,9,10,11]. In fact, similar issues and observations have been recognized during the development of CRISP-DM 2.0 [12]. In this way, eventually, the objectives of actionable knowledge discovery can be reached.

Even though the aim of actionable knowledge discovery needs to involve many aspects [8,2,9,10,11] besides interestingness, this paper only addresses the business interestingness issue. Obviously, it is important to develop business interestingness metrics so that business concerns can be reflected in real-world mining applications. In practice, there are ways to do it. For instance, in financial data mining, business evaluation metrics can be instantiated into *objective* [13,14] and *subjective* [6,15] measures as follows. *Profit*, *return*, *cost* and *benefit* [16,17] can be used to indicate a trading pattern's economic performance objectively. On the other hand, a trading pattern can be evaluated in terms of certain psychoanalytic factors specified by traders to measure its business significance. For example, "beat VWAP" is used by traders to measure whether a price-oriented trading rule is confident enough to "beat" Value-Weighted Average Price (VWAP) of the market. Therefore, besides *technical interestingness*, actionable knowledge mining needs to develop both *objective* and *subjective business interestingness* measures.

The above ideas actually indicate a view of the evolution of pattern interestingness development in data mining. We roughly categorize the initial work into the following efforts:

- studying interestingness metrics highlighting the significance and evaluation of technical subjective concerns and performance,
- developing a new theoretical framework for valuing the actionability of extracted patterns,
- involving domain and background knowledge into the search of actionable patterns.

Aiming at involving and satisfying business expectations in actionable knowledge discovery, this paper presents a framework of knowledge actionability. An extracted knowledge is *actionable* if it can satisfy not only technical concerns but also business expectations. Following this framework, we have developed a domain-driven, actionable knowledge discovery methodology [8,2,9,10,11]. This paper demonstrates the develop-

ment of particular business interestingness measures in mining social-security activity patterns associated with government debts [18,19,20]. The case studies show that the involvement of business expectations can greatly evidence and enhance the evaluation foundation of knowledge actionability and business interest of identifying patterns.

The remainder of this paper is organized as follows. In Section 1, a knowledge actionability framework is discussed which highlights the significant involvement of business expectations. In Section 2, we illustrate how to integrate business interestingness in the process of discovering activity patterns in social security data. Further, we present some discussion on the balance and resolution of the incompatibility between technical significance and business expectations in Section 3. We conclude this paper in Section 4.

1. Balancing Technical and Business Interestingness

Technically, the progress of pattern interestingness studies has experienced two phases. Recently, a typical trend is towards business interestingness, and the balance between technical and business interestingness so that the gap between academic findings and business expectations can be bridged.

1.1. Knowledge Actionability Studies Concentrating on Technical Significance

In the development of data mining methodologies and techniques, the understanding and modeling of knowledge actionability is a progressive process. In the framework of traditional data mining, the so-called actionability, $act()$, is mainly embodied in terms of technical significance. In general, technical interestingness, $tech\_int()$, measures whether a pattern is of interest or not in terms of a specific statistical significance criterion corresponding to a particular data mining method.

There are two steps of technical interestingness evolution. The original focus was basically on technical objective interestingness, $tech\_obj()$[13,14], which aims to capture the complexities and statistical significance of pattern structure. For instance, a coefficient was developed for measuring the objective interestingness of correlated stocks.

Let $X = \{x_1, x_2, \ldots, x_m\}$ be a set of items, DB be a database, and x an itemset in DB. Let e be interesting evidence discovered in DB through a modeling method M. For the above procedure, we have the following definition.

Definition 1. *Phase 1:* $\forall x \in X, \exists e$: $x.tech\_obj(e) \longrightarrow x.act(e)$

Recent work appreciates technical subjective measures, $tech\_sub()$ [6,21,15], which also recognize the extent to which a pattern is of interest to a particular user. For example, probability-based belief [4] is used to describe user confidence on unexpected rules [21,22].

Definition 2. *Phase 2:* $\forall x \in X, \exists e$: $x.tech\_obj(e) \wedge x.tech\_subj(e) \longrightarrow x.act(e)$

It is fair to say that the traditional data mining research framework is mainly focused on the technical significance-based methodology. Unfortunately, very few of the algorithms and patterns identified are workable in real-world mining applications. Business expectation are rarely cared for and most work concentrates on technical significance. In

Table 1. Pattern's performance.

Relationship Type	Explanation
$tech\_int() \Leftarrow biz\_int()$	The pattern e does not satisfy business expectation but satisfies technical significance
$tech\_int() \Rightarrow biz\_int()$	The pattern e does not satisfy technical significance but satisfies business expectation
$tech\_int() \Leftrightarrow biz\_int()$	The pattern e satisfies business expectation as well as technical significance
$tech\_int() \not\Leftrightarrow biz\_int()$	The pattern e satisfies neither business expectation nor technical significance

practice, from the actionability perspective, a pattern e may present one of the following four scenarios as listed in Table 1.

Therefore, the objective of actionable knowledge discovery is to mine patterns satisfying the relationship $tech\_int() \Leftrightarrow biz\_int()$. However, in real-world mining, it is often a kind of artwork to tune thresholds and balance the significance and the difference between $tech\_int()$ and $biz\_int()$. Quite often a pattern with significant $tech\_int()$ creates less significant $biz\_int()$. Contrarily, it often happens that a pattern with less significant $tech\_int()$ generates a much higher business interest, $biz\_int()$ [11]. Our experience advises us that new data mining methodologies and techniques should be studied to balance technical significance and business expectations, and bridge the gap between business development and academic research. To this end, in Section 3, we discuss some lessons that are helpful for the balance and to make the bridge between academic research and real-world development.

1.2. Knowledge actionability satisfying technical significance and business expectation

With the involvement of domain intelligence and domain experts, data miners realize that the actionability of a discovered pattern must be assessed by and satisfy domain user needs. To achieve business expectations, *business interestingness*, $biz\_int()$, measures to what degree a pattern is of interest to a business person in terms of social, economic, personal and psychoanalytic factors. Similar to $tech\_int()$, *business objective interestingness*, $biz\_obj()$, has recently been recognized by some researchers, say in profit mining [7] and domain-driven data mining [2], as part of $biz\_int()$. In this case, we reach Phase 3 of knowledge actionability studies.

Definition 3. *Phase 3:* $\forall x \in X, \exists e : x.tech\_obj(e) \wedge x.tech\_subj(e) \wedge x.biz\_obj(e) \longrightarrow x.act(e)$

The above review of knowledge actionability studies shows that *actionable knowledge* should satisfy both technical and business concerns [2]. Since the satisfaction of technical interestingness is the antecedent of actionability, we view *actionable knowledge* as that which satisfies not only technical interestingness, $tech\_int()$, but also user-specified business interestingness, $biz\_int()$. In fact, actionability should recognize technical significance of an extracted pattern that also permits users to specifically react to it to better service their business objectives.

Definition 4. *Knowledge Actionability*: Given a mined pattern e, its actionable capability $act(e)$ is described as its degree of satisfaction of both technical and business interestingness.

$$\forall x \in X, \exists e \;\; x.tech\_int(e) \wedge x.biz\_int(e) \longrightarrow x.act(e) \tag{1}$$

In real-world data mining, one also recognizes that *business subjective interestingness*, $biz\_sub()$, also plays an essential role in assessing $biz\_int()$. This leads to a comprehensive cognition of actionability, which we name as Phase 4. In this way, the above knowledge actionability framework can be further instantiated in terms of *objective* and *subjective* dimensions from both *technical* and *business* sides as follows.

Definition 5. *Phase 4:* $\forall x \in X, \exists e$: $x.tech\_obj(e) \wedge x.tech\_subj(e) \wedge x.biz\_obj(e) \wedge x.biz\_subj(e) \longrightarrow x.act(e)$

As we will discuss in Section 2, in evaluating activity patterns discovered in social security data, we define specific business measures *debt amount*, *debt duration*, *debt amount risk*, and *debt duration risk* to measure the risk and impact associated with an activity pattern or sequence.

On the basis of the above knowledge actionability framework, there are two sets of interestingness measures that need to be developed in actionable knowledge discovery. For instance, we say a mined association trading rule is (technically) interesting because it satisfies requests on *support* and *confidence*. Moreover, if it also beats the expectation of user-specified *market index return* then it is a generally actionable rule.

2. Case Study and Evaluation

In this section, due to limitations in space, we only focus on business interestingness related measures and evaluation. We introduce real-world mining applications that highlight the involvement and development of business expectations, $biz\_int()$. The measures we develop reflect not only the objective business concerns of identified patterns, but also the likelihood of their impacts on businesses.

In a social security network, a customer activity or activity sequence may be associated with governmental debts or non-debts [19]. Activity mining [18] can discover those activities that will likely result in debts, which can greatly support and improve risk-based decision making in governmental customer contacts, debt prevention and process optimization. For instance, the following data describes a set of debt-targeted activities, where letters A to Z represent different activities and \$ indicates the occurrence of a debt.

$< (DABACEKB\$), (AFQCPLSWBTC\$), (PTSLD\$), (QWRTE\$),$
$(ARCZBHY) \ldots >$

To support business-oriented evaluation, it is essential to build up effective business interestingness measures to quantify to what extent an activity or activity sequence leads to debt. Activity impact metrics also provide a means to assess the business interestingness of an identified activity pattern. The idea of measuring social security activity im-

pact is to build up quantitative measures in terms of debt statistics and the relationship between activity patterns and debt. Debt statistics describe the statistical features of a debt-targeted activity pattern. For instance, a frequent activity pattern $\{ACB \rightarrow \$\}$ can be mined from the above activity set. Suppose the total number of itemsets in this data set is $|\sum|$, where the frequency of pattern e is $|e|$, then we define debt statistics in terms of the following aspects.

Definition 6. The total debt amount $d\_amt()$ is the sum of all individual debt amounts $d\_amt_i(i = 1, \ldots, f)$ in f itemsets matching the pattern *ACB*. Then we get the *average debt amount* for the pattern *ACB*:

$$\overline{d\_amt}() = \frac{\sum_1^f d\_amt()_i}{f} \tag{2}$$

Definition 7. Debt duration $d\_dur()$ for pattern *ACB* is the *average duration* of all individual debt durations in f itemsets matching *ACB*. The debt duration $d\_dur()$ of an activity is the number of days a debt remains valid, $d\_dur() = d.end\_date - d.start\_date + 1$, where *d.end_date* is the date a debt is completed, *d.start_date* is the date a debt is activated. A pattern's *average debt duration* $\overline{d\_dur}()$ is defined as:

$$\overline{d\_dur}() = \frac{\sum_1^f d\_dur()_i}{f} \tag{3}$$

Furthermore, we can development risk ratios to measure to what extent a pattern may lead to debt.

Definition 8. A pattern's *debt amount risk*, $risk_{amt}$, is the ratio of the total debt amount of activity itemsets containing e to the total debt amount of all itemsets in the data set, denoted as $risk(e \rightarrow \$)_{amt} \in 0, 1$. The larger its value is, the higher the risk of leading to debt.

$$risk(e \rightarrow \$)_{amt} = \frac{\sum_1^{|ACB|} d\_amt()_i}{\sum_1^{|\sum|} d\_amt()_i} \tag{4}$$

Definition 9. A pattern's *debt duration risk*, $risk_{dur}$, is the ratio of the total debt duration of activity itemsets containing e to the total debt duration of all itemsets in the data set, denoted as $risk(e \rightarrow \$)_{dur}$. Similarly to the debt amount support measure, $risk(e \rightarrow \$)_{dur} \in 0, 1$, and the larger its value, the higher the risk of leading to debt.

$$risk(e \rightarrow \$)_{dur} = \frac{\sum_1^{|e|} d\_dur()_i}{\sum_1^{|\sum|} d\_dur()_i} \tag{5}$$

Table 2. Frequent debt-targeted activity patterns in an unbalanced activity set

Frequent sequences	SUP	CONF	LIFT	ZSCORE	$\overline{d\_amt}$ (cents)	$\overline{d\_dur}$ (days)	$risk_{amt}$	$risk_{dur}$
$C, R \rightarrow \$$	0.0011	0.7040	19.4	92.1	22074	1.7	0.034	0.007
$I, C \rightarrow \$$	0.0011	0.6222	17.1	87.9	22872	1.8	0.037	0.008
$C, D \rightarrow \$$	0.0125	0.6229	17.1	293.7	23784	1.2	0.424	0.058

We tested the above business measures in mining activity patterns in the Australian social security debt-related activity data from 1^{st} Jan to 31^{st} Mar 2006. The data involves four data sources, which are activity files recording activity details, debt files logging debt details, customer files recording customer profiles, and earnings files storing earnings details. Our experiments analyzed activities related to both income and non-income related earnings and debts. To analyze the relationship between activity and debt, the data from activity files and debt files were extracted. We extracted activity data including 15,932,832 activity records recording government-customer contacts with 495,891 customers, which lead to 30,546 debts in the first three months of 2006.

Table 2 illustrates three frequent activity patterns discovered from the above unbalanced activity dataset (Labels C, R, I, D are activity codes that used in business field, \$ means governmental debt. Frequent activity sequential pattern "$C, R \rightarrow \$$" indicates that if a customer undertakes activity C then R, it is likely that he/she will result in a governmental debt). In the table, "SUP", "$CONF$", "$LIFT$" and "$Z-SCORE$" stand for support, confidence, lift and $z-score$ of the rule. These three rules have high confidence and lift but low support. Interestingly the impact on debt of the first two rules is not as big as the impact of the third, which has the highest risk of leading to longer average duration and debt amount.

3. Towards Domain-Driven, Actionable Knowledge Discovery

The gap between academic research and real-world development in data mining cannot be bridged very easily due to many complicated factors in respective areas. However, as one of the grand challenges and focuses of next-generation KDD, actionable knowledge discovery is facing a promising prospect through emerging reseach on knowledge actionability and domain driven data mining. In this section, we briefly discuss the potential of aggregating business and technical interestingness, and discovering actionable knowledge based on a domain driven data mining methodology.

3.1. Aggregating Technical and Business Interestingness

The gap or conflict between the perspectives on interestingness of academia and business indicates different objectives of the two stakeholders. To fill the gap or resolve the conflict, a promising direction is to develop a hybrid interestingness measure integrating both business and technical interestingness. This can reduce the burden of requiring domain users to understand those jargons and merging the expectations of both sides into a single actionability measure. However, the potential incompatibility between technical significance and business expectation makes it difficult to set up weights for two parties

to be integrated simply. Therefore, a conventional weight-based approach [23] may not work well because it only presents a simply linear weighting of technical and business interestingness.

Rather than only considering the weights of different metrics, there may be other three directions of interest to us in combining technical and business interestingness.

Fuzzy weighting of individual interestingness measures. Fuzzily weighted hybrid interestingness measures can be developed and then tested on knowledge discovered in financial applications.

Multi-objective optimization. Taking different interestingness measures as multiple objectives, we then view the discovery of actionable knowledge as a process of multi-objective optimization.

Fuzzy weighting of patterns. This consists of mining and evaluating patterns in terms of technical and business interestingness separately. We then develop fuzzy aggregation mechanism to combine these two sets of patterns.

In the following, we briefly introduce the idea of developing fuzzy interestingness aggregation and ranking methods to combine $tech\_int()$ and $biz\_int()$ and re-rank the mined pattern set to balance its adjustment to both sides.

First, patterns are mined in a given data set using the same models with different evaluation criteria. In the technical experiments, patterns are selected purely based on technical interestingness. The identified patterns are then re-ranked by checking the satisfaction of business expectations by business users.

Second, both groups of patterns are fuzzified in terms of fuzzy systems defined by the data miners and business analysts respectively. The extracted patterns are then fuzzified into two sets of fuzzily ranked pattern sets in terms of fitness functions $\overline{tech\_int}()$ and $\overline{biz\_int}()$, respectively.

Third, we then aggregate these two fuzzy pattern sets to generate a final fuzzy ranking through developing fuzzy aggregation and ranking algorithms [24].

This final fuzzily ranked pattern set is recommended to users for their consideration. Although this strategy is a little bit fuzzy, it combines two perspectives of interestingness and balances individual contributions and diversities.

3.2. Towards Domain Driven Data Mining

Recently, research on theoretical frameworks of actionable knowledge discovery has emerged as a new trend. [25] proposed a high-level microeconomic framework regarding data mining as an optimization of decision "utility". [7] built a product recommender which maximizes net profit. In [26], action rules are mined by distinguishing all stable attributes from some flexible ones. Additional work includes enhancing the actionability of pattern mining in traditional data mining techniques such as association rules [6], multi-objective optimization in data mining [23], role model-based actionable pattern mining [27], cost-sensitive learning [28] and postprocessing [29], etc.

A more thorough and fundamental direction is to develop a practical data mining methodology for real-world actionable knowledge discovery, i.e., domain driven data mining. Contrasting a data-driven perspective, [8,2,9,10,11] proposed a domain-driven data mining methodology which highlights actionable knowledge discovery through involving domain knowledge, human cooperation and reflecting business constraints and

expectations. The motivation of domain driven data mining is to complement and enhance existing data mining methodologies and techniques through considering and involving real-world challenges. The fundamental idea of domain driven data mining is the paradigm shift from data-centered hidden pattern mining to domain-driven actionable knowledge discovery. Fundamentally, we highlight the significant involvement, development, support and meta-synthesis of the following four types of intelligence [30].

In-depth data intelligence. This is to let data tell stories about a business problem, rather than just interesting patterns. For instance, in financial data mining, in-depth rules from general trading patterns may disclose more workable knowledge in stock market data.

Human intelligence. This is to study how to involve human roles and knowledge in actionable knowledge discovery. For instance, studies can be on the dynamic involvement of humans and human knowledge in dynamic mining.

Domain intelligence. This is to study how domain knowledge and environment can be involved to enhance knowledge actionability. For instance, system support is studied on domain-specific organizational constraints and expectations in discovering actionable patterns.

Web intelligence. This is to study how Web resources and support can be utilized to strengthen actionable knowledge discovery.

Intelligence meta-synthesis. Finally, it is critical for us to synthesize all above intelligence into an integrative actionable knowledge discovery system. This involves the development of an appropriate infrastructure, an operational process, a communication language, knowledge repsentation and mapping, etc.

We believe the research on domain-driven actionable knowledge discovery can provide concrete and practical guidelines and hints for the corresponding theoretical research in a general manner.

4. Conclusions

Actionable knowledge discovery is widely recognized as one of major challenges and prospects of next-generation KDD research and development. With increasing number of enterprise data mining applications, the progress in this area may greatly benefit enterprise operational decision making. On the other hand, it is obvious that it is not a trivial task to identify knowledge of interest to business expectations. A typical problem is the involvement and handling of business interestingness in actionable knowledge discovery. This involves the representation and modeling of business interestingness, as well as the balance of technical significance and business expectations.

In this paper, we analyze the evolution of interestingness research and development in data mining. In particular, we highlight the significant phase of considering both technical and business interestingness from both objective and subjective perspectives. This phase addresses the requirements of actionable knowledge discovery and provides a framework for it. We illustrate how business expectations can be modelled through real-world applications of discovering activity patterns in social security data.

Mining knowledge that can satisfy user needs and support users to take actions to their advantage is not an easy task. Following the proposed high-level knowledge action-

ability framework, we believe the following efforts are promising for actionable knowledge discovery: (i) developing a domain-oriented general business interestingness framework, such as, cost-benefit or profit-risk metrics, (ii) developing synthesis frameworks for combining and balancing multiple technical and business interestingness objectives.

References

[1] Gur Ali, O.F., Wallace, W.A.: Bridging the gap between business objectives and parameters of data mining algorithms. Decision Support Systems, 21, 3-15 (1997)

[2] Cao, L., Zhang, C.: Domain-driven data mining—a practical methodology. Int. J. of Data Warehousing and Mining, 2(4): 49-65 (2006)

[3] Ghani, R., Soares, C.: Proc. of the KDD 2006 Workshop on data mining for business applications (2006)

[4] Silberschatz, A., Tuzhilin, A.: What makes patterns interesting in knowledge discovery systems. IEEE Transactions on Knowledge and Data Engineering, 8(6):970-974 (1996)

[5] Zhang, D. and Zhou, L.: Discovering golden nuggets: data mining in financial application, IEEE Transactions on SMC Part C, 34(4):513-522 (2004).

[6] Liu, B., W. Hsu, S. Chen, and Y. Ma: Analyzing Subjective Interestingness of Association Rules. IEEE Intelligent Systems, 15(5): 47-55 (2000)

[7] Wang, K., Zhou, S. and Han, J.: Profit Mining: From Patterns to Actions. EBDT2002

[8] Cao, L., Zhang, C.: Domain-driven actionable knowledge discovery in the real world. PAKDD2006, LNAI 3918, 821-830, Springer (2006)

[9] Cao, L. et al. Domain-Driven, Actionable Knowledge Discovery, IEEE Intelligent Systems, 22 (4):78-89, 2007.

[10] Cao, L., Luo, D., Zhang C.: Knowledge Actionability: Satisfying Technical and Business Interestingness, International Journal of Business Intelligence and Data Mining (IJBIDM), 2007.

[11] Cao, L., Zhang, C.: The evolution of KDD: Towards domain-driven data mining. International Journal of Pattern Recognition and Artificial Intelligence, 21(4): 677-692 (2007).

[12] CRISP-DM: www.crisp-dm.org.

[13] Freitas, A.: On objective measures of rule surprisingness. In J. Zytkow and M. Quafafou, editors, PKDD'98, 1-9 (1998)

[14] Hilderman, R.J., and Hamilton, H.J.: Applying objective interestingness measures in data mining systems. PKDD'00, 432-439 (2000)

[15] Silberschatz, A., Tuzhilin, A.: On Subjective Measures of Interestingness in Knowledge discovery. Knowledge Discovery and Data Mining, 275-281 (1995)

[16] Cao, L., Luo, C., Zhang, C.: Developing actionable trading strategies for trading agents, IAT2007, IEEE Computer Science Press (2007).

[17] Cao, L.: Multi-strategy integration for actionable trading agents, IAT2007, IEEE Computer Science Press (2007).

[18] Cao, L.: Activity mining: challenges and prospects. ADMA2006, 582-593, LNAI4093, Springer (2006)

[19] Cao, L., Zhao, Y., Zhang, C.: Mining impact-targeted activity patterns in unbalanced data. Technical report, University of Technology Sydney (2006)

[20] Cao, L., Zhao, Y., Zhang, C., Zhang, H.: Activity Mining: from Activities to Actions, International Journal of Information Technology & Decision Making, 7(2) (2008).

[21] Padmanabhan B. and Tuzhilin A.: Unexpectedness as a measure of interestingness in knowledge discovery, Decision and Support Systems 27: 303-318, 1999

[22] Padmanabhan, B., and Tuzhilin, A.: A belief-driven method for discovering unexpected patterns. KDD-98, 94-100

[23] Freitas, A.: A Critical Review of Multi-Objective Optimization in Data Mining– A Position Paper. SIGKDD Explorations, 6(2): 77-86 (2004)

[24] Cao, L., Luo, D., Zhang, C.: Fuzzy Genetic Algorithms for Pairs Mining. PRICAI2006, LNAI 4099, 711-720 (2006)

[25] Kleinberg, J. Papadimitriou, C., and Raghavan, P.: A Microeconomic View of Data Mining. Journal of Data Mining and Knowledge Discovery (1998)

[26] Tzacheva, A., Ras W.: Action rules mining. Int. J. of Intelligent Systems. 20: 719-736 (2005)

[27] Wang, K., Jiang, Y., Tuzhilin, A.: Mining Actionable Patterns by Role Models. ICDE 2006
[28] Domingos, P.: MetaCost: a general method for making classifiers cost-sensitive. KDD1999
[29] Yang, Q., Yin, J., Lin, C., Chen, T.: Postprocessing Decision Trees to Extract Actionable Knowledge. ICDM2003
[30] Cao, L., Zhang, C., Luo, D., Dai, R., Chew, E.: Intelligence Metasynthesis in Building Business Intelligence Systems. Int. Workshop on Web Intelligence meets Business Intelligence, Springer (2006)

This page intentionally left blank

Applications of Data Mining in E-Business and Finance
C. Soares et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-890-8-111

A Deterministic Crowding Evolutionary Algorithm for Optimization of a KNN-based Anomaly Intrusion Detection System

F. de TORO-NEGRO, P. GARCÌA-TEODORO, J.E. DIÁZ-VERDEJO, G.MACIÁ-FERNANDEZ.
Signal Theory, Telematics and Communications Department,
University of Granada, Spain

Abstract. This paper addresses the use of an evolutionary algorithm for the optimization of a K-nearest neighbor classifier to be used in the implementation of an intrusion detection system. The inclusion of a diversity maintenance technique embodied in the design of the evolutionary algorithm enables us to obtain different subsets of features extracted from network traffic data that lead to high classification accuracies. The methodology has been preliminarily applied to the Denial of Service attack detection, a key issue in maintaining continuity of the services provided by business organizations.

Key words. evolutionary algorithm, machine learning, intrusion detection.

Introduction

With the increased complexity of security threats, such as malicious Internet worms, denial of service (DoS) attacks, and e-business application attacks, achieving efficient network intrusion security is critical to maintaining a high level of protection. The efficient design of intrusion detection systems (IDS) is essential for safeguarding organizations from costly and debilitating network breaches and for helping to ensure business continuity. An IDS is a program that analyzes what happens or has happened in a computer network and tries to find indications that the computer has been misused. An IDS will typically monitor network traffic data passing through the network in order to generate an alert when an attack event is taking place. On the other hand, two different kinds of detection schemes can be applied to detect attacks in the data being monitored. Signature-based detection systems try to find attack signatures in the data monitored. Anomaly detection systems rely on the knowledge of what should be the normal behaviour of the monitored data to flag any deviation of the normal behaviour as an attack event, so they need a model of what normal behaviour is. Machine learning algorithms [1] can be trained with labelled network traffic data so that it can classify unknown traffic data captured from a computer

network with a certain accuracy. One type of these algorithms would consist of a binary classifier enabling us to separate two different groups of observations: normal traffic and malicious traffic (containing some kind of attack).

One way to proceed in a signature-based intrusion detection system is to process network traffic data into data flows and use some features from the data flows to signature the current network traffic data so an adequate response may be triggered when an attack is occurring (see Fig. 1). Finding the features that lead to a good classification performance is a major issue of such intrusion detection systems.

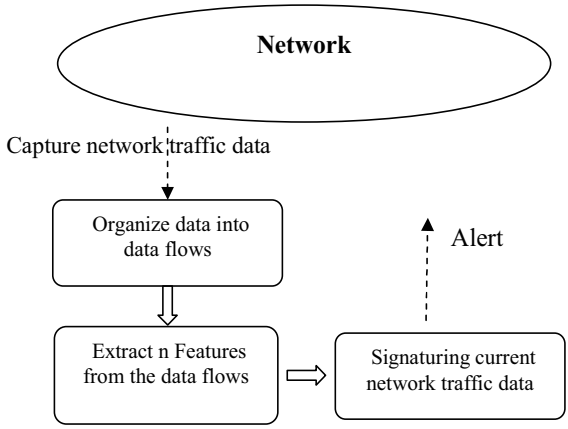

Fig. 1 A signature-based Intrusion Detection System

Besides maximizing the performance, the so-called feature selection stage [2] aims to simplify the complexity and improve the accuracy of the classification system eliminating irrelevant features.

Evolutionary algorithms (EAs) [3,4] are stochastic optimization procedures which apply a transformation process (crossover and mutation operators), inspired by the species natural evolution, to a set (population) of coded solutions (individuals) to the problem. These procedures have revealed great success in dealing with optimization problems with several solutions [5,6] due to their special ability to explore large search spaces and capture multiple solutions in a single run.

In this context, the present work investigates the use of EAs as a new alternative signature-based IDS approach to those reported in the literature. For that, EAs will be used to train a supervised machine learning algorithm (two constituents of what is known as "soft-computing") consisting of a binary K-Nearest Neighbour (KNN) classifier. The Evolutionary Algorithm used is based on deterministic crowding [9]. After the training process, which can be seen as a multisolution optimization problem, information is retrieved about the features leading to good classification performance when using a KNN binary classifier.

This work is organized as follows: in Section 2, the overall methodology proposed in this paper is described. Then, Section 3 shows some experimental work

carried out by using labelled network traffic data provided by DARPA [7]. Finally, Section 4 is devoted to discussing the main findings and conclusions.

1. Materials and Methods

The general framework proposed in this work is depicted in Fig. 2. First of all, network data (IP packets) are organized into data flows. Then, n features – previously defined- characterizing the data flows are extracted to obtain an n-dimensional feature vector representing the network traffic in a given time window. A binary classifier is trained to classify each feature vector as belonging to normal traffic or malicious traffic. The training of the classifier is thus formulated as an optimization problem addressed by an evolutionary algorithm. The optimization problem consists of maximizing the performance classification of the binary classifier. A candidate solution to the problem is an n-dimensional weight vector. Each component of the weight vector is a real number ranging between 0 and 1 and represents the degree of importance of each feature in the classification algorithm. The fitness of a given weight vector – necessary to guide the search of the evolutionary algorithm – is the performance classification obtained by the binary classifier when the feature vectors are multiplied by the said weight vector. The performance classification is obtained by comparing the real labels of the feature vectors to the decision label computed by the binary classifier.

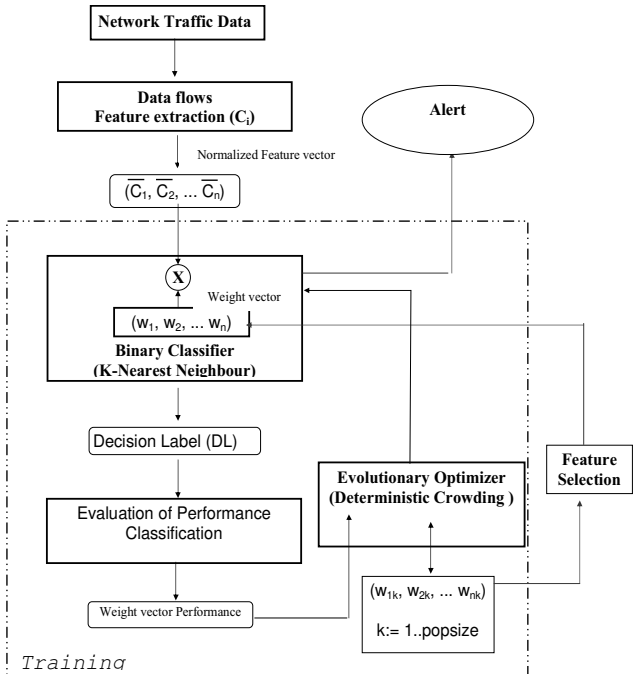

Fig. 2. Methodology for training a KNN algorithm for network traffic data classification by using an evolutionary algorithm for feature selection.

Due to the fact that the EA works with a population of k individuals (see Fig. 2) candidate solutions, different choices of weight vectors can be explored in a single iteration of the algorithm. If the necessary diversity maintenance mechanism [8] is incorporated into the EAs, the solutions found will be geometrically different from each other, providing flexibility to select the features to be considered in the intrusion detection system. This is of great importance due to the fact that the extraction of some features can be less time-consuming than others (i.e. they can be computed by using a smaller time window or they have less complex extraction).

A K-nearest neighbour (KNN) algorithm is chosen as binary classifier in this work due to the simplicity of its implementation. A KNN algorithm computes the class (decision label) an observation belongs to as the class of the majority of the k nearest labelled observations of a reference model. Thus, Euclidean distances in the feature space from the observation to all the members of the reference model have to be previously computed. Due to the fact that a KNN algorithm is a distance-based method, the feature vectors are normalized to have a value ranging between 0 and 1 by dividing each original feature value by its variance in the reference model.

A Deterministic Crowding [9, 10] technique has been embodied in the design of the evolutionary algorithm implemented in this work in order to improve the chances of obtaining diversified solutions from this natural multimodal optimisation problem. Deterministic crowding [9] (see Fig. 3) is chosen as diversity maintenance technique for two main reasons: (1) it shows a good performance in several comparative studies with some other methods [8, 10]; (2) In contrast with other techniques like Clearing [11] or fitness sharing [4], there is no need to determine any user parameter. The distance metric in step 08 (Fig.3) is defined in the weight vector search space (genotype distance) to encourage dissimilarity between features contained in the solutions. The algorithm used for training the KNN classifier is shown in Fig. 3.

Deterministic_Crowding Procedure

01 Create randomly Population P of Candidate Solutions (individuals) of Size *Popsize*
02 **While** (stop_condition) FALSE
03 $P^* = \varnothing$
04 **While** (Sizeof(P^*) $\neq$ *Popsize*)
05 Select two individuals p_1 and p_2 from P (without replacement)
06 Crossover p_1 and p_2 to obtain h_1 and h_2
07 Mutate h_1 and h_2 to obtain c_1 and c_2 (with mutation probability rate *pmut*)
08 **If**
$[\text{Distance}(p_1,c_1)+\text{Distance}(p_2,c_2)] \leq [\text{Distance}(p_1,c_2)+\text{Distance}(p_2,c_1)]$
09 **If** c_1 is better than p_1 **then** $P^* = P^* \cup \{c_1\}$ **else** $P^* = P^* \cup \{p_1\}$
10 **If** c_2 is better than p_2 **then** $P^* = P^* \cup \{c_2\}$ **else** $P^* = P^* \cup \{p_2\}$
Else
11 **If** c_1 is better than p_2 **then** $P^* = P^* \cup \{c_1\}$ **else** $P^* = P^* \cup \{p_2\}$
12 **If** c_2 is better than p_1 **then** $P^* = P^* \cup \{c_2\}$ **else** $P^* = P^* \cup \{p_1\}$
EndWhile
13 $P = P^*$
14 Evaluate the Performance of each candidate solution in P using the Diagnostic Scheme (K- nearest neighbourn classifier)
EndWhile

Fig.3. Evolutionary Algorithm based on deterministic crowding used for training the classifier

2. Case Study

For the purpose of testing the aforementioned classifier methodology, a database containing information about the traffic in network that was created by DARPA [7] for training purposes has been used. This database was built with simulated network traffic data containing normal traffic data and 22 different kinds of computer attacks that fall in one of the following groups:

- DoS (Denial of Service): the attacker targets some computing or memory resource and makes it too busy or full to handle legitimate requests, or denies legitimate user access to that resource, for example SYN flood, ping of death, smurf, etc.
- R2U (Remote to User): the attacker exploits some vulnerability to gain unauthorized local access from a remote machine, for example guessing password.
- U2R (User to Root): the attacker has access to a normal user account (obtained legitimately or otherwise) and using this is able to gain root access by exploiting a vulnerability hole in the system, for example buffer overflow attacks.
- PROBE (Probing): attacker scans the network to gather information or find known vulnerabilities. An attacker with a map of machines and services that are available on a network can use this information to look for weak points, for example through port scan.

There are 41 features present in the data set. The first 9 of these are "intrinsic" features, which describe the basic features of individual TCP data flows (TCP connections), and can be obtained from raw tcpdump output. The remaining features have been constructed as described in [12, 13]. Thus, features 10 to 22 are content-based features obtained by examining the data portion (payload) of a TCP data flow and suggested by domain knowledge. Features 23 to 41 are "traffic-based" features that are computed using a window. Features 23 to 31 use a two-second time window ("time-based"), and features 32 to 41 are constructed using a window of 100 TCP connections ("host-based"). The reasons for the different windows is that the DoS and PROBE attacks were shown to involve many TCP connections in a short time frame, whereas R2U and U2R attacks are embedded in the data portions of the TCP connections and normally involve a single TCP connection.

In this work, only the detection of attacks falling in the category of DoS attacks is addressed. On the other hand, only numerical features have been considered from the original feature set. In this way, only 6 out of the 9 "intrinsic" features have been used in this work (see Tables 1 and 2 for a complete list of features considered). As part of the pre-processing of the database, duplicate data were withdrawn. Training and evaluation were performed by a random selection of 500 feature vectors labelled as normal traffic, and 500 feature vectors classified as DoS attacks.

Table 1. Subset of DARPA database features used in this work (part I).

#	Name of Feature	Description
1	Duration	Duration of the connection
2	Source bytes	Bytes sent from source to destination
3	Destination bytes	Bytes sent from destination to source
4	Land	1 if connection is from/to the same host/port; 0 otherwise
5	Wrong fragment	Number of wrong fragment
6	Urgent	Number of urgent packets
7	Hot	Number of “hot” indicators
8	Failed logins	Number of failed logins
9	Logged in	1 if successfully logged in; 0 otherwise
10	# compromised	Number of “compromised” conditions
11	Root shell	1 if root shell is obtained; 0 otherwise
12	Su attempted	1 if “su root” command attempted; 0 otherwise
13	#root	Number of “root” accesses
14	#file creations	Number of file creations operations
15	#shells	Number of shells prompts
16	#access files	Number of operations on access control files
17	#outbound cmds	Number of outbounds commands in an ftp session
18	hot login	1 if the login belongs to the “hot” list; 0 otherwise
19	guest login	1 if the login is a “guest” login; 0 otherwise
20	Count	Number of connections to the same host as the current connection in the past two seconds
21	Srv count	Number of connections to the same host as the current connection in the past two seconds
22	Serror rate	% of connections that have SYN errors
23	Srv serror rate	% of connections that have SYN errors
24	Rerror rate	% of connections that have REJ errors
25	Srv rerror rate	% of connections that have REJ errors
26	Same srv rate	% of connections to the same service

Table 2. Subset of DARPA database features used in this work (part II).

#	Name of Feature	Description
27	diff srv rate	% of connections to the different services
28	srv diff host rate	% of connections to different hosts
29	dst host count	Count of connections having the same destination host
30	dst host srv count	Count of connections having the same destination host and using the same service
31	dst host same srv rate	% of connections having the same destination host and using the same service
32	dst host diff srv rate	% of different services on the current host
33	dst host same src port rate	% of connections to the current host having the same source port
34	dst host srv diff host rate	% of connections to the same service coming from different hosts
35	dst host serror rate	% of connections to the current host that have S0 error
36	dst host srv serror rate	% of connections to the current host and specified service that have an S0 error
37	dst host rerror rate	% of connections to the current host that have an RST error.
38	dst host srv rerror rate	% of connections to the current host and specified service that have an RST error.

The classification result can fall into one of the following cases: (1) The algorithm classifies the traffic as malicious and the traffic is in fact malicious (True Positive, TP); (2) The algorithm classifies the traffic as normal and the traffic is in fact normal (True Negative, TN); (3) The algorithm classifies the traffic as malicious but the traffic is normal (False Positive, FP); (4) The algorithm classifies the traffic as normal but the traffic is malicious (False Negative, FN).

Based on these cases, two different functions have been considered as classification performance indicators:

- Classification Accuracy (C): represents the ratio between the values of correctly classified and overall traffic.

$$C = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- Sensitivity (S): represents the ratio between the values of detected and total malicious traffic.

$$S = \frac{TP}{TP + FN} \tag{2}$$

The value of the measures C and S is calculated by applying the leave-one-out method [14]. In each cycle, one vector is selected from the database as the test element. This vector is classified with the rest of the individuals in the population serving as classification references. Then the following four counters are updated: True_Positive (TP), True_Negative (TN), False_Positive (FP) and False_Negative (FN). Finally C and S are calculated following Eq. (1) and Eq. (2). It is important to mention that only Classification Accuracy is taken into account to guide the search of the evolutionary algorithm.

The crossover operator used in the deterministic crowding procedure is a single-point real-coded operator. Two different types of mutation have been considered: uniform mutation – Eq. (3) – and Gaussian mutation – Eq. (4) –. Both are computed respectively as:

$$w_j^{'} = U(0,1) \tag{3}$$

$$\begin{aligned} \mu &= w_j \\ \sigma &= \sqrt{\min\left[(1-w_j), w_j\right]} \\ w_j^{'} &= N(\mu, \sigma) \end{aligned} \tag{4}$$

Then, the new weight components are obtained by applying mutation using a uniform distribution function and a normal distribution function, respectively. Each type of mutation is used with a probability of 0.5. The mutation rate probability has been set to 0.6.

The Evolutionary Algorithm ran for 400 iterations with 50 candidate solutions (individuals) with classifications accuracies ranging between 95% and 99%. Four of the found solutions (weight vectors) are shown in Table 3.

These level of high classification accuracies have also been obtained by other authors addressing Denial of Service attack detection in DARPA database. In this sense, [15] is an excellent review of the different machine learning algorithms applied to the database used in this work.

According to this study, the best performance (97.85% of classification accuracy) for Denial of Service computer attack detection is reached by using a k-means clustering [16]. Nevertheless, other algorithms reach similar performance (96.9% by using a classifier based on rules [17] and 97.5% by using a classifier based on decision-trees [18]). As can be seen, the algorithm used in this work obtains similar results. However, it has the important advantage of obtaining different subsets of weighted features to be used in the classifier. In this way, as an example, feature #12 is weighted nearly 0 in solution #1 but its weight is nearly 1 in solution #2 (both of them leading to similar classification accuracies and sensibilities). On the other hand, solutions with higher sensitivity are better than others with lower sensitivity. The retrieval of several solutions during the optimization of the classifier enables us to give a certain degree of flexibility for choosing the features to be used in the detection (for example those of easier extraction from the network traffic data). To the best of

the authors' knowledge, this approach to the design of intrusion detection systems has not been addressed in previous specialized work dealing with the use of machine learning algorithms for computer attack detection applications.

Table 3. Classification accuracy and Sensibility of five of the found solutions (weight vectors) to the problem of DARPA Denial of Service attack detection

#Feature	*Solution #1*	*Solution #2*	*Solution #3*	*Solution #4*
1	0.02	0.80	0.37	0.01
2	0.73	0.00	0.90	0.75
3	0.21	0.75	0.14	0.20
4	0.07	0.06	0.29	0.06
5	0.83	0.96	1.00	0.82
6	0.33	0.85	0.87	0.24
7	0.92	0.46	0.47	0.92
8	0.99	0.18	0.18	0.99
9	0.47	0.40	0.75	0.25
10	0.40	0.10	0.47	0.40
11	0.87	1.00	0.22	0.87
12	0.98	0.04	0.02	0.98
13	0.42	0.63	0.95	0.42
14	0.75	0.63	0.21	0.96
15	0.60	0.23	0.95	0.99
16	0.94	0.91	0.25	0.59
17	0.46	0.25	0.72	0.91
18	0.99	0.48	0.62	0.50
19	0.61	0.75	0.54	0.12
20	0.90	0.56	0.58	0.86
21	0.52	0.03	0.91	0.75
22	0.30	0.60	0.46	0.74
23	0.75	0.91	0.52	0.50
24	0.45	0.66	0.03	0.05
25	0.75	0.29	0.93	0.62
26	0.15	0.75	0.95	0.16
27	0.25	0.76	0.80	0.68
28	0.19	0.09	0.08	0.90
29	0.64	0.08	0.33	0.58
30	0.45	0.80	0.49	0.89
31	0.75	0.25	0.59	0.43
32	0.75	0.53	0.99	0.95
33	0.12	0.10	0.58	0.29
34	0.90	0.04	0.61	0.76
35	0.42	0.02	0.29	0.78
36	0.50	0.73	0.51	0.20
37	0.42	0.62	0.94	0.29
38	0.98	0.89	0.41	0.93
C (%)	*97.2*	*97.3*	*97.2*	*94.4*
S (%)	*98.1*	*98.4*	*98.1*	*99.0*

3. Summary

This work addresses the problem of attack detection by using a new methodology consisting of a K-Nearest Neighbour binary classifier which produces a decision label (malicious traffic or normal traffic) by processing a feature vector representing the network traffic during a given window time. The features are automatically weighted

by using an evolutionary algorithm in order to optimize the performance of the KNN Classifier. The retrieval of more than one solution during the optimization process gives a certain degree of flexibility to choose the features to be used in the detection process. The methodology has been preliminarily applied to the Denial of Service attack detection problem contained in the DARPA database and has been validated by using the leave-one-out method. As future work, the authors intend to explore the performance of this methodology on other network traffic databases and sets of features.

References

[1] Maheshkumar Sabhnani, Gürsel Serpen: Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context, Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications, pp. 209-215, Las Vegas, 2003.

[2] Liu, H. and Motoda, H., Feature Selection for Knowledge Discovery and Data Mining, Kluwer, 1998.

[3] A. E. Eiben and J.E. Smith, Introduction to Evolutionary Computing, Natural Computing Series, Springer, 2003.

[4] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning. New York: Addison Wesley, 1989..

[5] J.P. Li, M.E Balazs, G.T. Parks, P.J. Clarkson, A species conserving genetic algorithm for multimodal function optimisation, Evolutionary Computation, Vol. 10, No. 3 (2002) 207-234.

[6] F.de Toro, J.Ortega, E.Ros, S.Mota, B.Paechter; Parallel Processing and Evolutionary Algorithms for Multiobjective Optimisation; Parallel Computing; Vol. 30, No. 6, pp. 721-739, 2004

[7] The UCI KDD Archive, Information and Computer Science, University of California, Irvine, "Kdd cup 1999 data set", http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[8] B. Sareni and L. Krähenbühl, Fitness Sharing and Niching Methods Revisited, IEEE Transaction on Evolutionary Computation, Vol. 2, No. 3, 1998.

[9] S.W. Mahfoud, Crowding and Preselection Revised, Parallel Problem Solving from Nature 2, R. Manner, B. Manderick (Eds.), Elsevier Science Publishers, Amsterdam, pp. 27-36, 1992.

[10] S.W. Mahfoud, A Comparison of Parallel and Sequential Niching Methods, Proceedings of the Sixth International Conference on Genetic Algorithms, Morgan Kauffman, San Mateo, CA, 1995.

[11] A. Petrowski. "A Clearing Procedure as a Niching Method for Genetic Algorithms". In Proc. 1996 IEEE Int. Conf. Evolutionary Computation, Nagoya, Japan, pp.798-803, 1996.

[12] S. Stolfo, W. Lee, A. Prodromidis, and P. Chan, "Cost-based modeling for fraud and intrusion detection: Results from the JAM project", in Proc. DARPA Information Survivability Conference and Exposition, 2000, Vol. II, pp. 1130-1144, IEEE Computer Press.

[13] W. Lee, S.Stolfo and K.W. Mok, "A data mining framework for building intrusion detection models", in IEEE Symposium on Security and Privacy, 1999, pp. 120-132.

[14] D. Hand, Discrimination and Classification, Wiley & Sons, New York, 1981.

[15] M. Sabhnani, G. Serpen Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context, Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications (MLMTA 2003), Las Vegas, NV, June 2003, pages 209-215.

[16] R.O. Duda, and P.E.Hart, "Pattern Classification and Scene Analysis", New York: Wiley, 1973

[17] R. Agarwal, and M.V. Joshi, "PNrule: A New Framework for Learning Classifier Models in Data Mining", Technical Report TR 00-015, Department of Computer Science, University of Minnesota, 2000

[18] I. Levin, "KDD-99 Classifier Learning Contest LLSoft's Results Overview", SIGKDD Explorations SIGKDD, January 2000, Vol. 1 (2), pp. 67-75

Applications of Data Mining in E-Business and Finance
C. Soares et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-890-8-121

Analysis of Foreign Direct Investment and Economic Development in the Yangtze Delta and Its Squeezing-in and Out Effect

Guoxin WU[1], Zhuning LI[1] and Xiujuan JIANG[2]
[1] *School of Economics & Management, Shanghai Institute of Technology*
Shanghai, 200235, China
[2] *Glorious Sun School of Business and Management, Donghua University*
Shanghai, 200051, China
Emails: gxwu@sit.edu.cn, li_zhuning@hotmail.com, xjjiang@mail.dhu.edu.cn

Abstract. In recent years, the academic circle has been paying increasing attention to the economic development in the Yangtze Delta of China, particularly its continuous development driven by growing Foreign Direct Investment (FDI). This article studies, by way of quantitative analysis, the correlation between FDI and economic development in the Yangtze Delta, and on this basis, analyzes the squeezing-in and out effect of FDI on the regional economic development, and draw some the conclusion.

Keywords. Foreign Direct Investment, Yangtze Delta, squeezing-in and out effect, Economic development.

Introduction

Foreign Direct Investment (FDI) in the Yangtze Delta shows a continual upward trend in 2007. The FDI actually absorbed and utilized in China in 2006 was USD 69.47 billion, topping the record of USD 60 billion in 2004. In the same year, the GDP in the Yangtze Delta amounted to RMB 4749.433 billion with the FDI actually utilized and introduced totalling USD33.427 billion, which was a historical breakthrough. According to statistics issued by the relevant departments, the Yangtze Delta has become a strong magnetic field in attracting foreign investment. With such a historical background, this article attempts to explore the correlation between FDI and the economic growth in the Yangtze Delta and perform an in-depth analysis on the squeezing-in and out effect of FDI on investment in this Region.

1. The Correlation Between FDI and Economic Development in the Yangtze Delata

In measuring the importance of FDI to the economic growth of the Yangtze Delta, we used the ratio between the cumulative amount of FDI and GDP (namely, the FDI/GDP ratio), the rationale being that the economic growth in a certain area is usually reflected by the growth of GDP, and that the ratio between the cumulative amount of FDI and

the GDP of the Yangtze Delta is one of the important indicators measuring the contribution of FDI absorption and utilization in the Region to its economic growth.

In order to verify the correlation between FDI and economic growth in the Yangtze Delta from the quantitative perspective, this article uses the unary linear regression method to study the relationship between the actual FDI and the GDP growth in this Region. In order to reduce errors and make this article more scientific, changes of foreign exchange rate in different periods have been considered while studying the ratio between total amount of FDI and GDP in the Region.

The GDP in the Yangtze Delta has been undergoing steady and fast increase since 1990. It was USD 64.195 billion in 1990, and increased to USD 595.652 billion in 2006, representing an increase of 827.88%. The fast increase in the total output requires a certain amount of production factors to match it, which brings about a smooth process of social reproduction. As an integral part of the total amount of production factors, FDI promotes economic growth through the national economic cycle system in the Yangtze Delta. In 1990, the amount of actually absorbed FDI was USD 366 million in the Region, and it increased to USD 33.427 billion in 2006, representing more than a 90-fold growth (see Table 1).

Table 1. Relationship between Cumulative Actual FDI and GDP in the Yangtze Delta. Source of data: Relative statistics from the Almanac of Statistics of Shanghai, the Almanac of Statistics of Jiangsu Province and the Almanac of Statistics of Zhejiang Province in different years. FDI is the cumulative amount since 1990. GDP is converted into US dollars as per the average mean of foreign exchange rate between RMB and USD.[2][3][4][5]

Unit: hundred million USD

Item/Year	cumulative actual FDI	GDP in the Yangtze Delta
1990	3.66	641.95
1991	8.67	672.01
1992	37.52	836.90
1993	101.04	1114.09
1994	181.48	1008.99
1995	274.37	1334.33
1996	387.45	1569.90
1997	508.49	1770.70
1998	624.57	1917.56
1999	731.36	2065.04
2000	844.37	2312.18
2001	98.163	2555.98
2002	1167.19	2861.62
2003	1438.21	3393.42
2004	1691.80	4070.19
2005	1969.33	4940.19
2006	2303.60	5956.52

X refers to the cumulative amount of FDI, and variable Y refers to the GDP. The empirical formula between X and Y is established as per the statistics shown in Table 1. The scatter chart (Figure 1) shows that the distribution is approximately a straight line, which indicates that a linear relationship exists between the variables.

The regression equation is $y_c = a + bx$, where, a and b are the parameters to be determined and y_c is the estimate of y. Using the least squares method, we have $a = 645.08$ and $b = 2.12$.

The correlation between variable x and y and the reliability of the regression equation can be tested with the coefficient of determination R^2[1], which can be obtained through calculation as $R^2=0.9855$. This means that the cumulative amount of FDI can explain 98.55% of the variation of the GDP. The model selected gives a fairly good fit to the actual data. Meanwhile, the F value is 1017.46, much higher than the critical value, and the linear relationship of the model is significant at the 95% level. The regression analysis results are shown in Table 2.

Since the DW① statistic 0.64, which is below the critical lower-bound value 1.13, the model has first-order serial correlation②. The Lagrange multiplier test indicates that it also has higher-order serial correlation③. The White Heteroscedasticity Test results are as follows (see Table 3).

Its adjoint probability is 0.01<0.05, so the null hypothesis of variance homoscedasticity should be rejected, that is, the model has heteroscedasticity. In order to eliminate the serial correlation, the Cochrane-Orcutt Iterative Method was employed to conduct general differential conversion. The adjusted regression result is $y_c = 257.43 + 2.47x$. The test results indicate the elimination of serial correlation, and the White test shows no heteroscedasticity.

---

① One of the basic hypotheses of the classical linear regression model is that the residual u_t of PRF does not have serial autocorrelation. The DW test was developed by J. Durbin and G.S. Watson in 1951to test serial autocorrelation of residuals. It is only suitable for testing whether the residuals have first-order serial correlation.

$$D.W. = \frac{\sum_{t=2}^{n} (\tilde{e}_t - \tilde{e}_{t-1})^2}{\sum_{t=1}^{n} \tilde{e}_t^2}$$

The test criteria is: if $0<DW<d_L$, there exist positive autocorrelation; if $d_L<DW<d_U$, it cannot be determined if there is correlation； if $d_U<DW<4-d_U$， there is no autocorrelation;
if $4-d_U<DW<4-d_L$, it cannot be determined if there is autocorrelation; if $4-d_L<DW<4$, there exist negative autocorrelation。d_L and d_U are the critical values found in the DW distribution table according to n and k and the set α。

② When the residual u_t is only correlated with the value at time $t-1$,

$$E(u_t u_{t-1}) \neq 0 \qquad t=2, \ldots, n$$

the correlation is called first-order serial correlation.

③ When the value of u_t is not only correlated with the value at time $t-1$, but also correlated with the value at time $t-1$, $t-2$, …,

$$u_t = f(u_{t-1}, u_{t-2}, \ldots)$$

the correlation is called higher-order serial correlation.

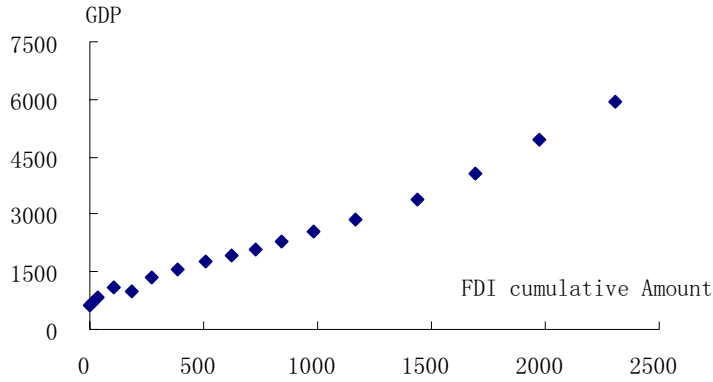

Figure 1. Scatter Chart of GDP and cumulative amount of FDI in the Yangtze River Delta.

Its adjoint probability is 0.01<0.05, so the null hypothesis of variance homoscedasticity should be rejected, that is, the model has heteroscedasticity. In order to eliminate the serial correlation, the Cochrane-Orcutt Iterative Method was employed to conduct general differential conversion. The adjusted regression result is $y_c = 257.43 + 2.47x$. The test results indicate the elimination of serial correlation, and the White test shows no heteroscedasticity.

Table 2. Results of regression analysis.

Constant and explanatory variable	Estimate of parameter	Standard error of the parameters	t statistics	Sig. (two-tailed)
C	645.0769	69.42580	9.291602	0.0000
X	2.116678	0.066358	31.89764	0.0000
Coefficient of Determination	0.985472	Mean of explained variable		2295.431
Adjusted coefficient of determination	0.984503	Standard deviation of explained variable		1533.251
Standard deviation of the regression equation	190.8696	Chichi information principle		13.45119
Residual sum of squares	546467.8	Shiwa information principle		13.54921
Logarithm of the likelihood function	-112.3351	F statistic		1017.46
DW statistic	0.638864	Probability of F statistic		0.000000

Table 3. Results of White heteroscedasticity test.

F-statistic	4.008951	Probability	0.046406
Obs*R-squared	6.008049	Probability	0.049587

It can be shown from the above analysis that the linear relation between x and y is significant, therefore, there is certain reliability in using the regression equation to make predictions, that is, there is an evident linear relation between the cumulative amount of FDI and the GDP of the Yangtze Delta, which means that there is a fairly close correlation between the FDI and GDP growth in the Yangtze Delta. Therefore, we can say that the FDI has promoted economic growth in the Yangtze Delta.

2. Analysis on the Squeezing In and Out Effect of FDI on Investment in the Yangtze Delta

The above analysis shows that FDI has promoted the economic development in the Yangtze Delta fairly well. However, with the increase of FDI scale, the marginal effect of FDI might decrease. Therefore, the appraisal of the squeezing-in and out effect of FDI on the domestic investment in the Yangtze Delta will be an indispensable part while appraising the effect of FDI on the economic growth in the Yangtze Delta. If investment from multinational companies squeezes in the domestic investment, the FDI promotes the new upstream or downstream investment flow in the region; or rather, FDI will have a squeezing-out effect on the regional investment, that is, it will replace certain domestic investment in the region, which will not contribute to the increase of total investment or capital in the region. Instead, it will squeeze out some of the investment in the region and thus bring about a certain negative external impact on the macro economy. Therefore, this article attempts to use a Total Investment Model and use the long-term coefficient of the squeezing-in and out effect of FDI to analyze this effect of FDI on the investment in the Yangtze Delta.

2.1. The Model and Method of Squeezing-in and Out

In order to evaluate the squeezing-in and out effect of FDI on the investment in the Yangtze Delta, a Total Investment Model[6] will have to be set up first. The total investment in a certain period equals domestic investment plus foreign investment in the region as shown in Eq. (1).

$$I_t = I_{d,t} + I_{f,t} \tag{1}$$

As for foreign investment, only the FDI is considered here, usually there is a time lapse between FDI inflows and the formation of actual investment as shown in Eq. (2).

$$I_{f,t} = \Phi F_t + \Phi_1 F_{t-1} + \Phi_2 F_{t-2} \tag{2}$$

Of the many factors influencing the domestic investment, we only select two variables: the prophase investment increase and the investment level as shown in Eq. (3).

$$I_{d,t} = \Phi^1_0 + \Phi^1_1 G_{t-1} + \Phi^1_2 G_{t-2} + \lambda_1 I_{t-1} + \lambda_2 I_{t-2} \tag{3}$$

Then, we change the above formula into the total investment formula as shown in Eq. (4).

$$I_t = \alpha + \beta_1 F_t + \beta_2 F_{t-1} + \beta_3 F_{t-2} + \beta_4 I_{t-1} + \beta_5 I_{t-2} + \beta_6 G_{t-1} + \beta_7 G_{t-2} + \varepsilon_t \tag{4}$$

where *I* is the Investment ratio (total investment/GDP ratio), where I_t, I_{t-1}, I_{t-2} are the investment ratio in the year of *t*, *t-1*, *t-2*; F is the ratio between FDI in flow and GDP, where, F_t, F_{t-1}, F_{t-2} are respectively the ratio in the year of *t*, *t-1*, *t-2*; G is the GDP growth rate, where G_{t-1}, G_{t-2} are respectively the GDP growth rate in the year of *t-1*, *t-2*; α is a constant; and ε is the serial incorrelation random error.

The following coefficient can be used to appraise the squeezing-in and out effect of FDI on domestic investment the in a relatively long period as shown in Eq. (5).

$$\hat{\beta} = \frac{\sum_{j=1}^{3} \beta_j}{1 - \sum_{j=4}^{5} \beta_j} \tag{5}$$

When $\beta_j (j = 1,2,\cdots,5)$ is significant, the value of $\hat{\beta}$ can be used to measure whether the FDI has squeezed the domestic investment of a county or a region in or out:

1. $\hat{\beta} = 1$, that is, in the long-term changes, a growth of 1% in the FDI/GDP will bring about 1% growth in the I/GDP, which shows that the investment of a multinational company is in parallel to the domestic investment. Under this circumstance, $\sum_{j=1}^{5} \beta_j = 1$.

2. $\hat{\beta} > 1$, that is, FDI has the squeezing-in effect on domestic investment in the long-term changes. That is, 1 unit of FDI brings about more than 1 unit of total investment. Under this circumstance, $\sum_{j=1}^{5} \beta_j > 1$.

3. $\hat{\beta} < 1$, that is, FDI has the squeezing-out effect on domestic investment in the long-term changes. That is, 1 unit of FDI brings about 1 unit decrease of total investment. That is to say, the FDI replaces domestic investment. Under this circumstance, $\sum_{j=1}^{5} \beta_j > 1$.

When $\hat{\beta} \neq 1$, FDI has external effect on the macro economy of the host country or region. The squeezing-in effect indicates positive external impact, and the squeezing-out effect indicates negative external impact.

2.2. Simulation results and analysis

The simulation analysis is done as per data in Table 4. The simulation results are divided into two situations:

Situation I: all 7 variables are used. The results of the regression analysis are shown in Results 1 in Table 5. The F Test is basically significant and the regression coefficient is not ideal. Where, $\hat{\beta} = 2.67$.

Situation II: Variable Gt-1 with the least reliability and significance level is removed. The results of the regression analysis are shown in Results 2 in Table 5. The simulation results are relatively ideal, and the reliability is relatively high. Where, $\hat{\beta} = 2.30$.

The results of the two kinds of simulation are both $\hat{\beta} > 1$, which shows that FDI has a squeezing-in effect on the investment in the Yangtze River Delta in the long term, that is, the utilization of FDI in the Yangtze River Delta has a positive effect from the investment perspective.

Table 4. Data on the investment equation variable. The investment ratio and GDP growth rate in the table are from the Almanac of Statistics of Shanghai, the Almanac of Statistics of Jiangsu Province and the Almanac of Statistics of Zhejiang Province in different years. The FDI/GDP ratio is calculated by the authors. GDP is converted into US dollars as per the average mean of foreign exchange rate between RMB and USD.

	It	Ft	Ft−1	Ft−2	It-1	It-2	Gt-1	Gt-2
1990	25.08	0.57	-	-	-	-	-	-
1991	26.22	0.75	0.57	-	25.08	-	16.48	-
1992	30.99	3.45	0.75	0.57	26.22	25.08	29.03	16.48
1993	38.66	5.70	3.45	0.75	30.99	26.22	39.08	29.03
1994	39.80	7.97	5.70	3.45	38.66	30.99	35.47	39.08
1995	41.64	6.96	7.97	5.70	39.80	38.66	28.13	35.47
1996	42.28	7.20	6.96	7.97	41.64	39.80	17.14	28.13
1997	40.03	6.84	7.20	6.96	42.28	41.64	12.46	17.14
1998	39.99	6.05	6.84	7.20	40.03	42.28	8.15	12.46
1999	37.93	5.17	6.05	6.84	39.99	40.03	7.70	8.15
2000	37.26	4.89	5.17	6.05	37.93	39.99	11.95	7.70
2001	38.17	5.37	4.89	5.17	37.26	37.93	10.53	11.95
2002	40.55	6.48	5.37	4.89	38.17	37.26	11.96	10.53
2003	45.50	7.99	6.48	5.37	40.55	38.17	18.59	11.96
2004	39.15	6.23	7.99	6.48	45.50	40.55	19.97	18.59
2005	46.42	5.62	6.23	7.99	39.15	45.50	21.34	19.97
2006	45.44	5.61	5.62	6.23	46.42	39.15	16.16	21.34

Table 5. Simulation results of the investment model and test of results. The numbers above the brackets are $\hat{\beta}$ values, the numbers in the brackets are t test values, * means significance at the 10% level and ** at the 5% level.

	Ft	Ft－1	Ft－2	It-1	It-2	Gt-1	Gt-2	Adjusted R2	F
Result 1	2.28	-2.78	-1.48	0.38	1.36	0.22	-0.03	0.299	1.79
	(2.24) *	(-2.28)*	(-0.90)	1.23)	2.09) *	1.24)	-0.14)		
Result 2	2.45	-2.75	-1.81	0.43	1.48	0.20		0.665	5.63
	(2.85) **	(-2.61) **	(-1.48)	(1.70)	(2.74) **	(2.17) *			

3. Empirical Analysis of the Squeezing-in Effect of FDI

The inflow of FDI has brought a tremendous amount of funding, advanced technology and managerial expertise to the Yangtze Delta, which has promoted its economic development and the adjustment of the industrial structure, increased employment opportunities and financial revenue, and also boosted the establishment of market economy mechanisms as well as the transformation of the business structure of state-owned enterprises in this Region. In general sense, in analyzing the impact of FDI on the economic growth of the host country (or region), we would generally consider the ratio between the FDI cumulative amount and the GDP of the host country (or region), and would also analyze factors conducive to the economic growth of the host country (or region). Such factors consist of a set of indicators, including capital formation indicators, industrial structure adjustment indicators, and employment generation indicators.

3.1. FDI Accelerates Fund Accumulation on the Yangtze Delta and Contributes to its Economic Growth

The absorption and utilization of FDI may increase the cumulative amount of funding in the Region, accelerate capital formation, and enhance its investment level. At the end of December 2006, the FDI absorbed and actually utilized amounted to USD 230.36 billion, which accounts for a significant proportion of the total fixed asset investment in the Region. Back in 1985, FDI only made up 1.92% of the total fixed asset investment in the Region. The percentage went up to 2.27% in 1990, representing an increase of 92.19%, and it jumped to 20.03% in 1994, showing a sharp increase of 780.72% as compared with the figure in 1990. Although the growth rate slightly dipped in the following years, it has always remained above 12% (see Table 6). It can be concluded that FDI plays a positive role in capital formation and accumulation in the Yangtze Delta.

Fund accumulation brought about by FDI is significantly conducive to the growth of the economy in the Yangtze Delta as a result of the investment multiplication effect. In the past few years, FDI in the Region has witnessed an annual average growth rate of 32.59% – a figure far higher than its GDP growth rate of 14.94%. [7]

Table 6. Proportion of FDI in fixed asset investment in the Yangtze Delta in years 2001-2006. The FDI amount is the actual amount of FDI inflow. The total amount of fixed assets is converted into US dollars as per the average mean of foreign exchange rate between RMB and USD.

year / item	2001	2002	2003	2004	2005	2006
Total amount of fixed assets in the Yangtze Delta (hundred million USD)	975.5	1160.3	1544.1	1594.0	2293.1	2706.7
FDI (hundred million USD)	137.2	185.6	271.0	253.6	277.5	334.3
Proportion (%)	14.1	16.0	17.5	15.9	12.1	12.3

3.2. FDI Generates Considerable Employment Opportunities for the Yangtze Delta

The squeezing-in effect of FDI brings about a positive influence on the employment situation in the Yangtze Delta, namely it generates many job opportunities in this Region. From the practice of utilizing FDI, it can be predicted that the employment opportunities that FDI creates directly in the Yangtze Delta will grow steadily. Meanwhile, auxiliary enterprises and other sectors generating FDI generate more employment opportunities in the Yangtze Delta. Therefore, FDI absorption and utilization play a significant role in alleviating the employment pressure in the Region.

Rationally utilizing labour resource is an important way of improving economic efficiency, and also a key element in enhancing productivity. Therefore, this article employs the production factor contribution method to calculate and analyze the contribution rate of FDI enterprise growth to employment so as to reveal the trend of change in the employment rate.

The Cobb-Douglas Production Function Model is as follows:[8][9][10]

$$Y = A \times K^{\theta} N^{1-\theta} \tag{6}$$

In Eq. (6), Y refers to the added value of FDI enterprises, K refers to the annual average fixed assets, N stands for annual average employed population, and θ is the elasticity coefficient. From Eq. (6), we have

$$\log\left(\frac{Y}{K}\right) = \log A + (1-\theta)\log\left(\frac{N}{K}\right) \tag{7}$$

The estimate of θ can be obtained using the Least Squares Method to estimate parameters in Eq. (7). Then, using the simplified model of Solow's Growth Velocity Equation, $y = (1-\theta) \times n + \theta \times k$, we can calculate the contribution rate of the growth of FDI enterprise output value on employment. In this formula, y refers to the added value of FDI enterprises, n stands for employment growth rate, and k stands for capital growth rate.

Using the above-mentioned method, we calculated the relative data in years 2000-2006, which yielded a fairly good result. The estimated formula is as follows:

$$\log\left(\frac{Y}{K}\right) = 2.6825 + 0.5689\log\left(\frac{N}{K}\right) \tag{8}$$

where R^2=0.9147 and the F test value equals 15.3666.

Analysis and significance test on the regression equation indicate that the correlation coefficient is close to 1, which testifies significant linear relationship between *log(Y/K)* and *log(N/K)*. The F test is also passed. The regression equation indicates a fairly good fit to the actual results.

Meanwhile, as $\log A = 2.6825,\ 1-\theta = 0.5689$, $\theta = 0.4311, A = 14.6223$, namely, Eq. (9) is

$$Y = 14.6223 \times K^{0.4311} \times N^{0.5689} \tag{9}$$

The formula deducted according to Solow's Growth Velocity Equation is

$$y_N = 0.5689 \times \frac{n}{y} \tag{10}$$

From the above formula, we can estimate the contribution rate of the growth in FDI enterprise output value to the employment situation in the Yangtze Delta in years 2000-2006. In years 2000-2006, with the rapid growth of output value of FDI enterprises, despite occasional fluctuations, the corresponding contribution rate to employment showed an upward trend. This demonstrates that FDI enterprises in the Yangtze Delta deploy personnel in accordance with market rules, which leads to a relatively low recessive unemployment rate, more indirect employment opportunities, and higher efficiency in utilizing labour resources.

4. Recommendations

4.1. Further Enhance the FDI Agglomeration Capability of the Yangtze Delta

With the continuous expansion of the FDI agglomeration scale in the Yangtze Delta, the FDI agglomeration capability of the Region will be continuously improved, and industrial clusters that are based on FDI will also be upgraded. Moreover, the regional agglomeration capability for the advanced manufacturing sector and modern service sector will also be gradually improved. These will accelerate industrialization in the Yangtze Delta, shorten the industrialization cycle of the Region, and speed up its modernization process. Meanwhile, it is worth noting that when the FDI agglomeration scale reaches a certain critical value, the demand and pressure on FDI transformation in the Region will also go up considerably.

4.2. Control Environment Problems in the Yangtze Delta Caused by FDI Absorption and Persist in Sustainable Development

With the expansion of FDI agglomeration, investment in pollution treatment, environment protection and improvement will also increase accordingly. Although there is, so far, no effective mechanism matching investment in pollution treatment to the FDI investment scale, strengthening environment supervision on FDI projects and increasing investment in urban environment improvement have become a shared concern among cities in the Yangtze Delta.

4.3. Advance Industrialization of New and High Technology in the Region to Form Core Competitiveness in the New and High-tech Industry

In the process of attracting and absorbing FDI, the introduction of advanced managerial concepts and new and high technology industries has vigorously boosted the economic development of the Yangtze Delta. However, enterprises in the Region are lacking in self-initiated innovation capability and core technology, and are weak in technology internationalization capability, which prevents the local economy from developing in a sustainable and healthy way. "In the process of economic globalization, some areas with economic and technological advantages play a key role in boosting the economic development and enhancing the international competitive status of the host countries and regions." It is clear that new technology industrialization is crucial to the economy, and is conducive to the enhancement of comprehensive national strength. It is also the driving force of regional economic growth.

4.4. Strengthen Adjustment and Integration of the Industrial Structure to Enhance the International Competitiveness of Regional Industry

Success in the regional economy in other countries proves that to be strongly competitive at an international level, an economic region must have core cities with complete service facilities, and form industrial clusters with clear hierarchical levels and division of labour. The Yangtze Delta hosts Shanghai – a metropolis aimed at developing itself into an international centre of economy, finance, trade, and navigation. It is also home to a series of distinctive industrial centers in the Jiangsu and Zhejiang provinces, whose industrial structure provides substantial room for mutual support and cooperation. Therefore, it can be predicted that with Shanghai as the locomotive and the modern industry and auxiliary facilities in Jiangsu and Zhejiang as the supporting force, the industrial cluster in the Yangtze Delta is bound to become internationally competitive after several years of construction and development. "To this end, we must adopt the concept of globalized operation, abolish local protectionism and local mentality", and build the Region into an organic part of the overall economic development of the Yangtze Delta, China and the world. This will definitely contribute to the formation of a complete industrial chain and consistency between the development strategy of multinationals in China and China's development strategy". Therefore, cities in the Yangtze Delta must adopt globalization strategies in order to adapt to the trend of global economic integration and to enhance their overall competitiveness in the global economic competition.

5. Conclusions

It can be seen from the above-mentioned analysis that FDI is conducive to the economic development of the Yangtze Delta. On the one hand, the Yangtze Delta can accumulate foreign funding through absorbing FDI. On the other hand, owing to the agglomeration effect, FDI can help attract the capital of local manufacturers, and thus vigorously boost local economic development. FDI has a squeezing-in effect on the investment in the Yangtze Delta, namely foreign investors have promoted the local investment in the Yangtze Delta in the process of investing in this Region themselves. It can be said that FDI in the Yangtze Delta makes more contribution to capital formation in the region than the investment made by local manufacturers. Generally speaking, FDI will not replace the local investment in the Region. On the contrary, it will be conducive to improving capital quality, promoting technology, upgrading industrial structure, and boosting auxiliary industry. Therefore, in the process of attracting and using FDI, the local enterprises in the Yangtze Delta will enhance cooperation with foreign-invested companies in technological research and development as well as product innovation so as to improve the upstream and downstream industrial correlation, making the Yangtze Delta the base location for production and R & D for multinational companies, and further elevating the international competitiveness of the Yangtze Delta.

Acknowledgment

This research was supported by the Fund for Candidates of Outstanding Young College Teachers of Shanghai under Grant 04YQHB161.

References

[1] Xie Wei'an. Microeconomics & Statistical Calculation Method, Shanghai: Tongji University Press, 1996:211～218 (in Chinese)
[2] Statistics Almanac of Shanghai Z.(several years) (in Chinese)
[3] Statistics Almanac of Jiangsu Province Z. (several years) (in Chinese)
[4] Statistics Almanac of Zhejiang Province Z.(several years) (in Chinese)
[5] World Investment Report Z. (several years) (in Chinese)
[6] Wu Guoxin. The Study on Problems and Relevance of FDI and Economic Growth in Yangtze Delta. International Business Research, 2006(6)
[7] Wu Guoxin. Reflections on Raising Economic Competitive Power of The Yangtze Delta. International Business Research, 2006(4)
[8] Wu Guoxin. A Positive Analysis of FDI's Impact on Employment Growth in Shanghai. Commercial Research, 2005(5)
[9] Lu Yuan.An empirical Analysis of Labor Deployment Efficiency in the Tertiary Industry of Shanghai. Shanghai Economic Review, 1998(7)
[10] Zhao Nonghua. The Experimental Analysis of the Improvement of Employment through the Development of the Tertiary Industry in Shanghai. Statistics Research, 2002(2)

Applications of Data Mining in E-Business and Finance
C. Soares et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-890-8-133

Sequence Mining for Business Analytics: Building Project Taxonomies for Resource Demand Forecasting

Ritendra DATTA [a], Jianying HU [b] and Bonnie RAY [b,c]
[a] *Department of Computer Science and Engineering*
The Pennsylvania State University, University Park, USA
e-mail: datta@cse.psu.edu
[b] *Mathematical Sciences Department*
IBM T.J. Watson Research Center, Yorktown Heights, USA
e-mail: jyhu@us.ibm.com
[c] *Data and Analytics Program Manager*
IBM China Research Lab, Beijing, China PRC
e-mail: bonnier@cn.ibm.com

Abstract. We develop techniques for mining labor records from a large number of historical IT consulting projects in order to discover clusters of projects exhibiting similar resource usage over the project life-cycle. The clustering results, together with domain expertise, are used to build a meaningful project taxonomy that can be linked to project resource requirements. Such a linkage is essential for project-based workforce demand forecasting, a key input for more advanced workforce management decision support. We formulate the problem as a sequence clustering problem where each sequence represents a project and each observation in the sequence represents the weekly distribution of project labor hours across job role categories. To solve the problem, we use a model-based clustering algorithm based on explicit state duration left-right hidden semi-Markov models (HsMM) capable of handling high-dimensional, sparse, and noisy Dirichlet-distributed observations and sequences of widely varying lengths. We then present an approach for using the underlying cluster models to estimate future staffing needs. The approach is applied to a set of 250 IT consulting projects and the results discussed.

Keywords. sequence mining, resource demand forecasting, Markov Models

Introduction

A good view into future resource needs is essential for driving profitability in a service-oriented businesses [1]. Large projects typically require multiple resources, each having different skills. The resource requirements are not static, instead varying over the life of a project as it enters different phases. Both lack of resources with the appropriate skills to carry out a project when needed as well as over-supply of resources who are under-utilized result in loss of profits to the business.

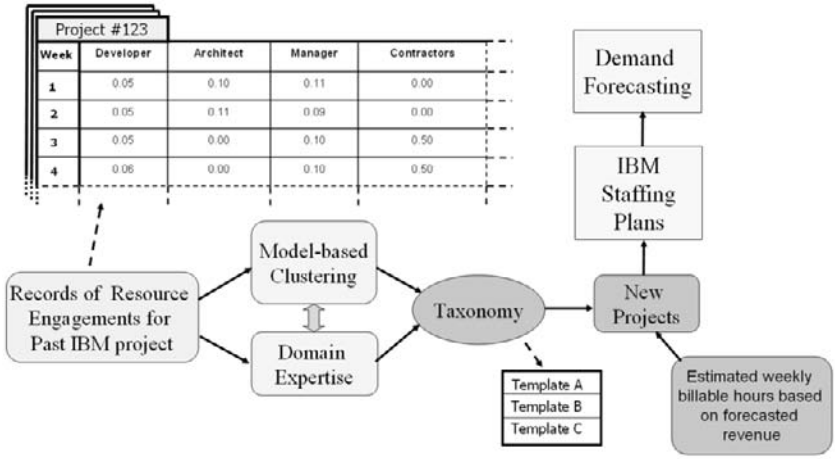

Figure 1. A high-level view of how the sequence clustering fits into overall workflow for demand forecasting.

One approach to predicting future resource demands is to create a project categorization scheme that links a set of project attributes captured in the early stages of negotiations with a client to typical resource requirements over the project life-cycle. In this paper, we present a model-based sequence clustering algorithm useful for finding groups of projects showing similar resource requirements over the project life cycle, and show how the resulting groups can be used to infer a project taxonomy. To automatically determine project groups, we use a hidden semi-Markov model (HsMM)-based clustering algorithm. Higher-level descriptions are associated with each cluster with the help of domain experts. The models describing each cluster are used to form templates representing typical staffing requirements for the different project types. These templates can then be used to generate forecasts of future staffing needs.

Three major contributions are given in this paper. First, we formalize the problem of project taxonomy building for workforce management as a sequence clustering problem. Second, we present a clustering algorithm that includes several new advances over past attempts at sequence modeling and clustering [2,3,4,5,6,7,8] to address challenges specific to the business problem. Third, we outline the use of the cluster model results for project-based resource demand forecasting. We demonstrate the effectiveness of the proposed approach on labor claim data from IBM Global Business Services (GBS) consulting engagements.

The remainder of the paper is structured as follows. In Section 1, we motivate formulation of the business problem as a sequence clustering problem and de“ne some notation. In Section 2, we present the HsMM-based sequence clustering algorithm. In Section 3, we outline use of the cluster results for project-based demand forecasting. Section 4 presents results of application to real data from a set of IBM Global Business Services consulting projects. We conclude with a discussion in Section 5.

1. Taxonomy Building as a Sequence Clustering Problem

Fig. 1 shows a high level workflow for using historical project labor data for resource demand forecasting. We concentrate here on the initial steps, i.e. processing and analyzing labor claim records to determine similarities between projects. We formalize the model-based clustering problem as follows. Suppose we have labor claim data for a set of m historical projects $\{P_1, P_2, \ldots, P_m\}$ completed by an organization, where the resources completing the labor are each labeled according to an agreed upon job role categorization scheme. Let the total number of resource types available within the organization be D. A project P_i has a duration of T_i (in weeks), and for each week j (relative to the starting week), the *proportion vector* representing the distribution of resources across job role categories is specified by $O_{i,j} \in \mathbb{R}^D$. Since resource distributions are represented as proportions, for each week i we have $\sum_{k=1}^{D} O_{i,j}(k) = 1,\ O_{i,j}(k) \in 0, 1$, where $O_{i,j}(k)$ (the k^{th} component of the vector) denotes the observed proportion of resource hours in job role category k in week j for project i. Thus the full specification of a project P_i, in terms of resource requirement distributions, is given by the sequence of proportion vectors, ordered by weeks: $P_i = (O_{i,1}, O_{i,2}, \ldots, O_{i,T_i}),\quad O_{i,j} \in \mathbb{R}^D,\ 1 \leq j \leq T_i$. Since durations of projects vary, T_i take different values for different projects. Therefore each P_i can be thought of as a variable-length multivariate sequence. Given that there are m such projects, we have a set of m multivariate sequences; we would like to find those sequences that exhibit similar behavior. More formally, the problem is to devise an algorithm for clustering variable-length multivariate sequences having some specific characteristics. In particular, we seek a *model-based* sequence clustering algorithm to capture the following characteristics of the business problem.

1. Projects are typically carried out in phases.
2. Temporal dependencies among project phases are common.
3. The observation vectors at each time period represent proportions, so appropriate statistical distributions are necessary for their modeling.
4. The proportion vectors may be sparse, i.e., only a small set of the resources may contribute to a project in a given week.
5. The number of clusters in the taxonomy is not known *a priori*.
6. There may be atypical weeks within projects, and atypical projects as a whole, both of which should be treated as outliers.

We wish to infer statistical models for each cluster, so that the resulting project taxonomy can be used to generate expected resource requirements given a selected project group. We call the model used to generate such a forecast a *staffing template*.

Note that alternative data mining algorithms, such as decision trees or itemset mining are not appropriate for this problem, as our objective is to group sets of projects that have no explicit labeling. The next section presents the sequence clustering algorithm.

2. HsMM-based Sequence Clustering

In this section, we provide motivation for the HsMM modeling choices used for analysis.

1. Proportion vectors are defined on a $(D-1)$-simplex, where D is the number of resources. Hence we use a Dirichlet distribution for modeling the observa-

tion vectors, a standard distribution for describing multivariate proportion data. For x in the $(D-1)$-simplex, the probability density function (p.d.f.) for the D dimensional Dirichlet distribution is defined as

$$f_d(x|b) = \frac{1}{B(b)} \prod_{i=1}^{D} x_i^{b(i)-1}, \quad x \in \mathbb{R}^{\mathbb{D}}, \tag{1}$$

where

$$B(b) = \frac{\prod_{i=1}^{D} \Gamma(b(i))}{\Gamma\left(\sum_{i=1}^{D} b(i)\right)} \tag{2}$$

and $\Gamma(\cdot)$ denotes the Gamma function.

2. Since phases within projects govern the resource requirements, we use a first-order hidden Markov model (HMM) [2] to characterize a project's transition through project phases over its life-cycle. An HMM is a finite-state probabilistic model governed by first-order state transitions.
3. Since project phases typically occur in a sequence, with the same phase not repeating itself, we restrict our model to (a) non-ergodic left-to-right HMMs, (b) having strict start and end states, with (c) state skipping being disallowed. These are essentially topological restrictions which yield a reasonable simplification to the model, without loss of information.
4. Because project phase durations may not necessarily follow a geometric distribution, which is implicit for a first-order HMM, we employ HMMs with state duration explicitly modeled by a Gamma distribution, allowing much more flexibility in the state durations. With explicit duration modeling, HMMs are no longer first-order models, and hence are referred to more accurately as hidden semi-Markov models (HsMM) [9]. The p.d.f. of the Gamma distribution is given by

$$f_g(x|\psi, \theta) = x^{\psi-1} \frac{e^{-x/\theta}}{\theta^{\psi}\Gamma(\psi)}, \quad x \in \mathbb{R}^{+}. \tag{3}$$

A schematic view of an N-state HsMM as described above is shown in Fig. 2. For clustering the project data under the assumption of an HsMM for each cluster, we use a modified version of the algorithm proposed in [4], designed to address challenges posed by the high dimensional, yet sparse, proportional nature of the observations, and the presence of noise both within a sequence and for a set of projects. The algorithm carries out **model construction** via robust estimation of HsMM parameters, with the number of HsMM states inferred iteratively using the Bayesian Information Criteria (BIC). The number of clusters is also automatically inferred using a modified BIC, referred to as **partition_BIC**. The steps in the main clustering algorithm are shown in Table 1.

The proposed model-based clustering algorithm extends the work of [4], primarily through additional focus on the modeling aspects of the work. The methods presented in [4] discretize the sequence data by performing an initial assignment of each observed time point to a state. An HsMM model is then applied to cluster the sequences of discretized data, rather than fitting a distribution to directly characterize the observed vectors, as we have done here using a Dirichlet distribution. Here, the characterization of

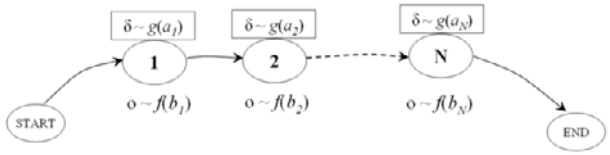

Figure 2. The left-to-right HsMM topology used for sequence modeling. Symbol δ denotes state/phase duration, $f(\cdot)$ denotes the Dirichlet p.d.f., and $g(\cdot)$ denotes gamma p.d.f.

Table 1. The proposed sequence clustering algorithm.

Assign all sequences to one cluster
Apply **model construction** to the one cluster
old_partition_BIC = $-\infty$
Compute **partition_BIC**
While **partition_BIC** $\geq$ **old_partition_BIC**
Compute individual **cluster_BIC** values
Split the weakest cluster in two using **hierarchical clustering**
Apply **model construction** to the two new clusters
While membership changes continues
Re-assign each sequence to its most likely cluster
Apply **model construction** to get a new set of clusters
old_partition_BIC = **partition_BIC**
Re-compute **partition_BIC**
Use **Monte Carlo simulation** to get likelihood thresholds
If a sequence has likelihood less than the threshold for its cluster
Reject the sequence as **noise/outlier**
Apply a final **model construction** to each cluster

each state by a parametric model is useful for generation of staffing templates, as discussed in the next section. However, parametric modeling necessitates the development of a dimension reduction procedure for estimating the Dirichlet distributions, as the number of observation dimensions having non-zero proportions is typically small relative to the total number of dimensions. We leave further discussion of the dimension reduction technique to another paper. Lastly, we generalize the state duration distribution to better reflect the characteristics of project durations in practice.

There are a number of reasons that motivate the complex nature of this algorithm for the application in question. First, the approach is completely independent of prior knowledge about the number of clusters that exist in the data, which is algorithmically determined. Second, the number of states in the HsMM for each cluster is also automatically determined, eliminating the need for the user to specify them. Both these properties are particularly attractive for the application in question, because it is very difficult to make conjectures about the size of a resource-based project taxonomy or the nature of each project cluster. The idea is to have a largely data-driven approach, with the only prior knowledge use being in the choice of statistical distributions for modeling. Finally, and most importantly, the choice of a model-based clustering algorithm as against simpler methods such as distance-based clustering is motivated by the fact that for our application, it is critical to get an insight into each cluster, so as to be able to make high-level in-

terpretations about each of them. The estimated model parameters for each cluster make it possible to draw such inferences.

3. Resource Demand Forecasting based on Cluster Results

In order to use the cluster results for resource demand forecasting, it is necessary to create a description of each cluster that reflects the typical resource requirements and project phase durations for the projects in that cluster. A straightforward way to obtain the typical resource distribution per phase for a cluster is to aggregate the hours in each resource category across all projects in the cluster for each identified project phase and compute the category distribution based on the total claimed hours for that phase. Note that this approach results in certain categories having very small, but non-zero, percentages because only a few projects in the cluster have hours claimed in that category for a phase. Here, we take advantage of the models generated for each cluster by using the estimated, reduced-dimension Dirichlet distributions for each state to directly compute the mean resource distribution for each phase.

To obtain estimates of project phase duration, we distinguish between two different scenarios. 1) Individual phase durations and resultant overall project duration are estimated based on the mean calculated directly from the gamma distribution for each corresponding state. 2) In the case the project duration is specified *a priori*, for example, because of requirements from the customer, the duration of each phase can be estimated by scaling the mean phase durations generated in 1) by the fixed total project duration. Further discussion of template generation from cluster results is given in [10], where the focus is on project clustering at the aggregate level, i.e., the problem of sequence clustering is not addressed.

Once the statistical cluster analysis and template creation are complete, the next challenge is to create an appropriate project taxonomy from the results. In general, this can be accomplished using a two-step process. 1) For each cluster, examine a set of project attributes (other than resource requirements) whose distribution of values suggests a name and description for each cluster. 2) Validate and refine cluster taxonomy labels and class descriptions through discussions with subject matter experts.

The objective of the first step is to identify unique characteristics represented by each cluster. The predominant attribute values within a cluster serve to provide alternative characterizations of projects, enabling linkage of a project's business attributes and a project's staffing requirements. This step may be accomplished formally, through building a classification model for the project groupings using business attributes as features, or informally, through discussion with subject matter experts. In Step 2, domain experts are used to validate the various project types to ensure that each discovered project type is both meaningful, from a practitioner's viewpoint, and distinct, meaning that groups identified as statistically distinct in fact represent true variations in resource distributions due to differences in the types of projects implemented.

The staffing templates can be used for generating project-based demand forecasts through application of the following steps.

1. Select a label for a new project instance from the created project taxonomy. This label is selected based on knowledge of the new project's business attributes and

the previously determined linkage between business attributes and the project taxonomy, as mentioned above.
2. Compute the duration of each project phase, either using a predetermined project duration or estimated project duration, as discussed above.
3. Estimate the number of project hours required for each project phase. This estimate can come directly from expert opinion, or be based on an established relationship between project revenue and project hours.
4. Distribute the expected hours per phase across the project job roles according to the established Dirichlet distributions.

Estimated resource demands and their start/end dates are aggregated across all project opportunities at a weekly level to achieve a total view of expected resource requirements over a specific time interval.

4. Results on IBM Data

To test the efficacy of the proposed method, we initially conducted a number of experiments on synthetic sequences and found the clustering results satisfactory. Given our business goal, we then tested how well the algorithm was able to produce meaningful resource-based project taxonomies for real project staffing data. We applied the method to a set of 250 historical SAP-related IBM projects, each lasting at least 10 weeks. SAP is an Enterprise Resource Planning application; SAP-related engagements represent a large portion of IBM's consulting projects and are thought to contain fairly predictable types of tasks, making their analysis easier. The resources on each project were labeled according to their primary job role using an IBM-defined taxonomy of job roles. Sixty-six different job roles were represented in these 250 projects, i.e., each observation vector was of dimension $n = 66$. The dimension reduction component of our algorithm selects informative dimensions and club the remainder into an extra dimension (since components add up to one). This component selected seven out of these sixty-six job roles for explicit modeling, plus an "other" category representing the aggregation of all remaining job roles. Based on the iterative clustering algorithm given in Table 1, we obtained a set of six project clusters. Fig. 3 shows the average percentage of resource hours in each of the eight job role categories relative to the total resource hours for the 250 projects. For each cluster, the graph shows estimates for each state in the constructed models, where the number of states are chosen automatically as part of modeling/clustering.

Roughly speaking, Cluster 1 can be said to represent projects consisting initially of Package Solution Integration Consultants (PSIC, Dimension 35) configuring and deploying particular SAP modules. In the second phase of the project, additional resources having different job roles, including a Project Manager (PM, Dimension 41) to coordinate the different activities and resources, are brought in and the role of the PSIC is reduced. These other roles include Application Developer (Dimension 4), Application Architecture (Dimension 2), Business Development Executive (Dimension 6), as well as Other. These types of projects are typical of many SAP engagements conducted by GBS. In contrast, Cluster 2 represents more of an initial roadmap/blueprint type of project, in which a project manager or partner is engaged as the primary resource in the initial stage of the project, while resources with additional skills are brought in only in the second phase, and the relative involvement of the PM is reduced substantially. Similar interpre-

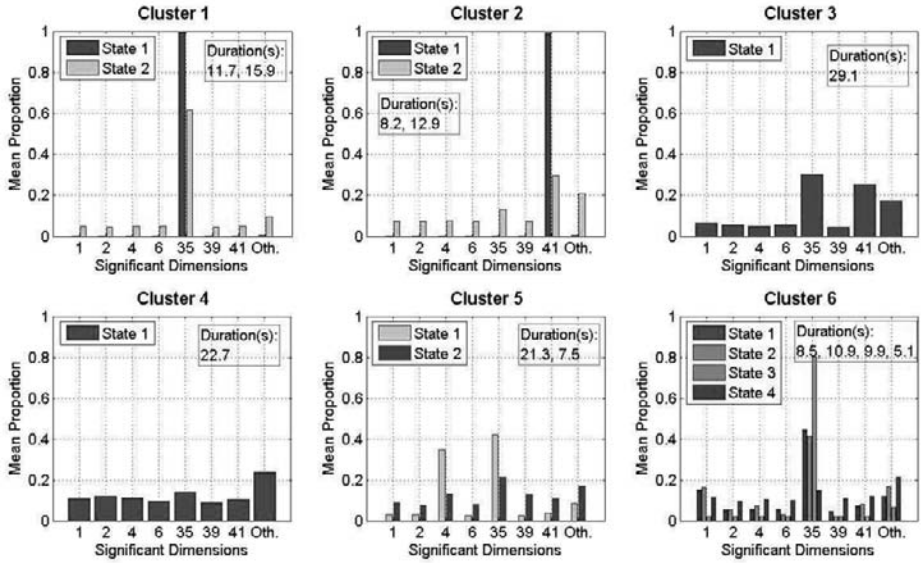

Figure 3. The six clusters generated with the IBM GBS data. For each cluster, and for each state, the mean values of proportions (Y-axis) for the significant resources (X-axis) are shown, along with the mean state durations, calculated from parameter estimates.

tations can be made for the other clusters. In discussion of these results with GBS domain experts, the identified clusters were found to be reasonable and representative of typical SAP engagements.

To create an appropriate project taxonomy from the results, we examined the distribution of additional project attributes (beyond resource distributions) for projects in each cluster, as discussed in Section 4. Attributes examined included the client's business sector (e.g. Industrial, Distribution), the service area most representative of the work (e.g., Customer Relationship Management, Supply Chain Management), the business unit of the project manager, the expected project revenue, the way the project was priced (e.g. Time and Materials, Fixed Price), etc. Information on the nature of the different projects within a cluster was also gleaned from the recorded project name and description. For the projects used in the experimentation, we were unable to find clear relationships between the project attributes mentioned above and project staffing patterns suggested by the sequence clusters. Thus, expert help was sought for further analysis. A summary of the observed staffing requirements in each cluster was used to discuss the results with GBS domain experts, who subsequently were able to relate the observed staffing patterns to project type names typically used within GBS. For example, Cluster 2 projects consisting primarily of project management and application architect were determined to represent the design phase of a project, commonly called a Phase 1/Blueprint project.

We do not give an example here to show the creation of staffing templates from these results and subsequent generation of staffing requirements for a newly identified project. Formal models for assigning a label to a new project instance were not developed in this case. See [10] for a detailed example of staffing template generation based on projects clustered at an aggregate level.

5. Conclusions

A sequence clustering-based approach to building resource-based project taxonomies has been proposed, which handles noise, sparse high-dimensional observations, explicit state duration modeling, and lack of knowledge about the number of states and clusters. The clustering algorithm has been applied to IBM GBS project data, and the obtained clusters have been found to be reasonable representations of project types based on resource engagements, making the sequence clustering framework an attractive approach to taxonomy building.

The proposed sequence clustering approach is quite general, with potential applications in a wide variety of other problem domains, including video sequencing, gene expression clustering [11], etc. For example, video shots can be represented by allocations or proportions of a quantized color set. This way, video shots can be represented by sequences of proportional vector, making our Dirichlet-based sequence clustering algorithm appropriate. Moreover, data noise and sparsity typically occur in such sequences as well. We leave application of the technique to this and other problems for future work.

References

[1] R. Melik, L. Melik, A. Bitton, G. Gerdebes, and A. Israilian, *Professional Services Automation: Optimizing Project and Service Oriented Organizations*, J. Wiley and Sons, 2001.

[2] L. Rabiner, *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceedings of the IEEE, 77(2):257–285, 1989.

[3] L. Rabiner and B.H. Juang, *Foundations of Speech Recognition*, Prentice Hall, 1993.

[4] J. Hu, B. Ray, and L. Han, *An Interweaved HMM/DTW Approach to Robust Time Series Clustering*, Proc. International Conference on Pattern Recognition, 145–148, 2006.

[5] C. Li, G. Biswas, M. Dale, and P. Dale, *Matryoshka: A HMM based Temporal Data Clustering Methodology for Modeling System Dynamics*, Intelligent Data Analysis, 6(3):281–308, 2002.

[6] S.Z. Yu and H. Kobayashi, *An Efficient Forward-Backward Algorithm for an Explicit-duration Hidden Markov Model*, IEEE Signal Processing Letters, 10(1):11–14, 2003.

[7] P. Smyth, *Clustering Sequences with Hidden Markov Models*, Proc. Neural Information Processing Systems, 648–654, 1997.

[8] T. Oates, L. Firoiu, and P.R. Cohen, *Using Dynamic Time Warping to Bootstrap HMM-based Clustering of Time Series*, Sequence Learning, Lecture Notes in Computer Science, 1828:35–52, 2000.

[9] S. E. Levinson, *Continuously Variable Duration Hidden Markov Models for Speech Analysis*, Proc. International Conf. Acoustics, Speech, Signal Processing, 11:1241–1244, 1986.

[10] J. Hu, B. Ray, and M. Singh, *Statistical Methods for Automated Generation of Services Engagement Staffing Plans*, IBM Journal of Research and Development, 51(3/4):281–293, 2007.

[11] A. Schliep, A. Schonhuth, and C. Steinhoff, *Using Hidden Markov Models to Analyze Gene Expression Time Course Data*, Bioinformatics, 19:255–263, 2003.

This page intentionally left blank

Applications of Data Mining in E-Business and Finance
C. Soares et al. (Eds.)
IOS Press, 2008

Author Index

This page intentionally left blank

This page intentionally left blank

This page intentionally left blank