Introduction to the Special Issue on Successful Real-World Data Mining Applications

Gabor Melli
PredictionWorks
6700 – 37th Ave SW
Seattle, WA, USA
98126
gmelli@predictionworks.com

Osmar R. Zaïane
Department of Computing Science, University of Alberta
Edmonton, Alberta, Canada
T6G 2E8
zaiane@cs.ualberta.ca

Brendan Kitts
Microsoft
One Microsoft Way
Redmond, WA, USA
98052
bkitts@excite.com

1. INTRODUCTION

Since its inception, the field of Data Mining and Knowledge Discovery from Databases has been driven by the need to solve practical problems [4]. From scaling to large databases and handling noisy and high-dimensional data to finding associational patterns in grocery store transaction data, data mining is a research area rich in application [1]. Despite its practical roots few case studies of data mining applications have been published. The industrial track of the annual SIGKDD conference has provided one such forum, but rarely do these papers present complete descriptions of deployed systems [2]. This special issue attempts to address the gap by showcasing the choices, strategies, and lessons learned from building a real-world data mining application. In a sense this collection is a follow-up to the first workshop on data mining case studies held during ICDM-2006 [3]. This issue however introduces several new papers. Of the 29 papers reviewed 10 papers were accepted. The papers come from a broad range of application areas including Customer Relationship Management, Medicine, Taxation, and Software Development.

2. CONTENTS

Data mining is today applied in a panoply of successful applications from business, marketing. medical research, financial securities, astronomy, surveillance, and even sports, to name a few. The papers in this special issue by no means represent all application domains. However the papers herein give a glimpse on current usage of data mining technology in real world systems. They concern mining patient records, tax returns, telecommunication data, consumer transactions, content of web pages, and software source code. These international papers give evidence of data mining utilized in a diverse and interesting range of successful applications and share experiences in surmounting hurdles when implementing and deploying real systems. The papers are not organized in any particular order as they pertain to distinct areas and techniques.

In "*Data Mining for Improved Cardiac Care*" R. B. Rao, S. Krishnan and R. S. Niculescu present a wonderful application of data mining to the improvement of quality of care and healthcare cost reduction surrounding cardiovascular disease. The application described is noteworthy because it appears to have saved lives. On a practical note the paper presents a novel use of external domain knowledge to work around the often encountered problem of incomplete and noisy data. R. Bharat Rao received first place in the 2005 Data Mining Case Studies Practice Prize for his diligent work in this area.

In "*Closing the Gap: Automated Screening of Tax Returns to Identify Egregious Tax Shelters*" DeBarr and Eyler-Walker describe the implementation of a data mining system in use at the Internal Revenue Service that identifies high-income individuals engaging in abusive tax shelters. The implementation has uncovered hundreds of millions of dollars in sheltered revenue, including elaborate sheltering operations. DeBarr and Eyler-Walker were second place winners in the recent 2005 Data Mining Case Studies Practice Prize.

In "*Blocking Objectionable Web Content by Leveraging Multiple Information Source*" Agarwal, Liu, and Zhang address the prevalent problem of identifying websites with objectionable content such as pornography. The paper presents an extended representation of the information encoded and related to a web page in order to improve classification and demonstrate its performance at AOL. Judging from the experimental results, the proposed solutions are promising.

In "*Championing an LTV Model for an LTC*" Freeman and Melli review some of the organizational hurdles that the implementation of a data mining system can encounter within a large corporation. The specific application is a predictive model for customer lifetime value at a large telecommunications company. The paper will be of particular interest to senior data miners that are also required to successfully champion a data mining project into fruition.

In "*Mining Source Code Elements for Comprehending Object-Oriented Systems and Evaluating Their Maintainability*" Kanellopoulos, Dimopulos, Tjortjis, and Makris describe the application of data mining to the process of the maintenance of object-oriented software. Developing reliable complex software systems is key to software engineering. The approach used involves the clustering of source code in order to recover knowledge about a software's structure and maintainability – much of a maintainers' time is typically spent for program comprehension. The paper includes a novel and comprehensive approach to empirical validation.

In "*Text mining for product attribute extraction*" Ghani, Probst, Liu, Krema and Fano investigate the application of data mining to the data preparation step in predictive modeling within the retail industry. One of the challenges faced by the practioner is this domain, whether the task is demand forecasting, assortment optimization or product recommendations, is the extraction of relevant attributes. Of current interest in particular is the extraction of implicit attributes.

In "*Service Quality Improvement Through Business Process Management based on Data Mining*" S-H. Ha and S-C. Park

present an application of a diverse set of data mining methods to the analysis of root factors for customer complaints. They describe an implementation for a life insurance company that pinpoints where complaints happened, the relationship among problems and the cause of problems. The problem is important and non-trivial because customer reactions are independent of the internal architecture of the business.

In "*Market Basket Recommendations for the HP SMB Store*" Singh, Thomas and Sepulveda present the application of market basket analysis to Hewlett-Packard's online store and call center for cross-sell and up-sell associations. The paper illustrates the commonly face challenge of performing a proof-of-concept within an extremely short time frame as a prerequisite for the funding of a system's implementation.

In "*Revenue Recovering with Insolvency Prevention on a Brazilian Telecom Operator*" Pinheiro, Evsukoff and Ebecken present the application of a data mining to the prediction of customers that may be associated with bad debt events, such as non-payment. This topic has received significant interest from both service companies and data mining vendors, but few case studies have been published to date. The paper covers several of the phases in the data mining process including a customer segmentation phase.

In "*Data-driven Modeling of Acute Toxicity of Pesticide Residues*" Lemke, Benfenati, and Mueller present an application of data mining to the analysis of chemical compounds toxicity to reduce the cost and time associated with the animal testing required by regulation. The paper interestingly presents both a theorical and analytical review of a solution to the modeling of high-dimensional noisy data.

3. CONCLUSION

In conclusion we hope that this special issue will inspire the reader to tackle even more ambitious initiatives and will also encourage them to share in their insights with others. Many data mining application areas remain undocumented. This collection illustrates how case studies may assist in the understanding of how whole data mining problems can be solved. Interesting applications, particularly those with direct social benefits, will help to raise awareness of the positive role that data mining can play in science and society.

4. ACKNOWLEDGMENTS

For their assistance with the review of papers our thanks go to Ajith Abraham, Sourav Bhowmick, Richard Bolton, Sanjay Chawla, Inderjit Dhillon, Guozhu Dong, Edmund Freeman, Marzena Kryszkiewicz, Ying Li, Flavia Moser, Simeon Simoff, Dan Simovici, Anthony Tung and Yabo Xu.

5. REFERENCES

[1] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy. editors. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press 1996

[2] R. L. Grossman, R. Bayardo, K. Bennet, and J. Vaidya, editors, Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2005), ACM Press, New York, 2005, ISBN 1-59593-135-X.

[3] B. Kitts, G. Melli and K. Rexer, editors, Proceedings of the First Workshop on Data Mining Case Studies collocated with International Conference on Data Mining (ICDM), 2005.

[4] G. Piatetsky-Shapiro, and W. Frawley. editors., Knowledge Discovery in Databases, AAAI/MIT Press, 1991

About the editors:

Gabor Melli is founder of PredictionWorks, a data mining and analytics company that specializes in the design and implementation of real-time predictive modeling solutions. He has worked in retailing, telecommunications, information retrieval, banking and software industry at companies such as Wal*Mart, Microsoft, Cingular, T-Mobile, and Washington Mutual attain from their databases. Gabor was the co-chair of the ICDM Workshop on Data Mining Case Studies and is the current Information Director for SIGKDD. He is a PhD candidate in Data Mining from S.F.U and has published several papers on data mining research and its application. (www.cs.sfu.ca /~melli)

Osmar R. Zaïane is an Associate Professor in Computing Science at the University of Alberta, Canada. He has research interests in novel data mining algorithms, web mining, text mining, image mining, and information retrieval. He has published more than 70 papers in refereed international conferences and journals, and taught on all six continents. Osmar Zaïane was the co-chair of the ACM SIGKDD International Workshop on Multimedia Data Mining in 2000, 2001 and 2002 as well as co-Chair of the ACM SIGKDD WebKDD workshop in 2002, 2003 and 2005. He is the Progam co-Chair for the IEEE International Conference on Data Mining ICDM'2007. (http://www.cs.ualberta.ca/~zaiane/)

Brendan Kitts is a Program Manager within Microsoft's AdCenter division. Brendan was formerly a Lecturer in Computer Science at the University of Technology, Sydney, where he was awarded tenure in 1994. Since then Brendan has held the position of Chief Scientist at a distinguished list of companies including Datasage Inc., Vignette Corporation, iProspect and Isobar Communications..