DDDM2007: Domain Driven Data Mining

Longbing Cao
University of Technology Sydney, Australia
lbcao@it.uts.edu.au

Philip S. Yu
IBM T.J. Watson Research center, USA
psyu@us.ibm.com

Chengqi Zhang, Yanchang Zhao
University of Technology Sydney, Australia
{chengqi,yczhao}@it.uts.edu.au

Graham Williams
Australian Taxation Office, Australia
Graham.Williams@togaware.com

ABSTRACT

Real-world data mining generally must consider and involve domain and business oriented factors such as human knowledge, constraints and business expectations. This encourages the development of a domain driven methodology to strengthen data-centered pattern mining. This report presents a review of the ACM SIGKDD Workshop on Domain Driven Data Mining (DDDM2007), held in conjunction with the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD07), which was held in San Jose, USA on 12 August, 2007. The aims and objectives of this workshop were to provide a premier forum for sharing innovative findings, knowledge, insights, experiences and lessons in tackling challenges met in domain driven, actionable knowledge discovery in the real world.

1. INTRODUCTION

In the last decade data mining has emerged as one of the most vivacious areas in information technology. Classic data mining is heavily dependent on data itself, and relies on data-centered methodologies. Existing approaches either view data mining as an autonomous data-driven trial-and-error process, or analyze business issues in an isolated and case-by-case manner. As a result, very often the knowledge discovered does not always generally satisfy real business needs.

As pointed out by SIGKDD panelists [5, 10] and recent research [8, 1, 11], actionable knowledge discovery is one of the grand challenges for the next generation of KDD research and development. There is at present a lack of methodologies, techniques, and applications which can effectively bridge the gap between academia and business [6]. We need to address business interestingness [9, 12] and successfully deliver knowledge [4] and patterns that are operationalisable and dependable. Domain driven data mining targets such challenges and objectives.

A high level of domain driven data mining framework has been discussed in [2, 3]. Domain driven, actionable knowledge discovery should involve, support and integrate the following intelligence and constraints: domain expert's role and computational capability, domain knowledge and intelligence, network and web intelligence, in-depth data intelligence, and the constrained environment and deliverables in specific domains. This naturally involves many general issues in domain driven data mining, including human-cooperated or centered [7] data mining and computing, the involvement and representation of domain knowledge and intelligence, the involvement and development of web and network mining and intelligence, processing mixed data and multi-data sources, in-depth data mining, constraint mining [7], business interestingness, combining technical significance with business expectations, actions in decision support, as well as issues of privacy, security, trust, dependability, and workability.

Certainly, it is a non-trivial effort to narrow gap between technical outputs and business expectations. It is promising but challenging to promote the paradigm shift from "data-centered pattern mining" to "domain-driven actionable knowledge discovery." A broad attention and effort in the research community, active interaction and collaboration between academia and industry, and innovation in data mining education are expected to contribute toward these goals. This report of the "2007 ACM SIGKDD International Workshop on Domain Driven Data Mining (DDDM2007)" summarizes one of the lines of effort addressing this field.

2. SUMMARY OF THE WORKSHOP

DDDM2007 has provided a premier forum for sharing findings, knowledge, insight, experience and lessons by:

- exploring next-generation data mining methodologies for actionable knowledge discovery, and identifying how KDD techniques can better contribute to critical domain problems in theory and practice;
- uncovering domain-driven data mining techniques identifying how KDD can better strengthen business intelligence in complex enterprise applications;
- encouraging the interaction between academia and industry, studying how KDD research results can be better delivered and accepted by business users;
- disclosing the applications of domain driven data mining identifying how KDD can be effectively deployed into solving complex practical problems; and

.

- identifying challenges and directions for future research and development in the dialogue between academia and business.

The DDDM2007 is organized by the Faculty of Information Technology, University of Technology, Sydney, Australia. The half-day workshop has competitively selected eight papers from five countries. These papers have addressed specific domain problems in crime, social security, blog, business, healthcare, finance, as well as theoretical issues. A common feature of contributed papers was that specific data mining techniques and approaches were developed and tested in specific domain problems. In the following, we briefly discuss each contribution.

The workshop started with crime detection for credit applications. In "Adaptive Communal Detection in Search of Adversarial Identity Crime", Clifton Phua, Vincent Lee, Kate Smith-Miles and Ross Gayler present an updated adaptive Communal Analysis Suspicion Scoring (CASS) algorithm. At pre-defined time intervals and by measuring current input size and previous output suspiciousness, CASS adaptively changes the appropriate parameter setting to trade off efficiency/speed and effectiveness/security. Their approach is validated with three sets of experiments on real credit applications.

A domain ontology can contribute to attribute selection and in the interpretation of mining results to make data mining better aligned with business understanding. In "Domain Ontology Driven Data Mining: A Medical Case Study", Yen-Ting Kuo, Andrew Lonie, Liz Sonenberg and Kathy Paizis explore the possibility of utilizing a medical domain ontology to categorize attributes for association rule mining Mined rules were reviewed by comparison to domain knowledge derived from a domain expert. It is claimed to deliver more meaningful results.

Effective and integrative use of patterns in distributed data sources may disclose combinational patterns that are informative and useful for business needs. In their work "Mining for Combined Association Rules on Multiple Datasets", Yanchang Zhao, Huaifeng Zhang, Fernando Figueiredo, Longbing Cao and Chengqi Zhang propose a new association rule method, named Combined Association Rules. It first extracts patterns in transactional data and specifies class labels with domain expert supervision. Then demographic patterns are attached to these patterns. The method was tested on real-world social security data for governmental debt recovery.

Steganalysis methods are proposed for detecting hidden information. A general steganalysis framework is investigated by Shen Ge, Yang Gao and Ruili Wang in "Least Significant Bit Steganography Detection with Machine Learning Techniques". It applies machine learning to detecting Least Significant Bit steganography hidden information in image data. Features derived from conventional methods are extracted, and then varying classifier methods based on them for detecting hidden information. They show better steganalysis with more effective classifier methods.

Identifying leading indicators is not easy, but critical for implementing business intelligence. This is addressed in "A Semi-automatic System with an Iterative Learning Method for Discovering the Leading Indicators in Business Processes". The co-authors Wei Peng, Tong Sun, Philip Rose and Tao Li propose a semi-automatic system and methods to iteratively discover leading indicators from real-time workflow events, equipment logs, and other metrics sources. The method utilizes domain knowledge to filter indicators, and enables incremental adjustment of underlying domain model through involving domain knowledge.

Customer and firm-specific information plays an important role in improving prediction accuracy. This is evidenced by Sai Zeng, Prem Melville, Christian A. Lang, Ioana Boier-Martin and Conrad Murphy in their paper "Predictive Modeling for Collections of Accounts Receivable". They use a supervised learning method to build models for predicting the payment outcomes of newly created invoices, thus enabling customized collection actions tailored for each invoice or customer. They illustrate the application in several firms.

The involvement of expert and business knowledge is essential in discovering new knowledge valuable for "smart" data mining algorithms. Based on real-life experience in enterprise data mining, in "Toward Knowledge-Driven Data Mining", Warwick Graco, Tatiana Semenova and Eugene Dubossarsky summarize a number of issues that may enhance the effectiveness of data mining if solved. They stress the necessary involvement of tools and techniques, expert knowledge, smart data, business knowledge and intelligence towards knowledge driven data mining.

Last but not least, blog-specific search and mining techniques are emerging. In "Blog Search and Mining in the Business Domain", Latent Semantic Analysis and Probabilistic Latent Semantic Analysis based probabilistic models are investigated for searching and mining business blogs. Various term weighting schemes and factor values are used for similarity search. Yun Chen, Flora S.Tsai, and Kap Luk Chan claim that domain driven data mining can better strengthen business intelligence in complex enterprise applications.

We hope the above contributions can promote the research and development of discovering actionable knowledge from complex domain problems, enhancing interaction and reducing the gap between academia and business, and driving a paradigm shift from *interesting hidden pattern mining* to *actionable knowledge discovery* in varying data mining domains.

3. CONCLUSION

In this paper, we present a brief review of the ACM SIGKDD Workshop on Domain Driven Data Mining (DDDM2007). DDDM2007 has brought together the researchers and experienced data miners in the real world who are interested in developing innovative methodologies, approaches, and enterprise applications for workable, dependable, and actionable knowledge discovery in the real life. Attendees have clearly realized the need for domain driven data mining, and informed of the promising efforts in developing corresponding techniques and applications. We hope such efforts can be encouraged and maintained through other professional activities in the community and industries.

Additional information regarding the DDDM2007 is available from the workshop website http://datamining.it.uts.edu.au/dddm/ and the Proceeding (ACM ISBN: 978-1-59593-846-6) published by ACM. Readers are also encouraged to refer to a special Trends & Controversies Department on "domain driven, actionable knowledge discovery" with IEEE Intelligent Systems magazine [1]. Extended versions of selected papers from the workshop plus other invitations will soon appear in

a volume of Springer's Lecture Notes in Computer Sciences Series. We expect broader and deeper studies in the area triggered by but not limited to this workshop.

3.1 Acknowledgements

This workshop is partially supported by Australian Research Council Discovery Grant (DP0773412, LP0775041, DP0667060). Yanchang Zhao is an Australian Post-Doctoral Industry fellow of the Australian Research Council Linkage grant.

4. REFERENCES

[1] Cao, L., Zhang, C., Yu, P., et al. Domain-Driven actionable knowledge discovery, *IEEE Intelligent Systems*, 22(4): 78-89, 2007.

[2] Cao, L., Zhang, C. The evolution of KDD: Towards domain-driven data mining. *Int. J. of Pattern Recognition and Artificial Intelligence*, 21(4): 677-692, 2007.

[3] Cao, L., Zhang, C. Domain-driven data mining, *Advances in Data Warehousing and Mining*, IGI Publisher, 2007.

[4] Domingos, P., Toward knowledge-rich data mining, *Data Mining and Knowledge Discovery: An International Journal*, 15(1): 21-28, 2007.

[5] Fayyad, U., Shapiro G., Uthurusamy R., Summary from the KDD-03 panel Data mining: the next 10 years, *ACM SIGKDD Explorations Newsletter*, 5(2): 191-196, 2003.

[6] 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.

[7] Han, J. Towards Human-Centered, Constraint-Based, Multi-Dimensional Data Mining, *An invited talk at Univ. Minnesota, Minneapolis*, Minnesota, 1999.

[8] Kriegel, H., Borgwardt, K., Kroger, P., Pryakhin, A., Schubert, M. and Zimek, A. Future trends in data mining, *Data Mining and Knowledge Discovery: An International Journal*, 15(1): 87-97, 2007.

[9] Liu, B., Hsu, W., Mun, L., and Lee, H. Finding Interesting Patterns Using User Expectations, *IEEE Transactions on Knowledge and Data Engineering*, 11(6): 817-832, 1999.

[10] Shapiro,G-P., Djeraba, C., Getoor, L., Grossman, R., Feldman R., and Zaki M. What Are The Grand Challenges for Data Mining? KDD-2006 Panel Report, *ACM SIGKDD Explorations Newsletter*, 8(2): 70-77, 2006.

[11] Shapiro, G.P., Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from "university" to "business" and "analytics", *Data Mining and Knowledge Discovery: An International Journal*, 15(1): 99-105, 2007.

[12] Tan, P., Kumar, V., Srivastava, J. Selecting the Right Interestingness Measure for Association Patterns, *Proceedings of SIGKDD02*, 15(1): 32-41, 2002.