Domain Driven Data Mining

Longbing Cao · Philip S. Yu · Chengqi Zhang · Yanchang Zhao

Domain Driven Data Mining

Springer

Longbing Cao
University of Technology,
Sydney
Fac. Engineering & Information Tech.
Centre for Quantum Computation
and Intelligent Systems
Broadway NSW 2007
Australia
lbcao@it.uts.edu.au

Philip S. Yu
Department of Computer Science
University of Illinois at Chicago
851 S. Morgan St.
Chicago IL 60607-7053
USA
psyu@cs.uic.edu

Chengqi Zhang
University of Technology,
Sydney
Fac. Engineering & Information Tech.
Centre for Quantum Computation
and Intelligent Systems
Broadway NSW 2007
Australia
chengqi@it.uts.edu.au

Yanchang Zhao
University of Technology,
Sydney
Fac. Engineering & lnformation Tech.
Centre for Quantum Computation
and Intelligent Systems
Broadway NSW 2007
Australia
yczhao@it.uts.edu.au

ISBN 978-1-4419-5736-8 e-ISBN 978-1-4419-5737-5
DOI 10.1007/978-1-4419-5737-5
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2009942454

To Sabrina Yue Cao and Bobby Yu Cao for their time, the peace and understanding they have given during writing this book.

Preface

Data mining has emerged as one of the most active areas in information and communication technologies (ICT). With the booming of the global economy, and ubiquitous computing and networking across every sector and business, data and its deep analysis becomes a particularly important issue for enhancing the soft power of an organization, its production systems, decision-making and performance. The last ten years have seen ever-increasing applications of data mining in business, government, social networks and the like.

However, a crucial problem that prevents data mining from playing a strategic decision-support role in ICT is its usually limited decision-support power in the real world. Typical concerns include its actionability, workability, transferability, and the trustworthy, dependable, repeatable, operable and explainable capabilities of data mining algorithms, tools and outputs.

This monograph, *Domain Driven Data Mining*, is motivated by the real-world challenges to and complexities of the current KDD methodologies and techniques, which are critical issues faced by data mining, as well as the findings, thoughts and lessons learned in conducting several large-scale real-world data mining business applications. The aim and objective of domain driven data mining is to study effective and efficient methodologies, techniques, tools, and applications that can discover and deliver actionable knowledge that can be passed on to business people for direct decision-making and action-taking.

In deploying current data mining algorithms and techniques into real-world problem-solving and decision-making, we have faced the crucial need to bridge the gap between academia and business, as well as addressing the gap between technical evaluation systems and real business needs. We have been confronted by the extreme imbalance between the large number of algorithms published versus the very few that are deployed in a business setting; the large number of patterns mined versus the few that satisfy business interests and needs; and many patterns identified versus the lack of recommended decision-support actions.

To bridge the above-mentioned gaps, and to narrow the extreme imbalance, it is crucial to amplify the decision-support power of data mining. Most importantly, it is

critical to enhance the actionability of the identified patterns, and to deliver findings that can support decision-making. These are the drivers of this book.

Our purpose is to explore the directions and possibilities for enhancing the decision-support power of data mining and knowledge discovery. The book is organized as follows. In Chapter one, we summarize the main challenges and issues surrounding the traditional data mining methodologies and techniques, and the trends and opportunities for promoting a paradigm shift from data-centered hidden pattern mining to domain-driven actionable knowledge delivery. Chapter two presents the domain-driven data mining methodologies. From Chapters three to five, we mainly extend the discussions about domain-driven data mining methodologies. In Chapter three, ubiquitous intelligence surrounding enterprise data mining is considered. Chapter four discusses knowledge actionability, while Chapter five summarizes several types of system frameworks for actionable knowledge delivery. Chapters six to eight present several techniques supporting domain-driven data mining. Chapter six introduces the concept of combined mining, leading to combined patterns that can be more informative and actionable. In Chapter seven, we discuss agent-driven data mining, which can enhance the power of mining complex data. Chapter eight summarizes the technique of post mining for enhancing knowledge power through post-processing of identified patterns. Chapters nine and ten illustrate the use of domain driven data mining in the real world. In Chapter nine, domain-driven data mining is used to identify actionable trading strategies and actionable market microstructure behavior patterns in capital markets. Chapter ten utilizes domain-driven data mining in identifying actionable combined associations and combined patterns in social security data. Chapter eleven lists some of the open issues and discusses trends in domain-driven data mining research and development. Chapter twelve lists materials and references about domain-driven data mining.

A typical trend in real-world data mining applications is to treat a data mining system as a problem-solving system within a certain environment. Looking at the problem-solving from the domain-driven perspective, many open issues and opportunities arise, indicating the need for next-generation data mining and knowledge discovery far beyond the data mining algorithms themselves. We realize that we are not at the stage for covering every aspect of these open issues and opportunities. Rather, it is our intention to raise them in this book for wider, deeper and more substantial investigation by the community.

We would like to convey our appreciation to all contributors, including Ms. Melissa Fearon and Ms. Jennifer Maurer from Springer US, for their kind support and great effort in bringing the book to fruition.

Chicago, USA
July 2009 *Philip S Yu*

Acknowledgements

Our special thanks to the following members: Dr Huaifeng Zhang, Dr Yuming Ou and Dr Dan Luo at the Data Sciences and Knowledge Discovery Lab (the Smart Lab), Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney, Australia for their support in time, experiments and discussions. We thank the Smart Lab for the environment, projects, funding, and support for the initiation of the research and development on domain driven data mining.

Our thanks go to our industry partners, in particular the Australian Commonwealth Government Agency Centrelink, the Capital Markets Cooperative Research Centre, the Shanghai Stock Exchange, and HCF Australia, for their partnership and contributions in terms of funding, data, domain knowledge, evaluation and validation in developing and applying domain driven data mining methodologies and techniques in the business problem solving.

Last but not least, we thank Springer, in particular, Ms. Melissa Fearon and Ms. Jennifer Maurer at Springer US for their kindness in supporting the publication of this monograph, and its sister book, *Data Mining for Business Applications*, edited by Longbing Cao, Philip S Yu, Chengqi Zhang and Huaifeng Zhang in 2008.

We appreciate all colleagues, contributors and reviewers for their kind contributions to the professional activities related to domain driven data mining (DDDM or D^3M), including the DDDM workshop series and special issues.

Contents