

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 & Information 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

© Springer Science+Business Media, LLC 2010

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden. The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

*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.

July 2009

Chicago, USA
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

1	Challenges and Trends	1
1.1	Introduction	1
1.2	KDD Evolution	2
1.3	Challenges and Issues	3
1.3.1	Issues of Traditional Data Mining Studies	4
1.3.2	Related Efforts on Tackling Traditional Data Mining Issues	6
1.3.3	Overlooking Ubiquitous Intelligence	7
1.3.4	Organizational and Social Factors	8
1.3.5	Human Involvement	9
1.3.6	Domain Factors	9
1.3.7	Knowledge Decision Power	10
1.3.8	Decision-Support Knowledge Delivery	10
1.4	KDD Paradigm Shift	11
1.4.1	Data-Centered Interesting Pattern Mining	11
1.4.2	From Data Mining to Knowledge Discovery	12
1.4.3	Multi-Dimensional Requirements on Actionable Knowledge Delivery	13
1.4.4	From Data-Centered Hidden Knowledge Discovery to Domain Driven Actionable Knowledge Delivery	15
1.4.5	D^3M : Domain Driven Actionable Knowledge Delivery	16
1.5	Towards Domain Driven Data Mining	17
1.5.1	The D^3M Methodology	18
1.5.2	Problem: Domain-Free vs. Domain-Specific	19
1.5.3	KDD Context: Unconstrained vs. Constrained	20
1.5.4	Interestingness: Technical vs. Business	22
1.5.5	Pattern: General vs. Actionable	23
1.5.6	Infrastructure: Automated vs. Human-Mining-Cooperated	24
1.6	Summary	25

2	<i>D</i>³<i>M</i> Methodology	27
2.1	Introduction	27
2.2	<i>D</i> ³ <i>M</i> Methodology Concept Map	27
2.3	<i>D</i> ³ <i>M</i> Key Components	28
2.3.1	Constrained Knowledge Delivery Environment	29
2.3.2	Considering Ubiquitous Intelligence	31
2.3.3	Cooperation between Human and KDD Systems	33
2.3.4	Interactive and Parallel KDD Support	34
2.3.5	Mining In-Depth Patterns	35
2.3.6	Enhancing Knowledge Actionability	36
2.3.7	Reference Model	37
2.3.8	Qualitative Research	38
2.3.9	Closed-Loop and Iterative Refinement	38
2.4	<i>D</i> ³ <i>M</i> Methodological Framework	40
2.4.1	Theoretical Underpinnings	40
2.4.2	Process Model	41
2.4.3	<i>D</i> ³ <i>M</i> Evaluation System	44
2.4.4	<i>D</i> ³ <i>M</i> Delivery System	46
2.5	Summary	47
3	Ubiquitous Intelligence	49
3.1	Introduction	49
3.2	Data Intelligence	49
3.2.1	What is data intelligence	49
3.2.2	Aims of involving data intelligence	50
3.2.3	Aspects of data intelligence	50
3.2.4	Techniques disclosing data intelligence	51
3.2.5	An example	52
3.3	Domain Intelligence	55
3.3.1	What is domain intelligence	55
3.3.2	Aims of involving domain intelligence	56
3.3.3	Aspects of domain intelligence	56
3.3.4	Techniques involving domain intelligence	57
3.3.5	Ontology-Based Domain Knowledge Involvement	57
3.4	Network Intelligence	59
3.4.1	What is network intelligence	59
3.4.2	Aims of involving network intelligence	59
3.4.3	Aspects of network intelligence	60
3.4.4	Techniques for involving network intelligence	60
3.4.5	An example of involving network intelligence	61
3.5	Human Intelligence	62
3.5.1	What is human intelligence	62
3.5.2	Aims of involving human intelligence	62
3.5.3	Aspects of human intelligence	63
3.5.4	Techniques for involving human intelligence	64

3.5.5	An example	64
3.6	Organizational Intelligence	65
3.6.1	What is organizational intelligence	65
3.6.2	Aims of involving organizational intelligence	66
3.6.3	Aspects of organizational intelligence	66
3.6.4	Techniques for involving organizational intelligence	67
3.7	Social Intelligence	67
3.7.1	What is social intelligence	67
3.7.2	Aims of involving social intelligence	68
3.7.3	Aspects of social intelligence	68
3.7.4	Techniques for involving social intelligence	69
3.8	Involving ubiquitous intelligence	69
3.8.1	The way of involving ubiquitous intelligence	69
3.8.2	Methodologies for involving ubiquitous intelligence	70
3.8.3	Intelligence Meta-synthesis of ubiquitous intelligence	71
3.9	Summary	72
4	Knowledge Actionability	75
4.1	Introduction	75
4.2	Why Knowledge Actionability	76
4.3	Related Work	77
4.4	Knowledge Actionability Framework	78
4.4.1	From Technical Significance to Knowledge Actionability ...	79
4.4.2	Measuring Knowledge Actionability	81
4.4.3	Pattern Conflict of Interest	83
4.4.4	Developing Business Interestingness	85
4.5	Aggregating Technical and Business Interestingness	87
4.6	Summary	90
5	<i>D</i>³<i>M</i> AKD Frameworks	93
5.1	Introduction	93
5.2	Why AKD Frameworks	94
5.3	Related Work	96
5.4	A System View of Actionable Knowledge Discovery	97
5.5	Actionable Knowledge Discovery Frameworks	101
5.5.1	Post Analysis Based AKD: PA-AKD	101
5.5.2	Unified Interestingness Based AKD: UI-AKD	102
5.5.3	Combined Mining Based AKD: CM-AKD	104
5.5.4	Multi-Source + Combined Mining Based AKD: MSCM-AKD	107
5.6	Case Studies	109
5.7	Discussions	110
5.8	Summary	112

6	Combined Mining	113
6.1	Introduction	113
6.2	Why Combined Mining	114
6.3	Problem Statement	117
6.3.1	An Example	117
6.3.2	Mining Combined Patterns	120
6.4	The Concept of Combined Mining	121
6.4.1	Basic Concepts	121
6.4.2	Basic Paradigms	123
6.4.3	Basic Process	124
6.5	Multi-Feature Combined Mining	126
6.5.1	Multi-Feature Combined Patterns	126
6.5.2	Pair Pattern	128
6.5.3	Cluster Pattern	129
6.5.4	Incremental Pair Pattern	129
6.5.5	Incremental Cluster Pattern	130
6.5.6	Procedure for Generating Multi-Feature Combined Patterns	131
6.6	Multi-Method Combined Mining	132
6.6.1	Basic Frameworks	132
6.6.2	Parallel Multi-Method Combined Mining	133
6.6.3	Serial Multi-Method Combined Mining	134
6.6.4	Closed-Loop Multi-Method Combined Mining	134
6.6.5	Closed-Loop Sequence Classification	136
6.7	Case Study: Mining Combined Patterns in E-Government Service Data	139
6.8	Related Work	139
6.9	Summary	142
7	Agent-Driven Data Mining	145
7.1	Introduction	145
7.2	Complementation between Agents and Data Mining	145
7.3	The Field of Agent Mining	147
7.4	Why Agent-Driven Data Mining	150
7.5	What Can Agents Do for Data Mining?	152
7.6	Agent-Driven Distributed Data Mining	154
7.6.1	The Challenges of Distributed Data Mining	154
7.6.2	What Can Agents Do for Distributed Data Mining?	154
7.6.3	Related Work	156
7.7	Research Issues in Agent Driven Data Mining	159
7.8	Case Study 1: F-Trade – An Agent-Mining Symbiont for Financial Services	160
7.9	Case Study 2: Agent-based Multi-source Data Mining	161
7.10	Case Study 3: Agent-based Adaptive Behavior Pattern Mining by HMM	162
7.10.1	System Framework	162

7.10.2	Agent-Based Adaptive CHMM	165
7.11	Research Resources on Agent Mining	167
7.11.1	The AMII Special Interest Group	167
7.11.2	Related References	168
7.12	Summary	168
8	Post Mining	171
8.1	Introduction	171
8.2	Interestingness Measures	172
8.3	Filtering and Pruning	174
8.4	Visualisation	176
8.5	Summarization and Representation	177
8.6	Post-Analysis	178
8.7	Maintenance	179
8.8	Summary	180
9	Mining Actionable Knowledge on Capital Market Data	181
9.1	Case Study 1: Extracting Actionable Trading Strategies	181
9.1.1	Related Work	181
9.1.2	What Is Actionable Trading Strategy?	182
9.1.3	Constraints on Actionable Trading Strategy Development	185
9.1.4	Methods for Developing Actionable Trading Strategies	189
9.2	Case Study 2: Mining Actionable Market Microstructure Behavior Patterns	196
9.2.1	Market Microstructure Behavior in Capital Markets	196
9.2.2	Modeling Market Microstructure Behavior to Construct Microstructure Behavioral Data	196
9.2.3	Mining Microstructure Behavior Patterns	199
9.2.4	Experiments	201
10	Mining Actionable Knowledge on Social Security Data	203
10.1	Case Study: Mining Actionable Combined Associations	203
10.1.1	Overview	203
10.1.2	Combined Associations and Association Clusters	203
10.1.3	Selecting Interesting Combined Associations and Association Clusters	205
10.2	Experiments: Mining Actionable Combined Patterns	207
10.2.1	Mining Multi-Feature Combined Patterns	208
10.2.2	Mining Closed-Loop Sequence Classifiers	213
10.3	Summary	215
11	Open Issues and Prospects	217
11.1	Open Issues	217
11.2	Trends and Prospects	218

12 Reading Materials	221
12.1 Activities on D^3M	221
12.2 References on D^3M	222
12.3 References on Agent Mining	223
12.4 References on Post-analysis and Post-mining	223
Glossary	225
Reference	233
References	233
Index	245