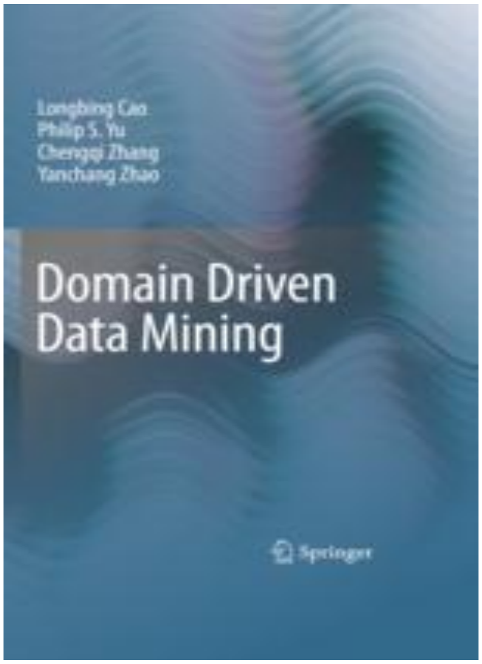
Domain Driven Data Mining

BY LONGBING CAO, PHILIP S. YU, CHENGQI ZHANG AND YANCHANG ZHAO.



REVIEWED BY
NORLAILA HUSSAIN
HELEN ZHOU

Data mining is a powerful paradigm of extracting information from data. It can help enterprises focus on important information in their data warehouse. Data mining is also known as *Knowledge Discovery in Databases* (KDD). It involves the extraction of hidden pattern to predict future trends and behaviors which allow businesses to make proactive, knowledge-driven decisions.

The current vast development in ubiquitous computing, cloud computing and networking across every sector and business has made data mining emerging as one of the most active areas in information and communication technologies (ICT) as data and its deep analysis becomes an important issue for enhancing the soft power of an organization, its production systems, decision making and performance.

However, there is a large gap has been identified by many studies between academic deliverables and business expectations, as well as between data miners and business analysts. The limited decision-support power of data mining in the real world has prevented it from playing a strategic decision-support role in ICT. The main concerns include the actionability, workability, transferability, and the trustworthy, dependable, repeatable, operable and explainable capabilities of data mining algorithms, tools and outputs.

Nevertheless, these challenges create opportunities for promoting a paradigm shift from data-centered hidden pattern mining to domain-driven actionable knowledge delivery. These real-world concerns and complexities of the KDD methodologies and techniques have motivated Cao et al. (2010) to propose domain driven data mining (D^3M) as effective and practical methodologies for actionable knowledge discovery in order to narrow down and bridge the gap between the academia and the business people. This proposal is elaborated in great length in their latest book “Domain Driven Data Mining”.

Domain driven data mining involves the study of effective and efficient methodologies, techniques, tools, and applications which can discover and deliver actionable knowledge that can be passed on to business people for direct decision-making and action-taking.

The book begins by highlighting the gap that exists between academia and business in Chapter 1. This gap includes the large numbers of algorithms published by academia versus only a few are deployed in a business setting. In addition, despite the large number of patterns mined or identified, only a few satisfy business needs and lack of recommended decision-support actions. They stressed that the algorithms models and resulting patterns and knowledge are short of workable, actionable and operable capabilities. The authors went on to summarize the main challenges and technical issues surrounding the traditional data mining and knowledge discovery methodologies.

To address the issues highlighted, the authors introduce the main components and methodological framework of D^3M methodologies in Chapter 2. Based on authors’ real world experiences and lessons learned in a capital market, significance results were discovered when domain factors are considered in data mining. An overall picture of D^3M focusing on the concept map, the key methodological components, the theoretical underpinning and the process model were outlined.

The discussions on domain-driven data mining methodologies are further elaborated in Chapter 3 to 5. The authors elaborated the importance of involving and consolidating relevant ubiquitous intelligence (i.e. data intelligence, human intelligence, domain intelligence, network and web intelligence, and organizational and social intelligence) surrounding data mining applications for actionable knowledge discovery and delivery. The definitions, aims, aspects and techniques for involving this ubiquitous intelligence into data mining are identified in Chapter 3.

A key concept in D^3M that is highlighted is *actionable knowledge discovery* (AKD). It involves and synthesizes domain intelligence, human intelligence and cooperation, network intelligence and in-depth data intelligence to define, measure, and evaluate business interestingness and knowledge actionability. The authors stressed the importance of AKD as an important concept for bridging the gap between technical-based approaches and business impact-oriented expectations on patterns discovered from data mining. This concept is elaborated in Chapter 4.

Four types of system frameworks for actionable knowledge delivery are then introduced in Chapter 5. The frameworks include *PA-AKD* (a two-step AKD process), *UI-AKD* (based on unified interestingness), *CM-AKD* (a multi-step AKD process), and *MSCM-AKD* (based on multiple data sources). The authors describe the flexibility of the proposed frameworks which can cover many common problems and applications and are effective in extracting knowledge that can be used by business people for immediate decision-making.

Chapters 6 to 8 outline several techniques supporting domain-driven data mining. Chapter 6 presents a comprehensive and general approach named *combined mining* for handling multiple large heterogeneous data sources targeting more informative and actionable knowledge. The authors describe this approach as a framework for mining complex knowledge in complex data where many mutative applications can be designed such as combined pattern mining in multiple data sources. They focus on providing general frameworks and approaches to handle multi-feature, multi-source and multi-method issues and requirements.

In Chapter 7, the authors introduce agent-driven data mining for D^3M. The basic concept, driving forces, technical means, research issues and case studies of agent-driven data mining are discussed. The authors suggest the interaction and integration between agents and data mining are necessary as agent technology can greatly complement data mining in complex data mining problems in situations such as data processing, information processing, user modeling and interaction, infrastructure and services.

Chapter 8 elaborates the technique of post analysis and post mining. This technique helps to refine discovered patterns and learned models and present useful and applicable knowledge to users. It uses visualization techniques which present the knowledge desired by the end users and which is easy to read and understand. The authors discuss interesting measures, pruning, selection, summarization, visualisation and maintenance of patterns.

To assist readers in understanding D^3M further, the authors continue to illustrate the use of domain driven data mining in the real world. In Chapter 9, the authors describe how domain-driven data mining is applied to identify actionable trading strategies and actionable market microstructure behavior patterns in capital markets. They elaborate some case studies in which this methodology has been used for smart trading, and mining for deeply understanding of exceptional trading behaviour on capital market data.

Chapter 10 utilizes domain-driven data mining in identifying actionable combined associations and combined patterns in social security data. It illustrates the use of domain driven data for better understanding government service quality, causes and effects of government service problems, customer behaviour and demographics, and government officer-customer interactions. The case study introduces several examples using the *MSCM-AKD* framework in identifying combined associations and combined associations clusters for debt prevention.

The final chapter summarizes some of the open issues and discusses trends in domain-driven data mining research and development. The authors highlighted several fundamental problems that need further investigation such as supporting social interaction and cognition in data mining and making data mining trustful and business-friendly. They also suggest the need for next-generation data mining and knowledge discovery that is far beyond the data mining algorithms as there are many open issues and opportunities arise when problem-solving is viewed from the domain-driven perspective.
Overall, the book is well-written and reading it has been an enjoyable one. The authors present interesting issues and opportunities for further exploration of data mining in the future. The main focus of the book is to demonstrate some new techniques to amplify the decision-support power of data mining and they have certainly succeeded.

THE BOOK:

CAO, PHILIP S. YU, CHENGQI ZHANG AND YANCHANG ZHAO (2010) DOMAIN DRIVEN DATA MINING, 1ST EDITION. 2010, XIII, 237 P. SPRINGER.
ISBN: 978-1-4419-5736-8

ABOUT THE REVIEWERS:

NORLAILA HUSSAIN
School of Engineering and Advanced Technology, Massey University, New Zealand. Contact her at:
N.Hussain@massey.ac.nz

HELEN ZHOU
School of Electrical Engineering, Manukau Institute of Technology, Auckland, New Zealand. Contact her at:
helen.zhou@manukau.ac.nz