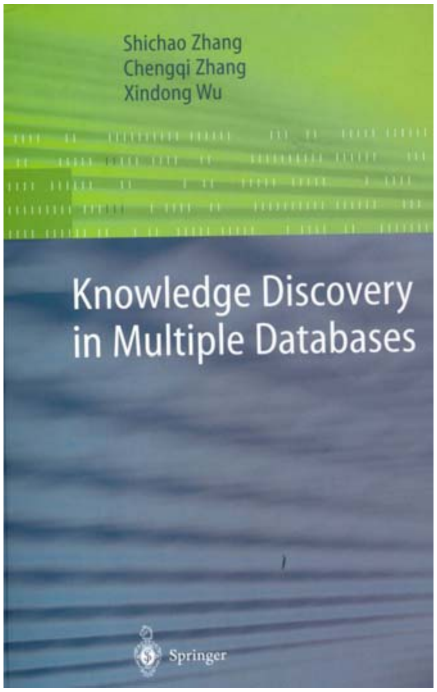
Knowledge Discovery in Multiple Databases

BY SHICHAO ZHANG, CHENGQI ZHANG, XINDONG WU, 2004. ISBN: 978-1-85233-703-2



REVIEW BY RAMESH K. RAYUDU*

In this digital world, we are inundated with terrabytes of data every day. The databases are being stored in distributed environment as companies become global and have offices all over the world. This lead to an interesting research problem in the field of data mining termed multi-database mining. Most solutions in this regard specify merging the databases into a single dataset. But this kind of merging can lead to many problematic issues such as data explosion, destruction of database distribution information, loss of data, and unnecessary inclusion of some data. In simple words, a single dataset may not reflect the real nature of multi-datasets.

"Knowledge Discovery in Multiple Databases" written by S. Zhang, C. Zhang and X. Wu is a book that addresses the issue of data-mining multiple databases. The book is a description of authors' research work and development of a new strategy termed local pattern analysis.

Local pattern analysis is claimed to discover useful patterns that cannot be mined in traditional multi-database mining techniques. The book discusses knowledge discovery principles at different levels of detail. Novice dataminers, researchers, academics and students will find the book helpful.

Functionally "Knowledge Discovery in Multiple Databases" is organised in two parts. The first half introduces knowledge discovery, multi-databases and some related research. The second half discusses the authors' techniques for pre-processing the data and identifying patterns from multi-databases. The authors provide a smooth transition from introducing a reader with multi-database mining (MDM) to their application and research. The authors also point out some very important shortcomings of earlier research and provide a good insight into the practical aspects of MDM.

Chapter 1 provides an introduction to data mining and discusses association rule mining in a format that keeps the reader interested in the topic. The chapter provided a good insight into mining aspect of both single and multi-databases. Authors' statement that dual-level applications present many challenges could be clearly identified from the reading the chapter.

The authors clearly identified the practical issues of a MDM process with emphasis on design of application independent database clustering. Identifying quality knowledge and resolving conflicts are important aspects of any MDM and challenge the boundaries of MDM research. Authors address these topics and also provide solutions. The chapter ends with authors' discussion on identifying the features of MDM and contrasting their research against each feature.

Knowledge discovery in databases (KDD) has been an active research area since mid-nineties and, since then, several books have been published in the area. The authors' also discuss this topic in their second chapter of the book. They start with the traditional approach of discussing the processing steps involved in KDD and continue on to discuss the latest research in each process. From this broad overview, the authors narrow their discussion on to association rule mining to discuss its effects on mining mono-databases. The final part of the chapter discusses the relevant research into MDM. Several algorithms including meta-learning and parallel datamining were discussed in appropriate detail so that a reader can understand the concepts.

Chapter 3 introduces authors' Local Pattern Analysis (LPA) strategy that is based on a competing model in sports. As each sport has a set of rules to choose its winners, LPA was developed to recognize patterns in multi-databases based on a dual-level multi-rule strategy. Through the strategy, they identify the three useful patterns: high-vote patterns, exceptional and global patterns.

To recognize patterns in multi-databases, the authors demonstrate the structure of a pattern and represent it in a multi-dimension space where each dimension is a selection factor. The details of the algorithm were avoided in the chapter and were dealt in the

*Massey University, New Zealand.
E-mail: r.k.rayudu@massey.ac.nz

subsequent chapters. In the final section of the chapter, the authors demonstrate and discuss the effectiveness of their algorithm LPA.

Chapter 4 is dedicated towards detection of quality knowledge in high-veridical data sources. The authors demonstrate the effectiveness of identifying quality data by considering one internal and six external databases as an example and then apply several existing techniques on the databases. They demonstrate that the technique discussed in the book can successfully detect the frequent itemsets while still preserving the distributive nature of the databases. The basic techniques that are necessary for identifying quality data has been discussed extensively in the chapter.

They developed several semantics to state that a veridical data-source combines the collected knowledge with a set of possibilities to obtain a higher level knowledge. The authors readily acknowledge that there may not be veridical data sources in the real-world but veridical properties can originate from other sources. Their developed framework was then applied to real-world databases. The application demonstrated that the stated algorithm can successfully identify data-sources with high success ratio based on their veridicality. The authors' claims that their algorithm works by distinguishing internal and external knowledge and the elimination of untrustworthy and fraudulent knowledge by veridicality analysis can be established through this chapter.

Chapter 5 is dedicated to identification of relevant databases for a datamining application. The authors use classification techniques to identify relevant databases. Classification is a challenging process and depends extensively on selected features related to an object being classified. The authors develop a new clustering algorithm that is application independent and utilizes MDM.

To search for a good classification from given multiple databases a two-step process is stated in this chapter. The first step is to design a procedure that generates a classification for a given threshold. The second step is to develop an algorithm that can search for a good classification based on a distance measure that measures the goodness of a database class. The developed algorithm was then applied to a set of databases and their results analyzed. The performance improvement is impressive and certainly advanced from other algorithms.

Decision making systems that use negative association rules to identify the mutually exclusive correlations among data items can create a problem when dealing with multi-databases. Identifying these conflicts and resolving them is an important aspect and is the topic of Chapter 6 of the book. The authors address the issue by introducing a local pattern synthesizing operator that can identify the local pattern set and resolves inconsistency using the weighted majority principle. The authors review some basic concepts of modal logic and construct a proof theory of the proposed logic. The last section of the chapter discusses on how to use the proposed logic framework to identify quality knowledge from multiple databases.

Chapters 7 and 8 discuss the detection of high-vote and exceptional patterns. To identify a high-vote pattern, the authors introduce a voting rate that is calculates voting of each branch. The measure of interest of a high voting pattern is stemmed from the relationship between the voting rate and the average voting rate (random pattern area). A fuzzy logic controller is then used to further identify the high-vote patterns.

While high-vote patterns show the commonality between different branches of a company, exceptional patterns depict the patterns that are unique to each branch. The authors' algorithm for identifying exceptional patterns was discussed in Chapter 8. The exceptional patterns were identified by a 'measure of interestingness' developed by the authors. After a good discussion of the algorithm, they demonstrate the algorithm using an example.

The major highlight of this book is the discussion of algorithms in relation to many practical problems and the functional hierarchy of a company. The intricacies and problems related to a multi-branch organization were discussed and related with the algorithms specified in the book. One such intricacy is the importance given to different branches in a company. The authors consider this importance and incorporate it into their model for synthesizing global patterns from local patterns.

As specified in Chapter 6 they use a weighting measure to achieve the synthesis. The resulting algorithm is then discussed in Chapter 9 of the book. The stated model synthesizes the association rules from multiple databases using a weighting process. The weighting process is elaborate and well discussed in the book. The algorithm developed is stated to synthesize rules from different databases and different databases can be mined concurrently.

The final chapter highlights the book's contributions and addresses the future issues related to the research.

The authors, through this book, provide a practical and logical approach to solving problems related to knowledge discovery in multi-database systems. Data mining multiple databases is a complex task and needs a proper direction. This book provides that direction. The authors addressed the problem well and discussed their solutions systematically.

One highlight of the book is the discussion on authors' creation of several MDM techniques and methodologies towards solving some practical problems. The primary focus of the book is to demonstrate some new techniques in mining multi-databases and the authors have certainly succeeded.