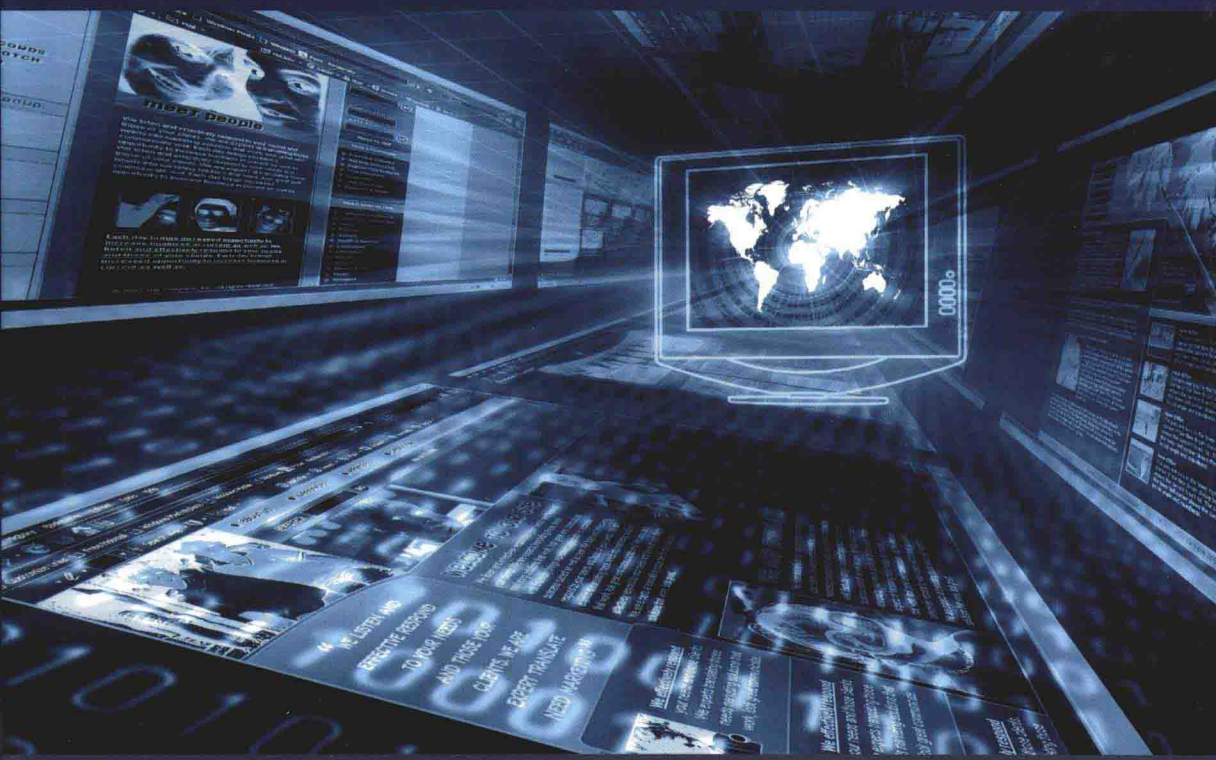
Data Mining and Data Warehousing

S.K. Mourya • Shalu Gupta



Alpha Science

Data Mining and Data Warehousing

S. K. Mourya
Shalu Gupta

Alpha Science International Ltd.
Oxford, U.K.

Data Mining and Data Warehousing
214 pgs. | 85 figs. | 7 tbls.

S. K. Mourya
Shalu Gupta
Department of Computer Science and Engineering
MGM's College of Engneering & Technology
Noida

ALPHA SCIENCE INTERNATIONAL LTD.
7200 The Quorum, Oxford Business Park North
Garsington Road, Oxford OX4 2JZ, U.K.

www.alphasci.com

Printed form the camera-ready copy provided by the Authors.

ISBN 978-1-84265-757-7

Printed in India

Data Mining and Data Warehousing

Dedicated

To our parents, grand parents
and
most beloved
Shilu
and our children

S. K. Mourya
Shalu Gupta

Preface

The explosion of information technology, which continues to expand data driven markets and business, has made data mining an even more relevant topic of study. Books on data mining tend to be either broad or introductory or focus on some very specific technical aspect of the field.

Data Mining and Data Warehousing in nine chapters explores in depth the core of data mining (classification, clustering and association rules) by offering overviews that include both analysis and insight. Written for graduate students from various parts of the country studying Data Mining courses. The book is an ideal companion to either an introductory Data Mining textbook or a technical Data Mining book.

Unlike many other books that mainly focus on the modeling part, this volume discusses all the important—and often neglected—parts before and after modeling.

The book is organized as follows. It is divided into nine chapters.

In **Chapter 1,** a brief introduction to data mining offers great promise in helping organizations uncover patterns hidden in their data that can be used to predict the behaviour of customers, products and processes, various forms of data preprocessing, data cleaning, missing values, noisy data etc. In this chapter various classifications and various issues of data mining have also been explained with examples to illustrate them.

Chapter 2 deals with explaining data preprocessing and the various other needs of data processing. Forms of data preprocessing, e.g., data cleaning, missing values, noisy data. In this chapter we have also discussed how to handle inconsistent data, data integration and transformation.

Chapter 3 deals with explanation how statistics measures are used in large databases through measuring of central tendency, measuring dispersion of data in data mining, some graphical techniques used in data analysis of continuous data, etc. Further in the chapter we have also discussed data cube approach (OLAP).

Chapter 4 deals with discovery of frequent patterns, association, and correlation relationships among huge amounts of data, how it is useful in selective marketing, decision analysis, and business management. A popular area of application is market basket analysis which studies the buying habits of customers by searching for sets of items that are frequently purchased together (or in sequence). Association rule mining that consists of first finding frequent item sets from which strong association rules in the form of A=>B are generated, has also been dealt with.

In **Chapter 5,** we have explained how classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. While classification predicts categorical labels (classes), prediction models deal with continuous-valued functions.

Chapter 6 explains that many clustering algorithms have been developed. These can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, model-based methods, methods for high-dimensional data (including frequent pattern–based methods), and constraint based methods. Some algorithms may belong to more than one category.

In **Chapter7,** we study a well-accepted definition of the data warehouse and see why more and more organizations are building data warehouses for the analysis of their data. In particular, we study the *data cube*, a multidimensional data model for data warehouses, architecture and design. Through this chapter we can study that clearly, the presence of a data warehouse is a very useful precursor to data mining, and if it is not available, many of the steps involved in data warehousing will have to be undertaken to prepare the data for mining.

Chapter 8 presents an overview of OLAP technology aggregation, efficient query facility and its multidimensional aspects. Such an overview is essential for understanding the overall data mining and knowledge discovery process.

Chapter 9 deals with various issues of privacy and security of data emerged at a relatively early stage in the development of data mining. This development is not at all surprising given that all activities of data mining revolve around data and many sensitive issues of accessibility or possible reconstruction of data records exist, along with backup and testing concerns.

S. K. Mourya
Shalu Gupta

Acknowledgements

"Few successful endeavors have ever been made by one person alone, and this book is no exception."

We would like to thank the people and organizations that supported us in the production of this book and the authors are greatly indebted to all. There are several individuals and organizations whose support demands special mention and they are listed in the following.

Our special thanks to MGM's family Mr. Kamalkishor N. Kadam (Chairman), Dr. Mrs. Geeta S. Lathkar (Director), Prof. Sunil Wagh (VP) for being our inspiration and enabling us to publish this book. Above all, we want to thank our colleagues of Computer Science and Engineering Department, MGM-Noida, our family and friends who supported and encouraged us in spite of all the time it took us away from them.

We also wish to thank Mr. N. K. Mehra, Director, Narosa Publishing House Pvt. Ltd., for his enthusiasm, patience, and support during our writing of this book. Our Production Editor, and his staff for their conscientious efforts regarding production, editing, proof-reading and design etc. also deserve our special thanks.

We would like to express our gratitude to the many people who saw us through this book and to all of the reviewers for their invaluable feedback.

S. K. Mourya
Shalu Gupta

Contents