

经 典 原 版 书 库

# 数据挖掘

概念与技术

(美) Jiawei Han 著  
伊利诺伊大学厄巴纳-尚佩恩分校

(加) Micheline Kamber Jian Pei 著  
西蒙-弗雷泽大学

Third Edition



# DATA MINING

Concepts and Techniques

(英文版·第3版)

*Data Mining*  
Concepts and Techniques (Third Edition)

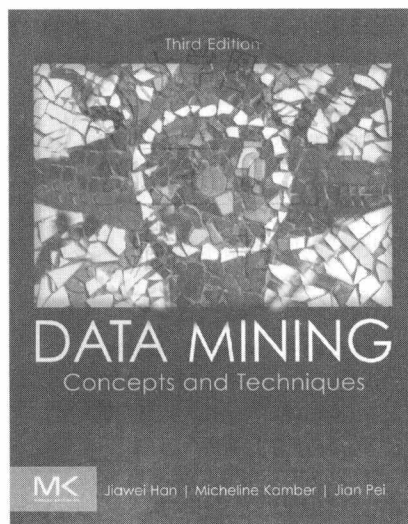
原 版 书

# 数据挖掘

概念与技术

(英文版·第3版)

*Data Mining*  
Concepts and Techniques (Third Edition)



(美) Jiawei Han 著 (加) Micheline Kamber Jian Pei 著  
伊利诺伊大学厄巴纳-尚佩恩分校 西蒙-弗雷泽大学



机械工业出版社  
China Machine Press

Jiawei Han, Micheline Kamber and Jian Pei: Data Mining: Concepts and Techniques, Third Edition (ISBN 978-0-12-381479-1).

Original English language edition copyright © 2012 by Elsevier Inc. All rights reserved.

Authorized English language reprint edition published by the Proprietor.

Copyright © 2012 by Elsevier (Singapore) Pte Ltd.

Printed in China by China Machine Press under special arrangement with Elsevier (Singapore) Pte Ltd. This edition is authorized for sale in China only, excluding Hong Kong, Macao SARs and Taiwan.

Unauthorized export of this edition is a violation of the Copyright Act. Violation of this Law is subject to Civil and Criminal Penalties.

本书英文影印版由 Elsevier (Singapore) Pte Ltd. 授权机械工业出版社在中国大陆境内独家发行。本版仅限在中国境内（不包括香港、澳门特别行政区及台湾地区）出版及标价销售。未经许可之出口，视为违反著作权法，将受法律之制裁。

封底无防伪标均为盗版

版权所有，侵权必究

本书法律顾问 北京市展达律师事务所

本书版权登记号：图字：01-2012-0222

图书在版编目（CIP）数据

数据挖掘：概念与技术（英文版·第3版）/（美）韩家炜，（加）坎伯（Kamber, M.），（加）裴健著. —北京：机械工业出版社，2012.3

（经典原版书库）

书名原文：Data Mining: Concepts and Techniques, Third Edition

ISBN 978-7-111-37431-2

I. 数… II. ①韩… ②坎… ③裴… III. 数据采集—英文 IV. TP274

中国版本图书馆 CIP 数据核字（2012）第 021422 号

机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码 100037）

责任编辑：迟振春

北京京师印务有限公司印刷

2012 年 3 月第 1 版第 1 次印刷

186mm×240mm·45.75 印张

标准书号：ISBN 978-7-111-37431-2

定价：118.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

客服热线：（010）88378991；88361066

购书热线：（010）68326294；88379649；68995259

投稿热线：（010）88379604

读者信箱：hzjsj@hzbook.com

# 出版者的话

文艺复兴以降，源远流长的科学精神和逐步形成的学术规范，使西方国家在自然科学的各个领域取得了垄断性的优势；也正是这样的传统，使美国在信息技术发展的六十多年间名家辈出、独领风骚。在商业化的进程中，美国的产业界与教育界越来越紧密地结合，计算机学科中的许多泰山北斗同时身处科研和教学的最前线，由此而产生的经典科学著作，不仅肇划了研究的范畴，还揭示了学术的源变，既遵循学术规范，又自有学者个性，其价值并不会因年月的流逝而减退。

近年，在全球信息化大潮的推动下，我国的计算机产业发展迅猛，对专业人才的需求日益迫切。这对计算机教育界和出版界都既是机遇，也是挑战；而专业教材的建设在教育战略上显得举足轻重。在我国信息技术发展时间较短的现状下，美国等发达国家在其计算机科学发展的几十年间积淀和发展的经典教材仍有许多值得借鉴之处。因此，引进一批国外优秀计算机教材将对我国计算机教育事业的发展起到积极的推动作用，也是与世界接轨、建设真正的世界一流大学的必由之路。

机械工业出版社华章公司较早意识到“出版要为教育服务”。自1998年开始，我们就将工作重点放在了遴选、移译国外优秀教材上。经过多年的不懈努力，我们与Pearson, McGraw-Hill, Elsevier, MIT, John Wiley & Sons, Cengage等世界著名出版公司建立了良好的合作关系，从他们现有的数百种教材中甄选出Andrew S. Tanenbaum, Bjarne Stroustrup, Brain W. Kernighan, Dennis Ritchie, Jim Gray, Afred V. Aho, John E. Hopcroft, Jeffrey D. Ullman, Abraham Silberschatz, William Stallings, Donald E. Knuth, John L. Hennessy, Larry L. Peterson等大师名家的一批经典作品，以“计算机科学丛书”为总称出版，供读者学习、研究及珍藏。大理石纹理的封面，也正体现了这套丛书的品位和格调。

“计算机科学丛书”的出版工作得到了国内外学者的鼎力襄助，国内的专家不仅提供了中肯的选题指导，还不辞劳苦地担任了翻译和审校的工作；而原书的作者也相当关注其作品在中国的传播，有的还专程为其书的中译本作序。迄今，“计算机科学丛书”已经出版了近两百个品种，这些书籍在读者中树立了良好的口碑，并被许多高校采用为正式教材和参考书籍。其影印版“经典原版书库”作为姊妹篇也被越来越多实施双语教学的学校所采用。

权威的作者、经典的教材、一流的译者、严格的审校、精细的编辑，这些因素使我们的图书有了质量的保证。随着计算机科学与技术专业学科建设的不断完善和教材改革的逐渐深化，教育界对国外计算机教材的需求和应用都将步入一个新的阶段，我们的目标是尽善尽美，而反馈的意见正是我们达到这一终极目标的重要帮助。华章公司欢迎老师和读者对我们的工作提出建议或给予指正，我们的联系方法如下：

华章网站：[www.hzbook.com](http://www.hzbook.com)

电子邮件：[hzjsj@hzbook.com](mailto:hzjsj@hzbook.com)

联系电话：(010) 88379604

联系地址：北京市西城区百万庄南街1号

邮政编码：100037



华章科技图书出版中心

# Foreword

Analyzing large amounts of data is a necessity. Even popular science books, like “super crunchers,” give compelling cases where large amounts of data yield discoveries and intuitions that surprise even experts. Every enterprise benefits from collecting and analyzing its data: Hospitals can spot trends and anomalies in their patient records, search engines can do better ranking and ad placement, and environmental and public health agencies can spot patterns and abnormalities in their data. The list continues, with cybersecurity and computer network intrusion detection; monitoring of the energy consumption of household appliances; pattern analysis in bioinformatics and pharmaceutical data; financial and business intelligence data; spotting trends in blogs, Twitter, and many more. Storage is inexpensive and getting even less so, as are data sensors. Thus, collecting and storing data is easier than ever before.

The problem then becomes *how to analyze* the data. This is exactly the focus of this Third Edition of the book. Jiawei, Micheline, and Jian give encyclopedic coverage of all the related methods, from the classic topics of clustering and classification, to database methods (e.g., association rules, data cubes) to more recent and advanced topics (e.g., SVD/PCA, wavelets, support vector machines).

The exposition is extremely accessible to beginners and advanced readers alike. The book gives the fundamental material first and the more advanced material in follow-up chapters. It also has numerous rhetorical questions, which I found extremely helpful for maintaining focus.

We have used the first two editions as textbooks in data mining courses at Carnegie Mellon and plan to continue to do so with this Third Edition. The new version has significant additions: Notably, it has more than 100 citations to works from 2006 onward, focusing on more recent material such as graphs and social networks, sensor networks, and outlier detection. This book has a new section for visualization, has expanded outlier detection into a whole chapter, and has separate chapters for advanced

methods—for example, pattern mining with top- $k$  patterns and more and clustering methods with biclustering and graph clustering.

Overall, it is an excellent book on classic and modern data mining methods, and it is ideal not only for teaching but also as a reference book.

Christos Faloutsos  
*Carnegie Mellon University*

# Foreword to Second Edition

We are deluged by data—scientific data, medical data, demographic data, financial data, and marketing data. People have no time to look at this data. Human attention has become the precious resource. So, we must find ways to automatically analyze the data, to automatically classify it, to automatically summarize it, to automatically discover and characterize trends in it, and to automatically flag anomalies. This is one of the most active and exciting areas of the database research community. Researchers in areas including statistics, visualization, artificial intelligence, and machine learning are contributing to this field. The breadth of the field makes it difficult to grasp the extraordinary progress over the last few decades.

Six years ago, Jiawei Han's and Micheline Kamber's seminal textbook organized and presented Data Mining. It heralded a golden age of innovation in the field. This revision of their book reflects that progress; more than half of the references and historical notes are to recent work. The field has matured with many new and improved algorithms, and has broadened to include many more datatypes: streams, sequences, graphs, time-series, geospatial, audio, images, and video. We are certainly not at the end of the golden age—indeed research and commercial interest in data mining continues to grow—but we are all fortunate to have this modern compendium.

The book gives quick introductions to database and data mining concepts with particular emphasis on data analysis. It then covers in a chapter-by-chapter tour the concepts and techniques that underlie classification, prediction, association, and clustering. These topics are presented with examples, a tour of the best algorithms for each problem class, and with pragmatic rules of thumb about when to apply each technique. The Socratic presentation style is both very readable and very informative. I certainly learned a lot from reading the first edition and got re-educated and updated in reading the second edition.

Jiawei Han and Micheline Kamber have been leading contributors to data mining research. This is the text they use with their students to bring them up to speed on

the field. The field is evolving very rapidly, but this book is a quick way to learn the basic ideas, and to understand where the field is today. I found it very informative and stimulating, and believe you will too.

Jim Gray  
*In his memory*



# Preface

The computerization of our society has substantially enhanced our capabilities for both generating and collecting data from diverse sources. A tremendous amount of data has flooded almost every aspect of our lives. This explosive growth in stored or transient data has generated an urgent need for new techniques and automated tools that can intelligently assist us in transforming the vast amounts of data into useful information and knowledge. This has led to the generation of a promising and flourishing frontier in computer science called *data mining*, and its various applications. Data mining, also popularly referred to as *knowledge discovery from data (KDD)*, is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories, or data streams.

This book explores the concepts and techniques of *knowledge discovery* and *data mining*. As a multidisciplinary field, data mining draws on work from areas including statistics, machine learning, pattern recognition, database technology, information retrieval, network science, knowledge-based systems, artificial intelligence, high-performance computing, and data visualization. We focus on issues relating to the feasibility, usefulness, effectiveness, and scalability of techniques for the discovery of patterns hidden in *large data sets*. As a result, this book is not intended as an introduction to statistics, machine learning, database systems, or other such areas, although we do provide some background knowledge to facilitate the reader's comprehension of their respective roles in data mining. Rather, the book is a comprehensive introduction to data mining. It is useful for computing science students, application developers, and business professionals, as well as researchers involved in any of the disciplines previously listed.

Data mining emerged during the late 1980s, made great strides during the 1990s, and continues to flourish into the new millennium. This book presents an overall picture of the field, introducing interesting data mining techniques and systems and discussing applications and research directions. An important motivation for writing this book was the need to build an organized framework for the study of data mining—a challenging task, owing to the extensive multidisciplinary nature of this fast-developing field. We hope that this book will encourage people with different backgrounds and experiences to exchange their views regarding data mining so as to contribute toward the further promotion and shaping of this exciting and dynamic field.

## Organization of the Book

Since the publication of the first two editions of this book, great progress has been made in the field of data mining. Many new data mining methodologies, systems, and applications have been developed, especially for handling new kinds of data, including information networks, graphs, complex structures, and data streams, as well as text, Web, multimedia, time-series, and spatiotemporal data. Such fast development and rich, new technical contents make it difficult to cover the full spectrum of the field in a single book. Instead of continuously expanding the coverage of this book, we have decided to cover the core material in sufficient scope and depth, and leave the handling of complex data types to a separate forthcoming book.

The third edition substantially revises the first two editions of the book, with numerous enhancements and a reorganization of the technical contents. The core technical material, which handles mining on general data types, is expanded and substantially enhanced. Several individual chapters for topics from the second edition (e.g., data preprocessing, frequent pattern mining, classification, and clustering) are now augmented and each split into two chapters for this new edition. For these topics, one chapter encapsulates the basic concepts and techniques while the other presents advanced concepts and methods.

Chapters from the second edition on mining complex data types (e.g., stream data, sequence data, graph-structured data, social network data, and multirelational data, as well as text, Web, multimedia, and spatiotemporal data) are now reserved for a new book that will be dedicated to *advanced topics in data mining*. Still, to support readers in learning such advanced topics, we have placed an electronic version of the relevant chapters from the second edition onto the book's web site as companion material for the third edition.

The chapters of the third edition are described briefly as follows, with emphasis on the new material.

**Chapter 1** provides an *introduction* to the multidisciplinary field of data mining. It discusses the evolutionary path of information technology, which has led to the need for data mining, and the importance of its applications. It examines the data types to be mined, including relational, transactional, and data warehouse data, as well as complex data types such as time-series, sequences, data streams, spatiotemporal data, multimedia data, text data, graphs, social networks, and Web data. The chapter presents a general classification of data mining tasks, based on the kinds of knowledge to be mined, the kinds of technologies used, and the kinds of applications that are targeted. Finally, major challenges in the field are discussed.

**Chapter 2** introduces the *general data features*. It first discusses data objects and attribute types and then introduces typical measures for basic statistical data descriptions. It overviews data visualization techniques for various kinds of data. In addition to methods of numeric data visualization, methods for visualizing text, tags, graphs, and multidimensional data are introduced. Chapter 2 also introduces ways to measure similarity and dissimilarity for various kinds of data.

**Chapter 3** introduces *techniques for data preprocessing*. It first introduces the concept of data quality and then discusses methods for data cleaning, data integration, data reduction, data transformation, and data discretization.

Chapters 4 and 5 provide a solid introduction to *data warehouses, OLAP* (online analytical processing), and *data cube technology*. **Chapter 4** introduces the basic concepts, modeling, design architectures, and general implementations of data warehouses and OLAP, as well as the relationship between data warehousing and other data generalization methods. **Chapter 5** takes an in-depth look at data cube technology, presenting a detailed study of methods of data cube computation, including Star-Cubing and high-dimensional OLAP methods. Further explorations of data cube and OLAP technologies are discussed, such as sampling cubes, ranking cubes, prediction cubes, multifeature cubes for complex analysis queries, and discovery-driven cube exploration.

Chapters 6 and 7 present methods for *mining frequent patterns, associations, and correlations* in large data sets. **Chapter 6** introduces fundamental concepts, such as market basket analysis, with many techniques for frequent itemset mining presented in an organized way. These range from the basic Apriori algorithm and its variations to more advanced methods that improve efficiency, including the frequent pattern growth approach, frequent pattern mining with vertical data format, and mining closed and max frequent itemsets. The chapter also discusses pattern evaluation methods and introduces measures for mining correlated patterns. **Chapter 7** is on advanced pattern mining methods. It discusses methods for pattern mining in multi-level and multidimensional space, mining rare and negative patterns, mining colossal patterns and high-dimensional data, constraint-based pattern mining, and mining compressed or approximate patterns. It also introduces methods for pattern exploration and application, including semantic annotation of frequent patterns.

Chapters 8 and 9 describe methods for *data classification*. Due to the importance and diversity of classification methods, the contents are partitioned into two chapters. **Chapter 8** introduces basic concepts and methods for classification, including decision tree induction, Bayes classification, and rule-based classification. It also discusses model evaluation and selection methods and methods for improving classification accuracy, including ensemble methods and how to handle imbalanced data. **Chapter 9** discusses advanced methods for classification, including Bayesian belief networks, the neural network technique of backpropagation, support vector machines, classification using frequent patterns, *k*-nearest-neighbor classifiers, case-based reasoning, genetic algorithms, rough set theory, and fuzzy set approaches. Additional topics include multiclass classification, semi-supervised classification, active learning, and transfer learning.

*Cluster analysis* forms the topic of Chapters 10 and 11. **Chapter 10** introduces the basic concepts and methods for data clustering, including an overview of basic cluster analysis methods, partitioning methods, hierarchical methods, density-based methods, and grid-based methods. It also introduces methods for the evaluation of clustering. **Chapter 11** discusses advanced methods for clustering, including probabilistic model-based clustering, clustering high-dimensional data, clustering graph and network data, and clustering with constraints.

**Chapter 12** is dedicated to *outlier detection*. It introduces the basic concepts of outliers and outlier analysis and discusses various outlier detection methods from the view of degree of supervision (i.e., supervised, semi-supervised, and unsupervised methods), as well as from the view of approaches (i.e., statistical methods, proximity-based methods, clustering-based methods, and classification-based methods). It also discusses methods for mining contextual and collective outliers, and for outlier detection in high-dimensional data.

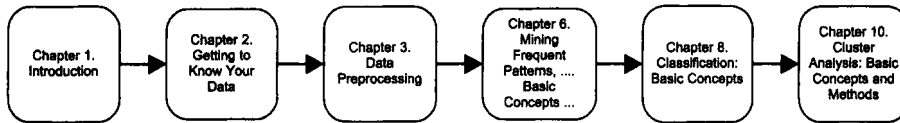
Finally, in **Chapter 13**, we discuss *trends, applications, and research frontiers* in data mining. We briefly cover mining complex data types, including mining sequence data (e.g., time series, symbolic sequences, and biological sequences), mining graphs and networks, and mining spatial, multimedia, text, and Web data. In-depth treatment of data mining methods for such data is left to a book on advanced topics in data mining, the writing of which is in progress. The chapter then moves ahead to cover other data mining methodologies, including statistical data mining, foundations of data mining, visual and audio data mining, as well as data mining applications. It discusses data mining for financial data analysis, for industries like retail and telecommunication, for use in science and engineering, and for intrusion detection and prevention. It also discusses the relationship between data mining and recommender systems. Because data mining is present in many aspects of daily life, we discuss issues regarding data mining and society, including ubiquitous and invisible data mining, as well as privacy, security, and the social impacts of data mining. We conclude our study by looking at data mining trends.

Throughout the text, *italic* font is used to emphasize terms that are defined, while **bold** font is used to highlight or summarize main ideas. Sans serif font is used for reserved words. Bold italic font is used to represent multidimensional quantities.

This book has several strong features that set it apart from other texts on data mining. It presents a very broad yet in-depth coverage of the principles of data mining. The chapters are written to be as self-contained as possible, so they may be read in order of interest by the reader. Advanced chapters offer a larger-scale view and may be considered optional for interested readers. All of the major methods of data mining are presented. The book presents important topics in data mining regarding multidimensional OLAP analysis, which is often overlooked or minimally treated in other data mining books. The book also maintains web sites with a number of online resources to aid instructors, students, and professionals in the field. These are described further in the following.

## To the Instructor

This book is designed to give a broad, yet detailed overview of the data mining field. It can be used to teach an introductory course on data mining at an advanced undergraduate level or at the first-year graduate level. Sample course syllabi are provided on the book's web sites ([www.cs.uiuc.edu/~hanj/bk3](http://www.cs.uiuc.edu/~hanj/bk3) and [www.booksite.mkp.com/datamining3e](http://www.booksite.mkp.com/datamining3e)) in addition to extensive teaching resources such as lecture slides, instructors' manuals, and reading lists (see p. xiv).



**Figure P.1** A suggested sequence of chapters for a short introductory course.

Depending on the length of the instruction period, the background of students, and your interests, you may select subsets of chapters to teach in various sequential orderings. For example, if you would like to give only a short introduction to students on data mining, you may follow the suggested sequence in Figure P.1. Notice that depending on the need, you can also omit some sections or subsections in a chapter if desired.

Depending on the length of the course and its technical scope, you may choose to selectively add more chapters to this preliminary sequence. For example, instructors who are more interested in advanced classification methods may first add “Chapter 9. Classification: Advanced Methods”; those more interested in pattern mining may choose to include “Chapter 7. Advanced Pattern Mining”; whereas those interested in OLAP and data cube technology may like to add “Chapter 4. Data Warehousing and Online Analytical Processing” and “Chapter 5. Data Cube Technology.”

Alternatively, you may choose to teach the whole book in a two-course sequence that covers all of the chapters in the book, plus, when time permits, some advanced topics such as graph and network mining. Material for such advanced topics may be selected from the companion chapters available from the book’s web site, accompanied with a set of selected research papers.

Individual chapters in this book can also be used for tutorials or for special topics in related courses, such as machine learning, pattern recognition, data warehousing, and intelligent data analysis.

Each chapter ends with a set of exercises, suitable as assigned homework. The exercises are either short questions that test basic mastery of the material covered, longer questions that require analytical thinking, or implementation projects. Some exercises can also be used as research discussion topics. The bibliographic notes at the end of each chapter can be used to find the research literature that contains the origin of the concepts and methods presented, in-depth treatment of related topics, and possible extensions.

## To the Student

We hope that this textbook will spark your interest in the young yet fast-evolving field of data mining. We have attempted to present the material in a clear manner, with careful explanation of the topics covered. Each chapter ends with a summary describing the main points. We have included many figures and illustrations throughout the text to make the book more enjoyable and reader-friendly. Although this book was designed as a textbook, we have tried to organize it so that it will also be useful to you as a reference

book or handbook, should you later decide to perform in-depth research in the related fields or pursue a career in data mining.

What do you need to know to read this book?

- You should have some knowledge of the concepts and terminology associated with statistics, database systems, and machine learning. However, we do try to provide enough background of the basics, so that if you are not so familiar with these fields or your memory is a bit rusty, you will not have trouble following the discussions in the book.
- You should have some programming experience. In particular, you should be able to read pseudocode and understand simple data structures such as multidimensional arrays.

## To the Professional

This book was designed to cover a wide range of topics in the data mining field. As a result, it is an excellent handbook on the subject. Because each chapter is designed to be as standalone as possible, you can focus on the topics that most interest you. The book can be used by application programmers and information service managers who wish to learn about the key ideas of data mining on their own. The book would also be useful for technical data analysis staff in banking, insurance, medicine, and retailing industries who are interested in applying data mining solutions to their businesses. Moreover, the book may serve as a comprehensive survey of the data mining field, which may also benefit researchers who would like to advance the state-of-the-art in data mining and extend the scope of data mining applications.

The techniques and algorithms presented are of practical utility. Rather than selecting algorithms that perform well on small “toy” data sets, the algorithms described in the book are geared for the discovery of patterns and knowledge hidden in large, real data sets. Algorithms presented in the book are illustrated in pseudocode. The pseudocode is similar to the C programming language, yet is designed so that it should be easy to follow by programmers unfamiliar with C or C++. If you wish to implement any of the algorithms, you should find the translation of our pseudocode into the programming language of your choice to be a fairly straightforward task.

## Book Web Sites with Resources

The book has a web site at [www.cs.uiuc.edu/~hanj/bk3](http://www.cs.uiuc.edu/~hanj/bk3) and another with Morgan Kaufmann Publishers at [www.booksite.mkp.com/datamining3e](http://www.booksite.mkp.com/datamining3e). These web sites contain many supplemental materials for readers of this book or anyone else with an interest in data mining. The resources include the following:

- **Slide presentations for each chapter.** Lecture notes in Microsoft PowerPoint slides are available for each chapter.

- ❖ **Companion chapters on advanced data mining.** Chapters 8 to 10 of the second edition of the book, which cover mining complex data types, are available on the book's web sites for readers who are interested in learning more about such advanced topics, beyond the themes covered in this book.
- ❖ **Instructors' manual.** This complete set of answers to the exercises in the book is available only to instructors from the publisher's web site.
- ❖ **Course syllabi and lecture plans.** These are given for undergraduate and graduate versions of introductory and advanced courses on data mining, which use the text and slides.
- ❖ **Supplemental reading lists with hyperlinks.** Seminal papers for supplemental reading are organized per chapter.
- ❖ **Links to data mining data sets and software.** We provide a set of links to data mining data sets and sites that contain interesting data mining software packages, such as IlliMine from the University of Illinois at Urbana-Champaign (<http://illimine.cs.uiuc.edu>).
- ❖ **Sample assignments, exams, and course projects.** A set of sample assignments, exams, and course projects is available to instructors from the publisher's web site.
- ❖ **Figures from the book.** This may help you to make your own slides for your classroom teaching.
- ❖ **Contents of the book in PDF format.**
- ❖ **Errata on the different printings of the book.** We encourage you to point out any errors in this book. Once the error is confirmed, we will update the errata list and include acknowledgment of your contribution.

Comments or suggestions can be sent to [hanj@cs.uiuc.edu](mailto:hanj@cs.uiuc.edu). We would be happy to hear from you.

# Acknowledgments

## Third Edition of the Book

We would like to express our grateful thanks to all of the previous and current members of the Data Mining Group at UIUC, the faculty and students in the Data and Information Systems (DAIS) Laboratory in the Department of Computer Science at the University of Illinois at Urbana-Champaign, and many friends and colleagues, whose constant support and encouragement have made our work on this edition a rewarding experience. We would also like to thank students in CS412 and CS512 classes at UIUC of the 2010–2011 academic year, who carefully went through the early drafts of this book, identified many errors, and suggested various improvements.

We also wish to thank David Bevans and Rick Adams at Morgan Kaufmann Publishers, for their enthusiasm, patience, and support during our writing of this edition of the book. We thank Marilyn Rash, the Project Manager, and her team members, for keeping us on schedule.

We are also grateful for the invaluable feedback from all of the reviewers. Moreover, we would like to thank U.S. National Science Foundation, NASA, U.S. Air Force Office of Scientific Research, U.S. Army Research Laboratory, and Natural Science and Engineering Research Council of Canada (NSERC), as well as IBM Research, Microsoft Research, Google, Yahoo! Research, Boeing, HP Labs, and other industry research labs for their support of our research in the form of research grants, contracts, and gifts. Such research support deepens our understanding of the subjects discussed in this book. Finally, we thank our families for their wholehearted support throughout this project.

## Second Edition of the Book

We would like to express our grateful thanks to all of the previous and current members of the Data Mining Group at UIUC, the faculty and students in the Data and



Information Systems (DAIS) Laboratory in the Department of Computer Science at the University of Illinois at Urbana-Champaign, and many friends and colleagues, whose constant support and encouragement have made our work on this edition a rewarding experience. These include Gul Agha, Rakesh Agrawal, Loretta Auvil, Peter Bajcsy, Geneva Belford, Deng Cai, Y. Dora Cai, Roy Cambell, Kevin C.-C. Chang, Surajit Chaudhuri, Chen Chen, Yixin Chen, Yuguo Chen, Hong Cheng, David Cheung, Shengnan Cong, Gerald DeJong, AnHai Doan, Guozhu Dong, Charios Ermopoulos, Martin Ester, Christos Faloutsos, Wei Fan, Jack C. Feng, Ada Fu, Michael Garland, Johannes Gehrke, Hector Gonzalez, Mehdi Harandi, Thomas Huang, Wen Jin, Chulyun Kim, Sangkyum Kim, Won Kim, Won-Young Kim, David Kuck, Young-Koo Lee, Harris Lewin, Xiaolei Li, Yifan Li, Chao Liu, Han Liu, Huan Liu, Hongyan Liu, Lei Liu, Ying Lu, Klara Nahrstedt, David Padua, Jian Pei, Lenny Pitt, Daniel Reed, Dan Roth, Bruce Schatz, Zheng Shao, Marc Snir, Zhaohui Tang, Bhavani M. Thuraisingham, Josep Torrellas, Peter Tzvetkov, Benjamin W. Wah, Haixun Wang, Jianyong Wang, Ke Wang, Muyuan Wang, Wei Wang, Michael Welge, Marianne Winslett, Ouri Wolfson, Andrew Wu, Tianyi Wu, Dong Xin, Xifeng Yan, Jiong Yang, Xiaoxin Yin, Hwanjo Yu, Jeffrey X. Yu, Philip S. Yu, Maria Zemankova, ChengXiang Zhai, Yuanyuan Zhou, and Wei Zou.

Deng Cai and ChengXiang Zhai have contributed to the text mining and Web mining sections, Xifeng Yan to the graph mining section, and Xiaoxin Yin to the multirelational data mining section. Hong Cheng, Charios Ermopoulos, Hector Gonzalez, David J. Hill, Chulyun Kim, Sangkyum Kim, Chao Liu, Hongyan Liu, Kasif Manzoor, Tianyi Wu, Xifeng Yan, and Xiaoxin Yin have contributed to the proofreading of the individual chapters of the manuscript.

We also wish to thank Diane Cerra, our Publisher at Morgan Kaufmann Publishers, for her constant enthusiasm, patience, and support during our writing of this book. We are indebted to Alan Rose, the book Production Project Manager, for his tireless and ever-prompt communications with us to sort out all details of the production process. We are grateful for the invaluable feedback from all of the reviewers. Finally, we thank our families for their wholehearted support throughout this project.

## **First Edition of the Book**

We would like to express our sincere thanks to all those who have worked or are currently working with us on data mining-related research and/or the DBMiner project, or have provided us with various support in data mining. These include Rakesh Agrawal, Stella Atkins, Yvan Bedard, Binay Bhattacharya, (Yandong) Dora Cai, Nick Cercone, Surajit Chaudhuri, Sonny H. S. Chee, Jianping Chen, Ming-Syan Chen, Qing Chen, Qiming Chen, Shan Cheng, David Cheung, Shi Cong, Son Dao, Umeshwar Dayal, James Delgrande, Guozhu Dong, Carole Edwards, Max Egenhofer, Martin Ester, Usama Fayyad, Ling Feng, Ada Fu, Yongjian Fu, Daphne Gelbart, Randy Goebel, Jim Gray, Robert Grossman, Wan Gong, Yike Guo, Eli Hagen, Howard Hamilton, Jing He, Larry Henschen, Jean Hou, Mei-Chun Hsu, Kan Hu, Haiming Huang, Yue Huang, Julia Itskevitch, Wen Jin, Tiko Kameda, Hiroyuki Kawano, Rizwan Kheraj, Eddie Kim, Won Kim, Krzysztof Koperski, Hans-Peter Kriegel, Vipin Kumar, Laks V. S. Lakshmanan, Joyce