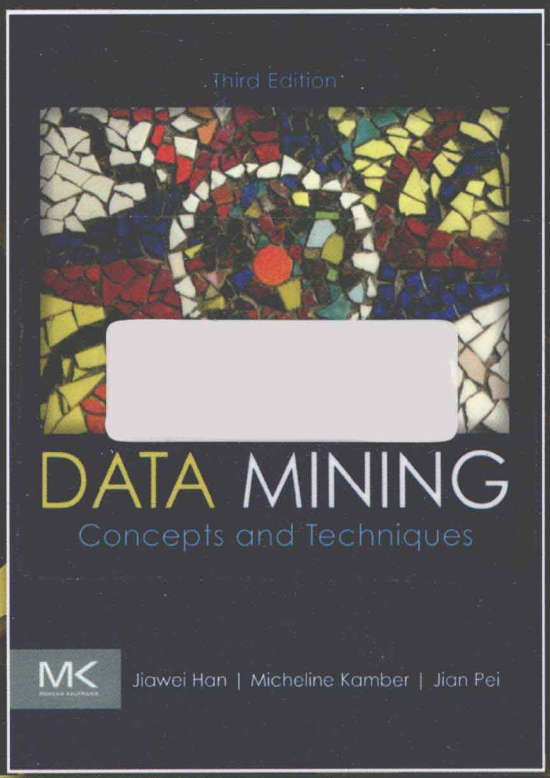
HZ BOOKS 华章教育
ELSEVIER
计 算 机 科 学 丛 书

原书第3版

数据挖掘
概念与技术

Jiawei Han Micheline Kamber Jian Pei 著
范明 孟小峰 译

Data Mining
Concepts and Techniques Third Edition



机械工业出版社
China Machine Press

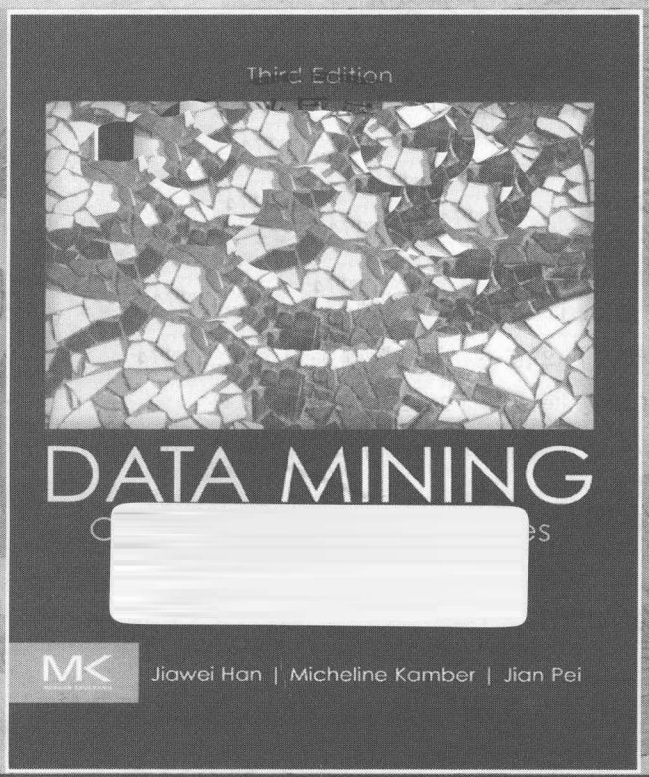
计 算 机 科 学 丛 书

原书第3版

数据挖掘
概念与技术

Jiawei Han Micheline Kamber Jian Pei 著
范明 孟小峰 译

Data Mining
Concepts and Techniques Third Edition



机械工业出版社
China Machine Press

本书完整全面地讲述数据挖掘的概念、方法、技术和最新研究进展。本书对前两版做了全面修订，加强和重新组织了全书的技术内容，重点论述了数据预处理、频繁模式挖掘、分类和聚类等的内容，还全面讲述了 OLAP 和离群点检测，并研讨了挖掘网络、复杂数据类型以及重要应用领域。

本书是数据挖掘和知识发现领域内的所有教师、研究人员、开发人员和用户都必读的参考书，是一本适用于数据分析、数据挖掘和知识发现课程的优秀教材，可以用做高年级本科生或者一年级研究生的数据挖掘导论教材。

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

Copyright © 2012 by Elsevier Inc. All rights reserved.

Authorized Simplified Chinese translation edition published by the Proprietor.

Copyright © 2012 by Elsevier (Singapore) Pte Ltd. All rights reserved.

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 SAR 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.

本书简体中文版由机械工业出版社与 Elsevier (Singapore) Pte Ltd. 在中国大陆境内合作出版。本版仅限在中国境内（不包括中国香港特别行政区及中国台湾地区）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：01-2012-0225

图书在版编目（CIP）数据

数据挖掘：概念与技术（原书第3版）/（美）韩家炜（Han, J.）等著；范明等译. —北京：机械工业出版社，2012.7

（计算机科学丛书）

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

ISBN 978-7-111-39140-1

Ⅰ. 数…　Ⅱ. ①韩…　②范…　Ⅲ. 数据采集　Ⅳ. TP274

中国版本图书馆 CIP 数据核字（2012）第 157938 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：盛思源

北京诚信伟业印刷有限公司印刷

2012 年 8 月第 1 版第 1 次印刷

185mm × 260mm · 31 印张

标准书号：ISBN 978-7-111-39140-1

定价：79.00 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

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，不仅擘划了研究的范畴，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
电子邮件：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

华章教育
华章科技图书出版中心

中文版序

Data Mining: Concepts and Techniques, Third Edition

We are pleased to see that our third edition has been translated into Chinese by Professor Fan and Meng. The first two editions were translated by them several years ago and have been well received among Chinese readers. In recent years, we have witnessed tremendous progress in the field of data mining research and applications internationally. As a promising new technology, data mining has attracted tremendous interest in the Far East as well. Numerous international and regional conferences on data mining and applications have appeared or held in this region. Many Chinese researchers have been playing an active role, contributing in both research and applications to the advances of this young field.

In this third edition, we have carefully selected and tailored the technical materials to be covered for the courses on data mining at both the undergraduate level and the first-year graduate level. We have updated and enhanced the existing chapters substantially with many new topics. Thus, we expect the publication of this edition in Chinese will help Chinese readers to learn and master the latest technology and put them into promising new applications.

With best regards,

(非常高兴地看到本书的第3版由范明和孟小峰教授翻译成中文。几年前，他们翻译了本书的前两版并被中文读者广泛接受。近年来，我们见证了数据挖掘研究和应用领域在世界范围内的巨大进展。作为一种具有良好发展势头的新技术，数据挖掘在远东也引起了极大兴趣。许多国际或地区性的数据挖掘和应用会议已经在该地区出现或召开。许多中国的研究者一直起着积极作用，为推动这个年轻领域的研究和应用做出了贡献。

在第3版中，我们对所包含的技术内容进行了精心挑选和剪裁，以便用于本科生和一年级研究生的“数据挖掘”课程。我们用许多新的主题，大幅度地更新和加强了已有的章节。因而，我们期望这个中文版将帮助中文读者学习和掌握这些最新技术，并将它们用于有希望的新应用。

谨致良好祝愿!)

Jiawei Han, Micheline Kamber, and Jian Pei

June 2012

试读结束：需要全本请在线购买： www.ertongbook.com

2001 年，Jiawei Han（韩家炜）和 Micheline Kamber 出版了数据挖掘领域具有里程碑意义的著作——本书的第 1 版。2006 年，他们又推出了本书的第 2 版。在这个龙年（2012 年），我们看到了本书的第 3 版，并且欣喜地看到该书增加了一位新的、年青的华人合著者 Jian Pei（裴健）。

数据挖掘是数据库研究、开发和应用最活跃的分支之一。这是很自然的事。数据库系统，特别是关系数据库系统的成功，使得我们有了强有力的事务处理工具。在计算机的帮助下，人们可以把传统的事务处理做得更好。不满足现状是社会前进的动力。人类当然不会仅仅满足于让计算机做事务处理。从信息处理的角度，人们更希望计算机帮助分析数据和理解数据，帮助他们基于丰富的数据做出决策。于是，数据挖掘（从大量数据中以非平凡的方法发现有用的知识）就成为一种自然的需求。正是这种需求引起了人们的关注，导致了数据挖掘研究和应用的蓬勃发展。

数据挖掘是一个多学科的交叉领域。这也是很自然的事。一方面，想要以非平凡的方法发现蕴藏在大型数据集中的有用知识，数据挖掘必须从统计学、机器学习、神经网络、模式识别、知识库系统、信息检索、高性能计算和可视化等学科领域汲取营养。另一方面，这些学科领域也需要从不同角度关注数据的分析与理解；数据挖掘也为这些学科领域的发展提供了新的机遇和挑战。今天，数据挖掘已经不再仅仅是数据库的研究者和开发者关注的问题，它已经成为统计学、机器学习等诸多领域的研究者和开发者的热点课题之一。这种学科交叉融合带来的良性互动，无疑促进了包括数据挖掘在内的诸学科的发展与繁荣。

自本书第 1 版问世已经过去了 11 年。在过去的 11 年中，Jiawei Han 教授多次来华讲学，我们先后翻译了本书的第 1 版和第 2 版。国内许多大学都纷纷开设数据挖掘课程，其中大部分学校都使用本书的英文版或中文版。我们高兴地看到数据挖掘的研究与应用在我国的蓬勃开展。许多学者和研究人员都对这个新兴的学科领域表现出了极大的兴趣，他们不仅来自数据库领域，而且包括统计学、人工智能、模式识别、机器学习等领域的研究人员。国内的学者和开发者在数据挖掘方面的研究与应用方面已经取得了许多令人鼓舞的成果。特别值得一提的是，近年来，数据库的顶级学术会议 SIGMOD、ICDE 和数据挖掘的顶级学术会议 KDD 都相继在国内举办。

过去的 11 年是数据挖掘研究与应用迅猛发展的 11 年：新的和改进的算法不断出现，所考察的数据类型日趋丰富，应用领域逐渐扩大。虽然所挖掘的基本知识类型并未增加很多，但是新的应用需要我们处理更加丰富的数据类型，如流、序列、图、时间序列、符号序列、生物学序列、空间、音频、图像和视频数据，因此需要新的技术。例如，流数据的关联、分类和聚类需要处理可能无限的数据，需要考虑数据的分布随时间的演变。Web 页面的分类不仅需要考虑页面本身的特征，而且还需要考虑页面的链接和被链接的页面的特征。

第 3 版对本书的前两版进行了全面修订，突出和加强了数据挖掘的核心内容，以足够的广度和深度涵盖该领域的核心内容。认识数据和数据预处理、数据仓库和 OLAP 技术、模式挖掘与关联分析、分类、聚类都分成两章。其中，前一章介绍基本概念和技术，后一章进一步讨论更高级的概念和方法。离群点检测单独成为一章，进行更深入的讨论。最后一章对数据挖掘研究与应用发展趋势进行了概述，把读者引向更深入的主题。与前两版相比，第 3 版

的组织更有利于教学。

如果说 11 年前本书的问世标志数据挖掘领域已见雏形，5 年前该书第 2 版的出版预示数据挖掘开始进入了成熟期，那么第 3 版的出版表明数据挖掘已经在向纵深发展，其最基本层面的内容已经趋于稳定，在计算学科的高年级本科生和研究生中广泛开展数据挖掘课程的教学已经是万事俱备。

Jiawei Han 教授早年就读于郑州大学，后赴美国留学，在威斯康辛大学获硕士和博士学位。他曾先后在美国西北大学、加拿大西蒙－弗雷泽大学任教，现在是美国伊利诺伊大学厄巴纳－尚佩恩分校计算机科学系的 Bliss 教授。Jiawei Han 教授是数据挖掘和数据库系统领域国际知名学者，ACM 和 IEEE 会士。他曾因在该领域的杰出贡献多次获奖，包括 ACM SIGKDD 创新奖（2004）、IEEE 计算机学会技术成就奖（2005）和 IEEE W. Wallace McDowell 奖（2009）。

徐华、叶阳东、姬安明、王静、李盛恩、李翠萍等参加了第 1 版的部分翻译工作，马玉书、董云海对第 1 版的部分译稿提出了很好的修改意见。第 2 版由范明和孟小峰翻译；译者的许多同事、朋友和学生，如眷红英博士和范宏建博士，阅读了第 2 版的部分译稿，并提出了一些建议和意见。第 3 版由范明和孟小峰翻译。译者的学生郭华平、李嘉、张亚亚和李晓燕参加了第 3 版的校对工作。

感谢本书的作者 Jiawei Han 教授。无论是第 1 版、第 2 版，还是第 3 版的翻译都得到了他的大力支持，他提供的方便使得本书的翻译工作能够在第一时间进行。Jiawei Han 教授还专门为第 2 版和第 3 版的中文版撰写了序言。

感谢机械工业出版社华章公司的编辑们，是他们的远见使得本书能够尽快与读者见面。

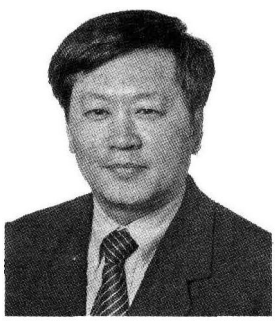
在第 3 版的翻译中，我们重新调整了部分术语的翻译。读过第 1 版、第 2 版的读者不难发现，第 3 版出现了许多的新术语，尚无固定译法。尽管我们力图为它们选择简洁、达意的中文术语，但仍然难免出现词不达意之处。译文中的错误和不当之处，敬请读者朋友指正。意见请发往 mfan@ zzu. edu. cn，我们将不胜感激。

我们将尽快向采用本书的教师提供讲稿和其他辅助支持。希望读者喜欢这本译著，希望这本译著有助于进一步推动我国的数据挖掘教学、研究和应用的深入开展。

范明　孟小峰

2012 年 6 月

范明 郑州大学信息工程学院教授，博士生导师。现为中国计算机学会数据库专业委员会委员、人工智能与模式识别专业委员会委员。长期从事计算机软件与理论教学和研究。主要讲授的课程包括程序设计、计算机操作系统、数据库系统原理、知识库系统原理、数据挖掘与数据仓库等。1989—1990 年曾访问加拿大 Simon Fraser 大学计算机科学系，从事演绎数据库研究。1999 年曾访问美国 Wright State 大学计算机科学与工程系，从事数据挖掘研究。当前感兴趣的研究方向包括数据挖掘和机器学习。先后发表论文 60 余篇。除本书外，还主持翻译了 Pang-Ning Tan、Michael Steinbach 和 Vipin Kumar 的《数据挖掘导论》。



孟小峰 博士，中国人民大学信息学院教授，博士生导师。现为中国计算机学会常务理事、中国计算机学会数据库专委会秘书长，《Journal of Computer Science and Technology》、《Frontiers of Computer Science》、《软件学报》、《计算机研究与发展》等编委。主持或参加过二十多项国家科技攻关项目、国家自然科学基金项目以及国家 863 项目、973 项目，先后获电子部科技进步特等奖（1996）、北京市科技进步二等奖（1998、2001）、中国计算机学会“王选奖”一等奖（2009）、北京市科学技术奖二等奖（2011）等奖励，入选“中创软件人才奖”（2002）、“教育部新世纪优秀人才支持计划”（2004）、“第三届北京市高校名师奖”（2005）。近 5 年在国内外杂志及国际会议发表论文 120 多篇，出版学术专著《Moving Objects Management: Models, Techniques, and Applications》（Springer）、《XML 数据管理：概念与技术》、《移动数据管理：概念与技术》（中国计算机学会学术著作丛书）等。获国家发明专利授权 8 项。近期主要研究领域为互联网络与移动数据管理，包括 Web 数据集成、XML 数据库系统、云数据管理、闪存数据库系统、隐私保护等。

第3版序

Data Mining: Concepts and Techniques, Third Edition

分析大量数据是必要的。甚至像“super crunchers”（超级电脑）这样流行的科技书也给出了从大量数据发现和得到直觉知识的非常好的事例。每个企业都从收集和分析数据中获益：医院可以从患者记录中识别趋势和异常，搜索引擎可以进行更好的秩评定和广告投放，环境和公共卫生部门可以识别数据中的模式和异常。这样的例子还有很多，如计算机安全和计算网络入侵检测、家用电器的能源消耗、生物信息学和药物数据的模式分析、财经和商务智能数据、识别博客中的趋势、唧喳（Twitter）等，不一而足。与数据传感器一样，存储设备价格越来越低，因此收集和存储数据比以前更加容易。

于是，问题变成如何分析数据。这恰是第3版的关注点。Jiawei、Micheline、Jian的教材全景式地讨论了数据挖掘的所有相关方法，从经典的分类和聚类主题，到数据库方法（例如，关联规则和数据立方体），到更新和更高级的主题（例如，SVD/PCA、小波、支持向量机）。

对于初学者来说，书中的阐述极其容易理解，对于高端读者也是如此。本书首先介绍基本概念，更高级的内容在随后的章节中。书中还使用了一些修辞疑问，这样做非常有助于吸引读者注意力。

我们已经使用前两版作为卡内基－梅隆大学数据挖掘课程的教材，并且准备继续使用第3版。新版内容有显著增加：值得注意的是，超过100篇引文引用2006年以来的工作，关注更近的研究，如图和社会网络、传感器网络，以及离群点检测。对于可视化，本书新增了一节；离群点检测扩充为一整章；而有些章被分开，以便介绍高级方法。例如，top-k模式等模式挖掘以及双聚类和图聚类。

总之，这是一本关于经典和现代数据挖掘方法的优秀专著，它不仅是一本理想的教材，而且也是一本理想的参考书。

Christos Faloutsos

卡内基－梅隆大学

我们被数据（科学数据、医疗数据、人口统计数据、金融数据和销售数据）所淹没。人们没有时间查看这些数据。人们的关注已经转到可贵的应付手段上。因此，我们必须找到有效方法，自动地分析数据、自动地对数据分类、自动地对数据汇总、自动地发现和描述数据中的趋势、自动地标记异常。这是数据库研究最活跃、最令人激动的领域之一。统计学、可视化、人工智能和机器学习方面的研究人员正在为该领域做出贡献。由于该领域非常广阔，很难把握它过去几十年的非凡进展。

六年前，Jiawei Han 和 Micheline Kamber 的原创性教科书将数据挖掘的内容组织在一起并呈现给读者。它预示了数据挖掘领域的创新黄金时代的到来。他们的书的新版反映了该领域的进展，一半以上的参考文献和历史注释都涉及当前的研究。该领域已经成熟，出现了许多新的、改进的算法；该领域已经拓宽，包含了更多数据类型，如流、序列、图、时间序列、地理空间、音频、图像和视频。我们不仅可以肯定这个黄金时代尚未结束（数据挖掘研究和商业兴趣正在继续增长），而且，这本数据挖掘的现代著作的面世是我们所庆幸的。

本书首先提供数据库和数据挖掘概念的简略介绍，特别强调数据分析。然后，逐章介绍分类、预测、关联和聚类等基础概念和技术。这些主题辅以实例，对每类问题均提供代表性算法，并对每种技术的应用给出注重实效的规则。这种苏格拉底式的表达风格具有很好的可读性，并且内容丰富。我已通过阅读第 1 版学到了许多知识，并且在阅读第 2 版时再次受益并更新了知识。

Jiawei Han 和 Micheline Kamber 在数据挖掘研究方面一直处于领先地位。这是一本他们用于培养自己的学生，以加快该领域发展的教材。该领域发展非常迅速，本书提供了一条学习该领域基本思想和了解该领域现状的快捷之路。我认为本书内容丰富、刺激，相信读者也会有同样的感触。

Jim Gray
Microsoft Research
美国加利福尼亚旧金山

前 言

Data Mining: Concepts and Techniques, Third Edition

社会的计算机化显著地增强了我们产生和收集数据的能力。大量数据从我们生活的每个角落涌出。存储的或瞬态的数据的爆炸性增长已激起对新技术和自动工具的需求，以帮助我们智能地将海量数据转换成有用的信息和知识。这导致称做*数据挖掘*的一个计算机科学前沿学科的产生，这是一个充满希望和欣欣向荣并具有广泛应用的学科。数据挖掘通常又称为*数据中的知识发现*（KDD），是自动地或方便地提取代表知识的模式；这些模式隐藏在大型数据库、数据仓库、Web、其他大量信息库或数据流中。

本书考察*知识发现*和*数据挖掘*的基本概念和技术。作为一个多学科领域，数据挖掘从多个学科汲取营养。这些学科包括统计学、机器学习、模式识别、数据库技术、信息检索、网络科学、知识库系统、人工智能、高性能计算和数据可视化。我们提供发现隐藏在*大型数据集*中的模式的技术，关注可行性、有用性、有效性和可伸缩性问题。因此，本书不打算作为数据库系统、机器学习、统计学或其他某领域的导论，尽管我们确实提供了这些领域的必要背景材料，以便读者理解它们各自在数据挖掘中的作用。本书是对数据挖掘的全面介绍。对于计算科学的学生、应用开发人员、行业专业人员以及涉及以上列举的学科的研究人员，本书应当是有用的。

数据挖掘出现于20世纪80年代后期，20世纪90年代有了突飞猛进的发展，并可望在新千年继续繁荣。本书全面展示该领域，介绍有趣的数据挖掘技术和系统，并讨论数据挖掘的应用和研究方向。写本书的重要动机是需要建立一个学习数据挖掘的有组织的框架——由于这个快速发展领域的多学科特点，这是一项具有挑战性的任务。我们希望本书有助于具有不同背景和经验的人交换关于数据挖掘的见解，为进一步促进这个令人激动的、不断发展的领域的成长做出贡献。

本书的组织

自本书第1版、第2版出版以来，数据挖掘领域已经取得了重大进展，开发出了许多新的数据挖掘方法、系统和应用，特别是对于处理包括信息网络、图、复杂结构和数据流，以及文本、Web、多媒体、时间序列、时间空间数据在内的新的数据类型。这种快速发展、新技术不断涌现使得在一本书中涵盖整个领域的广泛内容非常困难。因此，我们决定与其继续扩大本书的涵盖面，还不如让本书以足够的广度和深度涵盖该领域的核心内容，而把复杂数据类型的处理留给另一本即将面世的书。

第3版对本书的前两版做了全面修订，加强和重新组织了全书的技术内容，显著地扩充和加强处理一般数据类型挖掘的核心技术。第2版中讨论特定主题的章节（例如，数据预处理、频繁模式挖掘、分类和聚类）在这一版都被扩充，每章都分成两章。对于这些主题，一章囊括基本概念和技术，而另一章提供高级概念和方法。

第2版关于复杂数据类型的章节（例如，流数据、序列数据、图结构数据、社会网络数据和多重关系数据，以及文本、Web、多媒体和时间空间数据）现在保留给专门介绍*数据挖掘的高级课题*的新书。为了支持读者学习这些高级课题，我们把第2版的相关章节的电子版放在本书的网站上，作为第3版的配套材料。

第3版各章的简要内容如下（重点介绍新的内容）：

试读结束：需要全本请在线购买： www.ertongbook.com

第 1 章提供关于数据挖掘的多学科领域的导论。该章讨论导致需要数据挖掘的数据库技术的发展历程和数据挖掘应用的重要性。该章考察挖掘的数据类型，包括关系的、事务的和数据仓库数据，以及复杂的数据类型，如时间序列、序列、数据流、时间空间数据、多媒体数据、文本数据、图、社会网络和 Web 数据。该章根据所挖掘的知识类型、所使用的技术以及目标应用的类型，对数据挖掘任务进行了一般分类。最后讨论该领域的主要挑战。

第 2 章介绍一般数据特征。该章首先讨论数据对象和属性类型，然后介绍基本统计数据描述的典型度量。该章概述各种类型数据的数据可视化技术。除了数值数据的可视化方法外，还介绍文本、标签、图和多维数据的可视化方法。第 2 章还介绍度量各种类型数据的相似性和相异性的方法。

第 3 章介绍数据预处理技术。该章首先介绍数据质量的概念，然后讨论数据清理、数据集成、数据归约、数据变换和数据离散化的方法。

第 4 章和第 5 章是数据仓库、*OLAP*（联机分析处理）和数据立方体技术的引论。**第 4 章**介绍数据仓库和 OLAP 的基本概念、建模、结构、一般实现，以及数据仓库和其他数据泛化的关系。**第 5 章**更深入地考察数据立方体技术，详细地研究数据立方体的计算方法，包括 Star-Cubing 和高维 OLAP 方法。该章还讨论数据立方体和 OLAP 技术的进一步研究，如抽样立方体、排序立方体、预测立方体、用于复杂数据挖掘查询的多特征立方体和发现驱动的数据立方体的探查。

第 6 章和第 7 章介绍挖掘大型数据集中的频繁模式、关联和相关性的方法。**第 6 章**介绍基本概念，如购物篮分析，还有条理地提供了许多频繁项集挖掘技术。这些涵盖从基本 Apriori 算法和它的变形，到改进性能的更高级的方法，包括频繁模式增长方法，使用数据的垂直形式的频繁模式挖掘，挖掘闭频繁项集和极大频繁项集。该章还讨论模式评估方法并介绍挖掘相关模式的度量。**第 7 章**介绍高级模式挖掘方法。该章讨论多层和多维空间中的模式挖掘，挖掘稀有和负模式，挖掘巨型模式和高维空间数据，基于约束的模式挖掘和挖掘压缩或近似模式。该章还介绍模式探查和应用的方法，包括频繁模式的语义注解。

第 8 章和第 9 章介绍数据分类方法。由于分类方法的重要性和多样性，内容被划分成两章。**第 8 章**介绍分类的基本概念和方法，包括决策树归纳、贝叶斯分类和基于规则的分类。该章还讨论模型评估和选择方法，以及提高分类准确率的方法，包括组合方法和处理不平衡数据。**第 9 章**讨论分类的高级方法，包括贝叶斯信念网络、后向传播的神经网络技术、支持向量机、使用频繁模式的分类、k-最邻近分类、基于案例的推理、遗传算法、粗糙集理论和模糊集方法。附加的主题包括多类分类、半监督分类、主动学习和迁移学习。

聚类分析是第 10 章和第 11 章的主题。**第 10 章**介绍数据聚类的基本概念和方法，包括基本聚类分析方法的概述、划分方法、层次方法、基于密度的方法和基于网格的方法。该章还介绍聚类评估方法。**第 11 章**讨论聚类的高级方法，包括基于概率模型的聚类、聚类高维数据、聚类图和网络数据，以及基于约束的聚类。

第 12 章专门讨论离群点检测。本章介绍离群点的基本概念和离群点分析，并从各种监督力度（监督的、半监督的和无监督的）以及方法角度（统计学方法、基于邻近性的方法、基于聚类的方法和基于分类的方法）讨论离群点检测方法。该章还讨论挖掘情境离群点和集体离群点，以及高维数据中的离群点检测。

最后，在**第 13 章**我们讨论数据挖掘的趋势、应用和研究前沿。我们简略地介绍挖掘复杂数据类型，包括挖掘序列数据（例如，时间序列、符号序列和生物学序列），挖掘图和网络，以及挖掘空间、多媒体、文本和 Web 数据。这些数据挖掘方法的深入讨论留给正在撰

写的数据挖掘高级课题一书。然后，该章转向讨论其他数据挖掘方法学，包括统计学数据挖掘、数据挖掘基础、可视和听觉数据挖掘，以及数据挖掘的应用。讨论数据挖掘在金融数据分析、零售和电信产业、科学与工程，以及入侵检测和预防方面的应用。该章还讨论数据挖掘与推荐系统的联系。由于数据挖掘出现在我们日常生活的方方面面，所以我们讨论数据挖掘与社会，包括无处不在和无形的数据挖掘，以及隐私、安全和数据挖掘对社会的影响。我们用考察数据挖掘的发展趋势结束本书。

书中楷体字用于强调定义的术语，而黑体字用于突出主要思想。

本书与其他数据挖掘教材相比具有一些显著特点：它广泛、深入地讨论了数据挖掘原理。各章尽可能是自包含的，使得读者可以按自己感兴趣的次序阅读。高级章节提供了更大的视野，感兴趣的读者可以选读。本书提供了数据挖掘的所有主要方法，还提供了关于多维OLAP分析等数据挖掘的重要主题，这些主题在其他书中常常被忽略或很少提及。本书还维护了一个网站，其中包含大量在线资源，为教师、学生和该领域的专业人员提供支持。这些将在下面介绍。

致教师

本书旨在提供数据挖掘领域的一个广泛而深入的概览，可以作为高年级本科生或一年级研究生的数据挖掘导论。除了讲稿、教师指南和阅读材料列表等教学资源之外，本书网站（*www. cs. uiuc. edu/ ~ hanj/bk3* 或 *www. booksite. mkp. com/datamining3e*）还提供了一个样本课程安排。

根据授课学时、学生的背景和你的兴趣，你可以选取章节的子集，以不同的顺序进行讲授。例如，如果你只打算给学生讲授数据挖掘入门导论，可以按照图 P. 1 的建议。注意，根据需要，必要时可以省略其中某些节或某些小节。

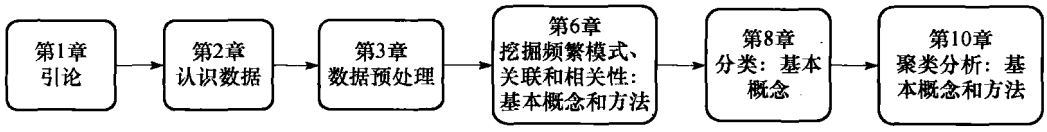


图 P. 1　入门导论课程的建议章节序列

根据学时和讲授范围，你可以有选择地把更多的章节增加到这个基本序列中。例如，对高级分类方法更感兴趣的教师可以首先增加“第 9 章　分类：高级方法”；对模式挖掘更感兴趣的教师可以选择包括“第 7 章　高级模式挖掘”；而对 OLAP 和数据立方体技术感兴趣的教师可以增加“第 4 章　数据仓库与联机分析处理”和“第 5 章　数据立方体技术”。

或者，你可以选择在两个学期的系列课程中讲授整本书，包括本书的所有章节，时间允许的话，加上图和网络挖掘这样的高级课题。这些高级课题可以从本书网站提供的配套材料选择，辅以挑选的研究论文。

本书的每一章都可以用做自学材料，或者用做数据库系统、机器学习、模式识别和数据智能分析等相关课程的专题。

每章后面都有一些习题，适合作为家庭作业。这些习题或者是用于测验对内容的掌握情况的小问题，或者是需要分析思考的大问题，或者是实现设计。有些习题也可以用做研究讨论课题。每章后面的文献注释可以用来查找包含正文中提供的概念和方法的来源、相关课题的深入讨论和可能的扩展的研究文献。

致学生

我们希望本书将激发你对年青，但正在快速发展的数据挖掘领域的兴趣。我们试图以清晰的方式提供材料，仔细地解释所涵盖的主题。每一章后面都附有一个小结，总结要点。全书包含了许多图和解释，以便使本书更加有趣和便于阅读。尽管本书是作为教材编写的，但是我们也试图把它组织成一本有用的参考书或手册，以有助于你今后在数据挖掘方面进行深入研究和求职。

为阅读本书，你需要知道什么？

- 你应当具有关于统计学、数据库系统和机器学习的概念和术语方面的知识。然而，我们尽力提供这些基础知识的足够背景，以便在读者对这些领域不太熟悉或者记忆有些淡忘时，也能够理解本书的讨论。
- 你应当具有一些程序设计经验。特别是你应当能够阅读伪代码，能够理解像多维数组这样的简单数据结构。

致专业人员

本书旨在涵盖数据挖掘领域的广泛主题。因此，本书是关于该主题的一本优秀手册。由于每一章的编写都尽可能独立，所以读者可以关注自己最感兴趣的课题。希望学习数据挖掘关键思想的应用程序员和信息服务管理人员可以使用本书。对于有兴趣使用数据挖掘技术解决其业务问题的银行、保险、医药和零售业的数据分析人员，本书也是有用的。此外，本书也可以作为数据挖掘领域的全面综述，有助于研究人员提升数据挖掘技巧，扩展数据挖掘的应用范围。

本书所提供的技术和算法是实用的，介绍的算法适合于发现隐藏在大型、现实数据集中的模式和知识，而不是挑选在小型“玩具”数据库上运行良好的算法。本书提供的每个算法都用伪代码解释。伪代码类似于程序设计语言 C，但也精心加以策划，使得不熟悉 C 或 C ++ 的程序员易于理解。如果你想实现算法，你会发现将我们的伪代码转换成选定的程序设计语言程序是一项非常简单的任务。

本书资源网站

本书网站的地址是 *www. cs. uiuc. edu/ ~ hanj/bk3*，另一个是 Morgan Kaufmann 出版社的网站 *www. booksite. mkp. com/datamining3e*。这些网站为本书的读者和对数据挖掘感兴趣的人提供了一些附加材料，资源包括：

- **每章的幻灯片**。提供了用微软的 PowerPoint 制作的每章教案。
- **高级数据挖掘的配套章节**。本书第 2 版的第 8 ~ 10 章涵盖了挖掘复杂的数据类型，这超出了本书的主题，对这些高级主题感兴趣的读者可从网站上获取。
- **教师手册**。本书习题的完整答案通过出版社的网站只向教师提供。
- **课程提纲和教学计划**。使用本书和幻灯片用于数据挖掘导论课程和高级教程的本科生和研究生，可以获取这些资源。
- **带超链接的辅助阅读文献列表**。补充读物的原创性文章按章组织。
- **到数据挖掘数据集和软件的链接**。我们将提供到数据挖掘数据集和某些包含有趣的数据挖掘软件包的站点的链接，如到伊利诺伊大学厄巴纳 - 尚佩恩分校 IlliMine 的链接（*http://illimine. cs. uiuc. edu*）。

- **作业、考试和课程设计样本**。一组作业、考试和课程设计样本将在出版社的网站上向教师提供。
- **本书的插图**。这可能有助于你制作自己的课堂教学幻灯片。
- **本书目录**。PDF 格式。
- **本书不同印次的勘误表**。欢迎读者指出本书中的错误。一旦错误被证实，我们将更新勘误表，并对你的贡献致谢。

评论或建议请发往 *hanj@ cs. uiuc. edu*。我们很高兴听到你的建议。

第 3 版致谢

我们向 UIUC 数据挖掘小组以前和现在的所有成员、伊利诺伊大学厄巴纳 - 尚佩恩分校计算机科学系的数据与信息系统实验室（DAIS）的教师和学生以及许多朋友和同事表达我们的诚挚谢意，他们始终不渝的支持使得我们在这一版的工作中受益匪浅。我们还希望感谢 UIUC 2010—2011 学年 CS412 和 CS512 课程的学生，他们仔细地通读了本书的初稿，找出了许多错误，提出了各种改进意见。

我们还希望感谢 Morgan Kaufmann 出版社的发行人 David Bevans 和 Rick Adams，感谢他们在我们写作本书时所表现出的热情、耐心和支持。我们感激该书的项目经理 Marilyn Rash 和她的团队，他们使得我们按期完稿。

我们对所有的评论者不胜感激，感谢他们的无价反馈。此外，我们感谢美国国家科学基金会、NASA、美国空军科学研究办公室、美国军事研究实验室、加拿大自然科学与工程研究委员会（NSERC），以及 IBM 研究院、微软研究院、Google、雅虎研究院、波音、HP 实验室和其他业界实验室，感谢他们在研究基金、合同和赠予方面对我们的研究的支持。这些研究加深了我们对本书所讨论课题的理解。最后，我们感谢我们的家人，感谢他们对该项目的全身心支持。

第 2 版致谢

我们向 UIUC 数据挖掘小组以前和现在的所有成员、伊利诺伊大学厄巴纳 - 尚佩恩分校计算机科学系的数据与信息系统实验室（DAIS）的教师和学生以及许多朋友和同事表示感谢，他们始终不渝的支持使得我们在第 2 版的工作中受益匪浅。这些人包括：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, Wei Zou。

Deng Cai 和 ChengXiang Zhai 对文本挖掘和 Web 挖掘两节，Xifeng Yan 对图挖掘一节，Xiaoxin Yin 对多重关系挖掘一节做出了贡献。Hong Cheng, Charios Ermopoulos, Hector Gonzalez, David J. Hill, Chulyun Kim, Sangkyum Kim, Chao Liu, Hongyan Liu, Kasif

Manzoor, Tianyi Wu, Xifeng Yan, Xiaoxin Yin 校阅了手稿的部分章节。

我们还希望感谢 Morgan Kaufmann 出版社的发行人 Diane Cerra，感谢她在本书写作期间的热情、耐心和支持。我们感激该书的项目经理 Alan Rose，感谢他不知疲倦和及时地与我们联系，安排出版过程的每个细节。我们对所有的评论者不胜感激，感谢他们的无价反馈。最后，我们感谢我们的家人，感谢他们对该项目的全身心支持。

第 1 版致谢

我们希望向曾经或正与我们一道从事数据挖掘相关研究和 DBMiner 项目，或者在数据挖掘方面向我们提供各种支持的所有人表示衷心感谢。这些人包括：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 Man Lam, James Lau, Deyi Li, George (Wenmin) Li, Jin Li, Ze-Nian Li, Nancy Liao, Gang Liu, Junqiang Liu, Ling Liu, Alan (Yijun) Lu, Hongjun Lu, Tong Lu, Wei Lu, Xuebin Lu, Wo-Shun Luk, Heikki Mannila, Runying Mao, Abhay Mehta, Gabor Melli, Alberto Mendelzon, Tim Merrett, Harvey Miller, Drew Miners, Behzad Mortazavi-Asl, Richard Muntz, Raymond T. Ng, Vicent Ng, Shojiro Nishio, Beng-Chin Ooi, Tamer Ozsu, Jian Pei, Gregory Piatetsky-Shapiro, Helen Pinto, Fred Popowich, Amynmohamed Rajan, Peter Scheuermann, Shashi Shekhar, Wei-Min Shen, Avi Silberschatz, Evangelos Simoudis, Nebojsa Stefanovic, Yin Jenny Tam, Simon Tang, Zhaohui Tang, Dick Tsur, Anthony K. H. Tung, Ke Wang, Wei Wang, Zhaoxia Wang, Tony Wind, Lara Winstone, Ju Wu, Betty (Bin) Xia, Cindy M. Xin, Xiaowei Xu, Qiang Yang, Yiwen Yin, Clement Yu, Jeffrey Yu, Philip S. Yu, Osmar R. Zaiane, Carlo Zaniolo, Shuhua Zhang, Zhong Zhang, Yvonne Zheng, Xiaofang Zhou, Hua Zhu。

我们还要感谢 Jean Hou, Helen Pinto, Lara Winstone, Hua Zhu，感谢他们帮助绘制本书的一些草图；感谢 Eugene Belchev，感谢他小心地校对了每一章。

我们还希望感谢 Morgan Kaufmann 出版社的执行总编辑 Diane Cerra，感谢她在本书写作期间的热情、耐心和支持；感谢本书的责任印制 Howard Severson 和他的同事，感谢他们尽职尽责的努力，使本书顺利出版。我们对所有的评论者不胜感激，感谢他们的无价反馈。最后，我们感谢我们的家人，感谢他们对该项目的全身心支持。

试读结束：需要全本请在线购买： www.ertongbook.com