TURING 图灵计算机科学丛书

PEARSON Addison Wesley

数据挖掘导论

Introduction to Data Mining

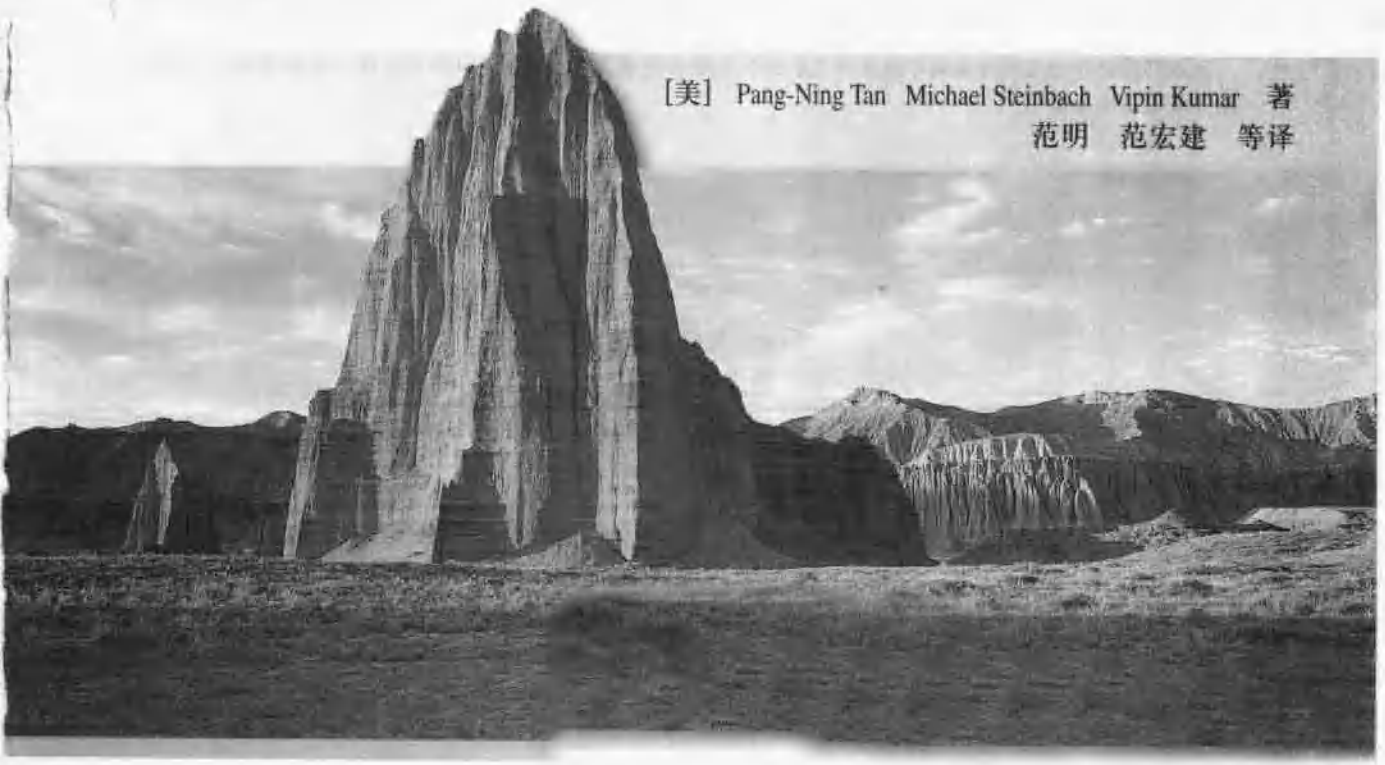
[美] Pang-Ning Tan Michael Steinbach Vipin Kumar 著
范明 范宏建 等译

人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵计算机科学丛书

数据挖掘导论

Introduction to Data Mining



[美] Pang-Ning Tan Michael Steinbach Vipin Kumar 著
范明 范宏建 等译

人民邮电出版社
POSTS & TELECOM PRESS

图书在版编目（CIP）数据

数据挖掘导论/（美）谭，（美）斯坦巴赫著；范明等译. —北京：人民邮电出版社，2006.5

（图灵计算机科学丛书）

ISBN 7-115-14698-5

Ⅰ. 数… Ⅱ. ①谭…②斯…③范… Ⅲ. 数据采集 Ⅳ. TP274

中国版本图书馆 CIP 数据核字（2006）第 032295 号

内 容 提 要

本书全面介绍了数据挖掘的理论和方法，旨在为读者提供将数据挖掘应用于实际问题所必需的知识。本书涵盖五个主题：数据、分类、关联分析、聚类和异常检测。除异常检测外，每个主题都包含两章：前面一章讲述基本概念、代表性算法和评估技术，后面一章较深入地讨论高级概念和算法。目的是使读者在透彻地理解数据挖掘基础的同时，还能了解更多重要的高级主题。此外，书中还提供了大量示例、图表和习题。

本书适合作为相关专业高年级本科生和研究生数据挖掘课程的教材，同时也可作为数据挖掘研究和应用开发人员的参考书。

图灵计算机科学丛书

数据挖掘导论

◆ 著 [美] Pang-Ning Tan Michael Steinbach Vipin Kumar

译 范 明 范宏建 等

责任编辑 杨海玲

◆ 人民邮电出版社出版发行 北京市崇文区夕照寺街 14 号

邮编 100061 电子函件 315@ptpress.com.cn

网址 http://www.ptpress.com.cn

北京鸿佳印刷厂印刷

新华书店总店北京发行所经销

◆ 开本：787×1092 1/16

印张：27.25

字数：694 千字 2006 年 5 月第 1 版

印数：1－4 000 册 2006 年 5 月北京第 1 次印刷

著作权合同登记号 图字：01-2005-5236 号

ISBN 7-115-14698-5/TP・5365

定价：49.00 元

读者服务热线：(010)88593802 印装质量热线：(010)67129223

版 权 声 明

Preface of the Chinese Edition

It is with great pleasure that we welcome the Chinese translation of our book by Professors Fan, Dr. Fan, *et al.*, who have previously translated several well-known statistics and data mining texts. Data mining is an area in computer science that aims to analyze the rapidly increasing amounts of business, scientific, and engineering data for knowledge and other profitable uses. The field has seen tremendous growth and development, with the great influx of scholars and researchers, not only from the Western countries but also from the Far East. We thank Professors Fan and Dr. Fan for their effort in doing the translation, which allows the book to reach a much broader audience among those students and researchers who are well-versed in the Chinese language. We hope that the readers of our book will find it to be both useful and engaging, and wish them the greatest success.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar
Michigan State University and the University of Minnesota, December 2005

中 文 版 序

我们非常高兴范明教授和范宏建博士等人将我们的书翻译成中文，他们在此之前翻译了几本关于统计学和数据挖掘方面的著名教材。数据挖掘是计算机科学的一个领域，其目的是通过分析快速增长的商业、科学和工程数据来获取知识和其他利益。我们已经目睹了这个领域的迅猛增长和发展，学者和研究人员大量涌入其中，他们不仅来自西方国家，而且来自远东地区。我们感谢范明教授和范宏建博士，他们的翻译成果使本书得以传播到更广的读者群，包括那些精通中文的学生和研究人员。我们期望读者会发现这是一部有用的和引人入胜的书籍。祝你们成功！

Pang-Ning Tan
Michael Steinbach
Vipin Kumar
2005 年 12 月于密歇根州立大学和明尼苏达大学

译 者 序

自从我和孟小峰等人翻译 J. Han 和 M. Kamber 的《数据挖掘：概念与技术》以来，我们高兴地看到数据挖掘的研究正在我国蓬勃开展。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，他们之中不仅有来自数据库领域的专家，而且不乏统计学、人工智能和模式识别、机器学习等领域的研究者。国内的学者和研究者在数据挖掘方面的研究已经取得了一些令人鼓舞的成果，并且正在逐渐与国际学术界同步。

数据挖掘的产生和发展一直是分析和理解数据的实际需求推动的。数据挖掘研究的进展也正是在于一直重视与其他领域研究者的合作。数据挖掘从工业、农业、医疗卫生和商业的需求取得动力，从统计学、机器学习等领域的长期研究与发展中汲取营养。我们相信，只要有理解数据的需求，就有推动数据挖掘研究与应用发展的动力；只要依靠多学科的团队，就能应对新的数据分析任务带来的挑战。

P. Tan、M. Steinbach 和 V. Kumar 的这本《数据挖掘导论》是继《数据挖掘：概念与技术》一书之后的另一本重要的数据挖掘著作。三位作者都是数据挖掘方面的研究者。Vipin Kumar 教授是数据挖掘和高性能计算领域的国际知名学者。本书原版在正式出版之前就已经被斯坦福大学、得克萨斯大学奥斯汀分校等众多名校采用。J. Han 教授也高度评价该书："这是一本全新数据挖掘的教材，值得大力推荐。它将成为我们的主要参考书。"

本书的特色在于，不需要读者具备数据库背景，只需要少量统计学或数学背景知识，而且取材涉及的学科和应用领域较多，实用性强，因此所适合的读者面较广。本书更加强调如何用数据挖掘知识解决各种实际问题，更加强调所挖掘的知识模式的评估。例如，就像我们能够从天空中的白云想象出各种动物和物体一样，每个聚类算法能够从几乎所有的数据集中发现聚类。如果数据集合中根本不存在自然的簇，所产生的聚类很难说具有实际意义。

全书共分 10 章。范明负责第 1~8 章的翻译，范宏建负责第 9 和 10 章的翻译。蒋宏杰、贾玉祥、许红涛和温箐笛也参加本书的最初翻译工作。全书的译文由范明负责统一定稿。在翻译的过程中，对发现的错误进行了更正，并得到原书作者的确认。

感谢 P. Tan、M. Steinbach 和 V. Kumar 为中文版撰写序言。感谢人民邮电出版社图灵公司的编辑们，他们在第一时间内引进本书，并组织翻译，使得中文版能够如此之快地与读者见面。

译文中的错误和不当之处，敬请读者朋友指正。意见和建议请发往 mfan@zzu.edu.cn。希望读者喜欢这本译著，希望这本译著有助于推动我国的数据挖掘研究与应用的深入开展。

范 明

2006 年 2 月于郑州大学

译 者 简 介

范明 郑州大学信息工程学院教授，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。当前感兴趣的研究方向包括数据挖掘、数据仓库和机器学习。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。先后发表论文40余篇。除本书外，近年来主持翻译的数据挖掘方面的著作还有 Jiawei Han 和 Micheline Kamber 的《数据挖掘：概念与技术》，Trevor Hastie、Robert Tibshirani 和 Jerome Friedman 的《统计学习基础：数据挖掘、推理与预测》。

范宏建 1999 年毕业于郑州大学计算机科学系，同年进入中国科学院软件研究所攻读硕士学位，次年赴澳大利亚墨尔本大学攻读博士学位，师从澳大利亚科学院院士 Kotagiri Ramamohanarao 教授。2004 年获墨尔本大学计算机科学博士学位。目前是墨尔本大学计算机科学与工程系的博士后。当前感兴趣的主要研究方向为机器学习和数据挖掘。先后在 PAKDD、RSFDGrC、IEEE GrC 和 Australian AI 等国际学术会议和 *IEEE Transactions on Knowledge and Data Engineering* 发表论文 10 余篇。

前　言

数据生成和收集技术的进步正在商业和科研领域产生海量数据集。数据仓库能够存储：企业销售和运作的详细情况，地球轨道卫星发送回地球的高分辨率图像和遥感数据，基因组实验对越来越多的有机体产生的序列、结构和机能数据。收集和存储数据的轻松简便，已经完全改变了人们对数据分析的态度：人们尽可能地收集各个时期和各种来源的数据。人们开始相信收集的数据肯定会有价值，或者为了当初收集它的目的，或者为了尚未知晓的某些理由。

数据挖掘领域冲破了当前数据分析技术在应对这些新型数据集提出的挑战中的种种局限性。数据挖掘并不是要取代其他分析领域，而是将它们作为其工作的基础。尽管数据挖掘的某些主题（如关联分析）是其独有的，但是，另一些主题（如聚类、分类和异常检测）则建立在其他领域在这些主题长期工作的基础之上。事实上，数据挖掘研究者们利用已有技术的自发性对增强和拓展这个领域以及推动它的快速发展起到了促进作用。

该领域的活力还表现在一直强调与其他领域的研究者合作。要迎接分析新类型数据所面临的挑战，抛开理解数据的人和数据所处的领域而简单地使用数据分析技术是不可行的。通常，组建多学科研究团队的高超技巧，已经成为数据挖掘项目（如创建新的独创性算法）取得成功的决定因素。正如历史上统计学的许多进展都是由农业、工业、医疗卫生和商业需求推动的一样，数据挖掘的许多进展也正在被这些领域的需求所推动。

本书源自 1998 年春季开始至今在明尼苏达大学为高年级学生和研究生开设的数据挖掘课程的讲义和教学幻灯片。在这些课程中开发的演示幻灯片和习题随着时间不断积累，成为本书的基础。数据挖掘的聚类技术综述最初是为准备该领域的研究而写的，它也成为本书一章的起点。随着时间的推移，又增加了关于数据、分类、关联分析和异常检测的几章。本书定稿后已在作者所在的学校（明尼苏达大学和密歇根州立大学）以及其他一些大学作为教材经过课堂试用。

在此期间，出现了许多数据挖掘方面的书籍，但是都不能完全满足我们学生的需要——他们主要是计算机科学专业的研究生和本科生，也包括来自工科和其他各学科专业的学生。他们的数学和计算机背景差异很大，但是都有一个共同目标：尽可能直接地学习数据挖掘，以便尽快地将其应用到各自的领域。因此，要求广泛数学和统计学预备知识的书对他们中的许多人都没有吸引力，需要坚实的数据库背景的书也有同样问题。为了满足这些学生需求而逐渐写成的本书，通过使用例子、关键算法的简洁描述和习题，尽可能直接把重点放在数据挖掘的主要概念上。

概述

具体而言，本书提供对数据挖掘的全面介绍，旨在使学生、教师、研究人员和专业人士容易理解并对他们有所帮助。本书所涵盖的领域包括数据预处理、可视化、预测建模、关联分析、聚类和异常检测。目标是讲述每个主题的基本概念和算法，从而为读者提供将数据挖掘应用于实际问题所需的必要背景。此外，本书也为有志于从事数据挖掘和相关领域研究的读者提供一个起点。

本书涵盖五个主题：数据、分类、关联分析、聚类和异常检测。除异常检测外，每个主题都有两章。对于分类、关联分析和聚类，前面一章讲述基本概念、代表性算法和评估技术，而后面较深入的一章讨论高级概念和算法。目的是使读者透彻地理解数据挖掘的基础，同时论述更多重要的高级主题。由于这种安排，本书既可用作为学习工具又可用作参考书。

为了帮助读者理解书中概念，我们提供大量示例、图表和习题。文献注释出现在每一章的结尾，是为那些对更高级的主题、重要的历史文献和当前趋势感兴趣的读者提供的。

致教师

作为一本教材，本书广泛适合于高年级本科生和研究生。由于学习这门课程的学生背景不同，他们可能不具备广博的统计学和数据库知识，因此本书只要求最低限度的预备知识——不需要数据库知识，并假定读者只有一般的统计学或数学背景。本书尽可能自成一体。统计学、线性代数和机器学习的必要基础知识已经结合到正文中。

由于讨论主要数据挖掘主题的各章也是自成一体的，因此主题的讲授次序相当灵活。核心题材在第2、4、6、8和10章介绍。尽管数据导论（第2章）应当最先讨论，但是基本的分类、关联分析和聚类（分别是第4、6、8章）可以以任意次序讲述。由于异常处理（第10章）与分类（第4章）和聚类（第8章）有一定的关系，这两章应当在第10章之前讲述。可以从高级的分类、关联分析和聚类（分别为第5、7、9章）中挑选不同的主题，以适合课程安排和教师与学生的兴趣。我们也建议教师用数据挖掘的实际项目和练习强化课程的教学。尽管这样做很耗费时间，但是实践性的作业可以大大提高这门课程的价值。

支持材料

本书的辅助材料可以在 Addison-Wesley 的 Web 网站（www.aw-bc.com/cssupport）上找到。提供给所有读者的支持材料如下：

- 课程幻灯片。
- 学生项目建议。
- 数据挖掘资源，如数据挖掘算法和数据集。
- 联机指南，使用实际的数据集和数据分析软件，为本书介绍的部分数据挖掘技术提供例子讲解。

其他支持材料（包括习题答案）只向采纳本书做教材的教师提供。评论和建议以及报告错误请通过 dmbook@cs.unm.edu 发给作者。

致谢

许多人都为本书做出了贡献。我们首先向我们的家人表示感谢，这本书是献给他们的。没有他们的耐心和支持，不可能写出本书。

我们要感谢明尼苏达大学和密歇根州立大学数据挖掘小组学生所做的贡献。Eui-Hong（Sam）Han 和 Mahesh Joshi 为最初的数据挖掘课程提供了帮助。他们编制的某些习题和演示幻灯片已经用在本书及其辅助幻灯片中。为本书的初稿提出建议或以其他方式做出贡献的数据挖掘小组中的学生包括 Shyam Boriah、Haibin Cheng、Varun Chandola、Eric Eilertson、Levent Ertöz、Jing Gao、Rohit Gupta、Sridhar Iyer、Jung-Eun Lee、Benjamin Mayer、Aysel Ozgur、Uygar Oztekin、Gaurav

Pandey、Kashif Riaz、Jerry Scripps、Gyorgy Simon、Hui Xiong、Jieping Ye 和 Pusheng Zhang。我们还要感谢明尼苏达大学和密歇根州立大学选修数据挖掘课程的学生，他们使用了本书的初稿，并提供了极富价值的反馈。我们特别感谢 Bernardo Craemer、Arifin Ruslim、Jamshid Vayghan 和 Yu Wei 的有益的建议。

Joydeep Ghosh（得克萨斯大学）和 Sanjay Ranka（佛罗里达大学）试用了本书的早期版本。我们也直接从得克萨斯大学下列学生那里获得了许多有用的建议：Pankaj Adhikari、Rajiv Bhatia、Frederic Bosche、Arindam Chakraborty、Meghana Deodhar、Chris Everson、David Gardner、Saad Godil、Todd Hay、Clint Jones、Ajay Joshi、Joonsoo Lee、Yue Luo、Anuj Nanavati、Tyler Olsen、Sunyoung Park、Aashish Phansalkar、Geoff Prewett、Michael Ryoo、Daryl Shannon 和 Mei Yang。

Ronald Kostoff（ONR）阅读了聚类一章的早期版本，并提出了许多建议。Musetta Steinbach 帮助发现了图中的错误。

我们要感谢明尼苏达大学和密歇根州立大学的同事，他们帮助创建了良好的数据挖掘研究环境。他们是 Dan Boley、Joyce Chai、Anil Jain、Ravi Janardan、Rong Jin、Oeorge Karypis、Haesun Park、William F. Punch、Shashi Shekhar 和 Jaideep Srivastava。我们还要向我们的数据挖掘项目的合作者表示谢意，他们是 Ramesh Agrawal、Steve Cannon、Piet C. de Groen、Fran Hill、Yongdae Kim、Steve Klooster、Kerry Long、Nihar Mahapatra、Chris Potter、Jonathan Shapiro、Kevin Silverstein、Nevin Young 和 Zhi-Li Zhang。

明尼苏达大学和密歇根州立大学的计算机科学与工程系为该项目提供了计算资源和支持环境。ARDA、ARL、ARO、DOE、NASA 和 NSF 为本书作者提供了研究资助。特别应该提到的是，Kamal Abdali、Dick Brackney、Jagdish Chandra、Joe Coughlan、Michael Coyle、Stephen Davis、Frederica Darema、Richard Hirsch、Chandrika Kamath、Raju Namburu、N. Radhakrishnan、James Sidoran、Bhavani Thuraisingham、Walt Tiernin、Maria Zemankova 和 Xiaodong Zhang 有力地支持了我们的数据挖掘和高性能计算研究。

与 Pearson Education 工作人员的合作令人愉快。具体地，我们要感谢 Michelle Brown、Matt Goldstein、Katherine Harutunian、Marilyn Lloyd、Kathy Smith 和 Joyce Wells。我们还要感谢 George Nichols 帮助绘图，Paul Anagnostopoulos 提供 LaTeX 支持。我们感谢 Pearson 邀请的审稿人：Chien-Chung Chan（阿克伦大学）、Zhengxin Chen（内布拉斯加大学奥马哈分校）、Chris Clifton（普度大学）、Joydeep Ghosh（得克萨斯大学奥斯汀分校）、Nazli Goharian（伊利诺伊理工学院）、J. Michael Hardin（阿拉巴马大学）、James Hearne（西华盛顿大学），Hillol Kargupta（马里兰大学巴尔的摩县分校和 Agnik 公司）、Eamonn Keogh（加利福尼亚大学里弗赛德分校）、Bing Liu（伊利诺伊大学芝加哥分校）、Mariofanna Milanova（阿肯色大学小石城分校）、Srinivasan Parthasarathy（俄亥俄州立大学）、Zbigniew W. Ras（北卡罗莱纳大学夏洛特分校）、Xintao Wu（北卡罗莱纳大学夏洛特分校）和 Mohammed J. Zaki（伦斯勒理工学院）。

目　录

第1章

绪　论

数据收集和数据存储技术的快速进步使得各组织机构可以积累海量数据。然而，提取有用的信息已经成为巨大的挑战。通常，由于数据量太大，无法使用传统的数据分析工具和技术处理它们。有时，即使数据集相对较小，由于数据本身的非传统特点，也不能使用传统的方法处理。在另外一些情况下，需要回答的问题不能使用已有的数据分析技术来解决。这样，就需要开发新的方法。

数据挖掘是一种技术，它将传统的数据分析方法与处理大量数据的复杂算法相结合。数据挖掘为探查和分析新的数据类型以及用新方法分析旧有数据类型提供了令人振奋的机会。本章，我们概述数据挖掘，并列举本书所涵盖的关键主题。我们从介绍需要新的数据分析技术的一些著名应用开始。

商务　借助POS（销售点）数据收集技术[条码扫描器、射频识别（RFID）和智能卡技术]，零售商可以在其商店的收银台收集顾客购物的最新数据。零售商可以利用这些信息，加上电子商务网站的日志、电购中心的顾客服务记录等其他的重要商务数据，更好地理解顾客的需求，做出更明智的商务决策。

数据挖掘技术可以用来支持广泛的商务智能应用，如顾客分析、定向营销、工作流管理、商店分布和欺诈检测等。数据挖掘还能帮助零售商回答一些重要的商务问题，如“谁是最有价值的顾客？”“什么产品可以交叉销售[1]或提升销售[2]？”“公司明年的收入前景如何？”这些问题催生了一种新的数据分析技术——关联分析（见第6、7章）。

医学、科学与工程　医学、科学与工程技术界的研究者正在快速积累大量数据，这些数据对获得有价值的新发现至关重要。例如，为了更深入地理解地球的气候系统，NASA已经部署了一系列的地球轨道卫星，不停地收集地表、海洋和大气的全球观测数据。然而，由于这些数据的规模和时空特性，传统的方法常常不适合分析这些数据集。数据挖掘开发的技术可以帮助地球科学家回答如下问题：“干旱和飓风等生态系统扰动的频度和强度与全球变暖之间有何联系？”“海洋表面温度对地表降水量和温度有何影响？”“如何准确地预测一个地区的生长季节的开始和结束？”

再举一个例子，分子生物学研究者希望利用当前收集的大量基因组数据，更好地理解基因的结构和功能。过去，传统方法只允许科学家在一个实验中每次研究少量基因。微阵列技术的最新突破已经能让科学家在多种情况下，比较数以千计的基因的特性。这种比较有助于确定每个基因的作用，或许可以查出导致特定疾病的基因。然而，由于数据的噪声和高维性，需要新的数据分

1. cross-sell，指根据顾客的兴趣推荐或显示相关商品以增加销售机会。——译者注
2. up-sell，指尝试向曾经购买的顾客销售价格更高的商品。——译者注

析方法。除分析基因序列数据外，数据挖掘还能用来处理生物学的其他难题，如蛋白质结构预测、多序列校准、生物化学路径建模和种系发生学。

1.1　什么是数据挖掘

数据挖掘是在大型数据存储库中，自动地发现有用信息的过程。数据挖掘技术用来探查大型数据库，发现先前未知的有用模式。数据挖掘还具有预测未来观测结果的能力，例如，预测一位新的顾客是否会在一家百货公司消费 100 美元以上。

并非所有的信息发现任务都被视为数据挖掘。例如，使用数据库管理系统查找个别的记录，或通过因特网的搜索引擎查找特定的 Web 页面，则是**信息检索**（information retrieval）领域的任务。虽然这些任务是重要的，可能涉及使用复杂的算法和数据结构，但是它们主要依赖传统的计算机科学技术和数据的明显特征来创建索引结构，从而有效地组织和检索信息。尽管如此，数据挖掘技术也已用来增强信息检索系统的能力。

数据挖掘与知识发现

数据挖掘是**数据库中知识发现**（knowledge discovery in database, KDD）不可缺少的一部分，而 KDD 是将未加工的数据转换为有用信息的整个过程，如图 1-1 所示。该过程包括一系列转换步骤，从数据的预处理到数据挖掘结果的后处理。

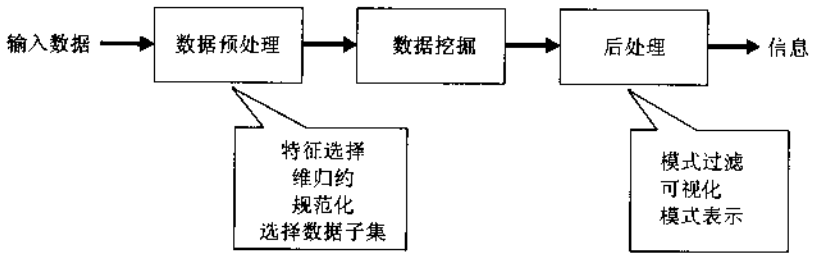


图 1-1　数据库中知识发现（KDD）过程

输入数据可以以各种形式存储（平展文件、电子数据表或关系表），并且可以驻留在集中的数据存储库中，或分布在多个站点上。**数据预处理**（preprocessing）的目的是将未加工的输入数据转换成适合分析的形式。数据预处理涉及的步骤包括融合来自多个数据源的数据，清洗数据以消除噪声和重复的观测值，选择与当前数据挖掘任务相关的记录和特征。由于收集和存储数据的方式可能有许多种，数据预处理可能是整个知识发现过程中最费力、最耗时的步骤。

“结束循环（closing the loop）”通常指将数据挖掘结果集成到决策支持系统的过程。例如，在商务应用中，数据挖掘的结果所揭示的规律可以与商务活动管理工具集成，使得可以进行和测试有效的商品促销活动。这样的集成需要**后处理**（postprocessing）步骤，确保只将那些有效的和有用的结果集成到决策支持系统中。后处理的一个例子是可视化（见第 3 章），它使得数据分析者可以从各种不同的视角探查数据和数据挖掘结果。在后处理阶段，还能使用统计度量或假设检验，删除虚假的数据挖掘结果。

1.2　引发数据挖掘的挑战

正如前面所提到的，当面临新的数据集提出的挑战时，传统的数据分析技术常常遇到实际困

难。下面是一些特定的挑战，它们引发了对数据挖掘的研究。

可伸缩 由于数据产生和收集技术的进步，数吉字节、数太字节甚至数拍字节的数据集越来越普遍。如果数据挖掘算法要处理这些海量数据集，则算法必须是可伸缩的（scalable）。许多数据挖掘算法使用特殊的搜索策略处理指数性搜索问题。可伸缩可能还需要实现新的数据结构，以有效的方式访问个别记录。例如，当要处理的数据不能放进内存时，可能需要非内存算法。使用抽样技术或开发并行和分布算法也可以提高可伸缩程度。

高维性 现在，常常遇到具有数以百计或数以千计属性的数据集，而不是数十年前常见的只具有少量属性的数据集。在生物信息学领域，微阵列技术的进步已经产生了涉及数千特征的基因表达数据。具有时间或空间分量的数据集也趋向于具有很高的维度。例如，考虑包含不同地区的温度测量的数据集。如果温度在一个相当长的时间周期内重复地测量，则维度（特征数）的增长正比于测量的次数。为低维数据开发的传统的数据分析技术通常不能很好地处理这样的高维数据。此外，对于某些数据分析算法，随着维度（特征数）的增加，计算复杂性迅速增加。

异种数据和复杂数据 通常，传统的数据分析方法只处理包含相同类型属性的数据集，或者是连续的，或者是分类的。随着数据挖掘在商务、科学、医学和其他领域的作用越来越大，越来越需要能够处理异种属性的技术。近年来，已经出现了更复杂的数据对象。这些非传统的数据类型的例子包括含有半结构化文本和超链接的 Web 页面集、具有序列和三维结构的 DNA 数据、包含地球表面不同位置上的时间序列测量值（温度、气压等）的气象数据。为挖掘这种复杂对象而开发的技术应当考虑数据中的联系，如时间和空间的自相关性、图的连通性、半结构化文本和 XML 文档中元素之间的父子联系。

数据的所有权与分布 有时，需要分析的数据并非存放在一个站点，或归属一个单位，而是地理上分布在属于多个机构的资源中。这就需要开发分布式数据挖掘技术。分布式数据挖掘算法面临的主要挑战包括：(1) 如何降低执行分布式计算所需的通信量？(2) 如何有效地统一从多个资源得到的数据挖掘结果？(3) 如何处理数据安全性问题？

非传统的分析 传统的统计方法基于一种假设-检验模式。换句话说，提出一种假设，设计实验来收集数据，然后针对假设分析数据。但是，这一过程劳力费神。当前的数据分析任务常常需要产生和评估数以千计的假设，因此希望自动地产生和评估假设导致了一些数据挖掘技术的开发。此外，数据挖掘所分析的数据集通常不是精心设计的实验的结果，并且它们通常代表数据的时机性样本（opportunistic sample），而不是随机样本（random sample）。而且，这些数据集常常涉及非传统的数据类型和数据分布。

1.3 数据挖掘的起源

为迎接前一节中的这些挑战，来自不同学科的研究者汇集到一起，开始着手开发可以处理不同数据类型的更有效的、可伸缩的工具。这些工作建立在研究者先前使用的方法学和算法之上，在数据挖掘领域达到高潮。特别地，数据挖掘利用了来自如下一些领域的思想：(1) 来自统计学的抽样、估计和假设检验，(2) 人工智能、模式识别和机器学习的搜索算法、建模技术和学习理论。数据挖掘也迅速地接纳了来自其他领域的思想，这些领域包括最优化、进化计算、信息论、