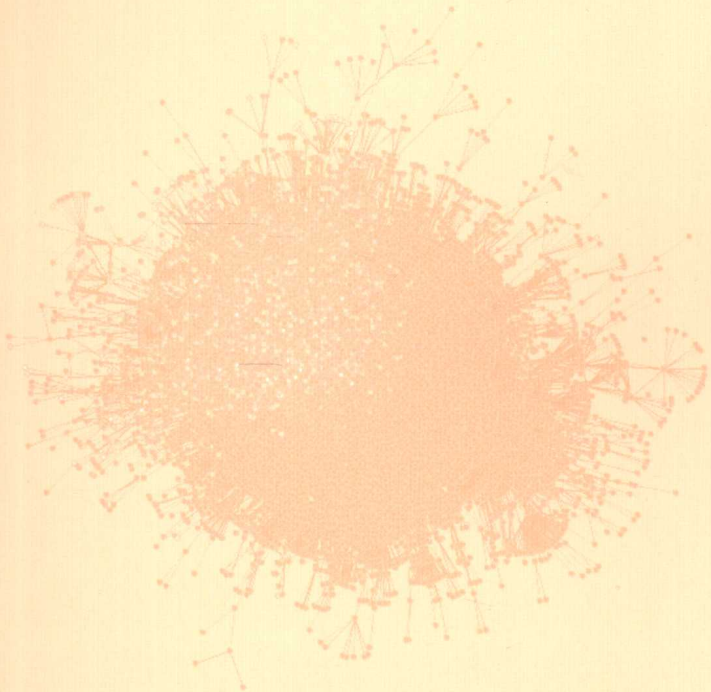
精品

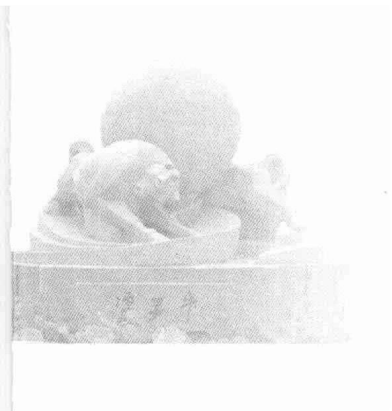
# 数据挖掘

第 2 版

朱　明　编著



中国科学技术大学出版社



中国科学技术大学精品教材

# 数据挖掘

SHUJU WAJUE

第 2 版

朱　明　编著

中国科学技术大学出版社

## 内 容 简 介

数据挖掘技术，又称为数据库知识发现，是20世纪90年代在信息技术领域开始迅速发展起来的计算机技术。作者结合自己近20年从事人工智能、机器学习、数据挖掘等方面的科研工作积累与教学经验，编著此书。

本书较全面系统地介绍了数据挖掘中常用和常见的数据挖掘方法，以及文本与视频数据挖掘方法。

本书的主要内容包括：数据挖掘基本知识、数据挖掘预处理方法、决策树分类及其他分类方法、关联知识挖掘方法、各种聚类分析方法，以及文本挖掘所涉及表示、分类和聚类等方法，还包括视频挖掘所涉及的视频镜头检测、字幕提取、视频摘要和视频检索等主要分析方法。

本书作为学习、掌握和应用数据挖掘方法和技术的综合指导书，是从事数据挖掘研究与应用人员，以及希望了解数据挖掘主要方法和技术的IT技术人员的良师益友；同时也是一本可用于大学高年级或研究生相关课程的教材和参考文献。

**图书在版编目(CIP)数据**

数据挖掘/朱明编著. —2版. —合肥：中国科学技术大学出版社，2008.11
（中国科学技术大学精品教材）
“十一五”国家重点图书
安徽省高等学校“十一五”省级规划教材
ISBN 978-7-312-02244-9

Ⅰ.数… Ⅱ.朱… Ⅲ.数据采集—高等学校—教材 Ⅳ.TP274

中国版本图书馆CIP数据核字(2008)第165028号

中国科学技术大学出版社出版发行
安徽省合肥市金寨路96号，230026
网址 http://press.ustc.edu.cn
安徽辉隆农资集团瑞隆印务有限公司
全国新华书店经销

开本：710×960 1/16 印张：31.5 插页：2 字数：580千
2008年11月第2版 2008年11月第2次印刷
印数：4001—7000册
定价：52.00元

# 总　序

2008年是中国科学技术大学建校五十周年。为了反映五十年来办学理念和特色，集中展示教材建设的成果，学校决定组织编写出版代表中国科学技术大学教学水平的精品教材系列。在各方的共同努力下，共组织选题281种，经过多轮、严格的评审，最后确定50种入选精品教材系列。

1958年学校成立之时，教员大部分都来自中国科学院的各个研究所。作为各个研究所的科研人员，他们到学校后保持了教学的同时又作研究的传统。同时，根据“全院办校，所系结合”的原则，科学院各个研究所在科研第一线工作的杰出科学家也参与学校的教学，为本科生授课，将最新的科研成果融入到教学中。五十年来，外界环境和内在条件都发生了很大变化，但学校以教学为主、教学与科研相结合的方针没有变。正因为坚持了科学与技术相结合、理论与实践相结合、教学与科研相结合的方针，并形成了优良的传统，才培养出了一批又一批高质量的人才。

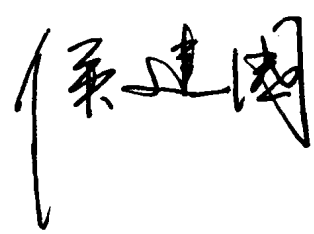
学校非常重视基础课和专业基础课教学的传统，也是她特别成功的原因之一。当今社会，科技发展突飞猛进、科技成果日新月异，没有扎实的基础知识，很难在科学技术研究中作出重大贡献。建校之初，华罗庚、吴有训、严济慈等老一辈科学家、教育家就身体力行，亲自为本科生讲授基础课。他们以渊博的学识、精湛的讲课艺术、高尚的师德，带出一批又一批杰出的年轻教员，培养了一届又一届优秀学生。这次入选校庆精品教材的绝大部分是本科生基础课或专业基础课的教材，其作者大多直接或间接受到过这些老一辈科学家、教育家的教诲和影响，因此在教材中也贯穿着这些先辈的教育教学理念与科学探索精神。

改革开放之初，学校最先选派青年骨干教师赴西方国家交流、学习，他们在带回先进科学技术的同时，也把西方先进的教育理念、教学方法、教学内容等带回到中国科学技术大学，并以极大的热情进行教学实践，使“科学与技术相结

合、理论与实践相结合、教学与科研相结合”的方针得到进一步深化，取得了非常好的效果，培养的学生得到全社会的认可。这些教学改革影响深远，直到今天仍然受到学生的欢迎，并辐射到其他高校。在入选的精品教材中，这种理念与尝试也都有充分的体现。

中国科学技术大学自建校以来就形成的又一传统是根据学生的特点，用创新的精神编写教材。五十年来，进入我校学习的都是基础扎实、学业优秀、求知欲强、勇于探索和追求的学生，针对他们的具体情况编写教材，才能更加有利于培养他们的创新精神。教师们坚持教学与科研的结合，根据自己的科研体会，借鉴目前国外相关专业有关课程的经验，注意理论与实际应用的结合，基础知识与最新发展的结合，课堂教学与课外实践的结合，精心组织材料、认真编写教材，使学生在掌握扎实的理论基础的同时，了解最新的研究方法，掌握实际应用的技术。

这次入选的50种精品教材，既是教学一线教师长期教学积累的成果，也是学校五十年教学传统的体现，反映了中国科学技术大学的教学理念、教学特色和教学改革成果。该系列精品教材的出版，既是向学校50周年校庆的献礼，也是对那些在学校发展历史中留下宝贵财富的老一代科学家、教育家的最好纪念。



2008年8月

# 前 言

随着数据库应用的普及，人们已陷入“数据丰富，知识匮乏”的尴尬境地。近年来网络的发展与普及，使得人类开始真正体会到数据海洋无边无际。面对如此巨大的数据资源，人们迫切需要新的数据分析方法和技术，以便能够利用信息技术发展的最新成果，来将这巨大的数据资源转换成有价值的信息与知识，并为我们制定科学决策提供必要的支持。

数据挖掘(Data Mining，简称 DM)，作为 20 世纪末刚刚兴起的数据智能分析技术，由于其所具有的广阔应用前景而备受关注。作为数据库与数据仓库应用研究而逐步形成的一个新兴的富有应用前景的领域，数据挖掘常常也被称为数据库知识发现(Knowledge Discovery from Database，简称 KDD)，它可以从数据库、数据仓库或其他数据源中，通过分析，自动抽取归纳出有价值的知识模式。

数据挖掘是一个多领域交叉的研究与应用领域，涉及的领域包括：数据库技术、人工智能、机器学习、神经网络、统计学、模式识别、信息检索、高性能计算等。本书主要介绍能够对大量数据进行挖掘处理的有关方法与技术，主要包括：数据挖掘基本知识、数据挖掘预处理方法、决策树分类及其他分类方法、关联知识挖掘方法、各种聚类分析方法，以及文本挖掘所涉及表示、分类和聚类等方法，还包括视频挖掘所涉及的视频镜头检测、字幕提取、视频摘要和视频检索等所涉及的主要分析方法。

数据挖掘作为一门新兴的学科，经过短短十几年的发展，就已表现出了强大生命力。广大从事数据库应用与决策支持以及模式识别、机器学习、智能计算等领域的科研人员迫切需要了解和掌握它。为此作者结合自己近 20 年从事人工智能、机器学习、数据挖掘等方面的科研工作积累与教学经验，并在本人 2002 年编著出版的《数据挖掘》一书基础上，结合国内外数据挖掘研究领域的最新研究成果，尤其借鉴中国科学院计算所、国防科学技术大学、复旦大学、北京邮电大学等国内知名科研单位在数据挖掘方面的研究成果，编写完成此书，以飨读者。

本书共分十章，主要介绍数据挖掘的基本概念、方法、技术与应用等方面的内容。前七章，主要介绍数据挖掘的主要方法；后三章则侧重介绍数据挖掘在文本挖掘和视频挖掘方面的具体应用。

第1章主要介绍了数据挖掘发展背景，数据挖掘定义及过程，数据挖掘主要功能，数据挖掘应用与发展趋势。

第2章主要介绍了数据预处理相关内容，主要包括：数据类型与质量的描述、数据清理方法、数据集成与变换方法以及数据归约方法。

第3章主要介绍了分类挖掘中最主要的一类方法：基于决策树的分类挖掘，具体内容包括：决策树基本方法、C4.5方法、CART方法、SLIQ方法和SPRINT方法；同时介绍了基于决策树的分类方法的简化与改进；最后介绍了分类模型的评估方法。

第4章主要介绍了分类挖掘中的其他方法，主要包括：贝叶斯分类方法、K近邻方法、人工神经网络方法、遗传进化方法、支持向量机方法、粗糙集方法和集成学习方法。

第5章主要介绍了关联挖掘的主要方法，具体包括：Apriori算法及其改进FP-tree算法、关联挖掘的三种并行算法和基于粒计算的关联挖掘方法。

第6章主要介绍了基于划分、层次、密度、网格和模型等五种聚类方法，同时还介绍了高维海量数据的聚类挖掘方法，以及基于蚁群算法的两种聚类挖掘方法。

第7章主要介绍了异类挖掘五种方法：基于统计、基于距离、基于偏差和基于密度的异常点检测方法，以及基于高维数据的异常点检测方法。此外还介绍了基于属性的异常点检测方法、时序异常点检测方法、空间异常点检测方法，以及时空异常点检测方法；最后介绍了数据流异常检测的两种方法。

第8章主要介绍了文本挖掘的主要内容，包括：文本分类方法、文本预处理方法和中文分词方法；此外还介绍了中文信息摘要方法、文本内容监管方法和文本信息检索方法。

第9章主要介绍了视频挖掘所涉及的内容，包括：四种镜头检测方法、关键帧提取方法、镜头快速切变检测方法和镜头渐变检测方法；此外还介绍了新闻视频挖掘所涉及的镜头检测、播音员镜头检测和新闻故事单元检测方法；同时介绍了广告检测所涉及的镜头切变、阈值选择和基于音频融合的广告检测方法；最后介绍了视频文本检测的四种方法。

第10章主要介绍了视频分析所涉及的慢镜头检测四种方法、视频摘要的实现方法，以及视频检索的相关方法。

本书在编写过程中，得到了中国科学技术大学信息学院吴刚院长的大力支持，在此表示感谢。

尽管作者付出诸多努力以求本书在内容上充实、全面，但是由于作者水平所限，加之数据挖掘领域的发展，尤其科研成果的快速更新，书中不足和错误之处在所难免，恳请广大读者和同行专家不吝赐教。

朱　明

2008年3月20日于合肥

# 目 次

# 第 1 章　数据挖掘导论

数据挖掘是 20 世纪 80 年代末开始逐步发展起来的一个新的研究领域，它是多个学科和技术相结合的产物。本章将简要介绍数据挖掘的发展背景、概念定义、主要方法及应用案例等内容。

## 1.1　数据挖掘的发展背景

随着数据库技术的迅速发展以及数据库管理系统的广泛应用，人们利用信息技术生产和搜集数据的能力大幅度提高，无数个数据库被用于商业管理、政府办公、科学研究和工程开发等领域，超级市场中的交易数据、加油站里的汽油销售数据、旅行社的旅游信息等等，均构成了数据库系统的信息来源。近年来，数据库所管理的数据量急剧增大，人们积累的数据越来越多。例如：美国 NASA 的地球观测系统(EOS)每小时向地面发回约 50 GB 的图像数据；美国沃尔玛零售系统每天会产生约 2 亿条交易数据。人们希望能够对其进行更高层次的分析，以便更好地利用这些数据。激增的数据背后隐藏着许多重要的信息，目前的数据库系统可以高效地实现数据的录入、查询、统计等功能，但无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势。缺乏挖掘数据背后隐藏的知识的手段，导致了“数据富有但知识贫乏”的现象。于是，一个新的挑战被提了出来：在这被称之为信息爆炸的时代，信息过量几乎成为人人需要面对的问题。如何才能不被信息的汪洋大海所淹没，从中及时发现有用的知识，提高信息利用率呢？要想使数据真正成为一个企业的资源，只有充分利用它为企业自身的业务决策和战略发展服务才行，否则大量的数据可能成为包袱，甚至成为垃圾。

建立在数据库系统上的计算机决策支持系统的出现，为进行高层次的数据决策分析提供了好的思路和方法。但由于决策支持系统在数据的采集、分析方法上的灵活性等方面存在局限性，使得人们不得不寻求更有效的途径去开拓数据决策分析的思路。人工智能为此作出了巨大贡献，人工智能经历了博弈、自然语言理解、知识工程等阶段，已经进入了机器学习的热点阶段。机器学习能够模拟人类的学习方式，通过对数据对象之间关系的分析，提取出隐含在数据中的模式，即知识。

正是由于实际工作的需要和相关技术的发展，利用数据库技术来存储管理数据，利用机器学习的方法来分析数据，从而挖掘出大量的隐藏在数据背后的知识，这些思想的结合最终形成了备受人们关注的研究领域：数据库中的知识发现(Knowledge Discovery in Databases，KDD)。其中，数据挖掘技术便是 KDD 中的一个最为关键的环节。

1995 年，在加拿大蒙特利尔召开了第一届知识发现和数据挖掘国际学术会议，数据挖掘一词被很快流传开来。人们将存储在数据库中的数据看作是形成知识的源泉，形象地将它们比喻成矿石。数据挖掘(Data Mining，DM)就是从大量的、不完全的、有噪声的、模糊的数据中，提取隐含在其中的、人们事先不知道的，但又是潜在有用的信息和知识的过程。

数据挖掘技术是人们长期对数据库技术进行研究和开发的结果。起初各种商业数据是存储在计算机的数据库中的，然后发展到可对数据库进行查询和访问，进而发展到对数据库的即时遍历。数据挖掘使数据库技术进入了一个更高级的阶段，它不仅能对过去的数据进行查询和遍历，并且能够找出过去数据之间的潜在联系，从而促进信息的传递。

推动数据挖掘的开发、应用和研究兴趣的重要技术因素分别是：

- 超大规模数据库的出现，例如商业数据仓库和计算机自动收集的数据记录；
- 先进的计算机技术，例如更快和更大的计算能力和并行体系结构；
- 对巨大量数据的快速访问，如 Web 搜索；
- 对这些数据挖掘相关的挖掘算法研究的深入。

商业数据库现在正在以空前的速度增长，并且数据仓库正在广泛地应用于各种行业；对计算机硬件性能越来越高的要求，也可以用现在已经成熟的并行多处理机的技术来满足；另外数据挖掘算法经过了这 10 多年的发展也已经成为一种成熟、稳定，且易于理解和操作的技术。

数据挖掘的进化历程如图 1.1 所示。从商业数据到商业信息的进化过程中，每一步前进都是建立在上一步的基础上的。从图 1.1 中可以看到，第四步进化是革命性的，因为从用户的角度来看，这一阶段的数据库技术已经可以快速地回答商

业上的很多问题了。

| 进化阶段 | 商业问题 | 支持技术 | 产品厂家 | 产品特点 |
| --- | --- | --- | --- | --- |
| 数据搜集（20 世纪 60 年代） | “过去五年中我的总收入是多少?” | 计算机、磁带和磁盘 | IBM,CDC | 提供历史性的、静态的数据信息 |
| 数据访问（20 世纪 80 年代） | “在新英格兰的分部去年三月的销售额是多少?” | 关系数据库(RDBMS)，结构化查询语言(SQL)，ODBC Oracle，Sybase，Informix，IBM，Microsoft | Oracle，Sybase，Informix，IBM，Microsoft | 在记录级提供历史性的、动态数据信息 |
| 数据仓库；决策支持（20 世纪 90 年代） | “在新英格兰的分部去年三月的销售额是多少？波士顿据此可得出什么结论?” | 联机分析处理(OLAP)、多维数据库、数据仓库 | Pilot，Comshare，Arbor，Cognos，Microstrategy | 在各种层次上提供回溯的、动态的数据信息 |
| 数据挖掘（正在流行） | “下个月波士顿的销售会怎么样？为什么?” | 高级算法、多处理器计算机、海量数据库 | Pilot，Lockheed，IBM，SGI,其他初创公司 | 提供预测性的信息 |

**图 1.1　数据挖掘的进化历程**

数据挖掘是一门交叉学科，它汇聚了数据库、人工智能、统计学、可视化、并行计算等不同学科和领域，因此近年来受到各界的广泛关注。

数据挖掘其实是一个逐渐演变的过程，电子数据处理的初期，人们就试图通过某些方法来实现自动决策支持，当时机器学习成为人们关心的焦点。机器学习的过程就是将一些已知的并已被成功解决的问题作为范例输入计算机，机器通过学习这些范例总结并生成相应的规则，这些规则具有通用性，使用它们可以解决某一类的问题。随后，随着神经网络技术的形成和发展，人们的注意力转向知识工程，知识工程不同于机器学习那样给计算机输入范例，让它生成出规则，而是直接给计算机输入已被代码化的规则，而计算机是通过使用这些规则来解决某些问题。专家系统就是这种方法所得到的成果，但它有投资大、效果不甚理想等不足。20 世纪 80 年代人们又在新的神经网络理论的指导下，重新回到机器学习的方法上，并