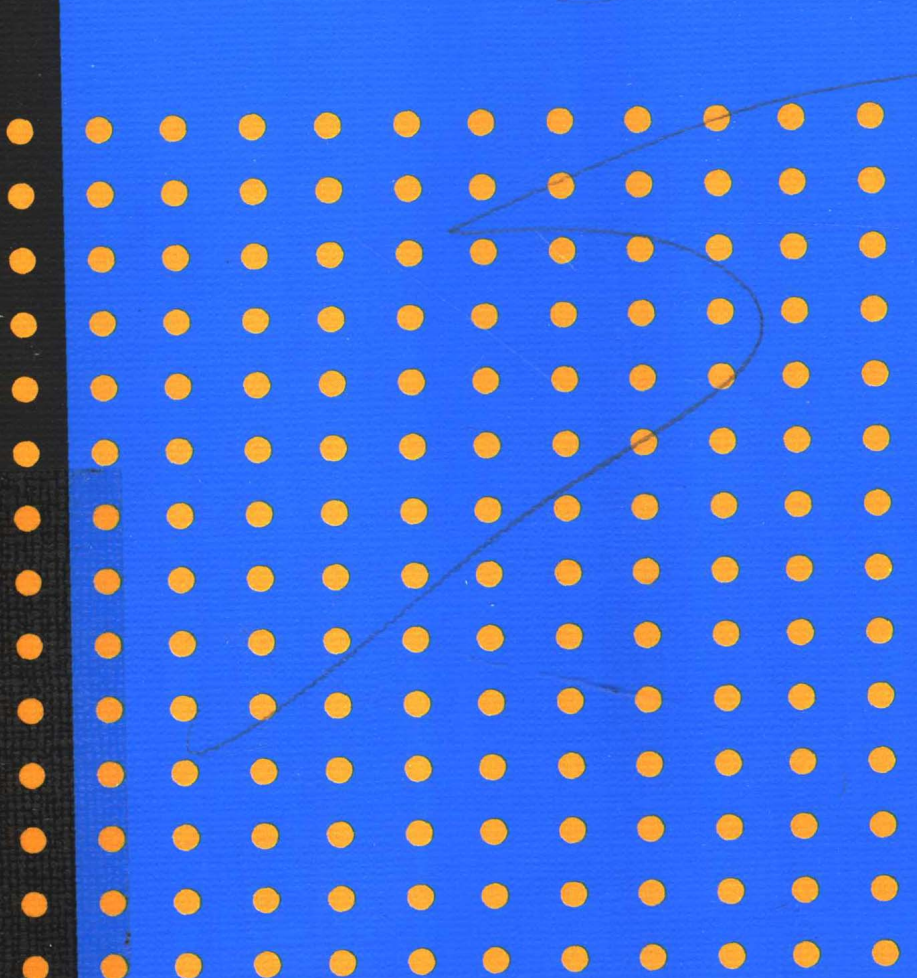
普通高等教育“十一五”国家级规划教材

重点大学计算机专业系列教材

数据挖掘原理与算法（第二版）

毛国君　段立娟　王实　石云　编著



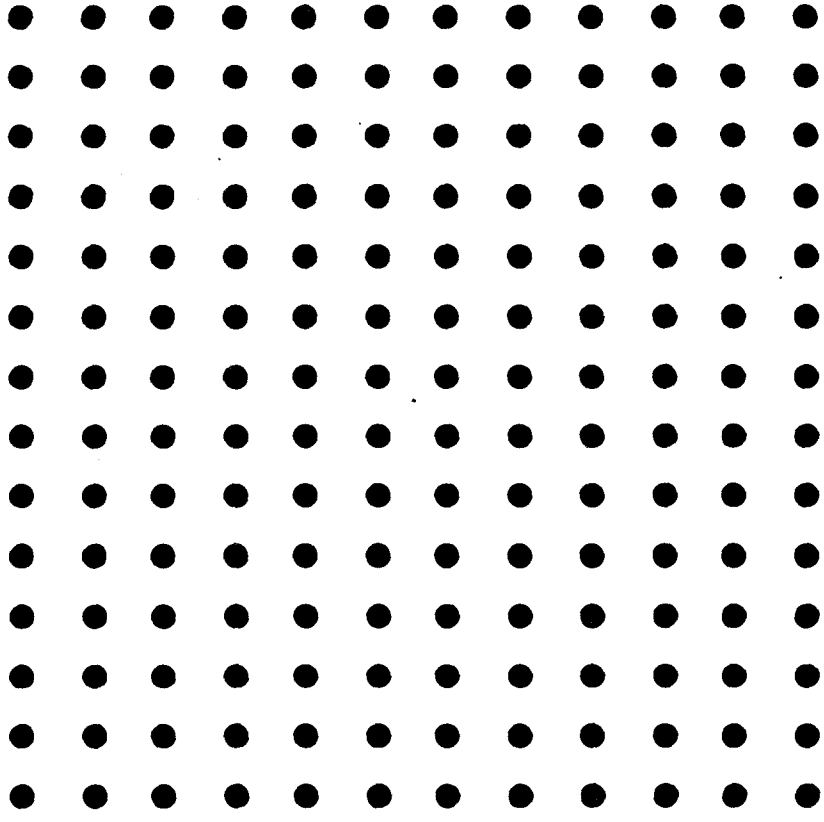
清华大学出版社

普通高等教育“十一五”国家级

重点大学计算机专业系列教材

数据挖掘原理与算法
(第二版)

毛国君　段立娟　王实　石云　编著



清华大学出版社
北京

内 容 简 介

本书是一本全面介绍数据挖掘和知识发现技术的专业书籍，它系统地阐述了数据挖掘和知识发现技术的产生、发展、应用以及相关概念、原理和算法，对数据挖掘中的主要技术分支，包括关联规则、分类、聚类、序列、空间以及 Web 挖掘等进行了理论剖析和算法描述。本书的许多内容是作者们在攻读博士学位期间的工作总结，一方面，对于相关概念和技术的阐述尽量先从理论分析入手，在此基础上进行技术归纳；另一方面，为了保证技术的系统性，所有的挖掘模型和算法描述都在统一的技术归纳框架下进行。同时，为了避免抽象算法描述给读者带来的理解困难，本书的所有典型算法都通过具体跟踪执行实例来进一步说明。

本书共分 8 章，各章相对独立成篇，以利于读者选择性学习。在每章后面都设置专门一节来对本章内容和文献引用情况进行归纳，它不仅可以帮助读者对相关内容进行整理，而且也起到对本内容相关文献的注释性索引功能。第 1 章是绪论，系统地介绍了数据挖掘产生的商业和技术背景，从不同侧面剖析了数据挖掘的概念和应用价值；第 2 章给出了知识发现的过程分析和应用体系结构设计；第 3 章对关联规则挖掘的原理和算法进行全面阐述；第 4 章给出分类的主要理论和算法描述；第 5 章讨论聚类的常用技术和算法；第 6 章对时间序列分析技术和序列挖掘算法进行论述；第 7 章系统地介绍了 Web 挖掘的主要研究领域和相关技术及算法；第 8 章是对空间数据挖掘技术和算法的分析和讲述。

本书可作为计算机专业研究生或高年级本科生教材，也可以作为从事计算机研究和开发人员的参考资料。作为教材，教师可以根据课时安排进行选择性教学。为了更好地让教师进行选择性教学，本书配有专门的教师用书，对内容的重点、难点和课时分配给出了对应的建议，对重要的和难度较大的习题进行了分析和解答。对于研究人员，本书是一本高参考价值的专业书籍。对于软件技术人员，可以把它当作提高用书或参考资料，一些算法可以通过改造用于实际的应用系统中。

图书在版编目(CIP)数据

数据挖掘原理与算法/毛国君等编著. —2 版. —北京：清华大学出版社，2007.12
(重点大学计算机专业系列教材)
ISBN 978-7-302-15876-9

Ⅰ. 数…　Ⅱ. 毛…　Ⅲ. 数据采集　Ⅳ. TP274

中国版本图书馆 CIP 数据核字(2007)第 119309 号

责任编辑：丁　岭　李　晔
责任校对：时翠兰
责任印制：何　芊
出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　**邮　　编**：100084
c-service@tup.tsinghua.edu.cn
社 总 机：010-62770175　**邮购热线**：010-62786544
投稿咨询：010-62772015　**客户服务**：010-62776969
印 刷 者：北京国马印刷厂
装 订 者：三河市溧源装订厂
经　　销：全国新华书店
开　　本：185×260　**印　　张**：21.75　**字　　数**：497 千字
版　　次：2007 年 12 月第 2 版　　**印　　次**：2007 年 12 月第1 次印刷
印　　数：1～4000
定　　价：29.00 元

INTRODUCTION

出版说明

随着国家信息化步伐的加快和高等教育规模的扩大，社会对计算机专业人才的需求不仅体现在数量的增加上，而且体现在质量要求的提高上，培养具有研究和实践能力的高层次的计算机专业人才已成为许多重点大学计算机专业教育的主要目标。目前，我国共有16个国家重点学科、20个博士点一级学科、28个博士点二级学科集中在教育部部属重点大学，这些高校在计算机教学和科研方面具有一定优势，并且大多以国际著名大学计算机教育为参照系，具有系统完善的教学课程体系、教学实验体系、教学质量保证体系和人才培养评估体系等综合体系，形成了培养一流人才的教学和科研环境。

重点大学计算机学科的教学与科研氛围是培养一流计算机人才的基础，其中专业教材的使用和建设则是这种氛围的重要组成部分，一批具有学科方向特色优势的计算机专业教材作为各重点大学的重点建设项目成果得到肯定。为了展示和发扬各重点大学在计算机专业教育上的优势，特别是专业教材建设上的优势，同时配合各重点大学的计算机学科建设和专业课程教学需要，在教育部相关教学指导委员会专家的建议和各重点大学的大力支持下，清华大学出版社规划并出版本系列教材。本系列教材的建设旨在“汇聚学科精英、引领学科建设、培育专业英才”，同时以教材示范各重点大学的优秀教学理念、教学方法、教学手段和教学内容等。

本系列教材在规划过程中体现了如下一些基本组织原则和特点。

1. 面向学科发展的前沿，适应当前社会对计算机专业高级人才的培养需求。教材内容以基本理论为基础，反映基本理论和原理的综合应用，重视实践和应用环节。

2. 反映教学需要，促进教学发展。教材要能适应多样化的教学需要，正确把握教学内容和课程体系的改革方向。在选择教材内容和编写体系时注意体现素质教育、创新能力与实践能力的培养，为学生知识、能力、素质协调发展创造条件。

3. 实施精品战略，突出重点，保证质量。规划教材建设的重点依然是专业基础课和专业主干课；特别注意选择并安排了一部分原来基础比较好的优秀教材或讲义修订再版，逐步形成精品教材；提倡并鼓励编写体现重点大学计算机专业教学内容和课程体系改革成果的教材。

4. 主张一纲多本，合理配套。专业基础课和专业主干课教材要配套，同一门课程可以有多本具有不同内容特点的教材。处理好教材统一性与多样化的关系；基本教材与辅助教材以及教学参考书的关系；文字教材与软件教材的关系，实现教材系列资源配套。

5. 依靠专家，择优落实。在制订教材规划时要依靠各课程专家在调查研究本课程教材建设现状的基础上提出规划选题。在落实主编人选时，要引入竞争机制，通过申报、评审确定主编。书稿完成后要认真实行审稿程序，确保出书质量。

繁荣教材出版事业，提高教材质量的关键是教师。建立一支高水平的以老带新的教材编写队伍才能保证教材的编写质量，希望有志于教材建设的教师能够加入到我们的编写队伍中来。

教材编委会

FOREWORD

前言

众所周知，数据库技术从20世纪80年代开始，已经得到广泛的应用。随着数据库容量的膨胀，特别是数据仓库以及Web等新型数据源的日益普及，人们面临的主要问题不再是缺乏足够的信息可以使用，而是面对浩瀚的数据海洋如何有效地利用这些数据。面对这一挑战，数据挖掘和知识发现技术应运而生，并显示出强大的生命力。数据挖掘和知识发现使数据处理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，而且能够找出过去数据之间的潜在联系，进行更高层次的分析，以便更好地解决决策、预测等问题。历经了十几年的发展，数据挖掘技术本身已经积累了一批有价值的理论和技术成果。同时，包括统计学、人工智能等在内相关学科的发展从某种程度上对数据挖掘技术的发展起到了极大的推动作用。根据麻省理工学院的《科技评论》评估，“数据挖掘”技术是对未来人类产生重大影响的十大新兴技术之一。毫不夸张地说，如今的“数据挖掘”已经成为计算机、信息科学以及相关领域的一个时髦名词，而且在诸如银行、电信、保险、交通、零售（如超级市场）以及天文学、分子生物学等领域得到应用。

诚然，要真正理解数据挖掘技术并不是一件容易的事。一方面，数据挖掘技术覆盖范围很广泛，需要从理论到应用、从概念到算法的完整过程来理解；另一方面，作为年轻的交叉研究领域，不同背景的研究人员（数据库、人工智能、数学等）可能提供不同的视角，而且本身仍在发展中。第一作者长期从事相关方面的教学工作，其中面临的问题之一就是教材的选择。由于目前相关书籍较少，而且侧重点不同，内容的完整性和科学性有待商榷。由于没有合适教材可用，在教学的初期不得不通过指定大量参考书或文献来解决，之后也采用补充讲义的形式来扩充。同时，对于一些软件工程师或工程硕士、在职硕士进修班等要求提高实践能力的人员来说，也需要在科学的理论（原理）框架下理解和掌握数据挖掘技术。基于这样的要求，本书的第一作者在多年各类教学和软件工程的实践基础上，对积累的素材进行了整理和加工，并且邀请段立娟博士、王实博士和石云博士

进行本书的编写。本书的许多内容是几位作者在攻读博士学位期间的工作总结。这些保证了本书的系统性、先进性和实用性。

本书可作为计算机专业研究生教材、高年级本科生选修教材,也可以作为从事计算机研究和开发人员的参考资料。为了保证内容的先进性和深度,对重点内容进行了重点阐述。本书内容相对全面,各章之间耦合度小。作为教材,教师可以根据学生类型、学时安排等进行选择性教学。作为参考书,读者可以根据自己的基础进行选择性学习或查阅。由于在每章后面都设置专门一节来对本章内容和文献引用情况进行归纳,它不仅可以帮助读者对相关内容进行整理,而且对读者,特别是研究人员,也起到文献的注释性索引功能。本书的所有典型算法都通过具体跟踪执行实例来进一步说明,这对于读者正确理解和应用算法是有益的。对于工程技术人员来说,这些算法完全可以通过理解基础上的改进或改造应用到实际工作中。

本书共分8章。第1章是绪论,系统地介绍了数据挖掘的概念、产生背景以及应用价值;第2章给出了知识发现的过程分析和应用体系结构设计,并对数据挖掘应用系统的主要功能部件和关键步骤进行了较为详尽的剖析;第3章全面阐述了关联规则挖掘的原理和算法,并对一些新的焦点问题(如多维、数量、约束关联规则挖掘)的最新成果尽可能地加以介绍;第4章给出分类的主要理论和算法描述;第5章讨论聚类的常用技术和算法;第6章对时间序列分析技术和序列挖掘算法进行论述;第7章系统地介绍了Web挖掘的主要研究领域和相关技术及算法;第8章是对空间数据挖掘技术和算法的分析和讲述。本书的第1～3章由毛国君执笔,第4～6章由段立娟执笔,第7章由王实执笔,第8章由石云执笔,全书由毛国君统稿。

特别感谢北京工业大学刘椿年教授和中国科学院高文和孙玉方研究员,因为他们作为作者们的导师,在作者们攻读博士学位期间对本书素材的积累提供了极大的帮助。本书也凝聚了北京工业大学硕士研究生徐启贵、鲁杰、尤春梅、邱洪君、罗春雨、孙岳、刘旭、杨霞玲、孙晓希、韩连华和一些本科学生的心血,他们在本书算法实例整理和验证等方面进行了很好的工作。此外,第一作者也感谢北京工业大学参加过相关课程学习的各类学生,它们的许多意见和文字更正,提高了本书的内容编排质量。同时第一作者也感谢所有作者及其家人,我们的合作是愉快的,家人的支持是保证本书顺利出版的必要前提。相信通过我们出色而有成效的工作将为读者提供一本有价值的专业书籍。本书对应的讲义和2005版,被许多高校作为研究生和本科生教材使用,经过近2年的尝试,使用者发现一些问题,并且提出了许多宝贵的意见。另外,本书得到了北京工业大学研究生课程建设经费的支持,这对本书的改版起到了重要的作用。对此,作者一并表示感谢。

作　者

2007年5月于北京

CONTENTS

目录

CHAPTER 1

绪　论　第1章

数据挖掘(Data Mining)是一个多学科交叉研究领域,它融合了数据库(Database)技术、人工智能(Artificial Intelligence)、机器学习(Machine Learning)、统计学(Statistics)、知识工程(Knowledge Engineering)、面向对象方法(Object-Oriented Method)、信息检索(Information Retrieval)、高性能计算(High-Performance Computing)以及数据可视化(Data Visualization)等最新技术的研究成果。经过十几年的研究,产生了许多新概念和新方法。特别是最近几年,一些基本概念和方法趋于清晰,它的研究正向着更深入的方向发展。

数据挖掘之所以被称为未来信息处理的骨干技术之一,主要在于它以一种全新的概念改变着人类利用数据的方式。20世纪,数据库技术取得了决定性的成果并且已经得到广泛的应用。但是,数据库技术作为一种基本的信息存储和管理方式,仍然以联机事务处理(On-Line Transaction Processing,OLTP)为核心应用,缺少对决策、分析、预测等高级功能的支持机制。众所周知,随着数据库容量的膨胀,特别是数据仓库(Data Warehouse)以及Web等新型数据源的日益普及,联机分析处理(On-Line Analytic Processing, OLAP)、决策支持(Decision Support)以及分类(Classification)、聚类(Clustering)等复杂应用成为必然。面对这一挑战,数据挖掘和知识发现(Knowledge Discovery)技术应运而生,并显示出强大的生命力。数据挖掘和知识发现使数据处理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,并且能够找出过去数据之间的潜在联系,进行更高层次的分析,以便更好地做出理想的决策、预测未来的发展趋势等。通过数据挖掘,有价值的知识、规则或高层次的信息就能从数据库的相关数据集合中抽取出来,从而使大型数据库作为一个丰富、可靠的资源为知识的提取服务。

特别需要指出的是,数据挖掘技术从一开始就是面向应用的。它不仅仅是面向特定数据库的简单检索查询应用,而是要对这些数据进行微观、中观乃至宏观的统计、分析、综合和推理,进而发现潜在的知识。这里所说的知识发现,不是要求发现放之四海而皆准的真理,也不是要去发现崭新

的自然科学定理和纯数学公式。所有发现的知识都是相对的,是面向特定领域的,同时还要能够易于被用户理解。

1.1 数据挖掘技术的产生与发展

1.1.1 数据挖掘技术的商业需求分析

数据挖掘之所以吸引专家学者的研究兴趣和引起商业厂家的广泛关注,主要在于大型数据系统的广泛使用和把数据转换成有用知识的迫切需要。20 世纪 60 年代,为了适应信息的电子化要求,信息技术一直从简单的文件处理系统向有效的数据库系统变革。20 世纪 70 年代,数据库系统的三个主要模式:层次、网络和关系型数据库的研究和开发取得了重要进展。20 世纪 80 年代,关系型数据库及其相关的数据模型工具、数据索引及数据组织技术被广泛采用,并且成为了整个数据库市场的主导。从 20 世纪 80 年代中期开始,关系型数据库技术和新型技术的结合成为数据库研究和开发的重要标志。从数据模型上看,诸如扩展关系、面向对象、对象-关系(Object-Relation)以及演绎模型等被应用到数据库系统中。从应用的数据类型上看,包括空间、时态、多媒体以及 Web 等新型数据成为数据库应用的重要数据源。同时,事务数据库(Transaction Database)、主动数据库(Active Database)、知识库(Knowledge Base)、办公信息库(Information Base)等技术也得到蓬勃发展。从数据的分布角度看,分布式数据库(Distributed Database)及其透明性、并发控制、并行处理等成为必须面对的课题。进入 90 年代,分布式数据库理论上趋于成熟,分布式数据库技术得到了广泛应用。目前,由于各种新型技术与数据库技术的有机结合,使数据库领域中的新内容、新应用、新技术层出不穷,形成了庞大的数据库家族。但是,这些数据库的应用都是以实时查询处理技术为基础的。从本质上说,查询是对数据库的被动使用。由于简单查询只是数据库内容的选择性输出,因此它和人们期望的分析预测、决策支持等高级应用仍有很大距离。

新的需求推动新的技术的诞生。随着信息技术的高速发展,数据库应用的规模、范围和深度不断扩大,已经从单台机器发展到网络环境。近年来由于数据采集技术的更新,如商业条码的推广、企业和政府利用计算机管理事务的能力增强,产生了大规模的数据。数以百万计的数据库系统在运行,而且每天都在增加。决策所面对的数据量在不断增长,即使像使用 IC 卡和打电话这样简单的事务也能产生大量的数据。随着数据的急剧增长,现有信息管理系统中的数据分析工具已无法适应新的需求。因为无论是查询、统计还是报表,其处理方式都是对指定的数据进行简单的数字处理,而不能对这些数据所包含的内在信息进行提取。人们希望能够提供更高层次的数据分析功能,自动和智能地将待处理的数据转化为有用的信息和知识。

数据挖掘的基础是数据分析方法。数据分析是科学研究的基础,许多科学研究都是建立在数据收集和分析基础上的。同时在目前的商业活动中,数据分析总是和一些特殊的人群的高智商行为联系起来,因为并不是每个人都能从过去的销售情况预测将来的发展趋势或作出正确决策的。但是,随着一个企业或行业业务数据的不断积累,特别是由于数据库的普及,人工去整理和理解如此大的数据源已经存在效率、准确性等问题。因此,

探讨自动化的数据分析技术，为企业提供能带来商业利润的决策信息就成为了必然。

事实上，可以将数据(Data)、信息(Information)和知识(Knowledge)看作是广义数据表现的不同形式。毫不夸张地说，人们对于数据的拥有欲是贪婪的，特别是计算机存储技术和网络技术的发展加速了人们收集数据的范围和容量。这种贪婪的结果导致了“数据丰富而信息贫乏(Data Rich & Information Poor)”现象的产生。数据库是目前组织和存储数据的最有效方法之一，但是面对日益膨胀的数据，数据库查询技术已表现出它的局限性。直观上说，信息或称有效信息是指对人们有帮助的数据。例如，在现实社会中，如果人均日阅读时间在30分钟的话，一个人一天最快只能浏览一份20版左右的报纸。如果你订阅了100份报纸，其实你每天也不过只阅读了一份而已。面对计算机中的海量数据，人们也处于同样的尴尬境地，缺乏获取有效信息的手段。知识是一种概念、规则、模式和规律等，它不会像数据或信息那么具体，但是它却是人们一直不懈追求的目标。事实上，在我们的生活中，人们只是把数据看作是形成知识的源泉。我们是通过正面的或反面的数据或信息来形成和验证知识的，同时又不断地利用知识来获得新的信息。因此，随着数据的膨胀和技术环境的进步，人们对联机决策和分析等高级信息处理的要求越来越迫切。在强大的商业需求的驱动下，商家们开始注意到有效地解决大容量数据的利用问题具有巨大的商机。学者们开始思考如何从大容量数据集中获取有用信息和知识的方法。因此，在20世纪80年代后期，产生了数据仓库和数据挖掘等信息处理思想。

1.1.2 数据挖掘产生的技术背景分析

任何技术的产生总是有它的技术背景的。数据挖掘技术的提出和普遍接受是由于计算机及其相关技术的发展为其提供了研究和应用的技术基础。

归纳数据挖掘产生的技术背景，下面一些相关技术的发展起到了决定性的作用：

- 数据库、数据仓库和Internet等信息技术的发展。
- 计算机性能的提高和先进的体系结构的发展。
- 统计学和人工智能等方法在数据分析中的研究和应用。

数据库技术从20世纪80年代开始，已经得到广泛的普及和应用。在关系型数据库的研究和产品提升过程中，人们一直在探索组织大型数据和快速访问的相关技术。高性能关系型数据库引擎以及相关的分布式查询、并发控制等技术的使用，已经提升了数据库的应用能力。在数据的快速访问、集成与抽取等问题的解决上积累了经验。数据仓库作为一种新型的数据存储和处理手段，被数据库厂商普遍接受并且相关辅助建模和管理工具快速推向市场，成为多数据源集成的一种有效的技术支撑环境。另外，Internet的普及也为人们提供了丰富的数据源。据说，在美国，电视普及达到5000万户大约用了15年，而Internet上网普及达到5000万户仅用了4年。而且Internet技术本身的发展，已经不光是简单的信息浏览，以Web计算为核心的信息处理技术可以处理Internet环境下的多种信息源。因此，人们已经具备利用多种方式存储海量数据的能力。只有这样，数据挖掘技术才能有它的用武之地。这些丰富多彩的数据存储、管理以及访问技术的发展，为数据挖掘技术的研究和应用提供了丰富的土壤。

计算机芯片技术的发展，使计算机的处理和存储能力日益提高。大家熟知的摩尔定律告诉我们，计算机硬件的关键指标大约以每18个月翻一番的速度在增长，而且现在看