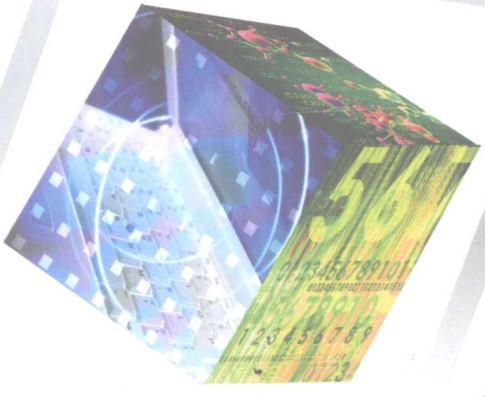
21世纪高等学校精品教材

谭建豪 章 兢 黄 耀 胡章谋 编著

数据挖掘技术



中国水利水电出版社
www.waterpub.com.cn

21世纪高等学校精品教材

数据挖掘技术

谭建豪　章　兢　黄　耀　胡章谋　编　著

中国水利水电出版社
www.waterpub.com.cn

内 容 提 要

本书较为系统地介绍了数据挖掘的基本概念、基本方法和基本技术以及数据挖掘的最新进展，并以较大篇幅叙述了数据挖掘在复杂工业系统中的应用情况。

本书深入而系统地阐述了数据挖掘的研究历史和现状、数据挖掘与数理统计的关系、数据挖掘技术（包括语义网络、智能体、分类、预测、复杂类型数据等基础概念和技术）、数据库系统及专家系统中的数据挖掘方式、数据挖掘的应用及一些具有挑战性的研究课题，对每类问题均提供了代表性算法和具体应用法则。全书共分7章，主要内容包括数据挖掘综述、从数理统计到数据挖掘、语义网络挖掘及其应用、智能体挖掘及其应用、分类挖掘及其应用、预测挖掘及其应用和复杂类型数据挖掘及其应用。

本书可作为高等院校自动化、电子信息、测控技术与仪表、电气工程、系统工程、机电工程等专业的本科生和研究生教材，也可作为相关专业工程技术人员的自学参考书。

本书配有免费电子教案，读者可以从中国水利水电出版社网站下载，网址为：http://www.waterpub.com.cn/softdown/，或与作者联系（tanjianhao96@sina.com.cn），获取更多教学资源。

图书在版编目（CIP）数据

数据挖掘技术/谭建豪等编著. —北京：中国水利水电出版社，2009

21世纪高等学校精品教材

ISBN 978-7-5084-6207-3

Ⅰ. 数… Ⅱ. 谭… 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212187号

书　　名	21世纪高等学校精品教材 **数据挖掘技术**
作　　者	谭建豪　章　兢　黄　耀　胡章谋　编　著
出版发行	中国水利水电出版社（北京市三里河路6号　100044） 网址：www.waterpub.com.cn E-mail：mchannel@263.net（万水） sales@waterpub.com.cn 电话：（010）63202266（总机）、68367658（营销中心）、82562819（万水）
经　　售	全国各地新华书店和相关出版物销售网点
排　　版	北京万水电子信息有限公司
印　　刷	北京蓝空印刷厂
规　　格	184mm×260mm　16开本　18.5印张　451千字
版　　次	2009年1月第1版　2009年1月第1次印刷
印　　数	0001—4000册
定　　价	35.00元

前　言

数据挖掘是一门综合性学科，借鉴了多门学科的概念、理论、方法和技术。这些学科包括数据库系统、专家系统、机器学习、统计学、模式识别、信息检索、人工神经网络（正文中简称神经网络）、支持向量机、遗传算法、模糊优化、高性能计算和数据可视化等。数据挖掘旨在发现隐藏在大型数据集中的知识模式，其基本问题是可行性、实用性、有效性和可伸缩性问题。

数据挖掘出现于20世纪80年代后期，随着国内外研究的深入，已经取得了一批有价值的成果，并展现出良好的发展势头。本书较系统地阐述了该领域目前的研究状况，介绍了几种典型的数据挖掘技术和系统，并讨论了数据挖掘的应用情况和研究方向。

与数据挖掘有关的一些书籍主要是关于数据库方面的。随着大规模数据库（或信息库）的广泛应用，人们已不满足于仅对数据或信息进行查询和检索，因为简单的查询和检索一般不能使用户直接获得带有结论性的信息。因此仅仅依靠查询和检索的手段，数据库蕴藏的知识是得不到充分发掘和利用的。

另外，人工智能、专家系统和机器学习方面的书籍对数据挖掘也有所论述。研究虽然取得了一定的进展，但是知识获取仍然是瓶颈问题，知识工程师从专家那里获取知识的方式仍然带有很强的个体性和随机性，没有统一的方法。

知识发现与数据挖掘是一门新的技术学科，其理论框架与技术原理尚处于研究的初级阶段，还没有形成完整的理论体系。尽管如此，仅从已有的研究成果看，也足以展示其广阔的研究领域和应用前景。本书拟在阐述知识发现与数据挖掘原理的基础上，探索将其用于复杂工业系统的原理、方法和技术。

目前，许多学校正在对传统的教学内容进行改革，自动化、电子信息、测控技术与仪表、电气工程、系统工程、机电工程等专业迫切需要较多与信息相关的知识。由于这些专业的学生在本科阶段已较为扎实地掌握了数据库系统的知识，在研究生阶段便转入人工智能理论与技术的学习。如何将数据挖掘技术及人工智能理论在一个大的框架内以统一的视角用于解决工程实际问题，是本书着力的方向。

本教材是为电子信息及相关专业编写的。作为一门技术基础课，它以“数据库技术”及“人工智能”课程为先修课。既要学时少，又要让学生对数据挖掘原理及其在复杂工业系统中的应用建立较全面的印象，同时还应该使学生学有所用，并为今后的发展打下基础，这是本书编写的指导思想。编者力求避免本书与先修课程内容重复，对书中必不可少的相关知识只作简单介绍。

本书深入而系统地阐述了数据挖掘的研究历史和现状、数据挖掘与数理统计的关系、数据挖掘技术（包括语义网络、智能体、分类、预测、复杂类型数据等基础概念和技术）、数据库系统及专家系统中的数据挖掘方式、数据挖掘的应用及一些具有挑战性的研究课题，并对每

类问题均提供代表性算法和具体应用法则。

全书共分7章,各章内容安排如下:

第1章,对数据挖掘进行综述。从技术角度和商业角度对数据挖掘进行定义,对数据挖掘与传统数据分析方法、数据挖掘和数据仓库、数据挖掘和在线分析处理(OLAP)、数据挖掘、机器学习和统计、软硬件发展对数据挖掘的影响之间的关系进行讨论;探讨数据挖掘所发现的知识类型、数据挖掘的功能、数据挖掘常用技术、数据挖掘中的数据仓库等内容;阐述数据挖掘系统的工作原理,其中包括数据挖掘系统结构和数据挖掘流程。

第2章,介绍从数理统计到数据挖掘的发展过程。阐述数据挖掘与数理统计的关系,对数理统计和数据库技术的结合进行讨论,由此说明数理统计在数据挖掘中的基础地位;重点讨论数理统计中的核心分析方法——回归分析法;就回归分析的基本概念、线性回归方程、线性相关的显著性检验、非线性回归分析、多元线性回归分析、一般情况下的线性回归分析进行论述;结合数据挖掘的特点,给出采用逐步回归分析法建立锻模设计准则的实例;就逐步回归分析的软件设计、锻模飞边尺寸设计准则的制定、锻模飞边金属消耗设计准则的制定等问题进行论述;最后,得出利用逐步回归分析软件建立的上述两类准则,并对结果进行分析,获得相关结论。

第3章,研究语义网络挖掘及其应用问题。阐述语义网络的概念,论述语义网络挖掘的原理;对课题基于AutoCAD的注塑模架设计专家系统进行详细讨论,以此说明语义网络挖掘在CAD系统中的应用情况。

第4章,研究智能体挖掘及其应用问题。阐述智能体概念,论述智能体挖掘算法;对课题基于智能对象和模糊推理的注塑模普通浇注系统进行详细讨论,以此说明智能体挖掘在CAD系统中的应用情况。

第5章,研究分类挖掘及其应用问题。阐述分类概念,论述决策树分类、贝叶斯分类、基于关联规则分类、基于数据库技术分类、基于支持向量机的分类、基于AIS模型分类等分类算法;对课题人工免疫算法及其在故障诊断中的应用进行详细的讨论,以此说明分类挖掘在解决复杂工程问题中的应用情况。

第6章,研究预测挖掘及其应用问题。阐述预测概念,论述技术(统计)预测、信息预测、拟合预测等预测挖掘方法,并与传统预测挖掘方法进行比较,介绍智能预测挖掘方法;对课题基于遗传算法的模糊优化算法及其在预测挖掘中的应用进行详细讨论,以此说明预测挖掘在CAD系统中的应用情况。

第7章,介绍复杂类型数据挖掘及其应用状况。探讨数据挖掘未来研究方向,论述网站数据挖掘、文本数据挖掘、语音数据挖掘、空间数据挖掘、图像数据挖掘等复杂类型数据挖掘;描述数据挖掘在超市布局、客户关系管理、天文数据分析、欺诈甄别等方面的应用情况;分析数据挖掘的技术、经济及社会因素。

本书可作为高等院校自动化、电子信息、测控技术与仪表、电气工程、系统工程、机电工程等专业的本科生和研究生教材,也可作为相关专业工程技术人员的自学参考书。

本书编者从事自动化专业的教学与科研十多年,积累了丰富的教学经验和可供参考的科研成果,这是本书得以成功编写的关键。

本书有关研究得到湖南省自然科技学术著作出版基金、国家自然科学基金(批准文号60634020)、湖南省自然科学基金(批准文号08JJ3132)的资助,中国水利水电出版社相关领导与编辑对本书的出版给予了大力支持,作者借此机会深表谢意。

在本书编写过程中,得到了鲁蓉蓉老师的鼎立支持和研究生陈文斌、刘小林、蒋海波、张伟刚、李丹、李晓光、唐莎、郭芙的大力帮助,在此表示衷心的感谢。

由于作者水平有限,书中不妥之处在所难免,恳请读者指正。

编　者

2008 年 9 月

目　录

第 1 章　数据挖掘综述

1.1　数据挖掘的研究历史和现状

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，随着人工神经网络（以下简称神经网络）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，它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。专家系统就是用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，并将其成果应用于处理大型商业数据库。

随着数据库技术的不断发展及数据库管理系统的广泛应用，数据库中存储的数据量急剧增大，在大量的数据背后隐藏着许多重要的信息，如果能把这些信息从数据库中抽取出来，将为公司创造很多潜在的利润，数据挖掘概念就是从这样的商业角度开发出来的。

数据仓库技术的发展与数据挖掘有着密切的关系。数据仓库的发展是推动数据挖掘技术发展的原因之一。但是，数据仓库并不是数据挖掘的先决条件，因为有很多数据挖掘可直接从操作数据源中挖掘信息。

确切地说，数据挖掘（Data Mining），又称数据库中的知识发现（Knowledge Discovery in Database - KDD），是指从大型数据库或数据仓库中提取隐含的、未知的、非平凡的及有潜在应用价值的信息或模式，它是数据库研究中的一个很有应用价值的新领域，融合了数据库、人工智能、机器学习、统计学等多个领域的理论和技术。

数据挖掘工具能够对将来的趋势和行为进行预测，从而很好地支持人们的决策。比如，经过对公司整个数据库系统的分析，数据挖掘工具可以回答诸如“哪个客户对我们公司的邮件推销活动最有可能作出反应？为什么？”等类似的问题。有些数据挖掘工具还能够解决一些很消耗人工时间的传统问题，因为它们能够快速地浏览整个数据库，找出一些专家们不易察觉的极有用的信息。

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。

我们以研究数据挖掘的历史可以发现，数据挖掘的快速增长和商业数据库的空前速度增长是分不开的，并且 20 世纪 90 年代较为成熟的数据仓库正同样广泛地应用于各种商业领域。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。

表1-1给出了数据进化的四个阶段，从中可以看到，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。

表1-1　数据进化的四个阶段

进化阶段	时间段	技术支持	生产厂家	产品特点
数据搜集	20世纪60年代	计算机，磁带等	IBM，CDC	提供静态历史数据
数据访问	20世纪80年代	关系数据库，结构化查询语言SQL	Oracle，Sybase，Informix，IBM，Microsoft	在记录中提供动态历史数据信息
数据仓库	20世纪90年代	联机分析处理，多维数据库	Pilot，Comshare，Arbor，Cognos，Microstrategy	在各层次提供回溯的动态的历史数据
数据挖掘	正在流行	高级算法，多处理系统，海量算法	Pilot，Lockheed，IBM，SGI，其他初创公司	可提供预测性信息

数据库中的“发现知识(KDD)”一词首次出现在1989年举行的第十一届国际联合人工智能学术会议上。到目前为止，由美国人工智能协会主办的KDD国际研讨会已经召开了8次，规模由原来的专题讨论会发展到国际学术大会，研究重点也逐渐从发现方法转向系统应用，注重多种发现策略和技术的集成，以及多种学科之间的相互渗透。1999年，亚太地区在北京召开的第三届PAKDD会议收到158篇论文。IEEE的Knowledge and Data Engineering会刊率先在1993年出版了KDD技术专刊。并行计算、计算机网络和信息工程等其他领域的国际学会、学刊也把数据挖掘和知识发现列为专题和专刊讨论。

目前，世界上比较有影响的典型数据挖掘系统有：SAS公司的Enterprise Miner，IBM公司的Intelligent Miner，SGI公司的SetMiner，SPSS公司的Clementine，Sybase公司的Warehouse Studio，RuleQuest Research公司的See5，还有CoverStory、EXPLORA、Knowledge Discovery Workbench、DBMiner、Quest等。

与国外相比，国内对DMKDD的研究稍晚，没有形成整体力量。1993年国家自然科学基金首次支持该领域的研究项目。目前，国内的许多科研单位和高等院校竞相开展知识发现的基础理论及其应用研究，这些单位包括清华大学、中科院计算技术研究所、空军第三研究所、海军装备论证中心等。其中，北京系统工程研究所对模糊方法在知识发现中的应用进行了较深入的研究，北京大学也在开展对数据立方体代数的研究，华中科技大学、复旦大学、浙江大学、中国科技大学、中科院数学研究所、吉林大学等单位开展了对关联规则开采算法的优化和改造，南京大学、四川联合大学和上海交通大学等单位探讨、研究了非结构化数据的知识发现及Web数据挖掘。

1.2　数据挖掘定义

1.2.1　技术角度的定义

数据挖掘(Data Mining)就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道但又是有用的信息和知识的过程。

与数据挖掘相近的同义词有数据融合、数据分析和决策支持等。数据挖掘的定义包括

几层含义:数据源必须是真实的、大量的、含噪声的;发现的是用户感兴趣的知识;发现的知识要可接受、可理解、可运用;并不要求发现放之四海而皆准的知识,仅支持特定发现的问题即可。

何谓知识?从广义上理解,数据、信息也是知识的表现形式,但是人们更把概念、规则、模式、规律和约束等看作知识。人们把数据看作是形成知识的源泉,好像从矿石中采矿或淘金一样。原始数据可以是结构化的,如关系数据库中的数据,也可以是半结构化的,如文本、图形和图像数据,甚至是分布在网络上的异构型数据。发现知识的方法可以是数学的,也可以是非数学的;可以是演绎的,也可以是归纳的。发现的知识可以用于信息管理、查询优化、决策支持和过程控制等,还可以用于数据自身的维护。因此,数据挖掘是一门交叉学科,它把人们对数据的应用从低层次的简单查询,提升到从数据中挖掘知识,提供决策支持。在这种需求牵引下,汇聚了不同领域的研究者,尤其是数据库技术、人工智能技术、数理统计、可视化技术、并行计算等方面的学者和工程技术人员,都投身到数据挖掘这一新兴的研究领域,形成新的技术热点。

这里所说的知识发现,不是要求发现普遍的真理,也不是要去发现崭新的自然科学定理和纯数学公式,更不是什么机器定理证明。实际上,所有发现的知识都是相对的,是有特定前提和约束条件,面向特定领域的,同时还要能够易于被用户理解。最好能用自然语言表达所发现的结果。

1.2.2　商业角度的定义

数据挖掘是一种新的商业信息处理技术,其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理,从中提取辅助商业决策的关键性数据。

简而言之,数据挖掘其实是一类深层次的数据分析方法。数据分析本身已经有很多年的历史,过去数据收集和分析的目的是用于科学研究,由于当时计算能力有限,对大数据量进行分析的复杂数据分析方法受到很大限制。现在,由于各行业业务自动化的实现,商业领域产生了大量的业务数据,这些数据不再是为了分析的目的而收集的,而是由于纯机会(Opportunistic)的商业运作而产生。分析这些数据也不再是单纯为了研究的需要,更主要是为商业决策提供真正有价值的信息,进而获得利润。但所有企业面临的一个共同问题是,企业数据量非常大,而其中真正有价值的信息却很少。因此,从大量的数据中经过深层分析,获得有利于商业运作、有竞争力的信息,与从矿石中淘金相似,“数据挖掘”也因此而得名。

因此,数据挖掘可以描述为:一种按企业既定业务目标,对大量的企业数据进行探索和分析,揭示隐藏的、未知的或验证已知的规律性,并进一步将其模型化的先进有效的方法。

1.2.3　数据挖掘与传统分析方法的区别

数据挖掘与传统的数据分析(如查询、报表、联机应用分析)的本质区别是数据挖掘是在没有明确假设的前提下挖掘信息、发现知识。数据挖掘所得到的信息应具有先前未知、有效和可实用三个特征。

先前未知的信息是指该信息是未曾预料到的,既数据挖掘是要发现那些不能靠直觉发现的信息或知识,甚至是违背直觉的信息或知识,挖掘出的信息越是出乎意料,就可能越有价值。在商业应用中最典型的例子就是一家连锁店通过数据挖掘发现了小孩尿布和啤酒之

间有着惊人的联系。

1.2.4　数据挖掘和数据仓库

在大部分情况下，数据挖掘都要先把数据从数据仓库中拿到数据挖掘库或数据集市中如图 1-1 所示。从数据仓库中直接得到进行数据挖掘的数据有许多好处，就如后面会讲到的，数据仓库的数据清理和数据挖掘的数据清理差不多，如果数据在导入数据仓库时已经清理过，那很可能在进行数据挖掘时就没必要再清理一次，而且所有数据不一致的问题都已经被解决了。

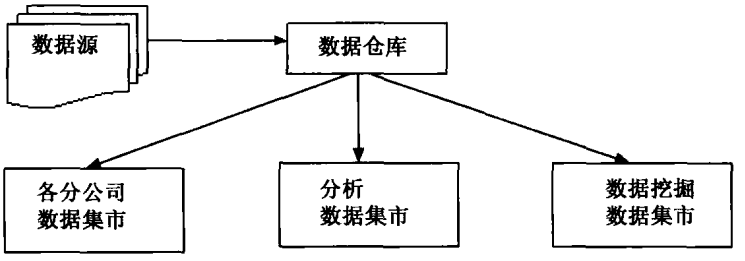


图 1-1　数据挖掘从数据库中得出

数据挖掘库，不一定非得是物理上单独的数据库，也可能是数据仓库的一个逻辑上的子集。但如果数据仓库的计算资源已经很紧张，最好还是建立一个单独的数据挖掘库，如图 1-2 所示。

当然，为了数据挖掘也不必非得建立一个数据仓库，数据仓库不是必需的。建立一个巨大的数据仓库，把各个不同源的数据统一在一起，解决所有的数据冲突问题，然后把所有的数据导到一个数据仓库内，是一项巨大的工程，可能要用大量时间，花上大量资金才能完成。要进行数据挖掘，可以把一个或几个事务数据库导到一个只读的数据库中，就把它当作数据集市，据此进行数据挖掘。

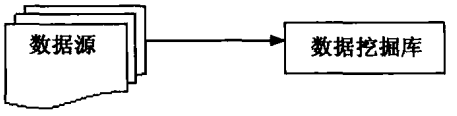


图 1-2　数据挖掘从事物数据库中得出

1.2.5　数据挖掘和在线分析处理

一个经常问到的问题是，数据挖掘和在线分析处理（OLAP）到底有何不同。通过下面的解释可以知道，它们是完全不同的工具，所基于的技术也大相径庭。

OLAP 是决策支持领域的一部分。传统的查询和报表工具是告诉用户数据库中都有什么（What happened），OLAP 则更进一步告诉用户下一步会怎么样（What next）及如果用户采取这样的措施又会怎么样（What if）。用户首先建立一个假设，然后用 OLAP 检索数据库来验证这个假设是否正确。比如，一个分析师想找到导致贷款拖欠的原因，他可能先做一个初始的假定，认为低收入的人信用度也低，然后用 OLAP 来验证他这个假设。如果这个假设没有被证实，他可能去查看那些高负债的账户，如果还不行，他也许要把收入和负债一起考虑，

一直进行下去，直到找到他想要的结果，或者放弃寻找。

也就是说，OLAP 分析师是建立一系列的假设，然后通过 OLAP 来证实或推翻这些假设，最终得到自己的结论。OLAP 分析过程在本质上是一个演绎推理的过程。但是如果分析的变量达到几十个或上百个，那么再用 OLAP 手动分析验证这些假设将是一件非常困难的事情。

数据挖掘与 OLAP 不同的地方是，数据挖掘不是用于验证某个假定的模式（模型）的正确性，而是在数据库中自己寻找模型。这在本质上是一个归纳的过程。比如，一个用数据挖掘工具的分析师想找到引起贷款拖欠的风险因素，数据挖掘工具可能帮他找到高负债和低收入是引起这个问题的因素，甚至还可能发现一些分析师从来没有想过或试过的其他因素，比如年龄等。

数据挖掘和 OLAP 具有一定的互补性。在采用通过数据挖掘得到的结论之前，用户也许要验证一下如果采取这样的行动会给公司带来什么样的影响，那么 OLAP 工具能回答这些问题。

在知识发现的早期阶段，OLAP 工具还有其他一些用途：可以帮用户探索数据，找到哪些是对一个问题比较重要的变量；发现异常数据和互相影响的变量。这都能帮用户更好地理解这些数据，加快知识发现的过程。

1.2.6　数据挖掘、机器学习和统计

数据挖掘利用了人工智能（AI）和统计分析的进步带来的好处，这两门学科都致力于模式发现和预测。

数据挖掘不是为了替代传统的统计分析技术。相反，它是统计分析方法学的延伸和扩展。大多数的统计分析技术都基于完善的数学理论和高超的技巧，预测的准确度还是令人满意的，但对使用者的要求很高。而随着计算机计算能力的不断增强，可以利用计算机强大的计算能力，通过相对简单和固定的方法完成同样的功能。

一些新兴的技术同样在知识发现领域取得了很好的效果，如神经元网络和决策树，在足够多的数据和计算能力下，它们几乎不用人的操作就能自动完成许多有价值的功能。

数据挖掘就是利用统计和人工智能技术的应用程序，把这些高深复杂的技术封装起来，使人们不用自己掌握这些技术也能完成同样复杂的工作，并且可以更专注于自己所要解决的问题。

1.2.7　软硬件发展对数据挖掘的影响

使数据挖掘这件事情成为可能的关键一点是计算机性能价格比的提高。在过去的几年里，磁盘存储器的价格几乎降低了 99%，这在很大程度上改变了企业界对数据收集和存储的态度。如果每兆的价格是 10 元，那存放 1TB 的费用是 10 000 000 元，但当每兆的价格降为 1 元时，存储同样的数据只需 1 000 000 元。

计算机计算设备的降价幅度同样非常显著，而且每一代芯片的诞生都会把 CPU 的计算能力提高一大步。内存 RAM 也一样，几年之内每兆内存的价格由上百元降到现在的不足 1 元。

在单个 CPU 计算能力大幅提升的同时，基于多个 CPU 的并行系统也取得了很大的进

步。目前大多数的服务器都支持多个 CPU,SMP 服务器簇甚至能让成百上千个 CPU 同时工作。

基于并行系统的数据库管理系统也给数据挖掘技术的应用带来了便利。如果有一个庞大而复杂的数据挖掘问题要求通过访问数据库取得数据,那么效率最高的办法就是利用一个本地的并行数据库。

1.3　数据挖掘研究内容

随着数据挖掘研究逐步走向深入,数据挖掘和知识发现的研究已经形成了三根强大的技术支柱:数据库、人工智能和数理统计。因此,KDD 大会程序委员会曾经由这三个学科的权威人物同时任主席。目前数据挖掘的主要研究内容包括基础理论、发现算法、数据仓库、可视化技术、定性定量互换模型、知识表示方法、发现知识的维护和再利用、半结构化和非结构化数据中的知识发现及网上数据挖掘等。

1.3.1　数据挖掘所发现的知识

数据挖掘所发现的知识最常见的有以下 5 类。

1. 广义知识(Generalization)

广义知识指类别特征的概括性描述知识。根据数据的微观特性发现其表征的、带有普遍性的、较高层次概念的、中观和宏观的知识,反映同类事物共同性质,是对数据的概括、精炼和抽象。

广义知识的发现方法和实现技术有很多,如数据立方体、面向属性的归约等。数据立方体还有其他一些别名,如“多维数据库”、“实现视图”、OLAP 等。该方法的基本思想是实现某些常用的代价较高的聚集函数的计算,诸如计数、求和、平均、最大值等,并将这些实现视图储存在多维数据库中。既然很多聚集函数需经常重复计算,那么在多维数据立方体中存放预先计算好的结果将能保证快速响应,并可灵活地提供不同角度和不同抽象层次上的数据视图。另一种广义知识发现方法是加拿大 SimonFraser 大学提出的面向属性的归约方法。这种方法以类似 SQL 语言表示数据挖掘查询,收集数据库中的相关数据集,然后在相关数据集上应用一系列数据推广技术进行数据推广,包括属性删除、概念树提升、属性阈值控制、计数及其他聚集函数传播等。

2. 关联知识(Association)

关联知识是反映一个事件和其他事件之间依赖或关联的知识。如果两项或多项属性之间存在关联,那么其中一项的属性值就可以依据其他属性值进行预测。最为著名的关联规则发现方法是 R. Agrawal 提出的 Apriori 算法。关联规则的发现可分为两步:第一步是迭代识别所有的频繁项目集,要求频繁项目集的支持率不低于用户设定的最低值;第二步是从频繁项目集中构造可信度不低于用户设定的最低值的规则。识别或发现所有频繁项目集是关联规则发现算法的核心,也是计算量最大的部分。

3. 分类知识(Classification & Clustering)

分类知识是反映同类事物共同性质的特征型知识和不同事物之间的差异型特征知识。最为典型的分类方法是基于决策树的分类方法。它从实例集中构造决策树,是一种有指导

的学习方法。该方法先根据训练子集(又称为窗口)形成决策树,如果该树不能对所有对象给出正确的分类,那么选择一些例外加入到窗口中,重复该过程一直到形成正确的决策集。最终结果是一棵树,其叶结点是类名,中间结点是带有分支的属性,该分支对应该属性的某一可能值。最为典型的决策树学习系统是ID3,它采用自顶向下不回溯策略,能保证找到一个简单的树。算法C4.5和C5.0都是ID3的扩展,它们将分类领域从类别属性扩展到数值型属性。

数据分类还有统计、粗糙集(RoughSet)等方法。线性回归和线性辨别分析是典型的统计模型。为降低决策树生成代价,人们还提出了一种区间分类器。最近也有人研究使用神经网络方法在数据库中进行分类和规则提取。

4. 预测型知识(Prediction)

预测型知识根据时间序列型数据,由历史的和当前的数据去推测未来的数据,也可以认为是以时间为关键属性的关联知识。

目前,时间序列预测方法有经典的统计方法、神经网络和机器学习等几种。1968年Box和Jenkins提出了一套比较完善的时间序列建模理论和分析方法,这些经典的数学方法通过建立随机模型,如自回归模型、自回归滑动平均模型、求和自回归滑动平均模型和季节调整模型等,进行时间序列的预测。由于大量的时间序列是非平稳的,其特征参数和数据分布随着时间的推移而发生变化。因此,仅仅通过对某段历史数据的训练,建立单一的神经网络预测模型,还无法完成准确的预测任务。为此,人们提出了基于统计学和基于精确性的再训练方法,当发现现存预测模型不再适用于当前数据时,对模型重新训练,获得新的权重参数,建立新的模型。也有许多系统借助并行算法的计算优势进行时间序列预测。

5. 偏差型知识(Deviation)

偏差型知识是对差异和极端特例的描述,用来揭示事物偏离常规的异常现象,如标准类外的特例、数据聚类外的离群值等。

所有这些知识都可以在不同的概念层次上被发现,并随着概念层次的提升,从微观到中观再到宏观,以满足不同用户不同层次决策的需要。

1.3.2　数据挖掘的功能

数据挖掘通过预测未来趋势及行为,做出前摄的、基于知识的决策。数据挖掘的目标是从数据库中发现隐含的、有意义的知识,主要有以下5类功能。

1. 自动预测趋势和行为

数据挖掘在大型数据库中自动寻找预测性信息,以往需要进行大量手工分析的问题如今可以迅速直接由数据本身得出结论。一个典型的例子是市场预测问题,数据挖掘使用过去有关促销的数据来寻找未来投资中回报最大的用户,其他可预测的问题包括预报破产及认定对指定事件最可能作出反应的种群。

2. 关联分析

数据关联是数据库中存在的一类重要的、可被发现的知识。若两个或多个变量的取值之间存在某种规律,就称为关联。关联可分为简单关联、时序关联和因果关联。关联分析的目的是找出数据库中隐藏的关联网。有时并不知道数据库中数据的关联函数,即使知道也是不确定的,因此关联分析生成的规则带有可信度。