滇一西一学一术一文一丛

数据挖掘及其应用研究

本书全面介绍了数据挖掘的背景信息、相关概念及其所使用的主要技术，针对关联规则数据挖掘，系统深入地描述了Apriori算法和FP-growth算法，并比较了各自的优缺点。本书基于关系代数的关联规则挖掘，讨论该算法的实现过程，并对该算法作复杂性分析，在聚类分析数据挖掘、序列模式挖掘等方面也有介绍。对网络入侵检测的拒绝服务攻击类型进行了序列模式挖掘和聚类分析，为进一步开发入侵检测系统提供决策依据。数据挖掘在农作物病虫害预警、居民消费结构分析、多媒体图像挖掘等方面都有应用。

郑继刚　著

云南大学出版社
Yunnan University Press

滇一西一学一术一文一丛

数据挖掘及其应用研究

郑继刚 著

云南大学出版社
Yunnan University Press

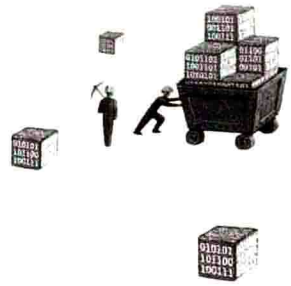
图书在版编目（CIP）数据

数据挖掘及其应用研究 / 郑继刚著. —昆明：云南大学出版社，2014
（滇西学术文丛. 第7辑）
ISBN 978-7-5482-1956-9

Ⅰ. ①数… Ⅱ. ①郑… Ⅲ. ①数据采集—研究 Ⅳ. ①TP274

中国版本图书馆CIP数据核字（2014）第046060号

责任编辑：徐　旻
装帧设计：刘　雨



滇 | 西 | 学 | 术 | 文 | 丛

数据挖掘及其应用研究

郑继刚　著

出版发行：云南大学出版社
印　　装：昆明市研汇印刷有限责任公司
开　　本：787mm × 1092mm　1/16
印　　张：7
字　　数：152千
版　　次：2014年5月第1版
印　　次：2014年5月第1次印刷
书　　号：ISBN 978-7-5482-1956-9
定　　价：20.00元

社　　址：昆明市翠湖北路2号云南大学英华园内
邮　　编：650091
电　　话：（0871）65031071　65033244
网　　址：http://www.ynup.com
E-mail：market@ynup.com

《滇西学术文丛》总序

蒋永文

保山学院，其前身为保山师范高等专科学校，地处气候宜人、风景秀丽、历史悠久、文化底蕴厚重的滇西重镇保山，建校已30余年，为滇西区域培养了上万名中小学教师和各行业优秀建设者，为祖国西部特别是边疆少数民族地区教育事业的繁荣昌盛与经济社会的可持续发展做出了重要贡献。2009年4月，学校被教育部批准为保山学院。这使我们站在了一个新的历史起点上，有了一个更为广阔的发展空间。

大学肩负着创造知识和传播知识的重任。学术是支撑大学的精髓，学科是构筑大学的基石，学者是大学精神的化身。教学与科研相统一是大学的基本理念。科研和教学是彼此促进的，在教学过程中，可以激发灵感，开阔思路，发现研究课题。而研究成果又可以丰富教学内容，促进教学质量的提高，二者相得益彰。为了给滇西地区提供更好的高等教育资源，保山学院必须建立一支热爱教育事业，业务素质过硬，高水平、高质量的教师队伍，为此，学校以重点学科建设为龙头，以形成科研特色，增强科研实力，提高效益为目标。学校近几年采取了资助科研立项、奖励科研成果、出版学术论文等措施，不断提高广大教师的教学水平和科研水平，已收到了较好的效果。为更好地为广大教师提供出版学术论著的园地，学校决定继续出版《滇西学术文丛》，出版学术水平较高的著作。相信《滇西学术文丛》的出版，一定会对保山学院科学研究的深入、学科建设和学科带头人、骨干教师的培养产生积极的影响。

辽阔的天空，允许大鹏展翅翱翔，也允许小鸟上下蓬蒿；广袤的大地，允许参天大树生长，也允许无名小草成长。我们是小鸟，我们是小草，这套丛书，远非成熟完美，作者水平也还需要不断提高。我们期待着批评和指教，相信我们会做得越来越好。

2013年5月

目　录

第1章 引 言

1.1 研究背景

人类社会进入信息时代，在计算机和服务器中保存的文件及数据数量成倍增长，用户期望从这些庞大的数据中获得最有价值的信息。计算机软件、硬件的快速发展使得数据采集和数据存储成为可能。尽管部门、各商业公司、科研院所积累了海量数据，但是这些数据中只有很少的一部分有用。信息用户面临着数据丰富而知识匮乏的问题，迫切需要能自动化、高效率地从海量数据中提取所需有用知识点的新型处理技术，在这样的需求背景下，数据挖掘技术应运而生。[①] 将传统数据分析方法和处理海量数据的复杂算法结合的数据挖掘技术，使从数据库中高效提取有用信息成为可能，为现今信息技术的发展奠定了基础。

关联规则(Association Rules)是数据挖掘技术中研究较为广泛并取得突破性研究进展的一个重要分支。例如，在一家超市里，有一个有趣的现象就是尿布和啤酒赫然摆在一起出售，但是这个奇怪的销售举措却使尿布和啤酒的销量都增加了。这是发生在美国沃尔玛连锁超市的真实案例，并一直广为流传，揭示了隐藏在尿布与啤酒背后的美国人的一种消费模式，丈夫们为孩子买尿布后又随手带回了自己喜欢的啤酒。[②] 可以看出，商家对海量交易数据进行挖掘和分析，能发现交易数据库中隐含的内在的有价值规律。

数据挖掘涉及多种学科和方法，有不同类型的分支。根据挖掘任务，可分为：序列模式发现、相似模式发现、混沌模式发现、依赖关系或依赖模型发现、异常和趋势发现、关联规则发现、数据汇总和聚类发现；根据挖掘对象，可分为：异质数据挖掘、网络数据挖掘、遗产数据挖掘、文本数据源挖掘、空间数据库挖掘、时态数据库挖掘、面向对象数据库挖掘、关系数据库挖掘和多媒体数据库挖掘；依据挖掘方法，可分为：数据库方法、聚类分析方法、现代数学分析方法、机器学习方法、神经网络方法、统

---

① J. Naisbitt, Megatrends. Ten New Direetions Transforming Our Lives. Warner Books, Rei Edition, 1988.

② Raskesh Agrawal, Tomasz Imielinski, Arun Swami. Database Mining: A performance perspective, IEEE Trans. Knowledge and Data Enginnering. May 1993: 914 - 925.

计方法、遗传算法方法、基于证据理论和元模型的方法、粗糙集方法和集成方法、近似推理和不确定性推理方法等；基于数据挖掘发现的知识，可分为：不确定性知识挖掘、预测型知识挖掘、异常型知识挖掘、差异型知识挖掘、广义型知识挖掘和关联型知识挖掘等。[①] 关联规则具有广泛的应用价值，应用在时间序列数据挖掘、空间数据挖掘、Web 数据挖掘、多媒体数据挖掘和不确定数据挖掘等方面。

1.1.1 空间数据挖掘

所谓空间数据，是指从地理信息系统(GIS)、遥感系统、多媒体系统、医学及卫星图像等各种应用系统中收集的、远超过人类大脑分析能力的数据。依据功能的不同，空间数据挖掘技术分为三种不同的类型：描述型、解释型和预测型。描述型空间挖掘技术将空间现象分布特征化，如空间聚类；解释型空间挖掘技术应用于处理空间关系，如处理空间对象与影响其空间分布的因素之间的关系；预测型空间挖掘技术则用于根据给定的属性预测另外的属性，如分类模型等。

空间数据挖掘的研究，目前主要集中在空间数据挖掘的体系结构方面，包括不确定性挖掘、并行数据挖掘、面向对象空间数据库的数据挖掘、统计分析和数据挖掘的协同效应及遥感图像挖掘、聚类分析挖掘、模糊空间关联规则挖掘、泛化时空数据挖掘等。主要是基于集合理论、统计和概率理论、机器学习、模仿生物学的基础上，运用地理信息科学的研究方法。

1.1.2 多媒体数据挖掘

多媒体数据，包括文本、文档、超文本、图像、图形、视频、声音及音频数据等，类型复杂。随着信息技术的快速发展，人们接触的数据形式日益丰富，多媒体数据大量涌现，形成许多海量的多媒体数据库。多媒体数据大多是非结构的、异构性的，拥有数十维甚至数百维的特征向量，因此，将非结构化、异构性数据转化为结构数据以及将特征向量降维成了多媒体数据挖掘的关键技术。[②]

一些研究者提出了多媒体数据挖掘系统原型 MDMP，将数据挖掘技术与多媒体数据库技术(如多媒体数据的建模表示、存储和检索技术等)进行有机结合，采用多媒体图像数据的多维分析、分类与聚类分析、关联规则挖掘等方法，广泛应用于卫星图片分析、医学影像诊断分析、地下矿藏预测等诸领域。

1.1.3 时序数据挖掘

时序数据挖掘以信息的时间特性研究为途径，深入探析事物的进化机制，揭示其

① 毛国君，段立娟，王实，等．数据挖掘原理与算法[M]．北京：清华大学出版社，2005：10－11.

② 郑继刚，谢芳．多媒体图像挖掘的关联规则挖掘[J]．红河学院学报，2009(5)：44－47.

波动的周期、振幅、趋势种类等内在规律，是一种有效的获取知识的途径。时序数据挖掘的主要技术包括相似搜索和趋势分析，在宏观经济预测、股市价格分析、市场营销及太阳黑子数量、河流流量、月降水量等多个领域得到了应用。

国内关于时序数据的研究较少，使用的方法和技术主要有人工神经网络技术，利用它预测和处理混沌观测时间序列能达到较高的精度。[①] 此外，还有通过对时序数据进行离散傅立叶变换将其从时域空间变换到频域空间，将时序数据映射为多维空间的点，在此基础上，有学者提出一种新的基于距离的离群数据挖掘算法。[②]

1.1.4 Web 数据挖掘

在互联网/网络技术的飞速发展及迅速普及背景下，各种信息均可在网上获得，这些信息具有数量巨大、分布广泛、动态性、全球化及多元性的特点。如何从庞大而杂乱的 Web 数据中找到有用的信息已成为当前的研究热点，Web 数据挖掘面临着其是世界上最大的数据库的难题。网络数据挖掘可分为四类：Web 使用挖掘、Web 结构挖掘、Web 内容挖掘和 Web 应用挖掘。

1.1.5 不确定数据挖掘

传统的数据挖掘加工位置已被精确定位，但在实际应用领域和实际应用过程中，由于测量仪器的局限性，测量数据不准确及不确定是不可避免的。数据的不确定性包括存在的不确定性和值的不确定性两种情况，存在的不确定性是指对象或元的存在与否不能确定，如关系数据库的某个元组和一个概率相关联表示这个元组存在的可信度；值的不确定性是指元组的存在是确定的，但它的具体值是不确定的。[③] 目前，不确定数据挖掘研究已成为热点，并且在关联规则、聚类分析、空间挖掘技术等方面都有突破，经典的 K - means 算法扩展到了 UK - means 算法，Apriori 算法扩展到了 UApriori 算法等。

① 臧洌．人工神经网络在混沌观测时序数据处理中的应用[J]．数据采集与处理，2001(4)：486 - 489.

② 郑斌祥，杜秀华，席裕庚．一种时序数据的离群数据挖掘新算法[J]．控制与决策，2002(3)：324 - 327.

③ 陆叶，王丽珍，张晓峰．从不确定数据集中挖掘频繁 co - location 模式[J]．计算科学与探索，2009，3(6)：656 - 664.

1.2 研究内容及意义

1.2.1 国内外研究综述

1.2.1.1 过去的研究

数据库知识发现第一次出现于1989年在底特律举行的第十一届国际联合人工智能会议上，其后到1995年在加拿大蒙特利尔召开的第一届KDD & Data Mining国际学术会议，再之后到每年都要召开一次的KDD & Data Mining国际学术会议，经过20余年的研究努力，数据挖掘技术取得了突破性进展和丰硕的研究成果，许多轻件公司都已经开发出数据挖掘产品，并在北美、欧洲等国家得到应用。①

数据挖掘是信息技术和数据库技术自然演变的结果。数据挖掘的发展历程经历了四个进化阶段，即数据搜集阶段、数据访问阶段、数据仓库阶段和决策支持阶段(详见表下表)。

数据挖掘研究的进化历程表②

进化阶段	支持技术	产品厂家	产品特点
数据搜集 (20世纪60年代)	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)	关系数据库、结构化 查询语言、ODBC	Oracle，Sybase， Informix，IBM，Microsoft	提供历史性的、动态的 数据信息
数据仓库、决策支持 (20世纪90年代)	联机分析处理、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)	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

1.2.1.2 数据挖掘研究的现状与成果

在国外，数据挖掘技术已在各个领域中广泛应用，典型应用如：IBM公司开发的AS(Advanced Scout)系统针对美国男子篮球职业联赛(National Basketball Association，简称NBA)的比赛数据，帮助教练优化战术组合；③ 生物学研究中用数据挖掘技术对DNA

① 陈娜．数据挖掘技术的研究现状及发展方向[J]．电脑与信息技术，2006(2)：46－49.

② 陆建江，张亚非，宋自林．模糊关联规则的研究与应用[M]．北京：科学出版社，2008：2.

③ 李菁菁，邵培基，黄亦潇．数据挖掘在中国的现状和发展研究[J]．管理工程学报，2004(3)：10－15.

进行分析；应用数据挖掘技术对银行或保险公司时有发生的诈骗行为进行预测；喷气推进实验室与加州理工学院合作开发的天文科学家 SKICAT 系统，可以帮助天文学家发现遥远的类星体，这是人工智能技术成功应用在天文学和空间科学的典型范例。

在学术研究中，信息处理、数据库、知识工程、人工智能及其他领域的国际学术期刊也纷纷开辟数据挖掘技术研究专栏或设置专刊，如电气电子工程学会(IEEE)的 Knowledge and Data Engineering 会刊于 1993 年出版数据挖掘技术专刊；在互联网上也有数据挖掘的电子出版物，其中就有权威的半月刊 Knowledge Discovery Nuggets；1997 年 10 月 7 日开始出版的在线周刊 DS *(DS 代表决策支持)，可用电子邮件免费订阅。①

与国外相比，国内的数据挖掘研究相对滞后，但已粗具整体研究实力。1993 年国家自然科学基金委员会首次资助数据挖掘领域的研究项目，而目前包括中国科学院计算技术研究所、清华大学、复旦大学、北京大学、中国人民大学等在内的科研院所和大学竞相开展数据挖掘及知识发现领域的基本理论及应用研究。在研究成果方面，清华大学周立柱教授领导的数据挖掘研究小组、复旦大学朱扬勇教授领导的数据挖掘工作组、中国科技大学蔡庆生教授领导的针对关联规则的研究小组、四川大学唐常杰教授领导的针对时间序列方面的数据挖掘研究小组，以及云南大学王丽珍教授领导的针对不确定数据挖掘的研究小组等，都取得了许多重要的研究成果；在数据挖掘算法研究领域成果丰硕的代表专家包括清华大学的陆玉昌教授及石纯一教授、北京科技大学的杨炳儒教授、中科院计算所的史忠值研究员、华东师范大学的周傲英教授、南京大学的周志华教授、武汉大学的李德仁院士等；国内该领域比较重要的会议有全国数据库学术会议(National DataBase Academic Conference，简称 NDBC)，权威的学术期刊有《软件学报》《计算机学报》《计算机研究与发展》等。②

数据挖掘的任务和挖掘方法的多样性给数据挖掘研究提出了许多具有挑战性的研究课题，在未来会形成一个更大的研究热潮。具体的研究热点可能会集中在以下几个方面：研究专用于知识发现的数据挖掘语言的正规化和规范化；探索可伸缩和可交互的数据挖掘方法，全面提高数据挖掘效率，尤其是在对超大规模的数据库挖掘方面；研究分布式环境下的数据挖掘技术，特别是在互联网上建立与传统数据库服务器配合以实现挖掘功能的数据挖掘服务器；探寻数据挖掘过程中的可视化方法，使用户能更好地理解和接受知识发现过程，便于在知识发现过程中实现人机交互；加强对图形图像数据、多媒体数据、文本数据等各种非结构化数据的挖掘；开发满足多种数据类型、噪声容限的挖掘方法，以解决异质数据集的数据挖掘问题；扩大数据挖掘应用范围，如应用于金融分析、犯罪侦查、生物医药研制等。

---

① 郑继刚，王边疆．数据挖掘研究的现状与发展趋势[J]．红河学院学报，2010，8(2)：45－48.

② 徐雪琪．基于统计视角的数据挖掘研究[D]．杭州：浙江工商大学，2007：5－6.

1.2.2 关联规则挖掘问题的研究现状与成果

例如，用户购买了台式计算机后，如何选择其他软、硬件，比如杀毒软件、打印机、数码相机、音响、扫描仪、摄像头等，关联规则能揭示这些产品潜在的相关影响。在1993年阿格拉瓦(R. Agrawal)等人提出关联规则的挖掘问题后，许多研究者对该问题进行了广泛的研究，目前，该问题研究的主要研究成果包括：多层关联规则挖掘算法、多值关联规则挖掘算法、多循环方式挖掘算法、分布/并行式挖掘算法、动态项目计数算法、增量式更新算法、基于概念格的关联规则挖掘算法等。

多层关联规则挖掘算法在每个抽象层上定义各自的最小支持度阈值、最小可信度阈值，在每一层挖掘关联规则。目前，已经出现的多层关联规则算法有汉(J. Han)等提出的ML_ T2L1算法及斯里肯特(R. Srikant)等提出的Cumulate算法。

多值关联规则挖掘算法，将多值属性的值离散划分为多个区间，每个区间作为一个属性。

多循环方式挖掘算法将整个数据挖掘过程分为若干层次，首先对各层次分别挖掘，待完成后再组合成最后的挖掘结果。这类算法包括阿格拉瓦(R. Agrawal)等人提出的Apriori、AprioriHybrid和AprioriTid；帕克(Park)等人提出的DHP；汉(J. Han)提出的FP-growth；Toivonen提出的抽样算法Sampling；赛维德尔(Savadere)等人提出的Partition等。其中Apriori算法和FP-growth算法是最有效和最有影响的算法。

分布/并行式关联规则挖掘算法处理的数据量非常庞大，有时数据跨地域分布。目前研究的算法多数是基于分布式处理器的模式，算法有PDM、CD、DD、FPM、HPA和IDD等。①

由伯利(Brin)等人提出的动态项目计数算法DIC，突破了层次算法的瓶颈(层次算法需要对数据库重复扫描多次，最大频繁项集长度约束扫描次数)。与Apriori算法仅在每次扫描完整个数据库之前确定新的候选项集不同，动态项目计数算法将数据库划分为若干被标记起始点的块，可以在任意起始点处的块添加新的候选项集。动态项目计数算法可对所有项集的支持度进行动态评估和支持，若一个项集的所有子集被确定是频繁的，则添加它作为新的候选项集。

增量式更新挖掘算法，主要有陈(D. W. Cheng)等提出的处理数据库中记录发生变化时的更新的更新挖掘算法FUP；并在此基础上，提出了不仅可以处理交易事务数的增加还可以处理交易事务数的删除或更改的FUP2算法。费尔德曼(Feldman)提出了一种在用户指定的最低支持度为绝对数且不变的条件下，只需考察所有真子集均为频繁项集而本身却不是频繁项集的Border算法的关联规则更新技术，但该算法为减少关联规则的更新代价仍需存储相关的频繁项集结果。此外，冯玉才等对关联规则的度量(可

① 张惠民，王晓卫，肖庆．数据库中关联规则的并行/分布式采掘技术[J]．装甲兵工程学院学报，2003，6(2)：38-41.

信度、支持度、提升度等)发生改变时的更新进行了深入研究，提出了 PIUA 和 IUA 算法。①

概念格也被运用到关联规则的挖掘算法中，古丁(Godin)等提出用概念格模型提取蕴含规则的算法；胡可云等在递增构造概念格算法的基础上，提出一种实现关联规则挖掘可视化且更有效的购物篮分析关联规则挖掘算法；王德兴提出利用剪枝概念格快速发现频繁闭项集算法。

1.2.3 研究的意义

目前有关关联规则的算法很多，大部分都是以经典 Apriori 算法为基础。这些关联规则挖掘算法的基本思想都是在基于频繁项集的发现算法基础之上的，即关键问题就是求满足用户指定的最小支持度和最小可信度的项集。为了提高挖掘关联规则的执行效率，大多数方法或降低搜索数据库的次数，或减少非相关项集的产生，但以此为代价却相对地增加了在其他方面的时间开销。针对上述的关联规则挖掘算法的缺点，研究和探讨更有效率的关联规则挖掘算法是本研究的最大动机。

提高数据挖掘的效率，算法的优劣往往起到决定性的作用，因此，改进现有的算法，并用实验数据验证算法的可靠性和推广性，具有十分重要的意义。将数据挖掘技术运用到网络入侵检测中，通过挖掘关联规则，进行聚类分析、序列模式分析来进行异常网络检测，从而可以根据入侵数据集建立误用检测模型和异常检测模型，为进一步开发入侵检测系统提供依据。

1.2.4 研究主要内容

对现有数据挖掘技术，尤其是关联规则数据挖掘技术进行了较为全面、深入、详尽的探讨。

(1)全面地分类、归纳和总结了现有关联规则挖掘技术，详细地分析了关联规则的典型挖掘算法，对各算法做了对比分析，客观地分析了各算法的优缺点，讨论了提高算法效率的各种优化技术。

(2)针对 Apriori 算法的不足，陈莉、焦李成提出了基于关系代数理论的关联规则挖掘算法，只需扫描一次数据库就能快速有效地挖掘出频繁项集，避免了往返多次扫描数据库，是一种高效、快捷的挖掘算法。进一步挖掘最大频繁项集和频繁闭项集，解决了产生大量冗余规则的问题。

(3)在实验平台上用 Matlab7.0 实现了 Apriori 算法和基于关系代数理论的关联规则挖掘算法，对两个算法的性能进行了比较。用公用机器学习数据集对该算法的有效性和准确性进行了验证，用教学评价数据集挖掘出频繁闭项集，解决了冗余规则问题。

---

① 冯玉才，冯剑林．关联规则的增量式更新算法[J]．软件学报，1998，9(4)：301－306.

1.3 数据挖掘概述

1.3.1 数据挖掘概念

数据挖掘技术(Data Mining，简称 DM)，或称从数据库中发现知识(Knowledge Discovery in Databases，简称 KDD)，定义为从数据库中发现潜在的、隐含的、先前不知道的有用的信息，又被定义为从大量数据中发现正确的、新颖的、潜在有用、并能够被理解的知识过程。[①] KDD 侧重于目的和结果，DM 侧重于处理过程和方法，KDD 是将未加工的数据转换为有用信息的整个过程，DM 是 KDD 不可或缺的一部分。

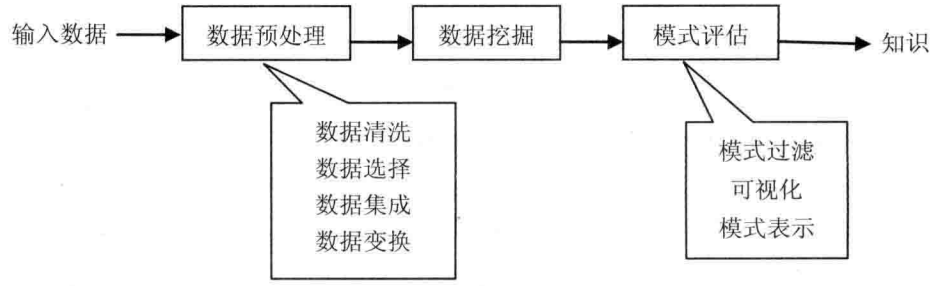


图 1-1 从数据库中发现知识(KDD)全过程

知识发现过程如图 1-1 所示，主要由以下步骤组成。

(1)数据预处理：将未加工的输入数据转换成适合分析的形式，为挖掘工作准备数据。数据清洗清除不一致和噪声数据；数据集成把多种数据源组合在一起；数据选择是从数据库中抽取与挖掘任务相关联的数据集；数据变换规范数据形式，以适合数据挖掘。由于收集和存储的数据形式多种多样，因此，数据预处理步骤在知识发现过程中可能是最费力、最耗时的步骤。

(2)数据挖掘：最基本的步骤，也是最重要的步骤，使用智能方法，自动、高效地发现有用知识，提取挖掘模式。

(3)模式评估：根据某种评价标准，识别表示知识的真正有用的模式，并确保只将有效的和有用的挖掘结果集成到专家系统中。

1.3.2 数据挖掘的起源

数据挖掘是一个交叉学科领域，受多个学科影响(如图 1-2 所示)，利用了来自以下一些领域的思想方法。

① Pang-Ning Tan，等，数据挖掘导论[M]. 范明等，译. 北京：人民邮电出版社，2006.

(1)人工智能的搜索算法、模式识别的建模技术和机器学习的学习理论。

(2)数据库技术和可视化技术，数据挖掘还吸收了其他领域的思想，这些领域包括进化计算和最优化、信息检索和信息论等。

(3)来自统计学的抽样、估计和假设检验。此外，一些其他的学科领域也起到重要的支撑作用，如并行计算技术、分布式技术等，这些技术在处理分布式数据集或海量数据集方面起到至关重要的作用。

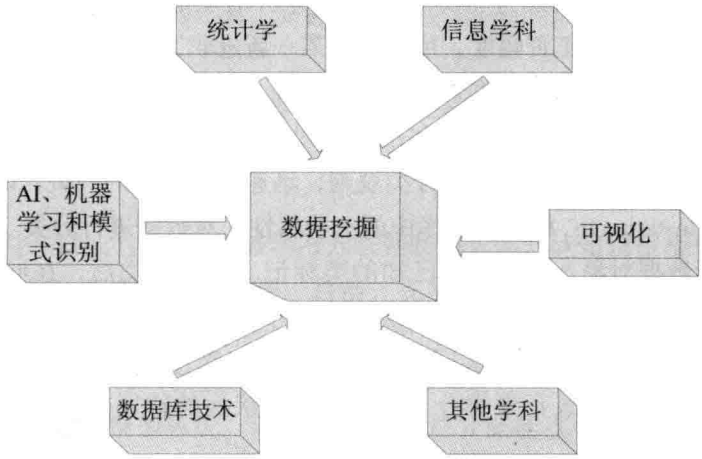


图1-2 数据挖掘受多学科影响

1.3.3 数据挖掘的主要问题

(1)不同类型的数据的多样性，期望有不同的数据挖掘系统。虽然已经广泛使用关系数据库，但其他数据库仍可能包含不常见的数据对象，而为不同数据对象的数据库开发一个通用的数据挖掘系统是不现实的，因此，对于不同的数据类型，应开发不同的数据挖掘系统。

(2)数据挖掘与用户交互问题。包括所挖掘的知识类型、临场即席挖掘、领域知识的使用和知识可视化。性能评价涉及数据挖掘算法的有效性、并行处理能力和可伸缩性。由于数据库的海量性和广泛分布性，要求挖掘算法必须是有效的和可伸缩的。

上述问题是数据挖掘目前面临的主要问题，也是未来数据挖掘技术研究及发展的主要趋势和挑战，未来的数据挖掘研究将集中在这些问题上。

1.3.4 数据挖掘的功能

数据挖掘主要用于从指定挖掘任务中挖掘用户所需的数据类型，下面依次介绍数据挖掘的主要功能。

(1)关联规则(Association Rule)：发现数据对象间的相互依赖关系，一个关联规则是 $X \Rightarrow Y$ 的形式，即 $A_1 \wedge A_2 \wedge \cdots \wedge A_m \Rightarrow B_1 \wedge B_2 \wedge \cdots \wedge B_n$ 的规则样式，$X \Rightarrow Y$ 表明满足 X