数据挖掘算法及其工程应用

章兢 张小刚 等编著

机械工业出版社
CHINA MACHINE PRESS

数据挖掘算法及其工程应用

章　兢　张小刚　等编著

机 械 工 业 出 版 社

随着数据库技术在工程领域中的广泛应用，对工程数据的后期分析和处理具有广泛的应用前景。本书以各类数据挖掘算法为核心，以智能数据分析技术的发展历程为主线，结合作者自身的研究和应用经验，详细阐述了数理统计、机器学习、软计算、关联挖掘和支持向量集等研究领域的成熟算法，并研究了各类方法在工业过程控制、水轮机调速智能监控、物流配送车辆路径优化等工程领域的实际应用，便于读者了解各种技术的应用对象、应用方法及应用效果。

本书内容丰富，论述简明，强调具体挖掘算法的分析和使用，力求实现数据挖掘技术从商业到工程领域应用的转变。可作为工科有关专业研究生和本、专科生的教学参考书，也可作为工程技术人员的自学读物。

图书在版编目（CIP）数据

数据挖掘算法及其工程应用/章兢等编著. —北京：机械工业出版社，2006.7

ISBN 7-111-19126-9

Ⅰ.数… Ⅱ.章… Ⅲ.数据采集 Ⅳ.TP274

中国版本图书馆 CIP 数据核字（2006）第 046577 号

机械工业出版社（北京市百万庄大街 22 号 邮政编码 100037）

责任编辑：贡克勤 版式设计：霍永明 责任校对：王 欣

封面设计：马精明 责任印制：洪汉军

北京汇林印务有限公司印刷

2006 年 6 月第 1 版第 1 次印刷

184mm×260mm·12.75 印张·310 千字

定价：25.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

本社购书热线电话（010）68326294

编辑热线电话（010）88379727

封面无防伪标均为盗版

前　言

近些年来，随着信息技术的飞速发展，在商务贸易和政府事务电子化、大规模工业生产过程中的智能监控和诊断、医疗领域的计算机诊断管理以及科学计算等应用领域，产生了不断增长的海量数据源。日益成熟的数据库技术和面向事务处理的信息管理系统，为这些数据的采集、存储和传输提供了稳定、可靠并且日益廉价的技术保证。在这种情况下，依靠传统方法对这些数据进行统计分析，已经不能应对信息技术带来的“数据灾难”难题。在上述背景下，一种新兴的自动信息提取技术，数据挖掘又称知识发现引起了学术界的广泛重视。如何有效地从海量数据中获取有用的知识信息，是数据挖掘研究所要解决的核心问题。

数据挖掘的研究领域涉及到数据库、机器学习、统计学等众多学科，现已经逐步扩展到不确定性推理技术、人工智能、高性能计算以及基于神经网络、模糊集理论、粗集理论、进化计算的软计算等研究领域，目前各种方向与技术之间相互融合，研究内容也纷繁复杂。从数据分析的角度讲，数据挖掘可分为关联挖掘、数据分类、数据聚类和预测等问题。根据数据类型对象的特点又可分为时序数据分析、空间数据挖掘、文本及 WEB 挖掘和多媒体数据库等研究方向。从具体算法实施角度讲，数据挖掘又有兴趣度评价、并行算法、增量算法及算法的复杂性等研究问题。

数据挖掘技术目前大都应用于商业领域。在工程应用系统中，大量的运行数据被实时数据库记录下来，其中蕴含了相关的人工控制经验、参数及管理优化信息等关键知识。研究使用数据挖掘技术从现场数据库中提取专家控制和决策知识，将为各类工程设计和综合优化开辟一个有效的分析设计工具。因此，研究数据挖掘技术在工程领域中的推广和研究正在逐步兴起。

向读者提供数据挖掘领域内的详细算法和问题解决过程中的实际经验，是本书撰写的宗旨。以各类数据挖掘应用算法为核心，以智能数据分析技术的发展历程为主线，结合作者自身的研究和应用经验，本书详细阐述了数理统计、机器学习、软计算、关联挖掘和支持向量集等研究领域的成熟算法，并研究了各类方法在工业过程控制、水轮机调速智能监控、物流配送车辆路径优化等工

程领域的实际应用。强调具体算法的分析和使用，实现从商业到工程领域应用的转变是本书的两大特色。

本书的编著得到教育部科学技术研究重点项目“复杂工业过程中信息融合与知识发现理论与应用研究”、湖南省重点科技项目“基于时序挖掘和信息融合技术的回转窑专家控制系统”、湖南省自然科学基金项目“数据挖掘技术在复杂工业过程监控中的应用研究”等科研项目的资助。

本书由章兢和张小刚等编著，徐雪松、谭建豪、王练红、何昭辉、张国云等分别参与了第1、2、3、5、6章部分的编写。全书是我们学术梯队共同研究成果的结晶，其中一些内容是本学术梯队中已毕业研究生、在研博士生及科研人员的工作成果。本书引用了些国内外该领域专家学者的著作、文章和研究报告等文献资料，他们的工作丰富了本书的内容，在此对他们表示衷心的感谢。

由于作者水平有限，错误和不足之处在所难免，恳请读者批评指正。

作　者

2006年3月

目　录

第 1 章　数据挖掘综述

自 20 世纪 60 年代以来，数据库和信息技术已经从原始的文件处理演化到复杂的、功能强大的数据库系统。自 70 年代以来，数据库系统的研究与开发已经从层次和网状数据库系统发展到开发关系数据库系统、数据建模工具、索引和数据组织技术。用户通过查询语言和用户界面，优化查询处理和事务管理，可以方便、灵活地访问数据。自 80 年代中期以来，数据库技术的特点是，广泛使用关系技术研究开发功能强大的数据库系统，同时使用了先进的数据模型，如扩充关系模型、面向对象模型、对象—关系模型和演绎模型。数据库的发展又促进了相关信息利用的发展，数据挖掘就是其中之一。

1.1　数据挖掘的概念和定义

随着计算机技术的迅猛发展以及网络的普及，许多行业如商业、企业、科研机构和政府部门等都有了更多的机会和便捷的方法与外界进行信息交流，数据库的规模、范围和深度都在快速不断扩大，从而积累了海量的、以不同形式存储的数据资料，同时在许多领域也建立了数据仓库。在这些海量数据中往往隐含着各种各样的信息，这些信息人们往往凭直觉与经验是难以发现的。如何从大量的数据中获得有价值的信息，采用传统的数据库技术已显得无能为力了，数据的迅速增加与数据分析处理方法滞后的矛盾越来越大，人们希望能够在对已有的大量数据分析的基础上进行科学研究、商业决策或企业管理，从而达到为决策服务的目的。数据挖掘（Data Mining，DM）就是为了满足这种需求而迅速发展起来的一种新的数据处理技术。它的实质是一种发现知识的应用技术，是一个提取有用信息的过程。自 20 世纪末提出以来，引起了许多专家学者的广泛关注，并应用到金融、零售业、工业过程、电力、医疗保健和政府决策等各个领域，取得了良好的社会效益和经济效益，具有广阔的开发和应用前景。

数据挖掘的发展历史虽然较短，但从 20 世纪 90 年代以来，它的发展速度很快，加之它是多学科综合的产物，目前还没有一个完整的定义，人们提出了多种数据挖掘的定义[1]，如 SAS 研究所（1997）：“在大量相关数据基础之上进行数据探索和建立相关模型的先进方法”；Bhavani（1999）：“使用模式识别技术、统计和数学技术，在大量的数据中发现有意义的新关系、模式和趋势的过程”；Hand et al（2000）：“数据挖掘就是在大型数据库中寻找有意义、有价值信息的过程”。

还有的定义认为：数据挖掘是从数据集合中自动抽取隐藏在数据中的那些有用信息的非平凡过程。这些信息的表现形式为：规则、概念、规律及模式等，可帮助决策者分析历史数据及当前数据，从中发现隐藏的关系和模式，进而预测未来可能发生的行为。

目前较通用的定义为：数据挖掘就是从大量的、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、人们事先不知道的、但又是潜在有用的信息和知识的过程。数据挖掘要解决的问题就是在庞大的数据中寻找有价值的隐藏信息，加以分析，并将

这些有意义的信息归纳成结构模式，提供给有关部门在进行决策时参考[2]。

1.2 数据挖掘的历史及发展

数据挖掘理论研究出现于20世纪80年代后期，90年代有了突飞猛进的发展。1989年8月在美国底特律召开的第11届国际人工智能联合会议的专题讨论会上首次出现KDD（Knowledge Discovery in Databases）这个术语。随后在1991年、1993年和1994年都举行KDD专题讨论会，汇集来自各个领域的研究人员和应用开发者，集中讨论数据统计、海量数据分析算法、知识表示、知识运用等问题。于是数据挖掘界于1995年召开了它的第一届知识发现与数据挖掘国际学术会议，KDD国际会议发展成为年会。数据挖掘研究界于1998年建立起一个新的学术组织ACM－S IGKDD，即ACM下的数据库中的知识发现专业组（Special Interested Group on Knowledge Discovery in Databases）。1999年ACM－SIGKDD组织了第五届知识发现与数据挖掘国际学术会议（KDD’99）。还有一些其他国际或地区性数据挖掘会议，如“知识发现与数据挖掘太平洋亚洲会议”（PAKDD），“数据库中知识发现原理与实践欧洲会议”（PKDD）和“数据仓库与知识发现国际会议”（DaWaK）[3]。

进入21世纪后，数据挖掘技术向纵深发展，与多种学科交叉、多种技术结合的数据挖掘理论及技术开始涌现。并行计算、计算机网络和信息工程等其他领域的国际学会、学刊把数据挖掘和知识发现列为专题和专刊讨论，甚至到了脍炙人口的程度。在Gartner Group的一次高级技术调查将数据挖掘和人工智能列为“未来三到五年内将对工业产生深远影响的五大关键技术”之首，并且还将并行处理体系和数据挖掘列为未来五年内投资焦点的十大新兴技术前两位。根据Gartner的HPC研究表明，“随着数据捕获、传输和存储技术的快速发展，大型系统用户将更多地需要采用新技术来挖掘市场以外的价值，采用更为广阔的并行处理系统来创建新的商业增长点。今后研究焦点将会集中到以下几个方面：

发现语言的形式化描述，即研究专门用于知识发现的数据挖掘语言，也许会像SQL语言一样走向形式化和标准化。

寻求数据挖掘过程中的可视化方法，使知识发现的过程能够被用户理解，也便于在知识发现的过程中进行人机交互。

研究在网络环境下的数据挖掘技术，特别是在互联网上建立DMKD服务器，并且与数据库服务器配合，实现WebMining。

加强对各种非结构化数据的开采，如对文本数据、图形数据、视频图像数据、声音数据乃至综合多媒体数据的开采。

交互式发现、知识的维护更新。

复杂工程过程的数据挖掘。

1.3 数据挖掘研究内容及功能

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以及网上数据挖掘等。数据挖掘所发现的知识最常见的有以下几类：

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，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 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）

它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）

它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。它是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，中间结点是带有分枝的属性，该分枝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。另外数据分类还有统计、粗糙集、神经网络、支持向量机等方法。

4. 预测型知识（Prediction）

它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。

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。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，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、到宏观，以满足不同用户不同层次决策的需要。

与上述研究内容相对应，数据挖掘主要有以下五类功能：

（1）分类　按照分析对象的属性、特征，建立不同的组类来描述事物。例如：银行部门根据以前的数据将客户分成了不同的类别，现在就可以根据这些来区分新申请贷款的客户，以采取相应的贷款方案。

（2）聚类　数据库中的记录可被化分为一系列有意义的子集，即聚类。聚类增强了人们对客观现实的认识，是概念描述和偏差分析的先决条件。聚类技术主要包括传统的模式识别方法和数学分类学。20 世纪 80 年代初，Mchalski 提出了概念聚类技术其要点是，在划分对象时不仅考虑对象之间的距离，还要求划分出的类具有某种内涵描述，从而避免了传统技术的某些片面性。

（3）关联规则和序列模式的发现　关联是某种事物发生时其他事物会发生的这样一种联系。例如：每天购买啤酒的人也有可能购买香烟，比重有多大，可以通过关联的支持度和可信度来描述。与关联不同，序列是一种纵向的联系。

（4）预测　数据挖掘自动在大型数据库中寻找预测性信息，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以及认定对指定事件最可能作出反应的群体。

（5）偏差的检测　数据库中的数据常有一些异常记录，从数据库中检测这些偏差很有意义。偏差包括很多潜在的知识，如分类中的反常实例、不满足规则的特例、观测结果与模型预测值的偏差、量值随时间的变化等。偏差检测的基本方法是，寻找观测结果与参照值之间有意义的差别对分析对象的少数的、极端的特例的描述，揭示内在的原因。

需要注意的是数据挖掘的各项功能不是独立存在的，而是互相联系发挥作用。

1.4　数据挖掘常用技术及工具

数据挖掘是一门涉及面很广的交叉学科，包括机器学习、数理统计、神经网络、数据库、模式识别、粗糙集、模糊数学等相关技术。根据数据挖掘的方法分，可粗分为：统计方法、机器学习方法、智能信息处理方法和数据库方法。

统计方法可细分为：回归分析（多元回归、自回归等）、判别分析（贝叶斯判别、费歇尔判别、非参数判别等）、聚类分析（系统聚类、动态聚类等）、探索性分析（主元分析法、相关分析法等）、支持向量机等。机器学习中，可细分为：归纳学习方法（决策树、规则归纳等）、基于范例的推理 CBR、贝叶斯网络等；智能信息处理方法中，可细分为：神经网络、模糊集及粗糙集、遗传算法、人工免疫方法；数据库方法主要是基于可视化的多维数据分析或 OLAP 方法，另外还有面向属性的归纳方法。

由于数据挖掘可以为企业构筑竞争优势，为社会带来巨大的经济效益，一些国际知名公司也纷纷加入数据挖掘的行列，研究开发相关的软件和工具。美国的 IBM 公司于 1996 年研

制了智能挖掘机，用来提供数据挖掘解决方案；SPSS 股份公司开发了基于决策树的数据挖掘软件即 SPSS CHAID；思维机器公司在 1997 年开发了 Darwin 数据挖掘套件；还有 Oracle 公司、SAS 公司和 Mapinfo 公司等都开发了相关产品。这些开发商和供应商的加入，使数据挖掘技术逐步走向实用化阶段，极大地推动了该技术的蓬勃发展。同时，数据挖掘的网站也日益增多，英国的网站 http://www.cs.bham.ac.uk，美国的 http://www.kdd.org，以及 Gregory Piatetsky-shapiro 在 1997 年创建的网站 http://www.kdnuggets.com 等都含有数据挖掘咨询、算法和技术、软件工具和开发商等相关的资源和信息。现许多公司和研究机构开发了一系列的工具，归纳如表 1－1 所示。

表 1－1　常用数据挖掘工具及其比较[2]

公　司　名	产　品　名	NN	DT	B	kM	kNN	S	Pred	TS	C	A
Angoss International Ltd.	KnowledgeSEEKER		Yes					Yes			
	Knowledge STUDIO	Yes	Yes		Yes			Yes		Yes	
Business Objects	Business Miner		Yes								
Cognos Incorporated	4Thought	Yes						Yes	Yes		
	Scenario		Yes								
HNC Software	DataBase Mining Marksman	Yes						Yes		Yes	
Informix Software Inc.	Red Brick Data Mine			Yes				Yes			
International Business Machines	Intelligent Miner	Yes	Yes			Yes		Yes	Yes	Yes	Yes
Accrue Software	Decision Series	Yes	Yes	Yes				Yes		Yes	Yes
Neural Ware	NeuralSIM	Yes						Yes			
Oracle Corp.	Darwin	Yes	Yes			Yes		Yes			
RightPoint Software	DataCruncher			Yes				Yes			
Salford Systems	CART		Yes					Yes			
SAS Institute	Enterprise Miner	Yes	Yes				Yes	Yes	Yes	Yes	Yes
SGI	MineSet		Yes	Yes	Yes			Yes		Yes	Yes
SPSS, Inc.	Answer Tree		Yes				Yes	Yes			
	Clementine	Yes	Yes					Yes	Yes	Yes	Yes
	Neural Connection	Yes					Yes	Yes			
Tandem Computers Incorporated	Object－Relational Technology										

（续）

公 司 名	产 品 名	NN	DT	B	kM	kNN	S	Pred	TS	C	A
Unica Technology	Pattern Recognition Workbench	Yes			Yes	Yes	Yes	Yes	Yes	Yes	
	Model 1	Yes	Yes	Yes	Yes		Yes	Yes			

注：NN = Neural Net（神经网络）；DT = Decision Tree（决策树）；B = Bayes（贝叶斯方法）；kM = k - Means（动态聚类）；kNN = k - Nearest Neighbor（最邻近算法）；S = Traditional Statistical Techniques（传统统计技术）；P = Prediction（预测）；TS = Time Series（时间序列）；C = Clustering（聚类方法）；A = Association（关联方法）。

1.5 数据挖掘应用热点

数据挖掘技术已经成功地应用于社会生活的方方面面，如政府管理决策、商业经营、科学研究和企业决策支持等领域，都可以采用数据挖掘技术解决一些问题。市场营销预测顾客的购买行为，划分顾客群体，使用交互式询问技术、分类技术和预报技术，更精确地挑选潜在的顾客；银行业侦测信用卡的欺诈行为，客户信誉分析；生产、销售和零售业预测销售额，决定库存量，批发点分布的规划和调度，物流管理。工业制造和生产领域也是一个十分有潜力的使用数据挖掘技术的市场，如质量控制、预测机器故障、挖掘影响生产力的关键因素等。就目前来看，将来的几个热点包括网站的数据挖掘（Web site data mining）、生物信息或基因（Bioinformatics/genomics）的数据挖掘、文本的数据挖掘（Textual mining）及工程应用中的数据挖掘。下面就这几个方面加以简单介绍。

1. 网站的数据挖掘

随着 Web 技术的发展，各类电子商务网站风起云涌，建立起一个电子商务网站并不困难，困难的是如何让您的电子商务网站有效益。要想有效益就必须吸引客户，增加能带来效益的客户忠诚度。电子商务业务的竞争比传统的业务竞争更加激烈，原因有很多方面，其中一个因素是客户从一个电子商务网站转换到竞争对手那边，只需点击几下鼠标即可。网站的内容和层次、用词、标题、奖励方案、服务等任何一个地方都有可能成为吸引客户，同时也可能成为失去客户的因素。而同时电子商务网站每天都可能有上百万次的在线交易，生成大量的记录文件（Logfiles）和登记表，如何对这些数据进行分析和挖掘，充分了解客户的喜好、购买模式，甚至是客户一时的冲动，设计出满足于不同客户群体需要的个性化网站，进而增加其竞争力，几乎变得势在必行。若想在竞争中生存进而获胜，就要比您的竞争对手更了解客户。

在对网站进行数据挖掘时，所需要的数据主要来自于两个方面：一方面是客户的背景信息，此部分信息主要来自于客户的登记表；而另外一部分数据主要来自浏览者的点击流（Click - stream），此部分数据主要用于考察客户的行为表现。但有的时候，客户对自己的背景信息十分珍重，不肯把这部分信息填写在登记表上，这就会给数据分析和挖掘带来不便。在这种情况之下，就不得不从浏览者的表现数据中来推测客户的背景信息，进而再加以利用。就分析和建立模型的技术和算法而言，网站的数据挖掘和原来的数据挖掘差别并不是特别大，很多方法和分析思想都可以运用。所不同的是网站的数据格式有很大一部分来自于点击流，和传统的数据库格式有区别。因而对电子商务网站进行数据挖掘所做的主要工作是数

据准备。目前，有很多厂商正在致力于开发专门用于网站挖掘的软件。

2. 生物信息或基因的数据挖掘

生物信息或基因数据挖掘则完全属于另外一个领域，在商业上很难讲有多大的价值，但对于人类却受益匪浅。例如，基因的组合千变万化，得某种病的人的基因和正常人的基因到底差别多大？能否找出其中不同的地方，进而对其不同之处加以改变，使之成为正常基因？这都需要数据挖掘技术的支持。

对于生物信息或基因的数据挖掘和通常的数据挖掘相比，无论在数据的复杂程度、数据量还有分析和建立模型的算法而言，都要复杂得多。从分析算法上讲，更需要一些新的和好的算法。现在很多厂商正在致力于这方面的研究。但就技术和软件而言，还远没有达到成熟的地步。

3. 文本数据挖掘

人们很关心的另外一个话题是文本数据挖掘。举个例子，在客户服务中心，把同客户的谈话转化为文本数据，再对这些数据进行挖掘，进而了解客户对服务的满意程度和客户的需求以及客户之间的相互关系等信息。从这个例子可以看出，无论是在数据结构还是在分析处理方法方面，文本数据挖掘和前面谈到的数据挖掘相差很大。文本数据挖掘并不是一件容易的事情，尤其是在分析方法方面，还有很多需要研究的专题。目前市场上有一些类似的软件，但大部分方法只是把文本移来移去，或简单地计算一下某些词汇的出现频率，并没有真正的分析功能。

4. 工程应用中的数据挖掘

数据挖掘技术从一开始就是面向应用的。随着现在各行业业务自动化的实现，冶金、采矿、电力、机械、生产、制造、过程控制等各工程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，也不是为了纯商业运作而产生。分析这些数据更主要是为决策提供真正有价值的信息，因此工程领域的数据挖掘应用需求日益明显，促使了数据挖掘技术及应用的更深一步的发展。

在工程领域中，积累的数据越来越多，如果工程领域的数据全部依靠手工分析显然是不现实的，因此必须提高自动化、智能化程度。更重要的是，有些情况下人难以到达现场，如航空、航天、深水作业等。这些都对信息的智能处理提出了要求。因此，数据挖掘对工程领域的意义是不言而喻的。近几十年来专家系统在工程领域得到广泛的应用，但是专家系统中知识库级知识一般是人工提取，归纳总结后用适合机器存储和应用的方式表示出来并灌输给机器，因此专家系统只是一个模式匹配系统，知识的获取成为影响专家系统的一个瓶颈。例如，20 世纪 60 年代由斯坦福大学开发的名为 DENDRAL 的专家系统能根据质谱仪给出的数据发现已知或未知的高分子化合物的分子结构；70 年代开发出的诊断和治疗传染性血液病的专家系统 MYCIN 和矿藏勘探的专家系统 PROSPECTOR；80 年代开发的为美国著名计算机公司 DEC 做 VAX 机硬件配置的专家系统 XCON 等[3]都难以逾越这个瓶颈。数据挖掘作为一种新的知识获取工具，将有效地改善专家系统的这种状态，对工程的智能化、信息化、自动化的发展起着不可估量的作用。

从工程角度讲，数据挖掘是一个需要经过反复的多次处理过程。数据挖掘的处理过程模型为数据挖掘提供了宏观指导和工程方法。合理的处理过程模型能将各个处理阶段有机地结合在一起，指导人们更好地开发及使用数据挖掘系统。从数据挖掘进入工程应用领域起，就

有人对数据挖掘的过程进行归纳和总结，提出了不同的数据挖掘处理过程模型。其中，Fayyad 等人给出的多处理阶段模型是一种通用模型，也是目前最广为接受的一种处理模型[4]，如图 1-1 所示。

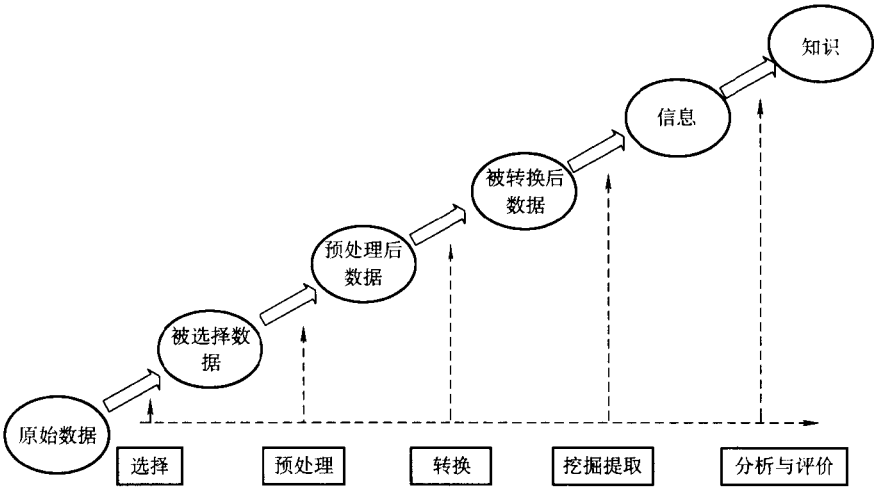

图 1-1　数据挖掘通用模型

Brachman 和 Anand 通过对数据挖掘实际工作的调查，发现用户与数据库的交互工作量很大，据此认为数据挖掘应该更着重于对用户进行知识发现的整个过程的支持，而不是仅仅限于在数据挖掘的一个阶段上，进而提出了以用户为中心的处理过程模型。Brachman 和 Anand 在他们开发的数据挖掘系统 IMACS（InteractiveMarketing Analysis and Classification System）中采用了这种以用户为中心的处理过程模型。George H. John 在其博士论文中给出另外一种数据挖掘处理过程模型[5]。该模型强调由数据挖掘人员和领域专家共同参与数据挖掘的全过程。文献［6］认为前述模型对知识发现过程中的反复学习和多目标学习支持不够，为此提出支持多数据集多学习目标的数据挖掘处理模型，该模型将数据和学习算法尽量分离，从而使得数据挖掘更适合实际工作的需要，并使得最终用户和数据挖掘人员之间的影响尽量小，以提高学习效率。在具体算法进行数据处理时，必须对数据进行简单的筛选和加工以剔除冗余数据。

上述 4 种处理模型的共同点是都要经过准备、预处理、算法设计、数据挖掘和后处理等共同的阶段，如图 1-2 所示。

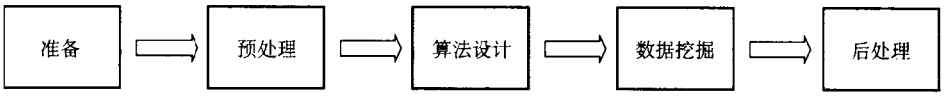

图 1-2　数据挖掘处理模型通用阶段

其中准备阶段包括问题定义、对象理解、数据收集等准备工作，搜索所有与业务对象有关的内部和外部数据信息，并从中选择出适用于数据挖掘应用的数据；预处理包括数据清理、压缩、变换等。研究数据的质量，为进一步的分析作准备，并确定将要进行的挖掘操作的类型；算法设计及数据挖掘为建立分析模型，才用合理算法进行数据抽取。这个分析模型

是针对挖掘算法建立的，建立一个真正适合挖掘算法的分析模型是数据挖掘成功的关键；后处理包括结果解释、输出、评价、分析、使用等，解释并评估结果。其使用的分析方法一般视数据挖掘操作而定，通常会用到可视化技术，将分析所得到的知识集成到业务信息系统的组织结构中去。

数据挖掘得到了广泛的应用，取得了良好的效果。

市场研究公司，如美国的 A.C. Nelson 和 Information Resources，欧洲的 GFK 和 Infracts Burke 等纷纷开始使用数据挖掘技术来处理迅速增长的销售和市场信息数据。商家的激烈竞争导致了市场的快速饱和，产品的迅速更新使得经营者对市场信息的需求格外强烈。利用数据挖掘对市场进行的有效预测，使这些市场研究公司获得巨大的效益。

英国广播公司（BBC）也应用数据挖掘技术来预测电视收视率，以便合理安排电视节目时刻表。信用卡公司 American Express 自采用数据挖掘技术后，信用卡使用率增加了 10% ~ 15%。

通用电器公司（GE）与法国飞机发动机制造公司（SNECMA），用数据挖掘技术研制了 CAS SIOPEE 质量控制系统，被 3 家欧洲航空公司用于诊断和预测波音 737 的故障，带来了可观的经济效益。

电力系统是一个非线性的工程系统，在运行过程中不断产生和积累大量的数据。在电力系统中，应用理论研究的方法已经解决了许多问题，但如果能应用数据挖掘技术，则可以更加充分地利用这些运行数据，揭示电力系统历年积累的数据背后蕴含的规则，找到解决问题更加合理的方法，同时还可以为决策提供根据有力的科学依据。这包括利用计算机通过对电力系统历史数据的学习归纳，建立起一个预测性模型，然后，根据当前已知的数据来预测未知的数据；用于对电力突发事故的处理过程进行分析，得出具有针对性的对策供专家参考；另外，对负荷进行分类，找出其中最显著的变化，以此为根据预测未来的负荷具有非常重要的意义。

美国钢铁公司和神户钢铁公司利用数据挖掘技术开发的 ISPA 系统，能分析产品性能规律和进行质量控制，取得了显著效果。我国宝钢在数据挖掘方面做了大量的研究，从基于数据挖掘技术的配矿系统的研究，到数据挖掘及其在宝钢质量控制中的应用研究，无论在理论研究，还是在生产应用上都取得很大的进展。而今，宝钢已经成功地开发了基于 SAS 系统的数据挖掘软件“实用数据挖掘系统 2.0（Practical Miner 2.0，简称 PM2.0）”，该软件现由美国 SAS 软件研究所（上海）有限公司、上海宏能软件公司以及上海宝钢计算机公司代理销售。PM2.0 是结合实际工作经验和数据挖掘理论而开发的软件，它包括了数据收集、数据筛选、数据采样、数据挖掘、知识优化一整套知识发现流程，由预处理阶段、数据挖掘阶段以及辅助阶段等三大模块，数据收集、数据抽样、数据预处理、可视化探索、聚类分析、模型建立、数据预测、优化设计、趋势分析、规范管理等 10 个主要功能以及系统帮助和退出系统等两个辅助功能组成。该系统是企业数据挖掘的一个有效工具，可以为企业进步提供有力的帮助。其主要特点是提供了从数据预处理，到数据呈现的整个数据挖掘过程；具有强大的数据预处理功能，模型建立功能以及目标优化功能；具有友好的界面，即使不熟悉数据挖掘技术的人员，也可以十分轻松地掌握该系统。