信息科学与技术丛书

纪希禹 主编

韩秋明 李微 李华锋 等编著

数据挖掘技术应用实例

- 数据挖掘理论知识
- 数据挖掘操作方法
- 七大热门领域的实际应用

机械工业出版社
CHINA MACHINE PRESS

信息科学与技术丛书

数据挖掘技术应用实例

纪希禹　主编

韩秋明　李　微　李华锋　等编著

机械工业出版社

本书在介绍数据挖掘技术理论和算法的基础上，通过不同领域的应用案例，来说明数据挖掘在实际应用中的具体操作方法，以期为读者提供一个更为广阔的视角。本书从理论、应用实例和数据挖掘的发展趋势，以及面临的机遇和挑战等方面，对数据挖掘技术进行了详细介绍，其中在应用实例部分分别介绍了数据挖掘技术在客户关系管理、市场营销、证券领域、电信领域、产品设计、军事领域以及 Web 数据挖掘等方面的应用。

本书可作为企事业单位信息管理部门以及其他各行各业的管理者、信息分析人员、数据统计人员、市场营销人员、研究与开发人员的参考资料，也可作为高等院校信息管理类、数据分析类、计算机类等相关专业的教材和参考书，还可作为高等院校毕业论文或毕业设计的参考资料。

图书在版编目（CIP）数据

数据挖掘技术应用实例/纪希禹主编. —北京：机械工业出版社，2009.4
（信息科学与技术丛书）
ISBN 978-7-111-26460-6

Ⅰ. 数… Ⅱ. 纪… Ⅲ. 数据采集 Ⅳ. TP274

中国版本图书馆 CIP 数据核字（2009）第 029935 号

机械工业出版社（北京市百万庄大街 22 号 邮政编码 100037）
策划编辑：车 忱
责任编辑：车 忱 常建丽
责任印制：邓 博

北京中兴印刷有限公司印刷
2009 年 4 月·第 1 版第 1 次印刷
184mm×260mm · 17 印张 · 420 千字
0 001—3 000 册
标准书号：ISBN 978-7-111-26460-6
定价：34.00 元

凡购本图书，如有缺页、倒页、脱页，由本社发行部调换
销售服务热线电话：（010）68326294 68993821
购书热线电话（010）88379639 88379641 88379643
编辑热线电话（010）88379753 88379739
封面无防伪标均为盗版

出 版 说 明

随着信息科学与技术的迅速发展，人类每时每刻都会面对层出不穷的新技术和新概念。毫无疑问，在节奏越来越快的工作和生活中，人们需要通过阅读和学习大量信息丰富、具备实践指导意义的图书来获取新知识和新技能，从而不断提高自身素质，紧跟信息化时代发展的步伐。

众所周知，在计算机硬件方面，高性价比的解决方案和新型技术的应用一直备受青睐；在软件技术方面，随着计算机软件的规模和复杂性与日俱增，软件技术不断地受到挑战，人们一直在为寻求更先进的软件技术而奋斗不止。目前，计算机在社会生活中日益普及，随着Internet延伸到人类世界的方方面面，掌握计算机网络技术和理论已成为大众的文化需求。由于信息科学与技术在电工、电子、通信、工业控制、智能建筑、工业产品设计与制造等专业领域中已经得到充分、广泛的应用。所以这些专业领域中的研究人员和工程技术人员越来越迫切需要汲取自身领域信息化所带来的新理念和新方法。

针对人们了解和掌握新知识、新技能的热切期待，以及由此促成的人们对语言简洁、内容充实、融合实践经验的图书迫切需要的现状，机械工业出版社适时推出了“信息科学与技术丛书”。这套丛书涉及计算机软件、硬件、网络和工程应用等内容，注重理论与实践的结合，内容实用、层次分明、语言流畅，是信息科学与技术领域专业人员不可或缺的参考书。

目前，信息科学与技术的发展可谓一日千里，机械工业出版社欢迎从事信息技术方面工作的科研人员、工程技术人员积极参与我们的工作，为推进我国的信息化建设作出贡献。

机械工业出版社

前　言

21 世纪是信息的世纪。信息已经和能源、材料一起成为支撑人类社会发展的三大要素，并显示出越来越重要的作用。21 世纪也是知识的世纪。以知识为主题的许多新研究对象的出现，如知识经济、知识产业、知识工人、知识管理、知识工程和知识网络等，丰富了理论研究的内涵，也推动了以网络为基础的信息技术向着更高层次发展。如何实现从信息到知识的转变呢？数据挖掘技术给出了答案。

随着计算机技术、网络技术以及通讯技术的迅速发展和普及，信息技术将我们领进入了全新的信息社会。但是随着信息技术的不断进步，人们收集数据的手段不断丰富，数据库、数据仓库容量的不断膨胀，以及 Web 等新型数据源的不断出现，各种各样的数据已经由不同的渠道汇成了浩瀚的海洋。据不完全统计，1993 年全球数据存储容量约为 2000TB，到 2000 年增加到大约 300 万 TB，到 2008 年，这一数字已经飙升至上亿 TB。人们面临的主要问题已经不再是没有充分的信息可选择，而是徜徉在如此庞大的信息之中，如何更为有效地利用它们，并且找到蕴含于这些信息之中的有价值的知识宝藏。当前的数据库系统无法发现隐藏在海量数据中的潜在联系和规则，不能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知识的手段，导致了人们面临“数据丰富而知识匮乏”的现象。因此，在人们需求的呼唤下，数据挖掘和知识发现技术应运而生，并得以在社会生活的各个领域蓬勃发展。数据挖掘推动信息分析处理技术上升到一个更高的阶段。比尔·盖茨曾预计，数据挖掘技术将是今后计算机技术发展的第二大方向。

数据挖掘技术是一个多学科交叉的综合研究领域。它融合了数据库技术、人工智能、机器学习、统计分析、模式发现、可视化技术、信息检索以及信号处理等多个学科领域的技术。不过也正因为它涉及的范围很广泛，发展的时间也不是很长，因此要真正理解数据挖掘的本质并不是一件容易的事情。目前市场上关于数据挖掘的教材和参考资料并不在少数，但是大部分要么过于偏重理论研究，与数据挖掘实际发展目标相违背；要么偏重算法应用，机器语言晦涩难懂，为广大数据挖掘技术学习者带来了不少麻烦；或局限于某一专门的领域，使其他领域的学习者虽然可以借鉴，但是由于不同领域数据来源不同，数据质量不同以及数据处理方法等等方面大相径庭，因此他们真正在自己的领域中应用操作起来有不少困难。我们经过研究后发现，只有在掌握了基础知识之后，再经过对实际应用案例的深入学习，才能又好又快地理解数据挖掘技术的内涵，把握数据挖掘技术的本质。因此，我们取长补短，在简单介绍了数据挖掘的基本概念和算法之后，通过对七个应用领域案例的介绍，详尽地说明了数据挖掘技术的应用情况。本书总共分为三部分。第一部分为第 1、2 章，为理论介绍部分。第二部分为第 3～9 章，比较翔实地介绍了数据挖掘技术在客户关系管理、市场营销、股票证券、电信领域、产品设计、军事研究以及 Web 挖掘等七个方面的应用实例。第三部分为第 10 章，简要介绍了数据挖掘的发展趋势。

本书的目的就是抛砖引玉，希望读者通过对本书的学习，了解和掌握数据挖掘技术的理念和算法，熟悉在各个领域应用的流程和分析方法，从而为以后的数据分析工作夯实基础。

本书可以作为高等院校信息管理类、数据分析类、计算机类等各相关专业的教材和参考书。也可作为高等院校毕业论文或毕业设计的参考资料。本书亦可作为企事业单位信息管理部门以及其他行业的管理者、信息分析人员、数据统计人员、市场营销人员、研究与开发人员以及数据挖掘爱好者的参考资料。

本书由纪希禹主编，韩秋明、李微、李华锋等编著。其中，纪希禹编写了第 1 章，韩秋明编写了第 2 章、第 3 章和第 10 章。李华锋编写了第 4 章和第 6 章。吴友蓉编写了第 5 章。李微编写了第 7 章、第 8 章和第 9 章。全书由韩秋明筹划并统稿。参加编写的人员还有邹素琼、冯强、曾德惠、许庆华、程亮、周聪、黄志平、胡松、邢永峰、邵军、边海龙、刘达因、赵婷、马鸿娟、侯桐、赵光明、李胜、李辉、侯杰、王红研、王磊、闫守红、康涌泉、李欢、蒋杼倩、王小东、张森、张正亮、宋利梅、何群芬、程瑶。

在本书的编写过程中参考了很多国内外的研究成果，也得到了四川大学公共管理学院情报学专业的老师和同学们的无私帮助和支持。河北大学日语系的赵玮、四川大学情报学专业的朱榕、石家庄信息工程职业学院的李倩、石家庄市商业银行的宋瑞鹏、北京神州泰岳软件股份有限公司的李忠浩、北京中科软科技股份有限公司的赵晴、中国移动通信沧州分公司的刘宇、河北大学信息管理与信息系统专业的田伟铮等，也为本书的出版做出了努力和贡献，在此一并表示感谢。

由于作者水平有限，再加上时间仓促，书中的错误与不足在所难免，敬请广大读者不吝批评指正。

为充分展现本书编写特点，帮助读者深刻理解本书编写意图与内涵，进一步提高对本书教学的使用效率，欢迎读者将图书使用过程中的问题与各种探讨、建议反馈给我们，我们会竭诚给您满意的答复。我们的 E-mail 是：china_54@tom.com。

作　者

目　录

第1章 绪 论

知识点：

- 数据挖掘的概念
- 数据挖掘的过程
- 数据仓库
- 联机分析处理（OLAP）

本章导读：

数据是无处不在的。当飞速增长的数据给我们带来方便和快捷的同时，也将我们推入浩瀚的数据海洋。目前数据的数量已经远远超出了人们分析和处理的能力，人们面对这些海量数据的时候，常常会感到无所适从，不知所措。那么，怎样才能不被数据的海洋淹没？怎样利用这些数据使它们为人们作贡献，为决策提供支持？怎样发现在数据海洋中蕴藏的丰富的知识？数据挖掘技术为我们提供了解决这些问题的好方法。本章将主要介绍数据挖掘的基本概念以及相关技术的基础，主要有数据挖掘、数据仓库、数据集市、元数据以及 OLAP 等。

1.1 数据挖掘的基本概念

近几年来，以互联网为代表的计算机信息技术的迅速普及，使人们的生活条件和社会环境发生了巨大的变化。人们生产、收集、存储、处理信息和数据的能力大幅提高。大量的数据库被广泛应用于企业管理、科学研究、电子商务、金融预测、商品零售、医药化工、政府办公以及工程开发等社会生活的各个领域，并且这一趋势仍将继续。人们积累的信息量已经达到了 TB 级，甚至 PB 级。信息爆炸是当今时代的一个重要特点，信息超载、信息过量几乎成为每个人都需要认真面对的问题。难道这些数据真的无法使用，成为所谓的数字垃圾，成为人们探求知识道路上的绊脚石了吗？其实不然，只要通过合理的思想引导，再通过适当的技术手段，使数据得到最大限度的利用，为决策和战略发展服务，就能使庞大的数据真正成为一种推动社会发展的生产力、一种孕育财富和知识的资源。否则，大量的数据不但不会为决策服务，不会为推动生产力发展作贡献，反而可能成为经济社会发展前进的包袱。那么如何面对海量信息而不被信息淹没？如何对这些数据进行更为深入的分析？如何发现在数据背后隐藏的重要信息？如何更好地利用这些数据，从而提取出有价值的知识？这些已经成为信息爆炸和知识化的时代给我们提出的关注热点和亟待解决的问题。因为当前的数据库系统无法发现隐藏在海量数据中的潜在联系和规则，不能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知识的手段，导致了人们面临的“数据丰富而知识匮乏”的现象。因此在需求的呼唤下，数据挖掘和知识发现技术应运而生，并得以在社会生活的各个领域蓬勃发

展，显示出其强大的生命力。微软总裁比尔·盖茨曾预计，数据挖掘技术将是今后计算机技术发展的第二大方向。

不错，数据挖掘技术正在以一种全新的理念改变着人类信息管理的方式。它融合了数据库技术、人工智能、机器学习、统计分析、模式发现、可视化技术、信息检索以及信号处理等多个领域的技术，使人们从单纯的对信息收集、整理、组织、存储、传播和利用，向信息重构、信息整合、知识创新等深层加工转变，使信息处理技术进入了一个更为高级的阶段。

1.1.1 啤酒与尿布

在数据挖掘的发展中，有一个案例是不得不提的，这就是由沃尔玛演绎的著名的“啤酒与尿布”。

在美国阿肯色州的一家大型超级市场里面，人们惊奇地发现这样一种现象，在货架上，尿布和啤酒竟然摆在一起，这在美国所有的超级市场里都是不曾有过的。这家超市就是世界著名商业零售连锁企业沃尔玛。它是拥有世界上最大数据仓库系统的企业之一。事情的经过是这样的。为了能够准确了解顾客在其门店的购买习惯，沃尔玛通过对其每一家门店的顾客购物行为进行货篮分析，想了解顾客经常一起购买的商品有哪些。在一项货篮关联分析中显示：“与尿布同时购买最多的商品居然是啤酒！”

这虽然是一个让人吃惊的分析结论，可这是数据挖掘技术对历史数据进行分析的结果，反映了数据内在的规律。那么这个结果符合现实情况吗？

于是，沃尔玛派出市场调查人员和数据分析师对这一数据挖掘结果进行调查分析。经过大量实际调查和分析，发现了购买这两种商品的主要客户几乎都是 25 岁到 35 岁、家中有婴儿的男性。由此，数据挖掘技术揭示了一个隐藏在“啤酒与尿布”背后的美国人的一种行为模式：在美国，年轻的太太们会经常叮嘱她们的丈夫在下班后去超级市场为小孩买尿布，而这些年轻的父亲中有 30%～40%的人同时也会为自己买一些啤酒，因此啤酒和尿布在一起被购买的机会还是很多的。

既然尿布与啤酒一起被购买的机会很多，于是沃尔玛就在其每一个门店内调整货架，将尿布与啤酒并排摆放在一起，结果使销售量大幅增长。

按照常规思维来说，尿布和啤酒完全沾不上边，那是什么让沃尔玛超市发现了这一个蕴含巨大商业利益的结论呢？这是因为他们在收集了其各个门店的详细交易数据之后，利用数据挖掘工具对这些数据进行了分析和挖掘，才发现了这一个让人感到意外但是具有价值的规律。这个案例真实地反映了数据挖掘的价值。

1.1.2 什么是数据挖掘

那么什么是数据挖掘呢？数据挖掘（Data Mining），就是从海量数据中挖掘出隐含在其中的矿藏——知识。数据挖掘这个术语首先出现在 1989 年 8 月美国底特律召开的第 11 届国际人工智能联合学术会议——“数据库中的知识发现”专题讨论会上，在随后的几年中，数据挖掘的专题讨论会继续举行。随着参会人数的增多和技术的发展成熟，从 1995 年开始，每年都要举办一次有关数据挖掘技术的国际性会议。除了数据挖掘技术的理论研究外，相当数量的数据挖掘产品和应用系统也随之出现，并且获得了一定的成功，得到了信息产业界的广泛

关注。

然而，数据挖掘从诞生发展至今虽已有约20年的历史，可目前却没有一个获得完全认同的定义。因为数据挖掘技术在不同领域有不同应用，各位学者和专家也分别从不同的角度进行定义。从数据库的角度来看，数据挖掘定义为从存储在数据库、数据仓库或者其他信息库中的大量数据中发现用户感兴趣的知识的过程。从统计学角度来看，数据挖掘是指通过分析目标数据集，来发现可理解的、有用的、经过整理归纳的数据，以及数据之间包含的可信的、以前未知的关系，并且将其通过可视化技术提供给数据拥有者的过程。从机器学习的角度来看，数据挖掘是指从数据中抽取隐含的、明显未知的和潜在的有用信息。

在不同的文献和应用领域也有一些其他的定义，比如Berry和Linoff认为，数据挖掘是通过自动或半自动化的工具对大量的数据进行探索和分析的过程，其目的是发现其中有意义的模式和规律。Ferruza认为，数据挖掘是在知识发现过程中，用来分辨出存在于数据中的未知关系和模式的一系列方法。Zekulin把数据挖掘定义为一个从大型数据库中提取出以前未知的、易理解的、可执行的信息，并且用它来为决策服务。还有学者认为，数据挖掘的概念应该分为狭义和广义两种。

一般认为，广义的数据挖掘又称数据库中的知识发现（Knowledge Discovery in Databases），简称知识发现（KDD）。它是从大量的、不完整的、有噪声的、模糊的和随机的数据中，提取隐含在其中的、人们事先不知道的、但又是可信的、潜在的和有价值的信息和知识的过程。这个概念包括以下几层含义：

（1）作为数据挖掘的数据源，其数据必须是海量的，含有噪声的。数据是指一个相关事实的集合，它是用来描述事物有关方面的信息。

（2）挖掘出来的模式是可理解的、易描述的、有用的。模式是指，对于数据源中的数据，可以用语言来描述其中数据的特性。只有当表达式E比列举出所有F_E中的元素的描述方法更为简便时，才可以称为模式。比如："如果成绩在91～100之间，则成绩优秀"这一描述明显要比"如果成绩为91、92、93、94、95、96、97、98、99或100，则成绩优秀"这种描述简便，则前者可称为一个模式，而后者就不能称为一个模式。

（3）通过数据挖掘发现的知识是用户感兴趣的。所谓兴趣，是指知识的可信度、新颖性、潜在作用性和可理解性的综合。可信度和潜在作用性指的是，从当前数据中通过数据挖掘所发现的模式必须有意义并具有一定的正确程度，否则数据挖掘就毫无用处。可信度可以通过新的数据来检验所发现模式的正确性，潜在作用性可以通过某些函数值来衡量。新颖性是指经过数据挖掘提取出的模式必须是新颖的，以前未知或不很明显的。模式是否新颖可以通过两个途径来衡量：其一是通过对比当前得到的数据和以前的数据或期望得到的数据，来判断该模式的新颖程度；其二是通过对比发现的模式与已有的模式的关系来判断新颖程度。可理解性是指，数据挖掘的一个目标就是将数据库中隐含的模式以容易被人理解的形式表现出来，从而帮助人们更好地了解数据库中所包含的信息。当然一个模式是否容易被人理解，这本身很难衡量，比较常用的方法是对其简单程度进行测试，从而判断其是否容易被人理解。

（4）通过数据挖掘发现的知识不是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，更不是什么机器定理证明。数据挖掘所发现的知识不是绝对的，是相对的，是有特定条件约束的，面向特定领域的。因为数据挖掘技术应用在社会生活的许多

领域，而每个领域和领域之间对数据的选择和评价标准是不尽相同的，因此所发现的模式和知识也是有很大区别的。

狭义的数据挖掘是一个利用各种分析工具在海量数据中发现模型和数据之间关系的过程，是知识发现过程中的一个步骤。一个完整的知识发现过程如图 1-1 所示。

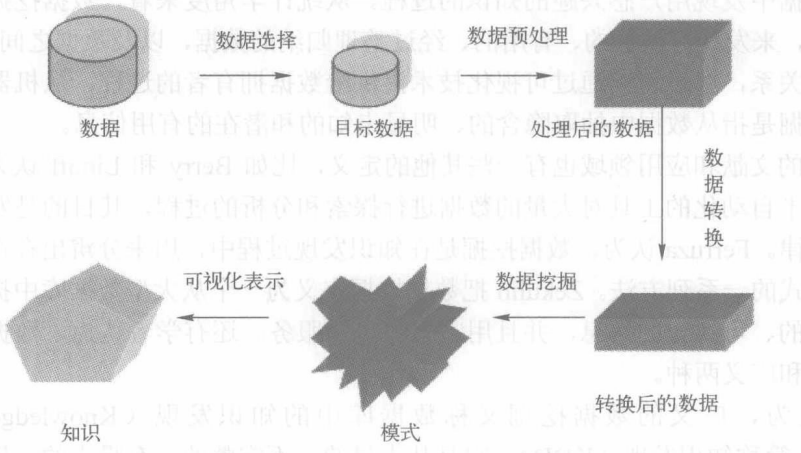

图 1-1 知识发现过程

从图 1-1 可见，数据挖掘只是知识发现过程中一个发现模式的子过程，并且是最核心的过程。

由于数据挖掘是一门融合了许多学科的交叉学科，受到了不同应用领域的研究者的关注，因此产生了不同的术语名称，主要有如下几种：“数据库中的知识发现”、“知识抽取”、“信息发现”、“智能数据分析”、“信息收获”、“数据考古”、“数据捕捞”以及“数据/模式分析”等。其中，最常用的术语是“知识发现”和“数据挖掘”。相对来讲，“数据挖掘”主要流行于统计领域（最早出现于统计文献中）、数据分析、数据库和管理信息系统领域；而“知识发现”则主要流行于人工智能和机器学习界。

1.1.3 数据挖掘的分类

数据挖掘是一门综合性学科，涉及的学科领域有很多，包括数据库技术、人工智能、机器学习、统计分析、模式发现、可视化技术、信息检索以及信号处理等。从不同的应用出发，可对数据挖掘进行不同的分类，比如，根据挖掘的数据库类型分类、根据采用的技术和方法分类、根据挖掘的知识类型分类、根据数据挖掘的应用领域分类等，如图 1-2 所示。

如图 1-3 所示为数据挖掘与其他学科的关系图。

1. 根据所挖掘的数据库类型分类

数据挖掘一般都要基于一定的数据源，这些数据源可能会是各种各样的数据库。如果是在关系数据库中的基础上进行的数据挖掘，则称之为关系型数据挖掘。如果是在面向对象数据库的基础上进行数据挖掘，则称之为面向对象型数据挖掘。类似地，针对何种类型的数据库，就有基于这一类型数据库的数据挖掘。由此可知，还有关系-对象型数据挖掘、事务型数据挖掘、多媒体数据挖掘、演绎数据挖掘、文本数据挖掘、空间数据挖掘、时间/序列数据挖

掘、Web 数据挖掘、数据仓库的数据挖掘等。

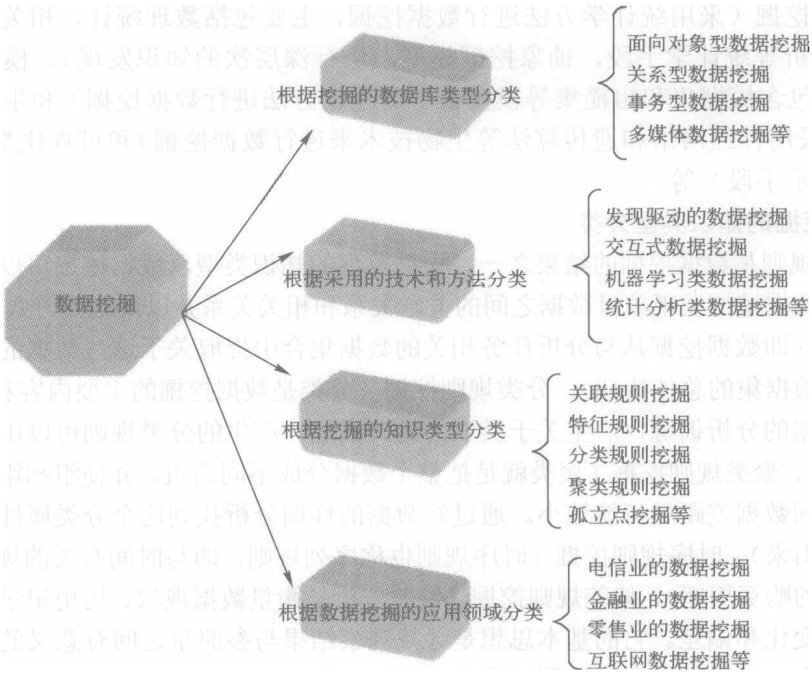

图 1-2 数据挖掘的分类

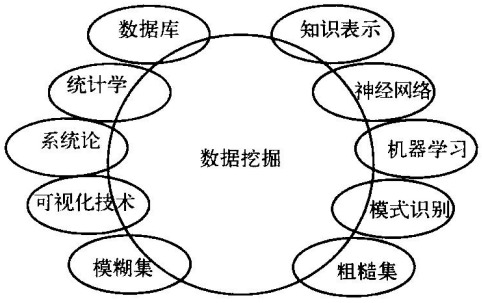

图 1-3 数据挖掘与其他学科的关系

2. 根据所采用的技术和方法分类

根据数据挖掘所采用的技术和方法，数据挖掘可以分为以下几类：有监督的数据挖掘（即采用自上而下的方法，当明确知道要检索的目标，有了很多已知模式的情况下，可以使用这种方法）、无监督的数据挖掘（即采用自下而上的方法，本质是让数据解释自己，这种方法是在数据中寻找模式，然后让使用者自己去判断哪些模式是有用的）、发现驱动的数据挖掘（在目标数据集上利用历史数据自动创建一个模型，来预测以后的行为。在前一步发现的基础上，通过把专家的领域知识应用到下一步的选择判断中，直到模型的形成。模型创建的过程就是数据挖掘的过程）、交互式数据挖掘（利用交互式的处理方式，逐渐

明确数据挖掘的目标，动态改变数据搜索方式，逐步加深数据挖掘过程）、机器学习类数据挖掘（通过采用归纳学习、决策树、类比学习等机器学习方法进行数据挖掘工作）、统计分析类数据挖掘（采用统计学方法进行数据挖掘，主要包括数理统计、相关分析、回归分析、聚类分析等统计学手段，抽象挖掘模型，进行深层次的知识发现）、模糊数学类数据挖掘（采用包含模糊集和粗糙集等模糊数学的理论方法进行数据挖掘）和生物技术类数据挖掘（主要采用神经网络和遗传算法等生物技术来进行数据挖掘）和可视化数据挖掘（采用可视化的表示手段）等。

3．根据挖掘的知识类型分类

发现各种规则是数据挖掘的结果之一。根据发现的知识类型，数据挖掘可以分为关联规则挖掘（通过数据挖掘发现项目数据之间的关联关系和相关关系并以规则的方式表现出来）、特征规则挖掘（即数据挖掘从与分析任务相关的数据集合中提取关于这些数据的特征式，特征式用来描述数据集的总体特征）、分类规则挖掘（分类是数据挖掘的主要内容和结果之一，通过对样本数据的分析训练，产生关于类别的精确描述，产生的分类规则可以用来对未来的数据进行预测）、聚类规则挖掘（聚类就是把整个数据分成不同的组，并使组和组之间差距尽可能地大，组间数据差距尽可能地小。通过对数据的样例分析找到这个分类属性值，并以规则的形式表现出来）、时序规则挖掘（时序规则也称序列规则，即与时间有关的规则，比如顾客对某一商品的购买周期）、偏差规则挖掘（偏差分析是衡量数据现状、历史记录以及数据标准之间的显著变化和偏差。它的基本思想是寻找观察结果与参照量之间有意义的差别）和孤立点挖掘（孤立点是指不符合一般数据模式的数据）等。

4．根据数据挖掘的应用领域分类

数据挖掘在实际中的应用主要是在商业领域（如产品设计、客户关系管理、市场营销以及决策支持）等、科学研究（如天文图形分析、卫星遥感器和 DNA 分子技术等）等很多方面。针对数据挖掘应用的各个细化的行业领域，可以分为电信业的数据挖掘、金融业的数据挖掘、零售业的数据挖掘、医疗卫生领域的数据挖掘、体育运动的数据挖掘和互联网数据挖掘等。

1.1.4 数据挖掘的特点和功能

传统的信息处理技术虽然也给信息处理带来了巨大的进步，作过巨大的贡献，并且现在仍在普遍使用，可是它不能像数据挖掘技术一样可以发现更深层次的知识。通过对数据挖掘的定义以及与传统的信息处理技术的比较，可以得出数据挖掘据有以下特点。

（1）海量性。要从数据中挖掘出规则，其数据必须是海量的、可以表示整个领域业务状况的。数据挖掘所处理的数据源一般是多个数据库经过数据预处理后形成的。

（2）复杂性。在建模方面，数据挖掘的重点大多放在“学习”上，对模型的复杂性和需要的计算量较为关注，而很少放在大样本的渐进推论上。数据挖掘技术有能力对复杂的数据关系进行建模，更适合解决复杂的问题。

（3）离散性。在实践中，算法涉及连续和离散变量的数据集是非常普遍的，统计学中的大多数变量分析方法是为连续变量设计模型的，但许多数据挖掘方法更适合离散变量的分析。实际中，一些基于规则的方法只能使用离散变量，挖掘之前需要将连续变量离散化。

（4）可用性。数据挖掘的目标在于发现知识，根据历史数据提取规则，管理和维护规则，并且将挖掘结果用于指导现在的行为和预测未来。因此挖掘的知识必须是可用的。

（5）动态性。数据挖掘出的规则也是随着社会的进步不断变化的，当前的规则只能反映当前的数据特征。由于数据的不断产生和更新，新数据不断加入进来，挖掘规则所用的数据与当前规则反映的情况吻合度会慢慢降低，因此，规则也需要动态更新。

（6）相对性。数据挖掘不是要发现放之四海而皆准的真理，不是要去发现新的自然科学定理和纯数学公式，也不是证明机器定理。它所发现的知识是相对的，是有特定条件约束的，面向特定领域的。

数据挖掘的任务就是从海量的历史数据中，挖掘出潜在的规则、模式和知识。它的任务就决定了它的功能。从大体上看，数据挖掘技术具有两大基本功能，即描述功能和预测功能。描述功能是指数据挖掘可以刻画数据库或数据仓库中数据的一般特性，发现数据间的联系。预测功能是指通过对已知数据的分析处理，在现有数据的基础上，预测未来的数据和某些发展趋势，为决策服务。具体来说，数据挖掘的功能主要包含分类分析、统计分析、概念/特征描述、关联分析、序列模式分析、孤立点分析和演变分析等。

1）分类分析。分类是寻找所描述数据或概念的模型或函数的过程。通过分析能够使用这些模型来预测数据中未知对象所属的类。这些模型基于对训练数据集的分析而得到，可以用多种形式表示，如分类规则、判定树、决策树或神经网络等。

2）统计分析。统计分析可以帮助找出与预测值相关的属性，根据相似数据的分析估算属性值的分布情况，对未来进行预测。

3）概念/特征描述。数据库中通常存放大量的细节数据，概念/特征描述就是用汇总的、简洁的、精确的方式对数据对象的概念和特征进行描述，概括这些数据的整体特征，使用户以简单而准确的方式来观察汇总的数据。这种数据描述可以提供一类数据的宏观概貌，或可将它与其他类相区别。

4）关联分析。关联分析用于发现大量数据中项集之间有意义的关联或相互关系，寻找给定数据集中数据项之间的有趣联系。关联规则的支持度和置信度是两个规则兴趣度的度量标准，它们分别反映发现规则的有用性和确定性。

5）序列模式分析。实时状态数据的存在需要在数据挖掘过程中加入时间因素。序列模式分析主要是通过对历史事件中频繁发生的事件序列进行分析，形成预测模式，来对未来行为进行预测。

6）孤立点分析和演变分析。数据库中可能包含一些数据对象与大部分数据对象的一般行为或模式不一致，这些不一致的数据就称为孤立点。大部分数据挖掘方法将孤立点视为噪声或异常数据丢掉，然而在一些应用中，罕见的事件可能比正常的事件包含更多潜在有用的知识。由此可见，从数据集合中检测这些孤立点并加以分析是十分有意义的。数据演变分析描述行为随时间变化的对象的规律或趋势。它包括趋势分析、相似性查找和周期性模式分析等方面。

1.2 数据挖掘的过程

数据挖掘是一个完整的、反复的人机交互处理过程，该过程需要经历多个相互联系的步骤。而且因为应用领域的分析目标需求不同，以及数据来源和含义的不同，其中的步骤也不会完全一样。一般来说，数据挖掘的过程主要包含五个阶段：① 数据准备。② 数据选择。③ 数据预处

理。④ 数据挖掘。⑤ 转换模型及模式评价。数据挖掘的基本步骤如图 1-4 所示。

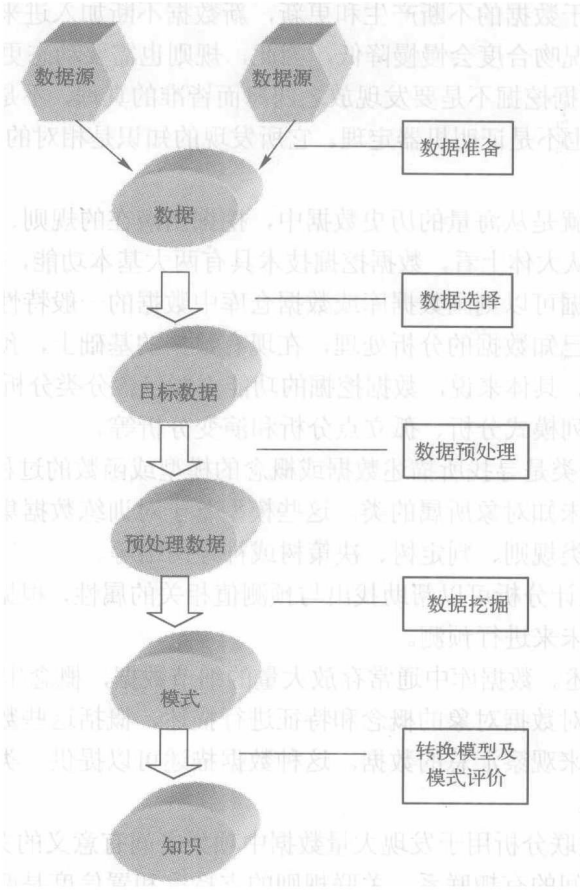

图 1-4 数据挖掘基本步骤

1.2.1 数据准备

数据是数据挖掘工作成功与否的基础，要进行数据挖掘，数据准备阶段必不可少。因为数据挖掘要处理的数据来自不同的数据源，数据量庞大，数据结构复杂，还有大量数据重复、歧义，并且里面夹杂着空缺数据、噪声数据、冗余数据等对数据挖掘有负面影响的数据。因此，数据准备在数据挖掘过程中起着至关重要的作用，是数据挖掘工作的基础。数据准备主要包含以下三个方面：

（1）确定项目目标，制定挖掘计划。数据挖掘的第一步就是要分析即将进行数据挖掘的部门的业务领域，熟悉相关的知识背景，了解业务内容，确定业务对象，从资源配置、技术、经济等方面作出项目的可行性分析和评价。然后根据部门的业务目标，通过与用户或者与用户团体的反复交流，确切了解此次数据挖掘需要处理的任务，并预测可能出现的问题和给出这些问题的解决途径，并制定一个针对数据挖掘结果的评价标准。最后结合部门和数据挖掘项目组的情况，生成一个数据挖掘项目计划。如图 1-5 所示是一个确定项目目标，制定挖掘计划的过程。