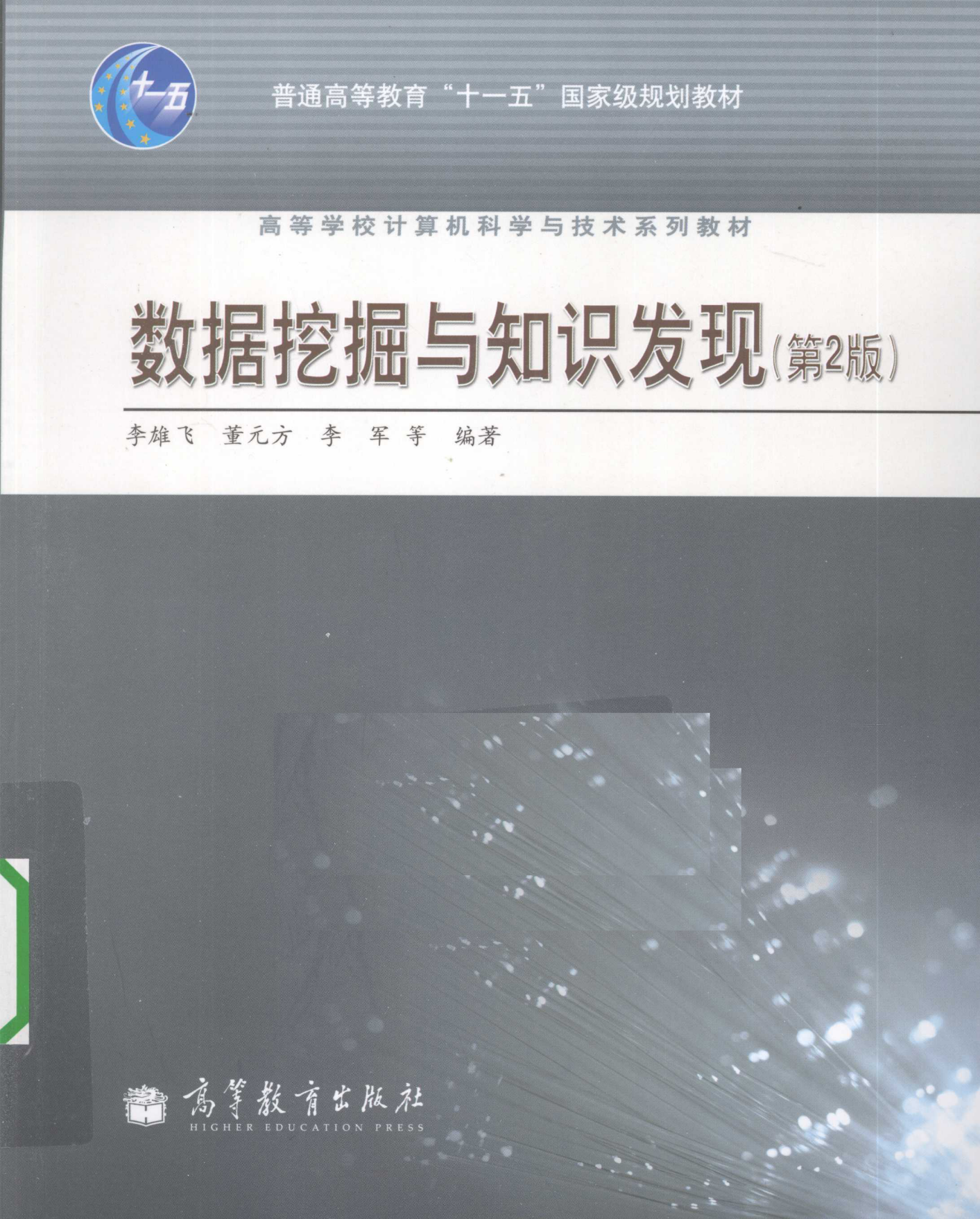
普通高等教育“十一五”国家级规划教材

高等学校计算机科学与技术系列教材

数据挖掘与知识发现(第2版)

李雄飞　董元方　李　军　等　编著



高等教育出版社
HIGHER EDUCATION PRESS

普通高等教育“十一五”国家级规划教材
高等学校计算机科学与技术系列教材

数据挖掘与知识发现

Shuju Wajue yu Zhishi Faxian

（第2版）

李雄飞　董元方　李　军　等　编著

高等教育出版社·北京
HIGHER EDUCATION PRESS　BEIJING

内容提要

本书是普通高等教育“十一五”国家级规划教材。全书共 12 章，第 1 章详尽地阐述了数据挖掘与知识发现领域中的一些基本理论、研究方法和技术标准，简单介绍了相关产品和工具，讨论了 KDD 与数据挖掘的概念、数据挖掘对象、知识发现过程、研究方法以及相关的研究领域和应用范围。第 2 章～第 9 章详细地介绍了关联规则、聚类分析、决策树、贝叶斯网络、人工神经网络、支持向量机、粗糙集、模糊集等数据挖掘模型与算法。第 10 章讨论了模型选择与模型评估方法。第 11 章和第 12 章简单介绍了数据预处理方法和数据挖掘技术标准、数据挖掘可视化技术和数据挖掘工具开发方法，并简单介绍了数据挖掘产品和工具。

本书可以作为计算机专业、信息类专业、管理类专业高年级本科生及研究生的教材或参考书，也可供有关人员学习参考。

图书在版编目（CIP）数据

数据挖掘与知识发现/李雄飞，董元方，李军等编著．—2版．—北京：高等教育出版社，2010.7
ISBN 978-7-04-030478-7

Ⅰ．①数… Ⅱ．①李… ②董… ③李… Ⅲ．①数据采集-高等学校-教材 Ⅳ．①TP274

中国版本图书馆 CIP 数据核字（2010）第 110033 号

出版发行	高等教育出版社	购书热线	010-58581118
社　　址	北京市西城区德外大街 4 号	咨询电话	400-810-0598
邮政编码	100120	网　　址	http:// www.hep.edu.cn
			http:// www.hep.com.cn
		网上订购	http:// www.landraco.com
经　　销	蓝色畅想图书发行有限公司		http:// www.landraco.com.cn
印　　刷	北京东君印刷有限公司	畅想教育	http:// www.widedu.com
		版　　次	2003 年 11 月第 1 版
开　　本	787×1092　1/16		2010 年 7 月第 2 版
印　　张	19.75	印　　次	2010 年 7 月第 1 次印刷
字　　数	430 000	定　　价	32.00 元

物料号　30478-00

第 2 版前言

所谓基于数据库的知识发现（Knowledge Discovery in Database，KDD），是指从大量数据中提取有效的、新颖的、潜在有用的、最终可被理解的模式的非平凡过程。数据挖掘是其中的一个重要步骤。KDD 一词首次出现在 1989 年举行的第 11 届美国人工智能协会（American Association for Artifical Intelligence，AAAI）学术会议上，此后 KDD 的研究逐步成为热点。目前，该领域成果已经应用到人类社会、经济、科技等生活的各个方面，相关的理论、标准和工具日趋成熟。

本书作为教育部普通高等教育“十一五”国家级规划教材，是对 2003 年版教材的全面修订。本次修订更强调理论和实践相结合，并把最新的数据挖掘理论和技术纳入其中。

与第 1 版相比，本书做了如下修改：

1. 大幅度修订了第 2 章和第 8 章

（1）改写第 1 版第 2 章数据预处理和数据仓库，并归入第 11 章数据预处理和可视化技术。缩小了数据仓库篇幅，新增了数据挖掘可视化技术，包括数据可视化、结果可视化和过程可视化等。

（2）把第 1 版第 8 章分类拆分成多章编写，丰富了相关内容。

① 第 4 章决策树：包括信息论基础、ID3 算法和 C4.5 算法等。

② 第 5 章贝叶斯网络：包括贝叶斯概率、贝叶斯学习和贝叶斯网络分类器。

2. 新增了第 7 章、第 10 章和第 12 章

（1）第 7 章支持向量机：包括现行 SVM、非线性 SVM、SVM 的 VC 维等。

（2）第 10 章模型选择与模型评估：包括过拟合问题、分类模型评估、聚类模型评估等。

（3）第 12 章数据挖掘工具与产品：包括数据挖掘标准、数据挖掘开源工具、数据挖掘产品等。

3. 删去了第 1 版第 10 章多媒体数据挖掘

全书贯穿两条主线：一条是从算法理论、技术标准到产品开发，另一条是从数据预处理、算法、模型评估到可视化技术。

本书可用做计算机专业、信息类专业、管理类专业本科及研究生的相关课程教材和教学参考书，也可供有关人员学习参考。教师在教学过程中可根据学时数、专业特点、课程性质等对教学内容做适当取舍。

本书由李雄飞、董元方、李军共同编著，王利民博士修编了第4章和第5章，赵炎同学参加了第12章的编写工作。

中国科学技术大学岳丽华教授审阅了原稿，提出了很多宝贵意见，在此表示衷心的感谢。

作　者

2010年5月于吉林大学

第 1 版前言

计算机技术和通信技术的迅猛发展将人类社会带入到了信息时代。在最近十几年里，数据库中存储的数据量急剧增大。例如，NASA 轨道卫星上的地球观测系统 EOS 每小时会向地面发回 50 GB 的图像数据；世界上最大的数据仓库之一，美国零售商系统 WalMart 每天会产生 2 亿左右的交易数据；人类基因组数据库项目已经搜集了数以 GB 计的人类基因编码数据；大型天文望远镜每年会产生不少于 10 TB 的数据，等等。大量的信息在给人们提供方便的同时也带来一系列问题。由于信息量过大，超出了人们掌握、理解信息的能力，因而给正确运用信息带来了困难。人们意识到隐藏在大规模数据背后的更深层次、更重要的内容能够描述信息的整体特征，可以预测事物的发展趋势。这些潜在信息在决策过程中具有重要的参考价值。为进一步提高信息的利用率，引发了一个新的研究方向：基于数据库的知识发现（Knowledge Discovery in Database）以及相应的数据挖掘（Data Mining）理论和技术的研究。

所谓基于数据库的知识发现是指从大量数据中提取有效的、新颖的、潜在有用的、最终可被理解的模式的非平凡过程。数据挖掘是整个 KDD 过程中的一个重要步骤，运用一些计算法从数据库中提取用户感兴趣的知识。KDD 一词首次出现在 1989 年，随后，很多学者在该领域开展研究工作。目前，关于数据挖掘与知识发现的研究工作已经被众多领域关注，如信息管理、商业、医疗、过程控制、金融等领域。作为大规模数据库中先进的数据分析工具，数据挖掘已经成为数据库及人工智能领域的研究热点之一。

数据挖掘和知识发现是一个涉及多学科的研究领域。数据库技术、人工智能、机器学习、统计学、粗糙集、模糊集、神经网络、模式识别、知识库系统、高性能计算、数据可视化等均与数据挖掘相关。本书全面系统地介绍了数据挖掘和知识发现领域的基本原理和研究方法，可以作为计算机科学与技术专业和信息科学方向高年级本科生和研究生的教材或参考书。第一章介绍了 KDD 与数据挖掘的概念、对象、过程、方法、相关领域和应用范围；第二章介绍了数据预处理和数据仓库技术，包括数据清理、数据约简、数据概念等级划分、多维数据模型等内容；第三章介绍粗糙集；第四章介绍模糊集；第五章介绍聚类分析，包括划分、层次、密度、网格、模型方法和孤立点分析等；第六章是关联规则，介绍关联规则基本模型和一些扩展模型；第七章介绍人工神经网络在知识发现中的运用；第八章是分类与预测，介绍决策树、贝叶斯分类、基于遗传算法的分类，讨论了分类精度和预测问题；第九章介绍了多媒体数据挖掘工作的有关进展。

1997年，吉林大学计算机学院的苑森淼教授建议作者在数据挖掘领域开展工作。几年来，作者在数据挖掘和知识发现领域先后承担了吉林省自然科学基金、国家自然科学基金等科研项目。在与研究生开展的讨论班中逐渐积累了本书的素材。在本书出版之际，向苑老师表示感谢。

特别感谢中国科学院计算技术研究所史忠植研究员，史老师在百忙中审阅了本书初稿，并在篇章总体结构和一些具体细节上给予指导，让作者受益匪浅。

本书由李雄飞、李军编著。宋海玉、李向群、陈鑫影、吴志辉和赵坤等参加了部分编写工作。由于水平有限，书中可能会有不足和遗漏，敬请读者和专家批评指正。

作　者

2003年5月于吉林大学

目　录

第1章 绪 论

1.1 引 言

科技的进步，特别是信息产业的发展，使人类社会步入一个崭新的信息时代。随着计算机应用的普及和数据库技术的不断发展，数据库管理系统的应用领域越来越广泛。条形码及信用卡的普及和使用，进一步加速了商业、金融、保险等领域的信息化进程。人们已经利用计算机取代了绝大部分手工操作。超市中的交易数据，加油站里的油料购买数据，旅行社中的旅行信息数据等均是数据库系统的信息来源。最近十几年中，数据库中存储的数据量急剧增大。例如，美国国家航空和航天局（National Aeronautics and Space Administration，NASA）轨道卫星上的地球观测系统（Earth Observing System，EOS）每小时会向地面发回 50 GB 的图像数据；世界上最大的数据仓库之一，美国零售商系统 WalMart 每天会产生 2 亿左右的交易数据；人类基因组数据库项目已经搜集了数以 GB 计的人类基因编码数据等。如此多领域的数据各自存放在相应的数据库中，致使数据库的规模日益扩大，已经达到数十兆字节，有的甚至更大。与此同时，大容量、高速度、低价格的存储设备也相继问世，管理大量数据的数据库管理系统以及各类数据仓库已经能够支持存储、检索如此规模的数据。但目前数据库系统所能做到的只是对数据库中已有的数据进行存取，通过这些数据获得的信息量仅占整个数据库信息量的一小部分，因为用来对这些数据进行分析处理的工具很少，而且有局限性。在信息时代，大量信息在给人们带来方便的同时，也带来了一系列问题，比如，信息量过大，超过了人们掌握、消化的能力；一些信息真伪难辨，给信息的正确运用带来困难；网络上的信息安全难以保障；信息组织形式的不一致性，增加了对信息进行有效、统一处理的难度等。另一方面，人们意识到，隐藏在这些数据之后的更深层次、更重要的信息能够描述数据的整体特征，可以预测发展趋势，这些信息在决策生成的过程中具有重要的参考价值。面对海量数据库和大量繁杂信息，如何才能从中提取有价值的知识，进一步提高信息的利用率，由此引发了一个新的研究方向：基于数据库的知识发现（Knowledge Discovery in Database，KDD）以及相应的数据挖掘（Data Mining）理论和技术的研究。

KDD 一词首次出现在 1989 年举行的第 11 届美国人工智能协会（American Association for Artifical Intelligence，AAAI）学术会议上，其后，在超大规模数据库（Very Large Database，VLDB）及其他与数据库领域相关的国际学术会议上也举行了 KDD 专题研讨会。1995 年在加拿大蒙特

利尔召开了第 1 届 KDD 国际学术会议（KDD'95），随后每年召开一次这样的会议。由 Kluwer Academic Publisher 出版，1997 年创刊的 Knowledge Discovery and Data Mining 是该领域中的第一本学术刊物。此后，KDD 的研究工作逐步成为热点。

知识发现和数据挖掘领域的研究工作适应市场竞争需要，它将为决策者提供重要的、潜在的信息或知识，从而产生不可估量的效益。目前，关于 KDD 的研究工作已经被众多领域所关注，如过程控制、信息管理、商业、医疗、金融等领域。美国政府开发的 Sequoia 2000 项目把 KDD 列为数据库研究领域中的重要课题之一。作为大规模数据库中先进的数据分析工具，KDD 的研究已经成为数据库及人工智能领域研究的一个热点。

1.2 KDD 与数据挖掘

1.2.1 KDD 定义

人们给 KDD 下过很多定义，内涵也各不相同，目前公认的定义是由美国 Microsoft Research Labs 的 Fayyad 等人提出的。所谓基于数据库的知识发现（KDD），是指从大量数据中提取有效的、新颖的、潜在有用的、最终可被理解的模式的非平凡过程。

数据：指一个有关事实 F 的集合，用以描述事物的基本信息。如学生档案数据库中有关学生基本情况的记录。一般来说这些数据都是准确无误的。

模式：语言 L 中的表达式 E，E 所描述的数据是集合 F 的一个子集 F_E。F_E 表明数据集中的数据具有特性 E。作为一个模式，E 比枚举数据子集 F_E 简单。如“如果分数在 81～90 之间，则成绩优良”可称为一个模式。

非平凡过程：KDD 是由多个步骤构成的处理过程，包括数据预处理、模式提取、知识评估及过程优化。所谓非平凡是指具有一定程度的智能性和自动性，而不仅仅是简单的数值统计和计算。

有效性（可信性）：从数据中发现的模式必须有一定的可信度，函数 C 将表达式映射到度量空间 M_C，c 表示模式 E 的可信度，$c=C(E, F)$。其中 $E \in L$，E 所描述的数据集合 $F_E \subseteq F$。

新颖性：提取出的模式必须是新颖的。模式是否新颖可以通过两个途径来衡量：一是通过当前得到的数据和以前的数据或期望得到的数据之间的比较结果来判断该模式的新颖程度；二是通过对比发现的模式与已有模式的关系来判断。通常用一个函数来表示模式的新颖程度 $N(E, F)$，该函数的返回值是逻辑值或是对模式 E 的新颖程度的一个判断数值。

潜在作用：指提取出的模式将来会实际运用，通过函数 U 把 L 中的表达式映射到度量空间 M_U，u 表示模式 E 的有作用程度，$u=U(E, F)$。

可理解性：发现的模式应该能够被用户理解，以帮助人们更好地了解和使用数据库中的信息，这主要体现在简洁性上。要想让一个模式更易于理解并不是一件很容易的事，需要对其简

单程度进行度量。用 s 表示模式 E 的简单度（可理解度），它也通过函数来反映，即 $s=S(E, F)$。

上述度量函数只是从不同角度进行模式评价，为方便起见，往往采用权值来进行综合评判。在某些 KDD 系统中，利用函数来求得模式 E 的权值 $i=I(E, F, C, N, U, S)$；在另外一些系统中，通过对求得的模式的不同排序来表示模式的权值大小。

综上所述，可以从 KDD 角度给知识下个定义：一个模式 E 对用户设定的阈值 I，如果 $I(E, F, C, N, U, S)>I$，则模式 $E\in L$ 可称为知识。

1.2.2 KDD 过程

KDD 是一个反复迭代的人机交互处理过程。该过程需要经历多个步骤，并且很多决策需要由用户提供。从宏观上看，KDD 过程主要由三个部分组成，即数据整理、数据挖掘和对结果的解释及评估。参见图 1.1 来解释其工作步骤。

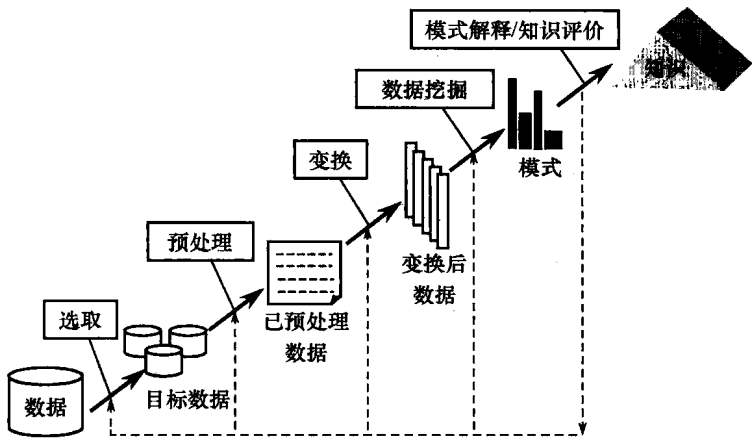


图 1.1 KDD 过程示意图

（1）数据准备：了解 KDD 应用领域的有关情况，包括熟悉相关的背景知识，搞清用户需求。

（2）数据选取：数据选取的目的是确定目标数据，根据用户的需要从原始数据库中选取相关数据或样本。在此过程中，将利用一些数据库操作对数据库进行相关处理。

（3）数据预处理：对步骤（2）中选出的数据进行再处理，检查数据完整性及数据一致性，消除噪声，过滤与数据挖掘无关的冗余数据，根据时间序列和已知的变化情况，利用统计等方法填充丢失的数据。

（4）数据变换：根据知识发现的任务对经过预处理的数据进行再处理，主要是通过投影或利用数据库的其他操作减少数据量。

（5）数据挖掘

① 确定 KDD 目标：根据用户的要求，确定 KDD 要发现的知识类型，因为对 KDD 的不同要求会在具体的知识发现过程中采用不同的知识发现算法，如分类、总结、关联规则、聚类等。

② 选择算法：根据步骤（5）确定的任务选择合适的知识发现算法，包括选取合适的模型和参数。同样的目标可以选用不同的算法来解决，这可以根据具体情况进行分析选择。有两种选择算法的途径，一是根据数据的特点不同，选择与之相关的算法；二是根据用户的要求，有的用户希望得到描述型的结果，有的用户希望得到预测准确度尽可能高的结果，不能一概而论。总之，要做到选择算法与整个 KDD 过程的评判标准相一致。

③ 数据挖掘：这是整个 KDD 过程中很重要的一个步骤。运用前面选择的算法，从数据库中提取用户感兴趣的知识，并以一定的方式表示出来（如产生式规则等）是数据挖掘的目的。

（6）模式解释：对在数据挖掘步骤中发现的模式（知识）进行解释。经过用户或机器评估后，可能会发现这些模式中存在冗余或无关的模式，此时应该将其剔除。如果模式不能满足用户的要求，就需要返回到前面的某些处理步骤中反复提取。例如，重新选取数据、采用新的数据变换方法、修改数据挖掘算法的某些参数值，甚至换另外一种挖掘算法，从而提取出更有效的模式。

（7）知识评价：将发现的知识以用户能理解的方式呈现给用户。这期间也包含对知识一致性的检查，以确信本次发现的知识不会与以前发现的知识相抵触。由于挖掘出来的知识最终是呈现给用户的，所以，应该以用户能够理解的最直观的方式作为最终结果。因此，知识发现工作还包括对模式进行可视化处理等。

在上述步骤中，数据挖掘占据非常重要的地位，它主要是利用某些特定的知识发现算法，在一定的运算效率范围内，从数据中发现出有关知识，决定了整个 KDD 过程的效果与效率。

1.2.3 数据库技术发展与数据挖掘

数据挖掘是 KDD 过程中的一个重要步骤，其中包括特定的数据挖掘算法。算法能在可接受的计算效率下，在 F 上产生一系列模式 E_i。有些文献中将 KDD 与数据挖掘混用。数据挖掘是在数据库技术中发展起来的，图 1.2 列举了数据库技术的演化历程。

20 世纪 60 年代，数据库和信息技术已经从原始的文件处理系统发展成为精密复杂、功能强大的数据库系统，这时的数据库系统是基于层次模型或网状模型的。到了 20 世纪 70 年代，关系数据库系统、数据建模工具、索引技术和数据组织技术取得了实质性进步。同时，用户通过查询语言、用户接口、优化查询进程和事务管理可以方便灵活地存取数据。以联机事务处理（On-Line Transaction Processing，OLTP）作为有效的模式，从根本上确立了关系数据库在数据存储、检索、管理大量数据中的主导地位。

进入 20 世纪 80 年代中期，一方面是关系数据库的黄金时期，另一方面对新型的强大的数据库系统的开发和研究也十分活跃。关系扩展、面向对象、面向关系、演绎模型等数据模型取得了相应进展。空间数据库、时态数据库、多媒体数据库、主动数据库、科学数据库、知识库和办公数据库等面向应用的数据库系统不断发展繁荣。分布式技术、多样数据、共享数据也得到了广泛深入的研究。异构数据库和基于 Internet 的全球信息系统在信息产业中占据重要地位。数据仓库也是这一时期的产物，它根据多维数据结构进行建模，包括数据清理、数据集成和联

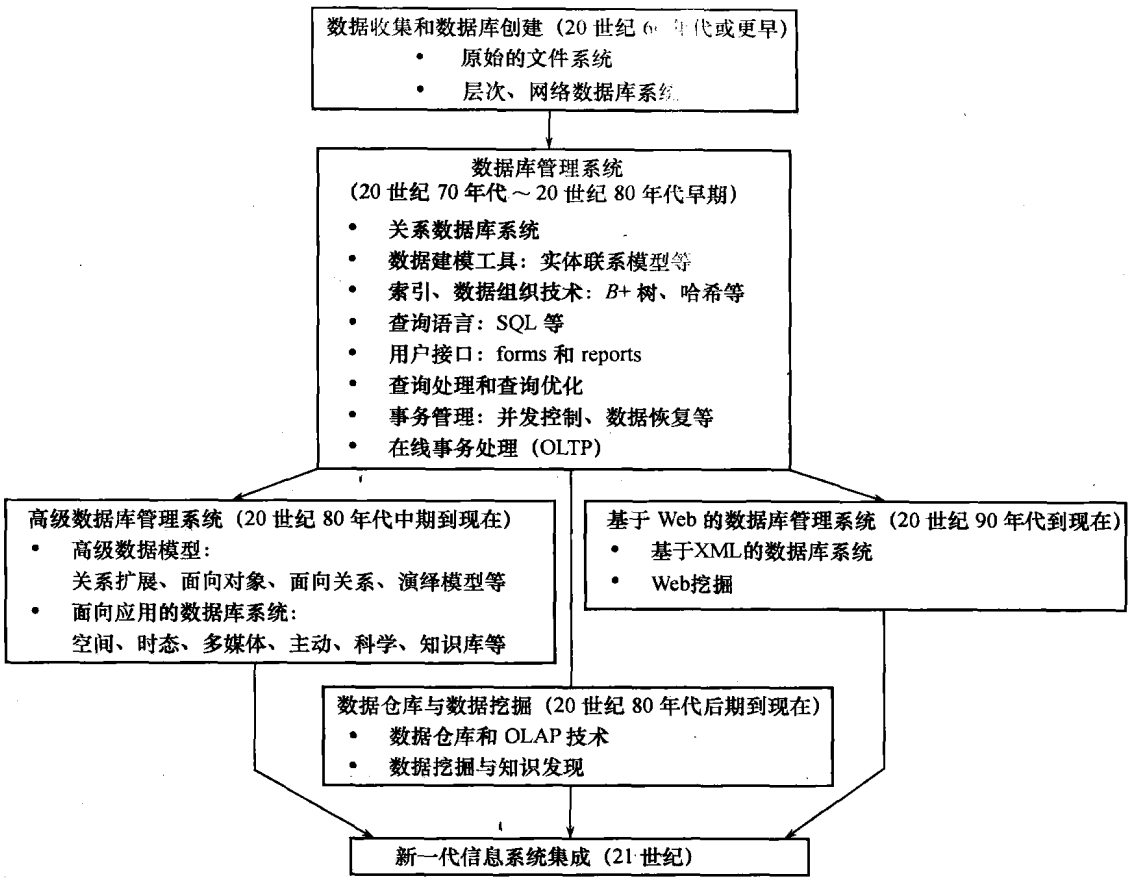


图 1.2 数据库技术的演化历程

机分析处理（On-Line Analytical Processing，OLAP）等。

OLAP 具有一定的多视角观察、分析、检索数据的能力，可以支持多维分析和决策，但仍然需要更深层次的分析。随着数据库技术的广泛应用，海量数据层出不穷，在给人们带来方便的同时也带来了一系列问题。面对海量数据库和大量繁杂信息，如何才能从中提取有价值的知识，进一步提高信息利用率，自然就把数据挖掘技术推上了历史舞台。

1.3 数据挖掘的对象与环境

1.3.1 数据与系统特征

KDD 和数据挖掘可以应用在很多领域中，它们具有如下一些公共特征：