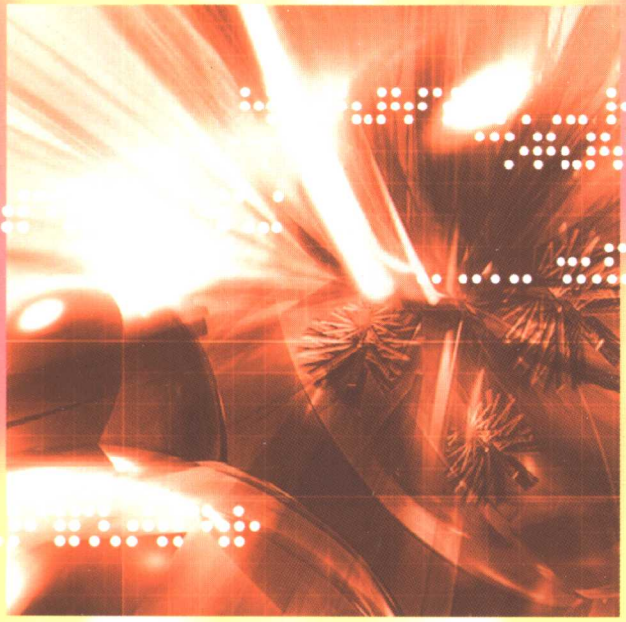
高等学校研究生系列教材

数据挖掘与知识发现

Data Mining and Knowledge Discovery

李雄飞　李军　编著



高等教育出版社
HIGHER EDUCATION PRESS

高等学校研究生系列教材

数据挖掘与知识发现

李雄飞　李　军　编著

高等教育出版社

内 容 提 要

本书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和研究方法。介绍了 KDD 与数据挖掘的概念、数据挖掘对象、知识发现过程、研究方法以及相关研究领域和应用范围。作为知识发现的数据预处理工作，简要叙述了数据清理、数据约简、数据概念等级分层、多维数据模型等内容。书中较详细地介绍了粗糙集、模糊集、聚类分析、关联规则、人工神经网络、分类与预测等数据挖掘方法，最后还简要介绍了多媒体数据挖掘工作的有关进展。

本书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，也可供有关科技人员学习参考。

图书在版编目（CIP）数据

数据挖掘与知识发现/李雄飞，李军编著. —北京：
高等教育出版社，2003.11
ISBN 7-04-013308-3

Ⅰ.数… Ⅱ.①李…②李… Ⅲ.①数据采集-高等学校-教材②知识信息处理-高等学校-教材
Ⅳ.①TP274②TP391

中国版本图书馆 CIP 数据核字（2003）第 099538 号

出版发行	高等教育出版社	购书热线	010-64054588
社　　址	北京市西城区德外大街 4 号	免费咨询	800-810-0598
邮政编码	100011	网　　址	http://www.hep.edu.cn
总　　机	010-82028899		http://www.hep.com.cn
经　　销	新华书店北京发行所		
印　　刷	中国农业出版社印刷厂		
开　　本	787×1092　1/16	版　　次	2003 年 11 月第 1 版
印　　张	15	印　　次	2003 年 11 月第 1 版
字　　数	320 000	定　　价	20.60 元

本书如有缺页、倒页、脱页等质量问题，请与当地图书销售部门联系调换。

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 50GB 的图像数据；世界上最大的数据仓库之一，美国零售商系统 Wal-Mart 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，简称 KDD），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、数据概念等级化分、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。

编　著

2003 年 5 月于吉林大学

目　录

第一章 绪 论

1.1 引 言

科技的进步，特别是信息产业的发展，把我们带入了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。条形码和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。超级市场中的交易数据，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，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 Wal-Mart 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）及相应的数据挖掘（Data Mining）理论和技术的研究。

基于数据库的知识发现（KDD）一词首次出现在 1989 年举行的第十一届 AAAI 学术会议上，其后，在 VLDB（Very Large Database）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一届 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 Publishers 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是适应市场竞争的需要的，它将为决策者提供重要的、前所未料的信息或知识，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 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 F 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 到 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，而绝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，通过函数 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：一是通过比较当前得到的数据和以前的数据或期望得到的数据来判断；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被实际运用，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。要想让一个模式更容易地被人们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），$s=S(E, F)$。

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现模式的权值大小。

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、数据挖掘和结果的解释评估。参见图 1.1 来解释其工作步骤。

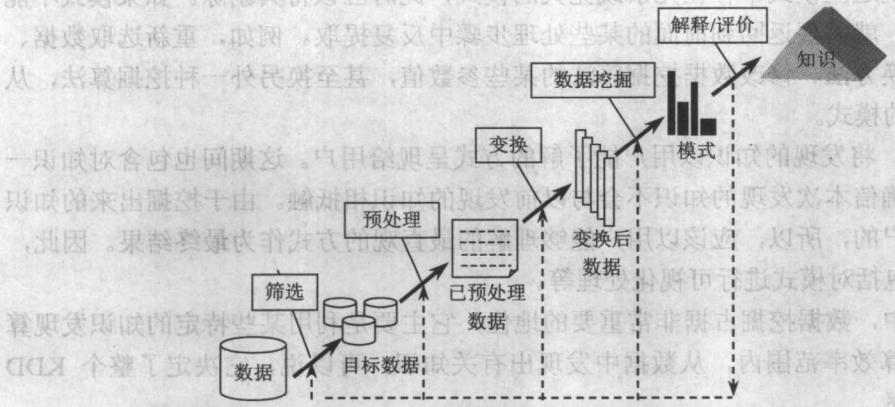


图 1.1 KDD 过程示意图

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，检查数据的完整性及数据一致性，消除噪声，滤除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. 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。

6. 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。

7. 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。

8. 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。

9. 知识评价：将发现的知识以用户能了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，可以说，它决定了整个 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 显示了数据库技术的发展历程。

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 70 年代，关系数据库系统、数据模型工具、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。以在线事务处理（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 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。扩展关系、面向对象、关系对象、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、数据集成和在线分析处理（OLAP）等。

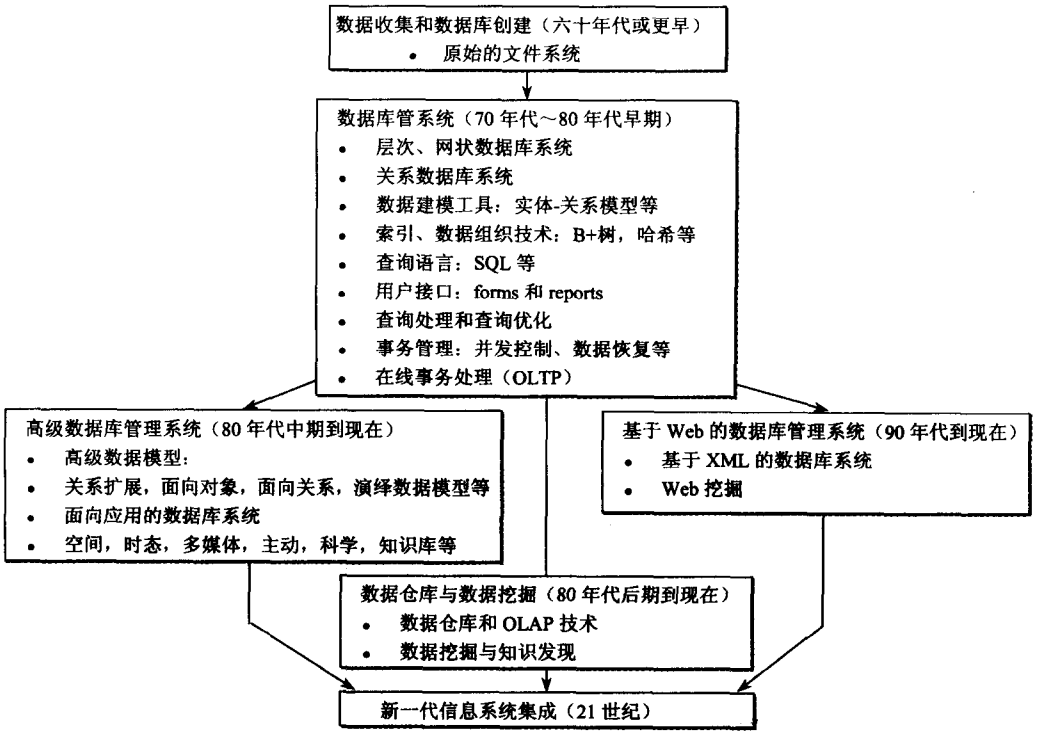


图 1.2 数据库技术的演化历程

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，大量信息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，自然就使数据挖掘技术成为深受关注的问题。

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不同的领域，而这些领域的数据与系统还是具有如下一些公共特征：

1. 海量数据集。
2. 数据利用非常不足。
3. 在开发知识发现系统时，领域专家对该领域的熟悉程度至关重要。
4. 最终用户专门知识缺乏。

为使知识发现系统更加有效，有几个软、硬件问题需要强调：第一，为使数据服务更加详尽，必须研究基础的体系结构、算法和数据结构；第二，解决存储管理中的新问题，开发有效的存储机制；第三，高层次的查询语言成为重要的研究课题；第四，描述多维对象的可视化工具在知识表示中将起重要作用。

1.3.2　数据结构

数据库中的数据可以采用多种形式，通常情况下，数字实体为第一类别，符号实体是第二类别。在第一类别中的数值量指数字、向量、2 维矩阵或多维数组等。符号实体用来描述定性的量（如黑暗、明亮等）。更进一步，描述某些概念等级时就会面对复合数据类型。图 1.3 就是这些数据形式的例子。

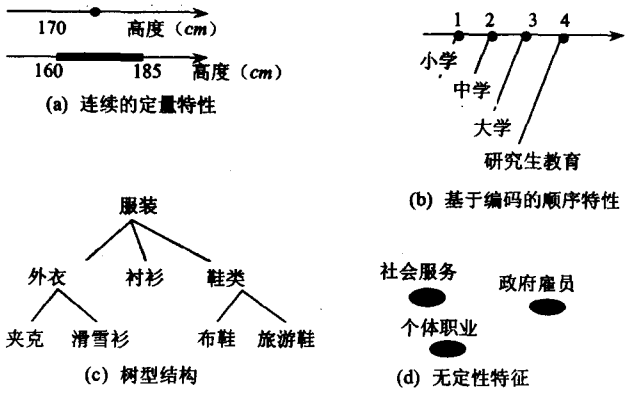


图 1.3　数据类型示例

重要的问题是在知识发现的观点上如何操作这些数据。某种情况下，将注意力放在比

较两部分数据上（计算它们之间的差距）；有时则更关注对象间的等级差异等。人们对数据的理解非常有限，因此要对数据进行抽象。比如，看气象图时，并不注意个别的温度，而是注意哪些区域气温高，哪些区域气温低。也就是说信息颗粒化（Granularity），将其描述成更高的抽象形式（集合）。通过信息颗粒化可以把大量数据压缩成单一的概念实体。集合压缩了元素，数据间隔包含了多个个体数据。比如描述汽车油耗，从不同的角度出发会涉及到不同的理论和方法。图 1.4 是关于汽车油耗的几种信息聚合模型。

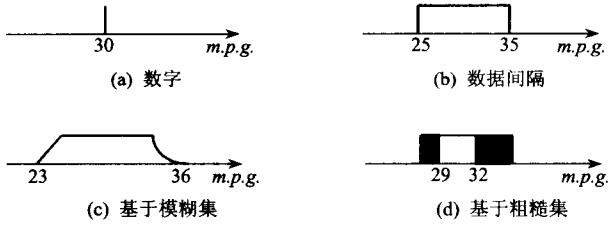


图 1.4 从不同角度出发的数据聚合

首先是用单一的数字描述，每加仑 30 英里。也可以用数字区间[25, 35]表示每加仑汽油可以行驶 25～35 英里。接下来改变边界的二值属性（“是”与“非”），引入模糊集和粗糙集的观点。单一的数字是最高级别的聚合，数字区间比模糊集和粗糙集的聚合级别要低，最低的聚合级别就是整个数据空间。数据挖掘是在整个数据库上发现聚合级别较高的知识。

1.3.3 数据库系统

数据挖掘的对象原则上可以是各种存储方式的信息。目前的信息存储方式主要包括关系数据库、数据仓库、事务数据库、高级数据库系统、文件数据和 Web 数据。其中，高级数据库系统包括面向对象数据库、关系对象数据库，以及面向应用的数据库（如空间数据库、时态数据库、文本数据库、多媒体数据库等）。

1. 关系数据库

一个数据库系统也称为数据库管理系统（DBMS），由一些相关数据组成，并通过软件程序管理和存储这些数据。DBMS 提供数据库结构定义，数据检索语言（SQL 等），数据存储，并发、共享和分布式机制，数据访问授权等功能。关系数据库由表组成，每个表有一个惟一的表名。属性（列或域）集合组成表结构，表中数据按行存放，每一行称为一个记录。记录间通过键值加以区别。关系表中的一些属性域描述了表间的联系，这种语义模型就是实体关系（ER）模型。关系数据库是目前最流行、最常见的数据库之一，为数据挖掘研究工作提供了丰富的数据源。

2. 数据仓库

数据仓库可以把来自不同数据源的信息以同一模式保存在同一个物理地点。其构成需要经历数据清洗、数据格式转换、数据集成、数据载入及阶段性更新等过程。严格地讲，数据仓库是面向问题的、集成的、随时间变化的、相对稳定的数据集，为管理决策提供支持。面向问题是指数据仓库的组织围绕一定的主题，不同于日复一日的操作和事务处理型的组织，而是通过排斥对决策无用的数据等手段提供围绕主题的简明观点。集成性是指数据仓库将多种异质数据源集成为一体，如关系数据库、文件数据、在线事务记录等。数据存储包含历史信息（比如，过去的 5～10 年）。数据仓库要将分散在各个具体应用环境中的数据转换后才能使用，所以，它不需要事务处理、数据恢复、并发控制等机制。

数据仓库根据多维数据库结构建模，每一维代表一个属性集，每个单元存放一个属性值，并提供多维数据视图，允许通过预计算快速地对数据进行总结。尽管数据仓库中集成了很多数据分析工具，但仍然需要像数据挖掘等更深层次、自动的数据分析工具。

3. 事务数据库

一个事务数据库由文件构成，每条记录代表一个事务。典型的事务包含惟一的事务标识（*trans_ID*），多个项目组成一个事务。事务数据库可以用额外附加的关联表记录其他信息，比如，在销售方面，事务交易日期、顾客 *ID* 及交易发生的部门等。更深层次的货篮数据（Market Basket）分析（如哪些商品经常同时销售等问题）只能利用数据挖掘思想来解决。

4. 面向对象数据库

面向对象数据库是基于面向对象程序设计的范例，其每一个实体作为一个对象。与对象相关的程序和数据封装在一个单元中，通常用一组变量描述对象，等价于实体关系模型和关系模型中的属性。对象通过消息与其他对象或数据库系统进行通信。对象机制提供一种模式获取消息并做出反应的手段。类是对象共享特征的抽象。对象是类的实例，也是基本运行实体。可以把对象类按级别分为类和子类，实现对象间属性共享。

5. 关系对象数据库

关系对象数据库的构成基于关系对象模型。为操作复杂的对象，该模型通过提供丰富数据类型的方法进一步扩展了关系模型。在关系查询语言中增加了新增类型的检索能力。关系对象数据库在工业和其他应用领域使用越来越普遍。与关系数据库上的数据挖掘相比，关系对象数据库上的数据挖掘更强调操作复杂的对象结构和复杂数据类型。

6. 空间数据库

空间数据库包含空间关系信息，比如，地理（地图）数据库、VLSI 芯片设计数据库、医学图像数据库和卫星图像数据库等。空间数据可以用包括 n 维位图、象素图等光栅格式表示（比如，2 维气象卫星图像数据可以用光栅格式表示，每一个象素记录一个降雨区域），也可以用向量形式表示（比如，道路、桥梁、建筑物等基本地理结构可以用点、线、多边形等几何图形表示为向量格式）。数据挖掘可以揭示地理数据中某种类型区域中的建筑物特征（比如，湖边建筑物特征等）。所以，对空间数据库的数据挖掘工作具有重要意义。

7. 时态数据库和时间序列数据库

这两种数据库均存储与时间有关的信息。时态数据库通常存储与时间属性相关的数据，这些属性可以是具有不同语义的时间戳。时间序列数据库存储随时间顺序变化的数据，比如股市中的变化数据等。数据挖掘技术可以用于发现对象演变特性或数据库中数据的变化趋势。时间可以是财政年、教学年、日历年等，也可以是年细分的季度或月。

8. 文本数据库

文本数据库是包含用文字描述的对象的数据库。这里的文字不是通常所说的简单的关键字，可能是长句子或图形，比如产品说明书、出错或调试报告、警告信息、简报等文档信息。文本数据库可以是无结构的（比如某些 WWW 网页），也可以是半结构的（比如一些邮件信息，HTML/XML 网页）。数据挖掘可以揭示对象类的通常描述，如关键字与文本内容之间的关联，基于文本对象的聚类等。

9. 多媒体数据库

在多媒体数据库中存储图像、音频、视频等数据。多媒体数据库管理系统提供在多媒体数据库中对多媒体数据进行存储、操纵和检索的功能，特别强调多种数据类型间（比如图像、声音等）的同步和实时处理。主要应用在基于图片内容的检索、语音邮件系统、视频点播系统。多媒体数据库挖掘、存储和检索技术需要集成标准的数据挖掘方法，还要构建多媒体数据立方体，运用基于模式相似匹配的理论等。

10. 异构数据库和遗产数据库

异构数据库由一组互连的自治的成员数据库组成。这些成员相互通信，以便交换信息和回答查询。一个成员数据库中的对象可以与其他成员数据库中的对象有很大差别。将它们的语义同化到整个异构数据库中十分困难。很多企业通过信息技术开发的长期历史（包括运用不同的硬件和操作系统）获得遗产数据库（Legacy Database）。遗产数据库是一组异构数据库，包括关系数据库、对象数据库、层次数据库、网状数据库、多媒体数据库、