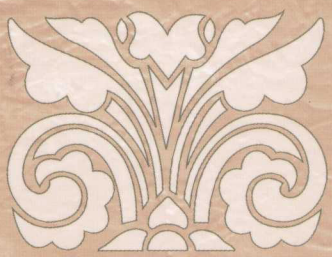
“十二五”普通高等教育本科国家级规划教材

高等院校信息管理与信息系统专业系列教材

数据挖掘技术与应用
（第2版）

陈 燕 编著



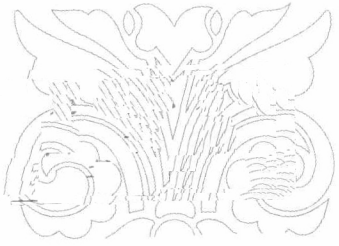
清华大学出版社

“十二五”普通高等教育本科国家级规划教材

高等院校信息管理与信息系统专业系列教材

数据挖掘技术与应用
（第2版）

陈燕　编著



清华大学出版社
北京

内 容 简 介

本书系统地阐述了数据挖掘产生的背景、技术、多种相关方法及具体应用，主要内容包括数据挖掘概述，数据采集、集成与预处理技术，多维数据分析与组织，预测模型研究与应用，关联规则模型及应用，聚类分析方法与应用，粗糙集方法与应用，遗传算法与应用，基于模糊理论的模型与应用，灰色系统理论与方法，基于数据挖掘的知识推理。

本书可作为管理科学与工程、信息科学与技术、应用数学等相关专业高年级本科生和研究生的数据仓库、数据挖掘及知识管理等相关课程的教材或参考资料，也可用来帮助相关的专业研究人员提升数据挖掘的技巧和开拓新的研究方向。

图书在版编目(CIP)数据

数据挖掘技术与应用 / 陈燕编著. --2 版. --北京：清华大学出版社，2016
高等院校信息管理与信息系统专业系列教材
ISBN 978-7-302-43249-4

Ⅰ. ①数… Ⅱ. ①陈… 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(2016)第 041529 号

责任编辑：白立军　徐跃进
封面设计：傅瑞学
责任校对：焦丽丽
责任印制：何　芊

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn,010-62795954
印 装 者：北京鑫海金澳胶印有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：16.25　　**字　　数**：383 千字
版　　次：2011 年 5 月第 1 版　2016 年 8 月第 2 版　　**印　　次**：2016 年 8 月第1 次印刷
印　　数：1～2000
定　　价：35.00 元

产品编号：068180-01

前　言

随着计算机应用技术和网络技术的普及，全社会的信息化程度不断提高，新的管理模式不断涌现，对信息系统的依赖程度越来越高。信息管理工程研究者和管理者面临严峻挑战：如何从海量、分散、复杂类型的数据海洋中，迅速找出有价值的和潜在有用的信息与知识？如何实现对多维数据的集中组织、分析与管理？数据仓库与数据挖掘可以为上述问题提供有效的解决方案。数据挖掘理论及方法研究与创新已经成为信息科学与管理工程领域最重要的研究方向之一。

笔者在数据仓库技术与数据挖掘模型方面潜心研究数十年。尤其近年来，通过国家自然科学基金(项目编号 71271034)，教育部、科技部和交通运输部，省市多个科研项目的资助，深入研究了数据挖掘的理论、技术与方法，获得多项科研成果。特别是面向交通运输、物流管理等特色领域，开展基于数据仓库与数据挖掘的创新性研究，取得了良好的社会效益与经济效益。

撰写本书的目的在于：利用数据仓库技术将异构的、多维的、具有复杂类型的多源数据整合到一个公共平台上进行统一组织与管理，在此基础上，采用多种数据挖掘方法与模型，实现从底层信息管理到高层知识管理全过程的信息深加工、挖掘与增值。

本书采用逐步演算和编程运行相结合的方式，力争使广大读者通过本书的学习能够快速掌握数据挖掘模型的理论、技术、方法及应用。全书共分为 11 章，包括数据挖掘概述，数据采集、集成与预处理技术，多维数据分析与组织，预测模型研究与应用，关联规则模型及应用，聚类分析方法与应用，粗糙集方法与应用，遗传算法与应用，基于模糊理论的模型与应用，灰色系统理论与方法，基于数据挖掘的知识推理。

本书主要由陈燕编写，屈莉莉、杨明、张琳、乔月英、吉飞、赵路、程澄、于莹莹、林博辞等参与完成部分章节中具体数据挖掘方法的应用算例和全书的核对工作。

本书自 2011 年出版以后，受到广大师生欢迎，此次再版，吸收了许多有益的建议，根据数据挖掘技术的发展，在保留第 1 版框架的基础上，对部分内容进行了修改、整理，希望广大师生一如既往地关注和喜欢本书。

本书旨在涵盖典型和有代表性的数据挖掘算法，但由于数据挖掘方法多种多样，还有许多数据挖掘模型需要进一步探讨。在编写过程中，笔者查阅了国内外大量文献资料，谨向书中提到的和参考文献中列出的学者表示感谢。如果由于我们工作的疏忽，致使本书中某处内容所参考的文献没有列出，在此向所涉及的作者深表歉意。同时，由于时间仓促和编者能力有限，书中难免存在一些不当之处，敬请广大读者批评指正。

陈　燕

2016 年 4 月

目　录

第1章 数据挖掘概述

本章阐述数据仓库和数据挖掘内涵并深入分析数据仓库、数据挖掘和联机分析处理三种技术之间的关系，给出了数据仓库系统的通用模式；提出了一种新颖的数据仓库系统中多维数据组织的形式化定义与描述方法；从数据挖掘系统的发展阶段、系统结构、相关技术、实现工具和应用领域等多个方面，概述了数据挖掘的理论、技术与方法。

1.1 数据仓库和数据挖掘定义与解释

1.1.1 数据仓库的定义与解释

数据仓库(Data Warehouse，DW)属于一种高层管理的新型数据库技术。将分散在诸多数据库系统(DataBase System，DBS)中的数据安全、平稳、有效地集成到一个公共信息平台模式下，这是数据仓库建立的基础。也就是说，在DBS趋于完善化的今天，其技术进一步发展的趋势是：建立基于DBS基础之上的DW，以实现DBS之上的高层管理、智能管理和知识管理，即实现数据挖掘与高层管理决策分析的最终目标。

数据仓库概念的提出者及相关技术的主要倡导者——美国著名信息工程学家Willian Inmon博士对数据仓库的解释是：数据仓库通常是一个面向主题的、集成的、相对稳定的、反映历史变化的数据的集合，用于支持经营管理中的决策制定过程。所谓面向主题，是指操作型数据库的数据组织面向事务处理任务，各个业务系统之间各自分离，而数据仓库中的数据是按照一定的主题域进行组织的。所谓集成，是指数据仓库中的数据是在对原有分散的数据库数据进行抽取、清理的基础上经过系统加工、汇总和整理得到的，必须消除源数据中的不一致性，以保证数据仓库内的信息是关于整个企业的一致的全局信息。

所谓相对稳定，是指数据仓库的数据主要供企业决策分析之用，所涉及的数据操作主要是数据查询，一旦某个数据进入数据仓库，一般情况下将被长期保留，也就是数据仓库中一般有大量的查询操作，但修改和删除操作很少，通常只需要定期地加载和刷新。所谓反映历史变化，是指数据仓库中的数据通常包含历史信息，系统记录了企业从过去某一时刻(如开始应用数据仓库的时刻)到目前的各个阶段的信息，通过这些信息，可以对企业的发展历程和未来趋势做出定量分析和预测。由于数据仓库涉及多元、多维的复杂数据，数据时间跨度大等多种特点，因此数据仓库是一个对多维异构数据一体化组织与管理的复杂过程。

1.1.2 数据挖掘的定义与解释

随着信息技术的发展与普及，大量的数据与信息的积累，如何从海量的数据中提取

有用的和有价值的信息，即知识，已成为信息技术研究的重要问题，数据挖掘技术应运而生。20 世纪 90 年代，以美国信息工程领域专家为代表，开始研究数据挖掘的理论与方法。

数据挖掘(Data Mining，DM)的概念最早是在 1995 年的美国计算机年会(ACM)上提出的，数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。

另一种比较公认的定义是 W. J. Frawley 和 G. Piatetsky-Shapiro 等人提出的，数据挖掘就是从大型数据库中的数据中提取人们感兴趣的知识。这些知识是隐含的、事先未知的、潜在的、有用的信息，提取的知识表示为概念(Concepts)、规则(Rules)、规律(Regulations)、模式(Patterns)等形式，后来专家们将这些形式的知识表达模式运用形式化定义来描述。

数据挖掘的一个重要过程就是从数据中挖掘知识，也称为数据库中知识发现(Knowledge Discovery in Databases，KDD)和知识提取、数据采掘等，并且可以在其过程中用于发现概念/类描述、分类、关联、预测、聚类、趋势分析、偏差分析和相似性分析及结果的可视化。

因此，可以将数据挖掘理解为：在庞大的数据库中寻找出有价值的隐藏事件，并利用人工智能、统计、预测的科学技术，将其数据进行科学有价值的提取和深入分析，找出其中的知识，并根据企业发展中的需求问题建立不同的挖掘模型，以此作为提供企业进行决策分析时的参考依据。

人们把原始数据视为形成知识的源泉，就像从矿石中采矿一样。原始数据可以是结构化的，如关系型数据库中的数据，也可以是半结构化的，如文本、图形、图像数据，甚至是分布在网络上的异构数据。发现知识的方法可以是数学的，也可以是非数学的；可以是演绎的，也可以是归纳的。发现了的知识可以用于信息管理、查询优化、决策支持、过程控制等，还可以用于数据自身的维护。数据挖掘的主要目标是：在众多复杂类型数据中找出“金块”，能在商务(企业)数据中找出提高销售量和效益的关键因素，并且也能通过数据挖掘找出影响企业效益增长的相关因素。因此，数据挖掘是一门广义的交叉学科，它汇聚了不同领域的研究者，尤其是数据库、人工智能、数理统计、可视化、并行计算等方面的学者和工程技术人员。

数据挖掘的概念随着其发展而不断得到充实，美国的一项研究报告将 DM 视为 21 世纪十大明星产业之一。数据挖掘已成为当今知识管理、商业智能领域最热门的话题之一。越来越多的企业通过对数据挖掘概念和技术的了解与应用，达到解决信息工程领域关键技术难题的目的。

数据挖掘的用途非常广泛。它可以应用在生产任务的预测与分析、生产效益的评估与分析、销售领域的预测分析、物流企业的货源预测与分析、交通肇事逃逸案的分析、超市的物品摆放、银行的贷款预测与决策分析、服装领域的职业服装号型归档、大型数据库的关联知识挖掘、企业绩效评估与分析等相关的领域中；也可以应用在更细致的研究中，比如：在金融行业出现的基于数据仓库贷款决策分析，可以将其银行和信用卡公司通过 DM 产品的相

关技术将庞大的顾客资料做筛选、分析、推演及预测，找出哪些是最有贡献的顾客，哪些是高流失率族群，或找出一个新的产品或促销活动可能带来的响应率，如何在合适的时间提供适当的产品及服务等挖掘功能。

数据挖掘技术从一开始就是面向应用的。它不仅是面向特定数据库的简单检索查询调用，而且要对这些数据进行微观、中观乃至宏观的统计、分析、综合和推理，以指导实际问题的求解，试图发现事件间的相互关联，甚至利用已有的数据对未来的活动进行预测。这样一来，就把人们对数据的应用，从低层次的末端查询操作，提高到为各级经营决策者提供决策支持。这种需求驱动力，比数据库查询更为强大。同时需要指出的是，这里所说的知识发现，不是要求发现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，所有发现的知识都是相对的，是有特定前提和约束条件、面向特定领域的，同时还要能够易于被用户理解，最好能用自然语言表达所发现的结果。

1.2 数据仓库系统的相关技术

数据仓库系统中主要包括数据仓库、数据挖掘、联机分析处理(On-Line Analysis Processing，OLAP)、KDD 和相关的数据集成、数据标准化、数据仓库建模技术、数据挖掘技术与方法、数据集市、可视化技术、自然语言解释、人机交互、知识发现与知识推理、网络集成技术等研究内容。

1.2.1 数据仓库系统相关技术之间的关系

1. 数据仓库与数据挖掘

数据仓库与数据挖掘作为决策支持新技术，近十年来发展迅速。数据仓库和数据挖掘二者相互结合共同发展，又相互影响促进，两者的联系概括如下：

数据挖掘(DM)和数据仓库(DW)是融合与互动发展的。对于数据挖掘，如果能同数据仓库协同工作，则可以简化数据挖掘过程的某些步骤，从而极大地提高数据挖掘的工作效率。数据仓库中的数据是经过预处理的，它清洗了原始数据中的不规范数据，统一了数据格式并做了一些必要的汇总，数据挖掘只需在此基础之上再做进一步的预处理。数据挖掘和数据仓库的协同工作，是数据挖掘专家、数据仓库技术人员和行业专家共同努力的成果，更是广大渴望从数据库“奴隶”到数据库“主人”转变的企业最终用户的通途。一方面，可以迎合和简化数据挖掘过程中的重要步骤，提高数据挖掘的效率和能力，确保数据挖掘中数据来源的广泛性和完整性；另一方面，数据挖掘技术已经成为数据仓库应用中极为重要和相对独立的方面和工具。若将数据仓库比作矿坑，DM 就是深入矿坑采矿的工作。毕竟 DM 不是一种无中生有的魔术，也不是点石成金的炼金术，若没有足够丰富完整的数据，是很难期待 DM 能挖掘出什么有意义的信息。要将庞大的数据转换成为有用的信息，必须先有效率地收集信息。随着科技的进步，功能完善的数据库系统就成了最好的收集数据的工具。数据仓库，简单地说，就是搜集来自其他系统的有用

数据存放在一个整合的存储区内。其实就是一个经过处理整合，且容量特别大的关系型数据库，用于存储决策支持系统(Decision Support System，DSS)所需要的数据，供决策支持或数据分析使用。从信息技术的角度来看，数据仓库的目标是在组织中，在正确的时间，将正确的数据交给正确的人。

数据挖掘和数据仓库的目的和过程不同。许多人对于 DW 和 DM 时常混淆，不知如何分辨。其实，数据仓库是数据库技术的一个新主题，利用计算机系统帮助我们操作、计算和思考，让作业方式改变，决策方式也跟着改变。数据仓库本身是一个非常大的数据库，它存储着由组织作业数据库中整合而来的数据，特别是由事务处理系统(On-Line Transaction Processing，OLTP)所得来的数据。将这些整合过的数据置放于数据仓库中，决策者则可以利用这些数据作决策；但是，这个转换及整合数据的过程，是建立一个数据仓库最大的挑战。因为将作业中的数据转换成有用的策略性信息是整个数据仓库的重点。综上所述，数据仓库应该具有这些数据：整合性数据(Integrated Data)、详细和汇总性的数据(Detailed and Summarized Data)、历史数据、解释数据的数据。从数据仓库挖掘出对决策有用的信息与知识，是建立数据仓库与使用数据挖掘的最大目的，两者的本质与过程不同。换句话说，数据仓库应先行建立完成，数据挖掘才能有效率地进行，因为数据仓库本身所含数据是干净(不会有错误的数据掺杂其中)、完备且经过整合的，因此两者关系可解读为数据挖掘是从数据仓库中找出有用信息的一种过程与技术。

一方面，数据仓库为数据挖掘提供了更好更广泛的数据源。数据仓库中集成和存储着来自异质信息源的数据，而这些信息源本身就可能是一个规模庞大的数据库。同时数据仓库存储了大量的、长时间的历史数据，可以用来进行数据的长期趋势分析，为决策者的长期决策行为提供支持。数据仓库中数据在时间轴上的纵深性是数据挖掘不能回避的难点问题之一。数据仓库为数据挖掘提供了新的支持平台。数据仓库的发展不仅为数据挖掘开辟了新的空间，并且对数据挖掘技术提出了更高的要求。作为数据挖掘的对象，数据仓库技术的产生和发展为数据挖掘技术开辟了新的战场，提出了新要求和挑战。数据仓库的体系结构努力保证查询和分析的实时性。数据仓库一般设计成只读方式，数据仓库的更新由专门一套机制保证，数据仓库对查询的强大支持使数据挖掘效率更高。数据仓库为更好地使用数据挖掘工具提供了方便。数据仓库的建立，应充分考虑数据挖掘的要求。用户可以通过数据仓库服务器得到所需要的数据，形成开采中间数据库，利用数据挖掘方法进行开采，获得知识。数据仓库为数据挖掘集成了企业内各部门全面的、综合的数据，数据挖掘要面对的是关系更复杂的企业全局模式的知识发现，数据仓库机制能够大大降低数据挖掘的障碍，一般进行数据挖掘要花大量的精力在数据准备阶段。数据仓库中的数据已经被充分收集起来进行了整理、合并，并且有些还进行了初步的分析处理。这样，数据挖掘的注意力能够更集中于核心处理阶段。另外，数据仓库中对数据不同粒度的集成和综合，能更有效地支持多层次、多种知识的开采。

另一方面，数据挖掘为数据仓库提供了更好的决策支持。高层决策要求系统能够提供更高层次的决策辅助信息，而基于数据仓库的数据挖掘能更好地满足高层战略决策的要求。数据挖掘对数据仓库中的数据进行模式抽取和知识发现，从数据仓库中揭示出对企业有潜

在价值的规律，形成知识，为知识管理提供内容，在知识管理中起到中流砥柱的作用。这些是数据仓库所不能提供的。数据挖掘对数据仓库的数据组织提出了更高的要求。数据仓库作为数据挖掘的对象，要为数据挖掘提供更多、更好的数据。其数据的设计、组织都要考虑到数据挖掘的要求。数据挖掘还为数据仓库提供广泛的技术支持。数据挖掘的可视化技术、统计分析技术等都为数据仓库提供了强有力的技术支持。

总之，数据仓库在纵向和横向都为数据挖掘提供了更广阔的活动空间。数据仓库完成数据的收集、集成、存储、管理等工作，为数据挖掘准备了经过初步加工的数据，使得数据挖掘能更专注于知识的发现。又由于数据仓库所具有的新特点，对数据挖掘技术提出了更高的要求。另一方面，数据挖掘为数据仓库提供了更好的决策支持，同时促进了数据仓库技术的发展。可以说，要充分发挥数据挖掘和数据仓库技术的潜力，就必须将二者有机地结合起来。

2. KDD 与数据挖掘的关系

KDD 是决策技术不可缺少的过程，也是数据仓库系统不可缺少的过程。Usama M. Fayyad 等专家对 KDD 定义为：它是识别有效的、新颖的、潜在的和最终可以理解模式的非平凡过程。经过数据挖掘之后的重要任务就是 KDD 的过程。曾经有的学者将数据挖掘、数据仓库、KDD 作为数据仓库系统的三部曲，缺一不可。有的学者认为数据挖掘和 KDD 是同一个概念，但有的学者认为它们之间存在差异。从技术角度看，数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的实际数据中，提取隐含的、先前未知的并有潜在价值的信息的非平凡过程。知识发现是从数据库中发现知识的全部过程，包括收集原始数据、数据清理、数据集成、数据仓库、数据选择、数据变换、数据预处理、数据挖掘、建立模型、模式评估、知识表示。数据挖掘是全部过程的一个特定的关键步骤，是指应用特定的算法从数据中提取模式。KDD 一般过程如图 1.1 所示。

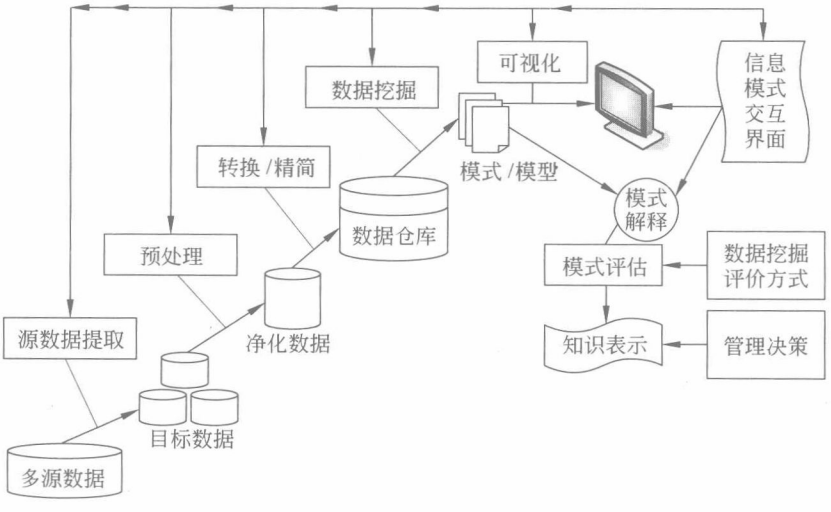


图 1.1　KDD 过程示意图

KDD 主要由以下步骤组成：

(1) 数据预处理　消除噪声或不一致数据；

(2) 数据组织与集成　多种数据源可以融合为一体进行异构数据的整合；

(3) 数据选择　从数据库中检索分析与任务相关的数据；

(4) 数据变换　将数据变换或统一成适合挖掘的形式，比如，有的要变成逻辑形式的数据，有的数据库要转化成逻辑数据库；

(5) 数据挖掘　按照主题要求，提出挖掘任务和基本步骤，使用智能手段，从大量数据(信息)中找出频繁出现的规律性事物，即提取数据模式；

(6) 模式评估　根据某种兴趣度度量，如支持度、可信度等，识别表示知识价值的模式；

(7) 知识表示　使用可视化和知识表示方法，展现与描述挖掘的信息和知识。

还有很多与数据挖掘和 KDD 相近或相关的术语，如数据分析(Data Analysis)、数据融合(Data Fusion)、数据的标准化/归一化、多智能体系统(Multi-Agent System，MAS)、决策支持系统、智能决策支持系统(Intelligent Decision Support System，IDSS)及群决策支持系统(Group Decision Support System，GDSS)等。

3. OLAP 与数据挖掘的关系

联机分析处理是针对特定问题的联机数据访问和分析。通过对信息(维数据)的多种可能的观察形式进行快速、稳定一致和交互性的存取，允许管理决策人员对数据进行深入观察。OLAP 委员会对联机分析处理的定义为：使分析人员、管理人员或执行人员能够从多种角度对从原始数据中转化出来的、能够真正为用户所理解的、并真实反映企业维特性的信息进行快速、一致、交互的存取，从而获得对数据更深入了解的一类软件技术。典型的 OLAP 系统体系结构如图 1.2 所示。

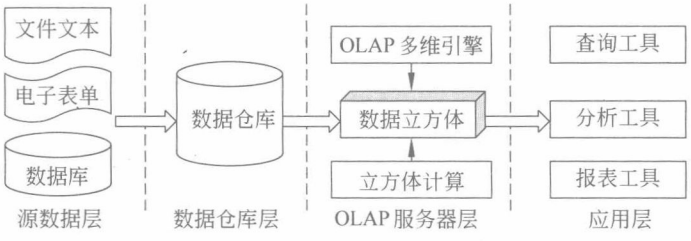


图 1.2　典型的 OLAP 系统体系结构

整个 OLAP 系统可采用 B/S 模式，大致分为四层：第一层是源数据层，存储了企业的业务细节数据。第二层是 OLAP 数据仓库层，数据抽取程序将源数据按主题进行归纳整理，存入 OLAP 数据库中，提供适合 OLAP 分析的详细、集成、准确的客户基础数据。第三层是 OLAP 服务器层，保存了分析所需要的客户聚集数据和相关的元数据，代理用户的分析请求，获取分析数据并返回给用户。第四层是应用层，让用户根据模型信息，提交分析请求，然后将获得的数据按用户需要的方式展现。

OLAP 和数据挖掘作为两种不同的数据分析工具，存在着许多不同之处：

(1) 是否主动进行数据分析，这是 OLAP 和数据挖掘最本质的区别。OLAP 是一种求

证性的分析工具，一般由客户预先设定一些假设，然后使用 OLAP 去验证这些假设，被动地进行数据分析；而数据挖掘是一种挖掘性的分析工具，它主要是利用各种挖掘算法主动地去挖掘大量数据中蕴含的规律和模式，主动地进行数据分析。

(2) 是否受到用户水平的约束，OLAP 是由用户驱动的，很大程度上受到用户水平的约束；而数据挖掘是由数据驱动的，系统能够根据数据本身的规律自动发掘潜在的模式，不受用户水平的约束。

(3) 从数据分析的深度来看，OLAP 位于较浅的层次；数据挖掘能从更深的层次上发现 OLAP 所不能发现的信息。

(4) 从分析的本质来看，OLAP 是首先建立一系列的假设，然后通过 OLAP 来证实或推翻这些假设从而得到结论，本质上是一个演绎推理的过程；而数据挖掘是依据数据特征采用不同的挖掘算法，在海量的数据中主动发掘模型，本质上是一个知识归纳的过程。

1.2.2 数据仓库系统模式

数据仓库能为 OLAP 和数据挖掘提供广泛和高质量的分析数据。

OLAP、数据挖掘和数据仓库的关系十分紧密。数据仓库的建立解决了依据主题进行数据存储的问题，提高了数据的存取速度；而 OLAP 分析与数据挖掘构成了数据仓库的表现层，将数据仓库中的数据通过不同的维和指标，灵活地展现出来，提高了数据的展现能力，进而提高了分析数据的能力与发现潜在知识的能力。

OLAP 对数据仓库具有很强的依赖性。没有数据仓库，OLAP 将很难实现；同样，在数据仓库选择主题时，也要参考 OLAP 分析的维度、指标，才能更好地为信息展示服务，为决策者进行业务分析提供依据。数据仓库与 OLAP 的关系是互补的，现代 OLAP 系统一般以数据仓库作为基础，即从数据仓库中抽取详细数据的一个子集并经过必要的聚集存储到 OLAP 存储器中供前端分析工具读取。在数据仓库应用中，OLAP 应用一般是数据仓库应用的前端工具，同时 OLAP 工具还可以和数据挖掘工具、统计分析工具配合使用，增强决策分析功能。

虽然数据仓库、OLAP 和数据挖掘是三种不同的信息技术，但其目标却都是辅助决策，所以它们之间存在着千丝万缕的联系。数据仓库拥有丰富的数据，但只有通过 OLAP 和数据挖掘才能使数据变成有价值的信息，才能体现出数据仓库的辅助决策功能，否则永远都是数据丰富而信息匮乏；反之，尽管 OLAP 和数据挖掘并不一定要建立在数据仓库的基础之上，但数据仓库却能提高两者的工作效率，让两者有更大的发展空间。对于 OLAP，无论其采用何种存储方式，数据最终都要转换成多维数据模型才能进行数据分析，而数据仓库中的星型模型和雪花模型都适用于 OLAP 的多维分析。

因此，在比较成熟的数据仓库系统中，数据仓库、OLAP 和数据挖掘往往融为一个以数据仓库为基础，与 OLAP 和数据挖掘相辅相成分析数据的模式。其中，数据仓库负责把所需要的数据按面向主题和有助于 OLAP 和数据挖掘分析的格式进行存储，并对原始数据进行预处理。OLAP 和数据挖掘则负责从不同的角度和层次对经过预处理的数据进行分析，挖掘出有用的模式。

通用的数据仓库系统如图 1.3 所示，其包括以下四部分。

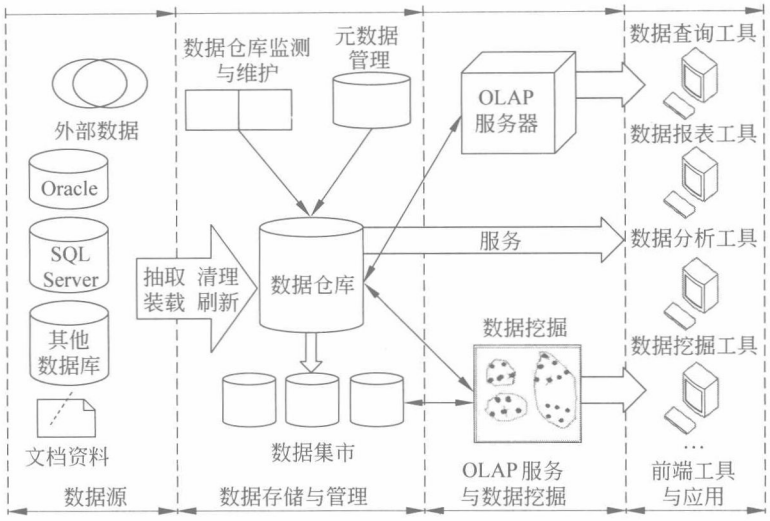


图 1.3　通用的数据仓库系统

(1) 数据源是数据仓库系统的基础，是整个系统的数据源泉。通常包括企业内部信息和外部信息。内部信息包括存放于关系数据库管理系统(Relational DataBase Management System，RDBMS)中的各种业务处理数据和各类文档数据。外部信息包括各类法律法规、市场信息和竞争对手的信息等。

(2) 数据的存储与管理是整个数据仓库系统的核心和关键。数据仓库的组织管理方式决定了它有别于传统数据库，同时也决定了其对外部数据的表现形式。要决定采用什么产品和技术来建立数据仓库的核心，则需要从数据仓库的技术特点着手分析。针对现有的业务系统数据，进行抽取、清理和有效集成，并按照主题进行组织。数据仓库按照数据的覆盖范围可以分为企业级数据仓库和部门级数据仓库(通常称为数据集市)。

(3) OLAP 服务器实现对需要分析的数据的有效集成，按多维模型予以组织，以便进行多角度、多层次的分析，并发现趋势。其具体实现可以分为关系 OLAP(Relational OLAP，ROLAP)、多维 OLAP(Multi-dimensional OLAP，MOLAP)和混合型 OLAP(Hybrid OLAP，HOLAP)。ROLAP 基本数据和聚合数据均存放在 RDBMS 之中，MOLAP 基本数据和聚合数据均存放于多维数据库中，HOLAP 基本数据存放于 RDBMS 之中，聚合数据存放于多维数据库中。

(4) 前端工具包括各种数据报表工具、数据查询工具、数据分析工具和数据挖掘工具等。其中基于 OLAP 和数据挖掘的前端工具分别是验证型工具和发掘型工具的代表。

综上所述，如果运用系统工程思想理解通用的数据仓库系统，应该将其划分为数据采集(子)系统、数据仓库(子)系统、数据挖掘(子)系统。数据采集(子)系统的主要内容包括数据采集对象的确立、数据集成技术与方法、数据预处理技术与方法、基于样本数据划分的通用数据挖掘模型系统、数据采集系统中的中间件技术等主要内容。数据仓库(子)系统的主要内容包括多维数据分析与组织、多维数据模型与结构、面向主体数据库(数据仓库)的