高等学校信息管理与信息系统专业系列教材

数据仓库与数据挖掘的原理及应用

李志刚　马　刚　编著

高等教育出版社
HIGHER EDUCATION PRESS

高等学校信息管理与信息系统专业系列教材

数据仓库与数据挖掘的原理及应用

李志刚　马　刚　编著

高等教育出版社

内容提要

本书详细阐述数据仓库与数据挖掘的基本原理，系统而全面地介绍数据仓库与数据挖掘的概念、作用、算法以及应用领域、相关学科和发展趋势，并着重讨论数据仓库和数据挖掘在企业管理中的应用及构建策略。基于 SQL Server 2005 介绍数据仓库与数据挖掘工具的操作和应用，并结合具体实例，阐述企业数据仓库和数据挖掘的实施过程。最后，以证券行业为对象提供一个数据挖掘的开发实例。本书的指导思想是在系统阐述基本知识和基本理论的基础上，强调实际应用能力的培养，充分体现数据仓库和数据挖掘技术作为解决实际问题的方法和工具的特点。

本书既可以作为信息系统、电子商务、管理科学与工程、计算机应用、软件工程等专业的本科高年级和研究生教材，又可以作为从事竞争情报、信息管理、知识管理、战略管理和软科学的研究人员的参考资料。

图书在版编目（CIP）数据

数据仓库与数据挖掘的原理及应用 / 李志刚，马刚编著. —北京：高等教育出版社，2008.2
ISBN 978 - 7 - 04 - 023014 - 7

Ⅰ. 数… Ⅱ. ①李…②马… Ⅲ. ①数据库系统 - 高等学校 - 教材②数据采集 - 高等学校 - 教材 Ⅳ. TP311.13 TP274

中国版本图书馆 CIP 数据核字（2008）第 000470 号

策划编辑	刘 艳	**责任编辑**	康兆华	**封面设计**	于文燕	**责任绘图**	郝 林
版式设计	张 岚	**责任校对**	金 辉	**责任印制**	陈伟光		

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	**免费咨询**	800 - 810 - 0598
邮政编码	100011	**网 址**	http://www.hep.edu.cn
总 机	010 - 58581000		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	北京奥鑫印刷厂	**畅想教育**	http://www.widedu.com
开 本	787×960 1/16	**版 次**	2008 年 2 月第 1 版
印 张	22.25	**印 次**	2008 年 2 月第 1 次印刷
字 数	410 000	**定 价**	27.80 元

物料号 23014 - 00

前　言

随着计算方法和信息技术的不断发展，大量数据的产生和收集导致信息爆炸。现代社会的竞争趋势要求对这些数据进行实时的和深层次的分析。虽然现在已经出现更强大的存储系统和检索系统，但是使用者发现在分析所拥有的信息方面变得越来越困难。数据仓库提供了容纳大量信息的场所，但它只有和数据挖掘技术相结合才能最终解决用户的困惑，使用户能够从繁杂的数据中找出真正有价值的信息和知识。数据仓库可以加强企业对信息的管理能力，数据挖掘可以改善企业的经营状况，使企业的决策制定过程更加科学化和快速，为企业带来巨大的收益，增强企业的竞争优势。

数据仓库和数据挖掘是20世纪90年代中期兴起的决策支持新技术，它们是基于大规模数据库的决策支持系统的核心。数据仓库是区别于数据库的一种新型数据存储形式，它是面向主题的、集成的、不可更新的、随时间不断变化的数据集合，用以支持经营管理中的决策制定。数据挖掘是从数据库中发现知识的核心技术，它从大量的数据中提取隐含的、人所未知的、可信而有效的知识。数据挖掘能够对数据进行再分析，以期获得更加深入的了解。它具有预测功能，可以通过已有数据预测发展趋势。数据仓库与数据挖掘技术相结合，与现代的管理决策方法相结合，就能够使数据仓库在组织机构的经营管理决策中发挥巨大的作用。

我国数据挖掘技术的研究始于20世纪90年代，经过十几年的发展，这一领域目前正处于蓬勃发展时期。但是由于数据仓库、数据挖掘技术都是数据处理与分析领域出现的新技术，大部分人把目光投向基于这两项技术的基础理论的研究，特别是具体技术和算法的实现，而忽略对数据仓库、数据挖掘理论与实践相结合的应用研究。因此，笔者在结合科研项目的基础上，对数据仓库、数据挖掘技术进行较系统的研究，并将研究成果应用于财经、证券投资等领域；同时笔者在研究过程中不断地学习，既对原有的理论和实践进行总结，又不断地将所学到的知识运用到实践中去，丰富了原有的理论。

本书详细阐述数据仓库与数据挖掘的基本原理，系统而全面地介绍数据仓库与数据挖掘的概念、作用、算法以及应用领域、相关学科和发展趋势，并着重讨论数据仓库和数据挖掘在企业管理中的应用及构建策略。基于SQL Server 2005介绍数据仓库与数据挖掘工具的操作和应用，并结合具体实例，阐述企业数据仓库和数据挖掘的实施过程。最后，以证券行业为对象提供一个数据挖掘的开发

实例。本书的指导思想是在系统阐述基本知识和基本理论的基础上，强调实际应用能力的培养，充分体现数据仓库和数据挖掘技术作为解决实际问题的方法和工具的特点。本书兼顾理论性与通俗性，注重理论联系实际，叙述时力求深入浅出，着重阐述理论的基本思路与方法的基本步骤。

本书的目的旨在向读者系统阐述数据仓库与数据挖掘的基本原理、方法和实用工具，介绍国内外的最新研究成果。全书共有11章，第1章介绍数据仓库的基本概念和知识；第2章介绍联机分析处理的基本理论；第3章介绍数据仓库的设计思想、方法和技巧；第4章介绍数据仓库的规划与开发；第5章介绍各种数据仓库工具的基本功能及SQL Server 2005数据仓库工具的应用；第6章介绍数据挖掘的概念和相关知识；第7章介绍数据挖掘的算法；第8章介绍文本挖掘、Web挖掘等数据挖掘新技术；第9章介绍数据挖掘的工具及其应用；第10章介绍数据仓库与数据挖掘的综合应用；第11章介绍基于数据挖掘的上市公司财务危机预警应用实例，使读者能结合具体应用进行上机操作，消化和理解所学的知识。

本书由李志刚负责全书的整体策划和最后统稿。编写任务的分工如下：第1章、第2章、第6章、第7章由马刚编写；第3章由李志刚、宛石锋编写；第4章由李志刚、黄艳编写；第5章由马刚、李志刚编写；第8章、第10章由李志刚编写；第9章由李志刚、彭易成编写；第11章由彭易成、李志刚编写。郭丰恺、聂运洁参加部分的文字和图形处理工作。

在本书的编写过程中，笔者借鉴国内外一些文献和网上资料，由于编写体例的限制未将其在文中一一注明，只在参考文献中列出，在此谨向各位学者表示由衷的敬意和感谢。由于数据仓库和数据挖掘技术发展迅速，尽管笔者付出艰苦的努力，但由于本人水平所限，疏漏甚至错误之处在所难免，恳请专家与读者批评指正。

李志刚

2007年12月

目　录

第1章

数据仓库概述

本章主要内容

本章主要介绍从数据库到数据仓库的发展过程；数据仓库的概念与特点；数据仓库的关键名词；数据仓库的数据组织；数据集市的概念与特点，数据仓库与数据集市之间的区别；数据仓库的体系结构以及操作数据存储系统。

随着计算机应用范围的扩大和网络计算的不断发展，企业和各种类型的组织机构提出新的要求，希望计算机能够越来越多地参与数据分析与决策制定的工作。传统的数据库技术主要面对单一的数据库资源，更适合事务型处理等操作，其分析型处理的能力较弱。数据仓库在这种情形之下应运而生，将事务操作型环境和分析型环境进行分离，划清了数据处理的分析型环境与事务操作型环境之间的界限，从而由原来的以数据库为中心的单一数据环境发展为以数据仓库为中心的一种新型体系化环境。数据仓库技术实现了“数据→信息→知识”的过程，为企业各级管理层提供了有力的决策支持。

1.1 从数据库到数据仓库

1.1.1 决策支持技术与数据库技术的发展

1. 决策支持技术的发展

客观世界是由物质、能量和信息这三大基本要素所组成的，人类社会生活的每分每秒都无法离开信息。从远古时代开始，人类就一直在同信息打交道：原始人通过结绳记事的方式进行信息的存储；古埃及的象形文字主要由当时的账房先生表示庶民欠法老谷子的多少；现今更是一个信息爆炸的时代，人类已经彻底“淹没”在数据的“海洋”之中了。正如 John Naisbitt 在《大趋势》一书中所说：

“我们正在被信息所淹没,但我们却由于缺乏知识而感到饥饿。”

任何技术的演进都是伴随着某种发展过程的,决策支持系统(decision support system,DSS)技术也经历了一个漫长而复杂的演化进程,而且它仍然在继续演化。

图1.1表明从20世纪60年代初期到1980年DSS技术的演化过程。

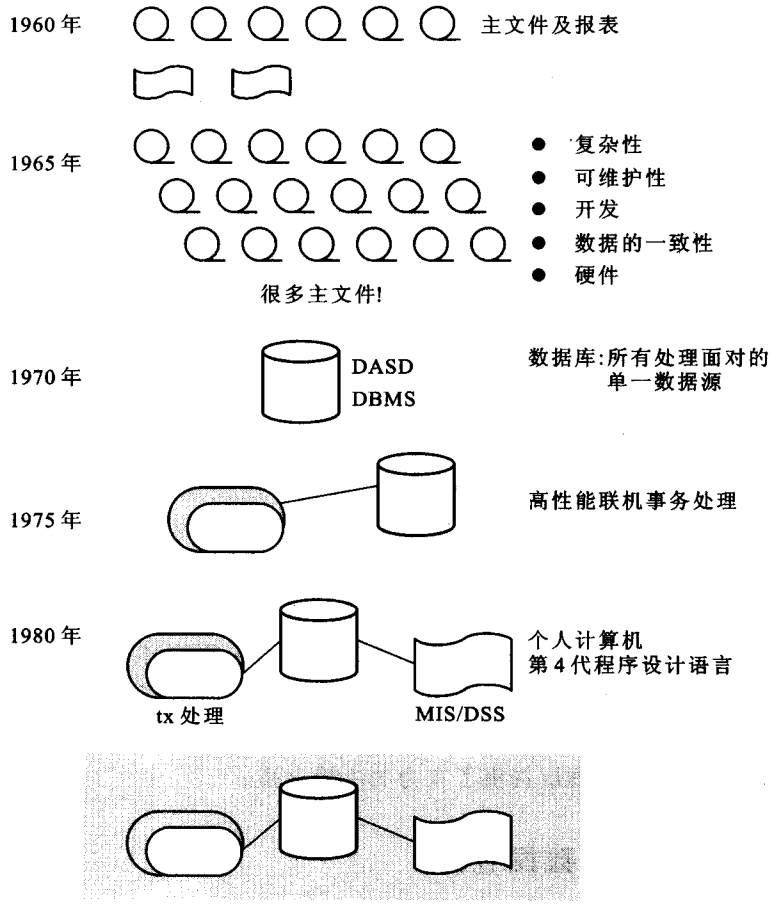

图1.1 体系化环境的早期演化过程

从时间的角度来看,决策支持技术的发展可以分为以下几个阶段。

(1) 20世纪50年代~20世纪60年代:数据处理阶段

数据处理(data processing)是计算机应用所涉及的最广泛的领域,约占70%。一个国家的现代化水平越高,数据处理的范围越广,数据量越大,数据处理所占的比例就越高。

(2) 20 世纪 60 年代 ~20 世纪 70 年代:管理信息系统阶段

随着 20 世纪五六十年代数据处理应用领域的不断扩大,20 世纪六七十年代西方发达国家兴起了管理信息系统(management information system,MIS)的热潮,而我国是在 20 世纪 70 年代末到 20 世纪 80 年代初,管理信息系统的研究和应用才逐步发展起来。

(3) 20 世纪 70 年代 ~20 世纪 80 年代:决策支持系统阶段

管理信息系统是在管理科学利用计算机的基础上发展起来的,它促使计算机的应用由数值计算领域拓宽到数据处理(非数值计算)领域,使计算机走向社会活动的各个方面。运筹学和系统工程学利用计算机发展形成了模型辅助决策系统,由于其所采用的模型主要是数学模型,其辅助决策能力主要体现在定量分析上。

20 世纪 70 年代初发展起来的决策支持系统把管理信息系统和模型辅助决策系统有机地结合起来,使得数值计算和数据处理融为一体,提高了辅助决策的能力。

20 世纪 70 年代中期,联机事务处理开始逐步取代数据库,更快速地访问数据成为可能。采用高性能联机事务处理方式,计算机可以完成先前无法完成的工作。

到 20 世纪 80 年代,一些更新颖的技术开始涌现,比如个人计算机(personal computer,PC)和第 4 代程序设计语言(fourth generation language,4GL)。终端用户开始扮演前所未有的角色——直接控制数据和系统,这超出了传统数据处理人员的工作范畴。随着 PC 与 4GL 技术的发展,一种新思想诞生了,即除了高性能联机事务处理外,可以对数据进行更多的处理,用来制定管理决策的处理过程可以实现了。在此之前,数据和技术不能一并用来导出详细的操作型决策。一种新的思想体系开始呈现,即单一数据库既能用于操作型的高性能联机事务处理,同时又可用于 DSS 分析与处理。

(4) 20 世纪 90 年代:智能决策支持系统综合决策支持系统阶段

智能决策支持系统(intelligent decision support system,IDSS)与综合决策支持系统(synthetic decision support system,SDSS)阶段的主要特征是模型技术、专家系统、数据仓库和数据挖掘技术的全方位集成,使得决策支持技术无论是在体系结构还是在信息处理能力方面都产生了较大的变化。

但是,IDSS 不能依据数据库中的大量数据进行学习和推理,知识并非由真正的学习而来。

而数据仓库(data warehouse,DW)、联机分析处理(on-line analytical processing,OLAP)、数据挖掘(data mining,DM)相结合构成的新型 DSS 具有较强的学习能力(从数据中获取新的有用的信息、知识与规则),传统 DSS 与新型 DSS 结合(DW + OLAP + DM + MB + DB + ES)形成更高级的 SDSS。