信息管理与信息系统专业
核心课程精品教材系列

DATA WAREHOUSE AND DATA MINING

数据仓库与数据挖掘

IMIS

周根贵 主编

浙江大學出版社

图书在版编目（CIP）数据

数据仓库与数据挖掘 / 周根贵主编. —杭州：浙江大学出版社，2004.8
（信息管理与信息系统专业核心课程精品教材系列）
ISBN 7-308-03831-9

Ⅰ. 数... Ⅱ. 周... 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. ①TP311.13②TP274

中国版本图书馆 CIP 数据核字（2004）第 081182 号

丛书策划 樊晓燕
封面设计 俞亚彤

责任编辑 杜玲玲
出版发行 浙江大学出版社
（杭州天目山路 148 号 邮政编码 310028）
（E-mail:zupress@mail.hz.zj.cn）
（网址：http://www.zjupress.com）
排　　版 杭州兴邦印务有限公司
印　　刷 富阳市育才印刷有限公司
开　　本 787mm×960mm 1/16
印　　张 19.25
字　　数 335 千
版 印 次 2004 年 8 月第 1 版 2006 年 4 月第 2 次印刷
印　　数 3001—4500
书　　号 ISBN 7-308-03831-9/TP·254
定　　价 28.00 元

信息管理与信息系统专业核心课程精品教材系列

编　委　会

序

关于信息化与后工业社会的话题已经被世人热烈地讨论了二十余年，正式的“信息管理与信息系统”专业在我国大学中的设立亦已是最近十来年的事。而近三年来，该专业却是我国大学本科乃至专科专业中发展和普及最快的专业之一。它表明了一个重要的事实：信息化在我国的新型工业化进程中正在扮演一个极其重要的角色。我国社会、经济的迅速发展对于既掌握信息技术，同时又拥有管理知识，能够运用信息技术于管理实践的复合型人才的需求正日益高涨。

信息管理与信息系统是一门具有交叉性的复合型学科，它融合了计算机科学、信息技术、管理学、经济学、系统科学、运筹学、组织行为学等学科的知识，它强调运用定性与定量相结合的方法及相关学科的研究手段，深入研究并解决各类社会系统中的信息管理问题。该专业直接以满足信息化建设人才与新型复合型管理人才的需求为目标，培养具有现代管理科学理论知识，具备较强的计算机及网络技术运用能力的，适合在经济管理部门、各类企事业单位从事信息系统建设和管理以及从事相应科学研究等工作的综合型高级专门人才。1998 年教育部调整学科专业目录，把原有的管理信息系统、经济信息管理、科技信息、图书情报检索、信息学及林业信息管理等专业改为“信息管理与信息系统”专业，作为管理学门类中“管理科学与工程”一级学科之下的

一个二级学科，使之在培养目标、内容和方向上均得到了进一步的凝练和提升。此后，由于社会、经济信息化进程的加速，该专业得到了快速的催生和发展。

众所周知，信息技术是当代发展最快的技术之一，相应的管理应用及经济、法律体系亦正面临着深刻的变化，因此，相关的教材亦面临技术基础多变和快速更新的挑战。为促进和支持该专业的发展，浙江大学出版社及时组织了有关专家在充分讨论和酝酿的基础上，精心组织出版了这套“信息管理与信息系统专业核心课程精品教材”。本人作为教育部高校管理科学与工程类教学指导委员会成员，参与了全国高校“信息管理与信息系统专业核心课程及其教学大纲”的修订工作。在组织本系列教材的编写时，亦强调了与教育部教学指导委员会规范化要求的一致与契合。本系列教材的编著者均有长期从事该专业教学和科研的经验，他们的前期工作经验使本系列教材的质量有了可靠的保证。相信这套教材的出版能为高校信息管理与信息系统专业教学水平的提高和规范化发展起到积极的推进作用。

吴晓波

2004 年 8 月于求是园

前　言

计算机网络与数据库技术的迅速发展和广泛应用，使得各行各业的管理工作进入了一个崭新的时代。广大基层管理人员摆脱了繁重的制表业务和数据处理工作，管理工作进一步规范化。各种在线事务处理信息系统的建立，对各种日常业务处理提供了有效的支持。然而，面对当今竞争日趋激烈与瞬息万变的市场，各级管理人员迫切需要根据组织的现状和历史数据做出判断和决策。他们希望能够从组织的信息系统中获取有效的、一致的决策支持信息，及时准确地把握市场变化的脉搏，做出正确有效的判断和抉择。概括地说，数据处理的重点应该从传统的业务处理扩展到在线分析处理，并从中得到面向各种主题的统计信息和决策支持信息。

数据仓库和数据挖掘技术就是针对上述问题而产生的一种技术解决方案，它是基于大规模数据库的决策支持系统环境的核心。数据仓库是面向主题的、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过程；而数据挖掘是从大量的数据中提取出隐含的、以前不为人所知的、可信而有效的知识。它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，可通过已有数据预测未来。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，就能使数据仓库在组织的经营管理决策中发挥巨大的作用。

本书的编写旨在提供数据仓库和数据挖掘领域的一个广博的且也是深入的概览。全书共分 10 章。第 1 章主要介绍数据仓库和数据挖掘的产生背景、应用和发展；第 2 章介绍数据仓库的组成和结构等技术与开发的基本概念以及第 3 章介绍数据仓库的技术管理；第 4 章详细介绍数据仓库的查询工具 OLAP；第 5 章介绍数据仓库的应用与开发工具 SQL Server；第 6 章介绍数据挖掘和知识发现的基本概念，以及数据挖掘的方法与技术；第 7 章详细介绍统计类的数据挖掘技术；第 8 章介绍知识类的数据挖掘技术；第 9 章进一步介绍 21 世纪各类新的数据挖掘技术；第 10 章代表性地介绍一些数据仓库与数据挖掘的综合应用场合和实例。

本书第 1 章由浙江工业大学周根贵编写；第 2,4,5 章由浙江工业大学黄洪编写；第 3,6,7,8 章由杭州电子科技大学陈朵玲编写；第 9 章由浙江科技学院李崇岩编写；第 10 章由浙江科技学院顾忠伟编写。全书最后由周根

贵教授修改、统稿。

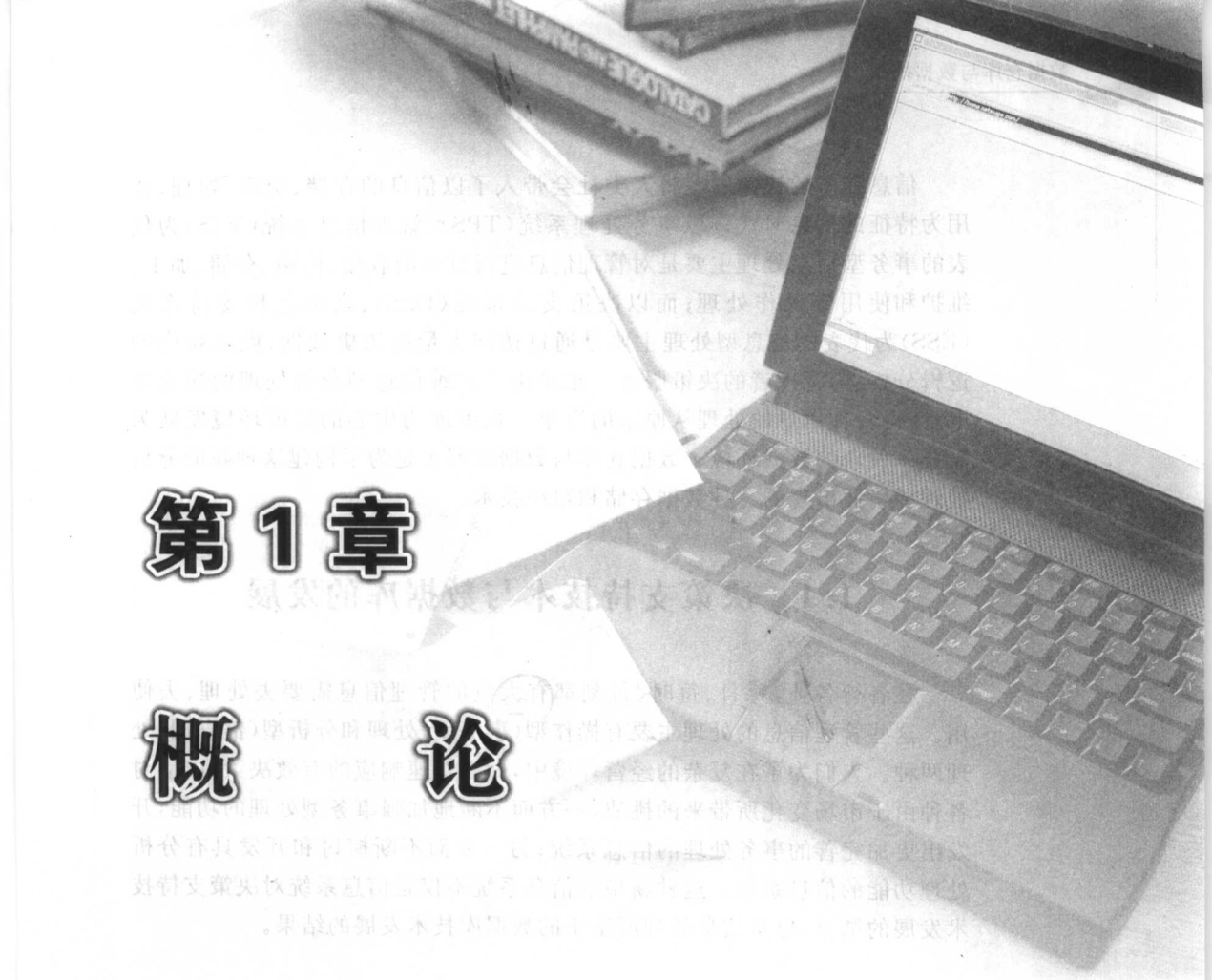
在本书的编写过程中，得到浙江工业大学朱艺华教授和杭州电子科技大学陈畴镛教授的大力支持与帮助，在此表示衷心的感谢。

我们在编写本书的过程中，尽可能做到深入浅出，力求概念正确，理论联系实际。由于数据仓库与数据挖掘是一个新的领域，发展非常迅速，加之我们水平有限，书中一定存在许多不足之处，恳切希望各位读者批评指正。

作　者

2004 年 5 月

目　录



第1章

概　论

学习目标

- 决策支持技术的发展
- 数据库技术的发展
- 数据仓库的应用和发展
- 数据挖掘的应用和发展

信息技术的快速发展将人类社会带入了以信息的存储、交换、处理、使用为特征的信息时代。以事务处理系统(TPS)、管理信息系统(MIS)为代表的事务型信息处理主要是对管理信息进行日常的收集、传递、存储、加工、维护和使用等操作处理;而以决策支持系统(DSS)、高级经理支持系统(ESS)为代表的信息型处理主要是通过访问大量的历史数据,做出相应的逻辑分析,为管理者的决策服务。正是由于这种信息型分析处理的快速发展,使得管理信息的处理从原来的以单一数据库为中心的数据环境发展为需要一种新的数据环境。数据仓库与数据挖掘正是为了构建这种新的分析处理环境而出现的一种数据存储和组织技术。

1.1 决策支持技术与数据库的发展

在各行各业,每日、每时、每刻都有大量的管理信息需要去处理、去使用。这些管理信息的处理主要有操作型(事务型)处理和分析型(信息型)处理两种。人们为了在复杂的经营环境中,做出快速响应的有效决策,以应对各种由于市场变化所带来的挑战,一方面不断地加强事务型处理的功能,开发出更加完善的事务处理的信息系统;另一方面不断探讨和开发具有分析处理功能的信息系统。这种新型的信息系统不仅是信息系统对决策支持技术发展的结果,也是信息处理所基于的数据库技术发展的结果。

1.1.1 决策支持技术的发展

数据处理(EDP)是电子计算机应用中最广泛的领域,约占70%。一个国家的现代化水平越高,数据处理的面越宽、量越大,数据处理所占的比例就越高。随着20世纪50年代到60年代数据处理领域应用的成功,20世纪60年代到70年代西方国家兴起了管理信息系统(MIS)的热潮。我国是在20世纪70年代末到80年代初才兴起了管理信息系统的应用。管理信息系统是在管理科学利用计算机后发展起来的,它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,使计算机走向社会和家庭。具体地说,管理信息系统是一个由人、计算机等组成的能进行数据的收集、传递、存储、加工、维护和使用的系统,它主要支持事物型的数据处理和信息管理。

20世纪70年代初,运筹学和系统工程工作者利用计算机形成了模型

辅助决策系统。由于采用的模型主要是数学模型，所以其辅助决策的能力主要表现在定量分析上，从而发展起把管理信息系统和模型辅助决策系统结合起来的决策支持系统(DSS)。DSS 主要是进行分析处理，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。它将数据、复杂的分析模型和用户友好的软件集成在一起，形成能够很好地支持各种复杂决策问题的信息系统，其目的是辅助管理决策。

另一方面，随着计算机技术的迅猛发展，20 世纪 60 年代末兴起了一个新研究领域——专家系统(Expert System，ES)，它是对 20 世纪 50 年代人工智能的进一步发展。专家系统是利用专家的知识在计算机上进行推理，达到专家解决问题的能力。1968 年 E. A. Feigenhanm 等人研制了 DENDRAL 专家系统，用来帮助化学家推断分子结构。1974 年 E. H. Shortliffe 等人研制了 MYCIN 专家系统，用来诊断和治疗感染性疾病。同一时期，人们还研制出不少其他专家系统。专家系统的出现使人工智能走上了实用化阶段。

专家的知识表现为产生式规则和语义网络等形式。知识的推理是采用符号逻辑中的假言推理。在搜索知识的时候，采用了深度优先或启发式搜索方法。专家系统也是一种很有效的辅助决策系统，它是利用专家的知识，特别是经验知识，经过推理得出辅助决策信息。对于专家知识，不规定它是数值的，更多的是不精确的定性知识。因此，专家系统辅助决策的方式属于定性分析。

专家系统和决策支持系统几乎是同时兴起，并沿着各自的道路发展起来的，它们都能起到辅助决策的作用，但辅助决策的方式完全不同。专家系统辅助决策的方式属于定性分析，决策支持系统辅助决策的方式属于定量分析。如果把这两者结合起来，辅助决策的效果将会大大改善，即达到定性辅助决策和定量辅助决策相结合。这种专家系统与决策支持系统相结合而形成的系统称为智能决策支持系统(IDSS)，它是决策支持系统的发展方向。

决策支持系统和专家系统的结合，并不是那样容易实现的，因为它们自成体系，要结合它们将有一些技术难题需要解决。专家系统结构中核心的部分由推理机、知识库和动态数据库三部分组成。知识库存放大量的专家知识；推理机完成对知识的搜索和推理；动态数据库存放已知的事实和推出的结果。专家系统中的动态数据库不同于决策支持系统中的数据库，相对来说，决策支持系统中的数据库是静态数据库。两系统中各部件之间的接口以及两系统的集成是形成智能决策支持系统的关键。

因此，不仅决策支持技术本身的发展对数据环境提出了更高的要求，而且决策支持技术与专家系统、人工智能的结合更要求能提供一种新的数据环境，为决策分析提供必要的数据源。只有当模型技术、专家系统以及这种新的数据环境的全方位的有机集成，才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上都产生了较大的变化，形成了人们熟悉而期望的智能决策支持系统。

1.1.2 数据库技术的发展

数据库(Database)一词起源于20世纪50年代。当时美国因战争需要，把各种情报集中在一起，存放在计算机中，称为Information Base或Database。数据库技术是研究数据库结构、存储、设计和使用的一门软件科学，于20世纪60年代中期产生，经过短短30年它已从第一代的网状、层状数据库，第二代的关系数据库系统，发展到第三代以面向对象模型为主要特征的数据库(尽管其在学术上和技术上都尚不够成熟，但1990年DBMS功能委员会发表的"第三代数据库宣言"已标志了第三代数据库的出现)，再到目前的数据仓库和数据集市等几个阶段。

数据模型是数据库系统的核心和基础。因此，数据库发展阶段划分应以数据模型的进展为主要依据和标志。数据模型根据其应用不同，可分为两大类或两个层次：①概念数据模型；②结构数据模型。其中概念数据模型只强调信息特征和语义，是现实世界到信息世界的第一层抽象，而用于划分数据库发展阶段的是用于机器世界的第二层抽象，即结构数据模型。层次数据库系统和网状数据库系统的结构数据模型(以下均简称为数据模型或模型)都是在20世纪60年代后期研究和开发的。它们从体系结构、数据库语言到数据存储管理均具有共同特征，可称为第一代数据库系统(见图1.1，图1.2)。

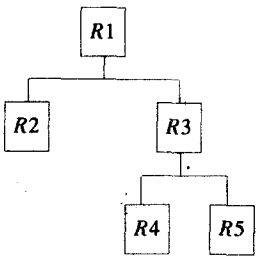


图1.1 层次模型

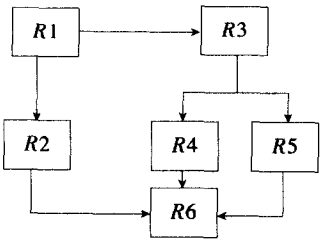


图1.2 网状模型