高等院校本科应用型经管专业规划教材

GaoDeng YuanXiao BenKe YingYongXing JingGuan ZhuanYe GuiHua JiaoCai

数据仓库与数据挖掘导论

李於洪／主 编

ShuJu CangKu Yu ShuJu WaJue DaoLun

经济科学出版社
Economic Science Press

高等院校本科应用型经管专业规划教材

数据仓库与数据挖掘导论

李於洪　主编

经济科学出版社

图书在版编目（CIP）数据

数据仓库与数据挖掘导论/李於洪主编．—北京：经济科学出版社，2012.8

高等院校本科应用型经管专业规划教材

ISBN 978-7-5141-2237-4

Ⅰ.①数… Ⅱ.①李… 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（2012）第 180740 号

责任编辑：纪晓津
责任校对：刘欣欣
版式设计：齐 杰
责任印制：王世伟

数据仓库与数据挖掘导论

李於洪 主编

经济科学出版社出版、发行 新华书店经销

社址：北京市海淀区阜成路甲 28 号 邮编：100142

总编部电话：88191217 发行部电话：88191537

网址：www.esp.com.cn

电子邮件：esp@esp.com.cn

北京欣舒印务有限公司印装

787×1092 16 开 17.5 印张 310000 字

2012 年 8 月第 1 版 2012 年 8 月第 1 次印刷

ISBN 978-7-5141-2237-4 定价：35.00 元

（图书出现印装问题，本社负责调换。电话：88191502）

内 容 简 介

本书为数据仓库与数据挖掘的基础教程，是作者多年来从事数据仓库与数据挖掘课程教学经验的梳理和总结。为了增强内容的直观性和可理解度，全书以大量图、表、实例融入其中。全书共分为四篇14章。第一篇为导引，共分2章：用实例和实例分析引导学生理解数据仓库与数据挖掘的概念内涵及其产生背景。第二篇为数据仓库，共分5章：详细介绍了数据仓库的体系结构及其组成部分的功能；从商业需求的角度介绍了数据仓库维度建模方法和联机分析处理操作；介绍了元数据在数据仓库建设中的重要性、分类方法与作用。第三篇为数据挖掘，共分4章：通过浅显易懂的语言及实例，深入浅出地介绍了关联分析方法、神经网络算法、决策树算法和聚类分析方法。第四篇为实验与工具，共分3章：提供了数据仓库实验、神经网络建模实验、决策树与关联分析实验，强化培养学生的应用能力。

本书可作为普通高等院校计算机专业、软件工程专业、信管专业等其他相关专业的教材，也可作为数据仓库与数据挖掘方面的培训教材，对于希望了解或学习数据仓库与数据挖掘知识的自学人士，本书具有较强的可读性。

前　言

人类运用现代信息技术高效、完美地实现了日常事务处理的计算机操作，基础是借助数据库技术、电子表格技术等对目标数据所进行的集成。当这样的数据库、电子表格不断激增时，一些跨地区或跨国经营的大型公司首先关注到了来自不同业务部门统计数据由于时间基准不统一、数据抽取算法不同、数据抽取级别不同等原因，导致对同一问题的分析会产生不同的结果。例如，销售部门的统计数据表明：去年的销售额上升了13%；而财务部门的统计数据表明：去年的销售额上升了9%。对公司管理层的决策制定带来了困难。于是，如何利用这些数据为公司高层管理者决策提供一个分析环境的问题就被提了出来。人们开始注意到，经过多年积累的日常事务处理的历史数据是公司高层进行决策分析的重要数据源。这就是数据仓库思想的由来。

而数据挖掘通常被认为是对数据仓库中数据进行分析的一系列方法与技术。

数据仓库与数据挖掘是人类智慧在信息技术应用领域的深层拓展，它为细织的高层管理者提供了一套比较完整的决策分析环境、数据组织体系结构和分析的理论与方法。在许多计算机类专业、信管类专业、软件类专业的培养计划中，数据仓库与数据挖掘是一门必修的专业课。

作者从事数据仓库与数据挖掘课程的教学工作7年，发现找到一本合适的《数据仓库与数据挖掘》入门教材十分困难。为此，作者于两年前开始着手在参考已有的各类《数据仓库与数据挖掘》、《数据仓库》、《数据挖掘》教材的基础上，根据自己的教学经验，

编写了《数据仓库与数据挖掘导论》教材，希望能写出一部既能反映数据仓库体系结构，又能将商业分析需求的建模方法与之相联系；既从理论、方法和算法上阐述数据挖掘的分析过程及其结果的意义，又能通过实验与工具进行建模，并对实验结果进行分析；既便于教学，又可以自学，概念清楚、实例完整、强化学生应用能力的数据仓库与数据挖掘教材。

本书与其他同类教材相比有以下特色：

1. 定位于数据仓库与数据挖掘的入门教材，注重基础性

自21世纪初我国高校开始陆续在计算机专业、信管类专业、软件类专业研究生课程中开改《数据仓库》、《数据挖掘》两门课程以来，至2005年，许多高校本科专业都相继开设了《数据仓库与数据挖掘》课程。由于是比较新的专业课程，人们还是普遍缺乏对数据仓库与数据挖掘基础知识的系统了解。作者在阅读不同行业有关数据仓库的研究文献时就发现，一些学者对数据仓库构建目标的认识是有偏差的，甚至理解是错误的；在数据库、数据仓库、数据集市三者关系上往往出现认识混淆。所以，作者将《数据仓库与数据挖掘导论》定位于阐述其基本概念、体系结构、基本理论和基础分析方法。

2. 理论阐述与实验及工具相结合，注重应用性

作者认为，不仅是专业技术人员需要学习数据仓库与数据挖掘的知识与技术，每一个在特定组织中从事管理工作或未来可能从事管理工作的人都应该了解数据仓库与数据挖掘的决策应用分析思想与过程，因为构建的数据仓库与数据挖掘环境是为管理层决策分析服务的。因此，本教材的内容力求浅显易懂，用实例诠释基本概念和基本原理，并提供实验及工具强化学生应用能力的培养。

3. 激发学生进一步学习和思考，注重引导性

数据仓库与数据挖掘是一门融理论性与实验性为一体的课程，而且理论和技术一直在不断出新。例如，实验工具在快速发展变化中，任何一本教材都只能为学生提供个别的数据仓库与数据挖掘平台工具进行实验。为了弥补教材内容的局限性对学生学习内容的影

响，作者在本教材中采取了以下措施。

第一，与大多数《数据仓库与数据挖掘》教材的作用不同，作者在相关章节的课后作业中安排了一些讨论题，例如，要求学生以小组为单位，共同完成 ETL 实验工具的调研，包括了解其运行环境要求、主要功能、适用领域、局限性等。还安排了一些撰写相关研究领域文献综述的作业。

第二，实验与工具课对学生明确提出了数据建模与分析的操作要求，要求每个学生自己选择建模数据并将分析结果向指导老师解释后，方可写实验报告。避免了所有学生都是同一个实验结果的情况。而且作者在带学生实验中也发现，有些实验结果难以分析或无法分析，其实这也是一个值得探讨的问题，对激发学生进一步思考很有宜处。在这种情况下，可以要求学生重新选择实验数据再做一次，看看结果如何。

本教材的第 12 章和第 13 章两个实验由作者的同事、浙江科技学院经管学院管理科学与工程系的教师陈金来编写，在此表示感谢。

欢迎各位教师和读者提出保贵意见和建议，作者将在今后的修改中进一步完善。

感谢浙江科技学院对我们进行《数据仓库与数据挖掘》课程的教学改革给予的立项支持，这本书中的一些内容体现了该教改项目的研究成果；感谢浙江科技学院经管学院对本书出版给予的大力支持和资助，使得本书能够顺利出版并惠及学生的教学；感谢浙江科技学院经管学院选修《数据仓库与数据挖掘》课程的信息管理与信息系统专业的学生，7 年来，他们的课堂发言、作业、讨论题、选做题提供了极富价值的反馈，为本书内容的系统性组织提供了有益的参考；感谢浙江科技学院经管学院实验中心的同事，他们帮助创建了良好的数据仓库与数据挖掘的实验环境；感谢经济科学出版社为本书出版所做的工作。

作者

2012 年 5 月 10 日

目　录

第一篇　导　引

第二篇 数据仓库

第三篇 数据挖掘

第四篇 实验与工具

第一篇　导　　引

第1章　数据仓库概念与内涵

学习要求

通过本章的学习，了解数据仓库产生的原因，并通过实例理解数据仓库（Data Warehousing，DW）的决策分析目标与作用，以及数据仓库的四个基本特征：数据仓库的数据是面向主题的、集成的、不可更新的、随时间不断变化的；弄清数据仓库与数据集市（Data Market，DM）的关系，以及独立数据集市和从属数据集市的内涵；了解粒度的概念。

《数据仓库与数据挖掘》缘于信息技术的自然演化。早期的数据库技术在数据收集、存储、检索实践中逐步完善，已成为技术成熟的现代日常事务处理必不可少的工具。面对日益增多的、分布在不同系统平台数据库中的数据，人们开始思考能否从纷繁复杂、大量积聚的数据环境中得到有用的决策信息，从而为企业的生存和发展提供正确的决策。数据仓库与数据挖掘就是为制定企业管理方面的重要决策而提出的解决方案。

1.1　数据仓库概念

伴随数据库技术的普遍应用，承载信息的数据随时间推移而不断增长。例如，对于一家制造业企业来说，可能有存储了多年的生产数据、销售数据、财务数据、市场数据、人事数据等，所有这些数据从结构上看是相对独立的，而且很有可能因初建时间不同，这些数据分布在不同的系统平台数据库上，甚至在文件系统中，存储形式多样。显然，这是不利于企业管理者在决策中进行全面分析和查询的。如果我们针对决策者的需求，对这些数据进行结构上的重组，按更方便决策分析的要求去设计，就能让企业的这些宝贵资源——原始数据实现其真正的信息价值。数据仓库的目标就源于此。

所以说，数据仓库就是管理者非常需要的、用来提供企业战略决策信息的系统环境。这种新的环境与以往支持的日常事务处理的数据库操作环境是分离

的，其决策分析数据源抽取自已有的企业日常事务处理数据库及文件系统，此外，还有外部数据。

1.1.1 数据仓库的产生

企业是从1960年开始建立和使用数据库系统的，日常工作中的订单处理、存货盘点、客户服务、付款接收、赔偿处理等现代商业行为都已完全依赖于数据库系统。20世纪90年代以来，企业的商业活动变得越来越复杂，尤其是跨国公司在全球不同的国家和地区拓展业务，加剧了竞争，致使企业的管理者渴望得到更多的信息来帮助企业及时做出正确的决策，从而提高竞争力。尽管现有的数据库系统的确提供了大量的信息来支持每天的日常事务处理业务，它在数据共享、数据与应用程序的独立性、维护数据的一致性和完整性及数据的安全保密性等方面提供了有效的手段，但管理者需要的是可以用来进行战略决策的信息（见表1-1）。当数据库与分析型应用结合时，就出现了许多问题。

表1-1 战略信息的特征

综合性	必须有一个独立的、从企业整体来看的视角
数据完整性	信息必须是准确的，必须符合商业规则
可用性	必须是通过直观方法容易获得的，对于分析工作是有用的
可靠性	每个商业因素都必须有且只有一个值
及时性	信息必须是在规定时间内准备好的，待用

资料来源：[美] Paulraj Ponniah. 数据仓库基础. 段云峰等译. 北京：电子工业出版社，2004，3.

第一，需要访问大量的企业内部数据和外部数据。由于企业各部门的操作型数据库管理系统往往是在长达几十年的企业信息化进程中相互独立地逐步建成的，例如，库存管理系统、订单处理系统、财务管理系统等，这些操作型日常事务处理系统常常依赖于不同的数据库管理系统平台，常见的有Oracle、Sybase、SQL Server、Excel等。一个公司过去使用的技术越多，公司里的数据就越互不相关——由于各种数据分散在各种相互分离的系统、多平台和不同结构中，它们在数据的定义及组织方式上都可能不同，因此，要想获得决策所需的综合性信息，必须熟悉不同的系统环境和数据定义，理解数据的具体含义和相互间的关系。显然，对决策信息获取人员的素质不但要求非常高，而且信息的抽取和综合分析也很不方便。

第二，日常事务处理数据库中的大量数据都是细节性数据和当前数据。例