普通高等教育“十一五”国家级规划教材

高等院校重点推荐教材

数据仓库与数据挖掘原理及应用

（第二版）

王丽珍　周丽华　陈红梅　肖　清　编著

科学出版社
www.sciencep.com

普通高等教育“十一五”国家级规划教材

高等院校重点推荐教材

数据仓库与数据挖掘原理及应用

（第二版）

王丽珍　周丽华
陈红梅　肖　清　编著

科学出版社

北京

内 容 简 介

本书全面深入地介绍了数据仓库、联机分析处理和数据挖掘的基本概念、基本方法和应用技术。全书分成三篇：数据仓库与OLAP篇的主要内容包括：数据仓库的基本概念、体系结构、模型设计、开发方法、ETL、元数据和数据集市，OLAP的基本概念、基本操作、数据模型和OLAP的实现及准则；数据挖掘与空间数据挖掘篇的主要内容包括关联分析方法、聚类分析技术、分类与预测方法、异常检测算法以及空间数据挖掘技术等；工具与实例篇介绍了数据挖掘工具及可视化、Cognos公司的BI主要产品和企业数据仓库系统构建。

本书可作为高等院校计算机软件与应用、信息科学等专业的学生学习数据仓库、OLAP及数据挖掘技术的实用教程或参考书，也可供从事数据仓库、数据挖掘研究、设计、开发等工作的科研、工程人员参考。

图书在版编目(CIP)数据

数据仓库与数据挖掘原理及应用/王丽珍等编著. —2版—北京：科学出版社，2009

（普通高等教育“十一五”国家级规划教材·高等院校重点推荐教材）

ISBN 978-7-03-025400-9

Ⅰ.数… Ⅱ.王… 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（2009）第150969号

责任编辑：鞠丽娜/责任校对：柏连海 王万红

责任印制：吕春珉/封面设计：三函设计

科学出版社 出版

北京东黄城根北街16号

邮政编码：100717

http://www.sciencep.com

新蕾印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2009年9月第 一 版 开本：787×1092 1/16

2009年9月第一次印刷 印张：19 1/4

印数：1—3 000 字数：456 500

定价：30.00元

（如有印装质量问题，我社负责调换<环伟>）

销售部电话 010-62134988 编辑部电话 010-62138978-8002

前　言

进入信息社会以来，信息技术经历了从计算机主机的信息集中处理方式到个人计算机（PC）的信息分布处理形式的转变；从单一的计算机操作系统到计算机互联网络操作的改变；从客户机/服务器（Client/Server）计算体系到多层体系结构计算模式的转变；从单一数据库到大型数据仓库和从局域网到 Internet 全球网的改变。现代信息技术的发展和现代科学技术的进步，使人类迈入新的时期——信息化时代。

信息处理技术的发展使得各类数据、信息急剧增长，给数据的传输、存储带来了许多新的问题，特别是由于各类不同事务产生大量不同类型的数据，这些数据分别被许多各个时期建立的应用系统所使用。人们希望能够看到所有数据和信息的综合情况，而这些数据和信息有许多不能被统一描述，不能被现有应用系统综合使用。针对这一问题，人们设想专门为业务的统计分析建立一个数据中心，它的数据来自联机的事务处理系统、异构的外部数据源、脱机的历史业务数据等，这个数据中心就是数据仓库。数据仓库技术的应运而生，成为信息技术领域非常热门的话题之一。

数据仓库技术的提出建立了一种体系化的数据存储环境，将分析决策所需要的大量数据从传统的操作环境中分离出来，使分散、不一致的操作数据转换成集成、统一的信息。企业内不同单位、不同角色的成员都可以在此单一的环境之下，通过运用其中的数据与信息，发现全新的视野和新的问题，产生用于决策的新分析方法。作为决策支持系统的重要组成部分，数据仓库为决策支持系统提供了分析决策所需的数据；OLAP 的产生，则进一步增强了决策支持系统快速、一致和交互性的分析能力，它利用存储在数据仓库中的数据完成各种分析操作，并以直观易懂的形式将分析结果展现给决策分析人员；而数据挖掘是从大量数据中提取或“挖掘”知识，从而实现从“数据→信息→知识”的过程，为企业的管理阶层提供各种层次的决策支持。

本书第二版对数据仓库、OLAP、数据挖掘的原理、技术、工具和应用做了全面深入的介绍和分析，对数据仓库、OLAP 和数据挖掘的发展及应用前景也进行了细致深入的讨论。全书按“基本概念→基本原理→实际应用”的组织思路分成三篇，分别是数据仓库与 OLAP 概念、原理和技术篇，数据挖掘与空间数据挖掘技术篇和工具及实例介绍篇。作者根据教学实践和有关反馈信息，对第一版中的内容做了如下几个方面的修改。

第一篇的内容由 7 章调整为 4 章，将原来的第 3、4、至第 5 章合并为一章，

更为紧凑地介绍数据仓库的模型设计和建立过程。调整原第 7 章内容为第 1 章的一小节，使第 1 章的内容更为完整。对数据仓库中的 ETL 和元数据以及 OLAP 的实现等内容做了进一步的补充和完善。

第二篇的内容由 3 章调整为 6 章，将关联挖掘和聚类分析独立成章，新增加异常检测和空间数据挖掘技术两章，并修改、补充和进一步完善了相关技术和算法。

对第三篇的修改主要是补充新工具和更新实例。在保留原有工具的基础上，补充了 Weka 数据挖掘工具，并结合作者参与开发的数据仓库建设新项目，更新原“移动通信业务数据仓库系统”为“企业数据仓库系统构建”。

本书第二版继续保持了第一版的组织特点，采用引言→主体内容→小结→习题的结构形式。每章后面的习题适于作为课后作业。这些习题或者归纳成小问题，用于测试对内容的掌握；或者归纳成大问题，需要分析思考甚至查阅资料来完成。在内容的介绍上，除理论联系实际外，还采用了大量的图示及实例，使该书具有较强的可读性和可理解性，因此，凡具有一定数据库基础知识的人，都能学懂本书的内容。

讲授这门课程一般需 54 学时左右。本书内容阐述深入浅出，因此本书既可作为课堂教学的教材，也可作为自学或进行研究探讨的参考书。

本书的写作过程也是学习、研讨、提高的过程。书的再版过程中，作者对国内外的大量资料进行了再次的归纳和整理，并认真学习了新的数据仓库、OLAP 和数据挖掘工具，对参与及主持开发的新的数据仓库系统进行了全面的分析和总结。

本书编写过程中，作者对书中内容进行了反复研讨，且有些章节可能由某位老师执笔，而由另一位老师修改，因此难以严格划分每个人之工作量。就执笔而言，其分工如下：第 1、2、4、10、12、13 章由王丽珍执笔，第 3、7、9 章由周丽华执笔，第 5、6、8 章由陈红梅执笔，第 11 章由肖清执笔。

在本书的编写过程中，云南大学研究生夏勇、胥玲芳、李剑彬、张晓峰、陆叶等为本书的完成做了大量的辅助性工作。本书的出版得到了科学出版社的大力支持，另外还得到国家自然科学基金（项目编号：60463004）的资助，在此表示衷心的感谢。

由于作者水平有限，书中错漏和不妥之处在所难免，恳请读者批评指正。

作　者

2009 年 3 月

目　录

第一篇　数据仓库与 OLAP

第二篇　数据挖掘与空间数据挖掘

第三篇　工具与实例

第一篇　数据仓库与 OLAP

第 1 章　数据仓库基本概念

计算机技术的迅速发展使得处理数据成为可能，这就推动了数据库技术的极大发展，但是面对不断增加如潮水般的数据，人们不再满足于数据库的查询功能，提出了深层次问题：能不能从数据中提取蕴藏于其中的知识为决策服务。就数据库技术而言已经显得无能为力了，这就急需有新的方法和技术来处理这些海量般的数据。在这种情况下，数据库逐步发展到了数据仓库。世界上最早的数据仓库是 NCR 公司为全美、也是全世界最大的连锁超市集团 Wal-Mart 在 1981 年建立的，而最早将数据仓库提升到理论高度进行分析并提出数据仓库这个概念的则是著名学者 W.H.Inmon，他对数据仓库所下的定义是：数据仓库是一个面向主题的、集成的、稳定的、随时间变化的数据的集合，用于支持管理决策过程。由此可见，数据仓库是一个综合的解决方案，主要用来帮助企业有关主管部门和业务人员做出更符合业务发展规律的决策。

1.1　从数据库到数据仓库

传统数据库以及联机事务处理（on-line transaction processing，OLTP）在日常的管理事务处理中获得了巨大的成功，但是对管理人员的决策分析要求却无法满足。因为，管理人员常常希望能够通过对组织中的大量数据进行分析，了解业务的发展趋势。而传统数据库只保留了当前的业务处理信息，缺乏决策分析所需要的大量的历史信息。为满足管理人员的决策分析需要，就需要在数据库的基础上产生适应决策分析的数据环境——数据仓库（data warehose）。

1.1.1　蜘蛛网问题

在市场经济的激烈竞争中，信息对于企业的生存和发展起着至关重要的作用。企业对信息的需求是多方面的，为了避免企业中各部门或各用户间的冲突和简化用户的数据视图，一种称作“抽取程序”的方法被广泛地应用。比如，市场部人员通常只关心企业的销售、市场策划方面的信息，而不注重企业的研发、生产等其他环节。因此，将销售、市场策划方面的信息抽取出来单独建立部门级的数据库很有必要，这样可以提高数据的访问效率。在部门级数据的基础上可能还要被继续执行抽取程序，以建立个人级的数据库。比如，专门负责制作公司财务报表的数据人员，常常需要从财务部门的数据库系统中抽取数据。又如，部门经理可能经常抽取常用的数据到本地，有针对性的建立个人级数据库就显得尤为重要。

随着数据的逐层抽取，很可能最终导致系统内的数据间形成了错综复杂的网状结构，如图 1.1 所示，人们形象地称为“蜘蛛网”。一个大型的公司每天进行上万次的数据

抽取很普遍。这种演变不是人为制造的，而是自然演变的结果。企业的规模越大，“蜘蛛网”问题就越严重。

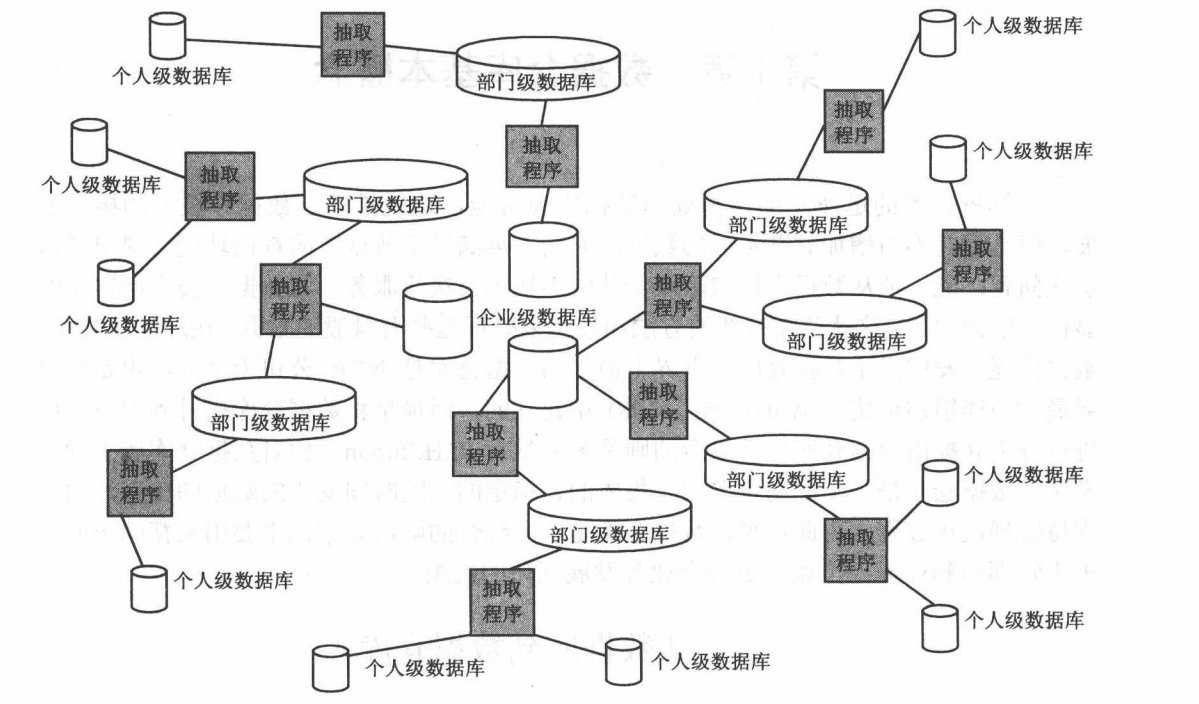

图 1.1　企业中存在的“蜘蛛网”现象

虽然网上的任意两个节点的数据可能归根结底是从一个原始库中抽取出来的，但其数据没有统一的时间基准，因而错综复杂的抽取与访问将产生很多的问题，主要有以下几个方面。

1. 数据分析的结果缺乏可靠性

图 1.2 中展示了某企业的市场部和计划部对项目 I 是否具有市场前景的分析过程和结果。市场部认为“项目 I 的市场前景很好”，而计划部却得到截然相反的结果——“项目 I 没有市场前景”。作为企业的最终决策者，将如何根据这样的结论进行决策呢？

为什么分析同一个企业数据库中的数据，却得到截然相反的结论呢？

首先，两部门可能抽取数据的内容不同。比如，市场部抽取的是项目 I 在大客户中的应用情况，而计划部抽取的是项目 I 在普通客户中的应用情况。

其次，可能两部门抽取数据的时间不同。如市场部在星期日晚上提取分析所需的数据，而计划部在星期三下午就抽取了数据。有任何理由相信对某一天抽取的数据样本进行分析与对另一天抽取的数据样本进行的分析可能相同吗？当然不能！企业内的数据总是在变的。

再次，引用外部信息的不同。分析项目的发展趋势常常需要引入企业外部的信息，比如报刊信息、国家的政策等。市场部门引用的外部信息来源可能与计划部门不同，而

外部信息自然是仁者见仁，智者见智，这也可能是导致最终分析结果不同的原因。

最后，分析程序的差异。市场部门使用的分析程序可能与计划部门不同，分析的内容和指标也可能不同。

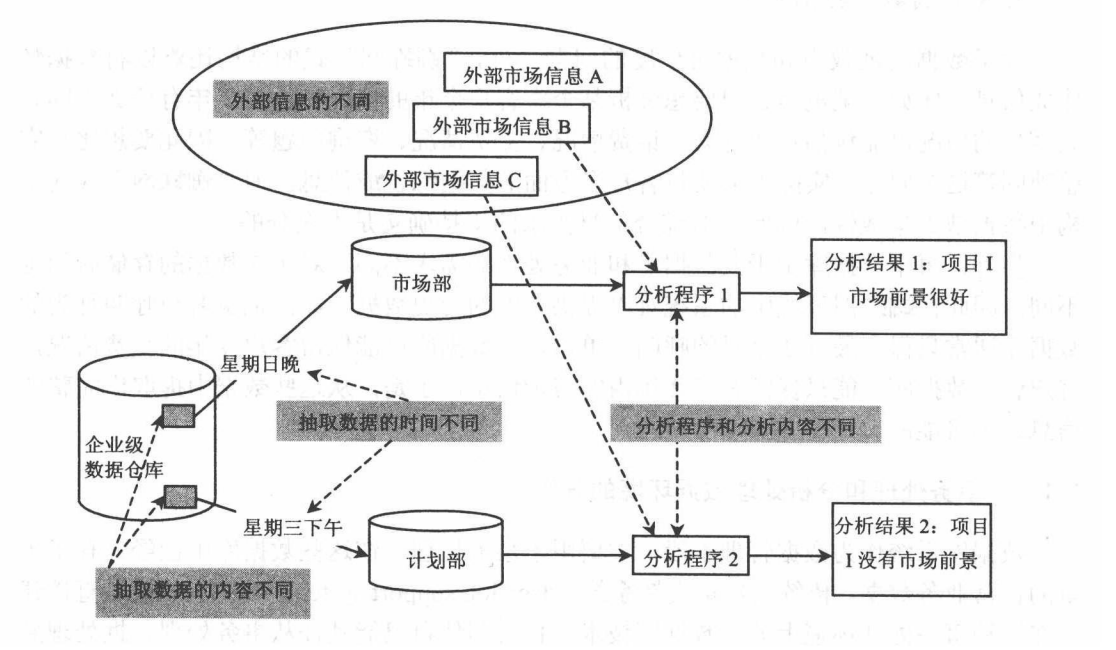

图 1.2　两个分析结果的差异

2. 数据处理的效率很低

数据分析的结果缺乏可靠性并不是蜘蛛网问题中唯一的主要问题。在一个大型企业中，不同级别的数据库可能使用不同类型的数据库系统，对于拥有巨型数据量的企业级数据库可能使用 IBM DB2，而对于部门级和个人级的中小型数据库可能使用 SQL Server。各种数据库的开发工具和开发环境不同，当需要在整个企业范围内查询数据时，数据处理的低效率将是不容忽视的。

如果一个大型企业的决策领导需要一份关于公司整体运营情况的报表，通常需要动用大量的人力和物力才能达到。首先，定位报表需要的数据，即确定报表涉及的内容分布在哪个数据库的哪个位置，然后调动各个部门的程序员/分析员对应用进行分析、设计和编码。

由于数据分散在各个数据库中，因此需要编写的程序很多。由于企业中使用的数据库类型很多，因此可能需要使用多种技术来实现。可见，面对企业中存在的蜘蛛网现象，为产生一份关于公司整体运营情况的报表，将动用大量的人力、物力和时间才能完成。

如果低效率的过程是一次性的，那么为生成报表花费大量的资源也是可取的。换句话说，如果生成第一份企业报表需要大量资源，生成所有后继报表可以建立在第一份企业报表基础之上，那么不妨为生成第一份报表付出一些代价，但是事实并非如此。

除非事先知道未来的企业报表需求，并且除非这些需求影响到第一张报表的建造，

每个新的企业报表总是要花费同前面差不多的代价。

因此，数据处理的低效率是蜘蛛网问题所面临的又一个问题。

3. 难以将数据转化成信息

除了数据处理效率和数据可信度的问题之外，“蜘蛛网”式的结构还难以将数据转化成信息。比如，某电信公司要想分析某个大客户今年的情况和过去 3 年有什么不同？大客户的情况可能包括呼叫行为、话费情况、交费情况、咨询问题等。因此要想比较完整地回答这个问题，实际上需要将客户多方面的数据综合成信息。但“蜘蛛网”式的结构中数据缺乏集成性，因此，对综合信息需求的支持确实是不充分的。

另外，每个数据库由于其数据量和业务处理的需求不同，对历史数据的存储时间也不同，因此在蜘蛛网环境中的系统难以提供完整的历史数据。如，记录客户呼叫行为的数据库通常只保留最近 3 个月的呼叫话单，财务数据库可能保留客户今年的交费情况，客户咨询数据库可能只保留客户 2 年内的咨询信息，于是，从这些数据中提取出完整的信息是不可能的。

1.1.2　事务处理和分析处理数据环境的分离

数据库系统作为数据管理手段，主要用于事务处理。在这些数据库中已经保存了大量的日常业务数据。传统的决策支持系统（decision support system，DSS）一般是直接建立在这种事务处理环境上的。数据库技术一直力图使自己能胜任从事务处理、批处理到分析处理的各种类型的信息处理任务。尽管数据库在事务处理方面的应用获得了巨大的成功，但它对分析处理的支持一直不能令人满意，这也正是产生“蜘蛛网”问题的原因之所在。因此，要解决“蜘蛛网”问题，必须将用于事务处理的数据环境和用于分析处理的数据环境分离开。

这样，数据处理被分为事务型处理和分析型处理两大类。事务型处理以传统的数据库为中心进行企业的日常业务处理。比如电信部门的计费数据库用于记录客户的通信消费情况，银行的数据库用于记录客户的账号、密码、存入和支出等一系列业务行为。

分析型处理以数据仓库为中心分析数据背后的关联和规律，为企业的决策提供可靠有效的依据。比如，通过对超市近期数据进行分析可以发现近期畅销的产品，从而为公司的采购部门提供指导信息。又如，对高校大学生就业信息进行分析的结果及结论，可以有效的指导学校制定招生计划和合理设置专业等。

事务处理的使用人员通常是企业的具体操作人员，处理的数据通常是企业业务的细节信息，其目标是实现企业的业务运营；而分析处理的使用人员通常是企业的中高层的管理者，或者是从事数据分析的工程师。决策分析数据环境包含的信息往往是企业的宏观信息而非具体的细节，其目的是为企业的决策者提供信息支持，并最终指导企业的商务活动。事务处理和信息分析数据环境的划分如图 1.3 所示。

事务处理和信息分析数据环境的分离划清了数据处理的分析型环境与事务型环境之间的界限，从而由原来以单一数据库为中心的数据环境发展为以数据库为中心的事务处理系统和以数据仓库为基础的分析处理系统。企业的生产环境，也由以数据库为中心的环境发展为以数据库和数据仓库为中心的环境。