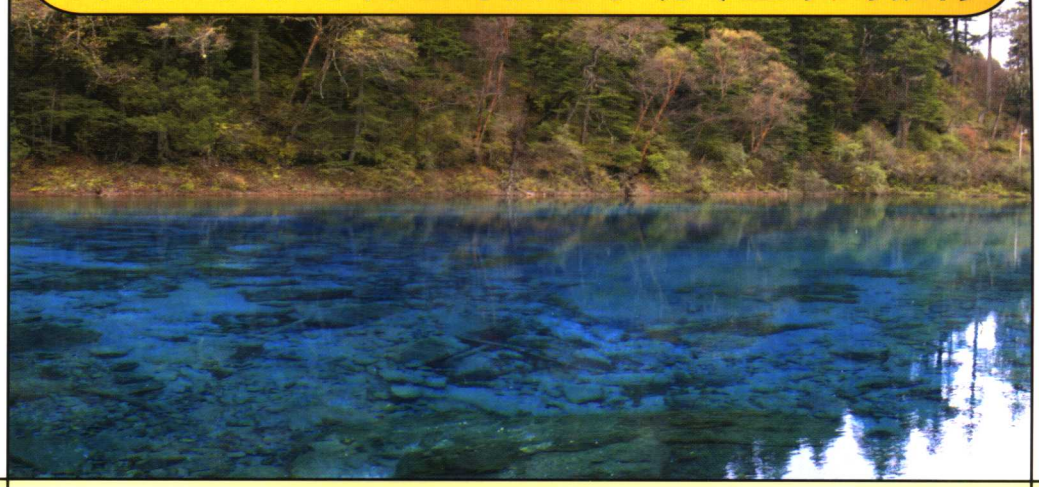
高等院校信息管理与信息系统专业系列教材



数据仓库与数据挖掘教程

陈文伟 编著

清华大学出版社

高等院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教程

陈文伟 编著

清华大学出版社
北 京

内 容 简 介

数据仓库与数据挖掘都是从数据资源提取信息和知识进行辅助决策。由于数据资源丰富，数据仓库与数据挖掘辅助决策效果十分显著。

本书系统介绍数据仓库原理、联机分析处理、数据仓库设计与开发、数据仓库的决策支持应用，数据挖掘原理、信息论的决策树方法、集合论的粗糙集方法、关联规则、公式发现、神经网络、遗传算法、文本挖掘与 Web 挖掘，以及数据仓库与数据挖掘的发展。

本书对数据仓库的系统介绍，在于突出决策支持的本质。对数据挖掘的各类方法均介绍了它们的理论基础和实现方法，并通过例子进行了说明。

本书的特点是从数据仓库和数据挖掘的兴起与演变来说明它们的本质，通过实例来解释它们的原理，这样便于读者学习和掌握，适于本科生和研究生使用。

图书在版编目(CIP)数据

数据仓库与数据挖掘教程/陈文伟编著．—北京：清华大学出版社，2006.8
(高等院校信息管理与信息系统专业系列教材)
ISBN 7-302-13154-6

Ⅰ．数…　Ⅱ．陈…　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．①TP311.13　②TP274

中国版本图书馆 CIP 数据核字(2006)第 059604 号

出 版 者：清华大学出版社　　**地　　址**：北京清华大学学研大厦
http://www.tup.com.cn　　**邮　　编**：100084
社 总 机：010-62770175　　**客户服务**：010-62776969
责任编辑：范素珍
印 装 者：北京鑫海金澳胶印有限公司
发 行 者：新华书店总店北京发行所
开　　本：185×260　**印张**：18.5　**字数**：443 千字
版　　次：2006 年 8 月第 1 版　2006 年 8 月第1 次印刷
书　　号：ISBN 7-302-13154-6/TP・8320
印　　数：1～3000
定　　价：25.00 元

出版说明

20世纪三四十年代，一直摸索着前进的计算技术与刚走向成熟的电子技术结缘。这一结合，不仅孕育了新一代计算工具——电子计算机，还产生了当时谁也没有料到的巨大效应：电子计算机——这种当初为计算而开发出来的工具，很快就超出计算的范畴，成为“信息处理机”的代名词；人类开始能够高效率地开发并利用信息；信息对人类社会的作用得以有效地发挥，并逐步超过材料和能源，成为人类社会的重要支柱；信息产业急剧增长，信息经济高度发展，社会生产力达到了新的高度；人们的信息化意识不断加强，人类在信息资源方面更加激烈地竞争，社会发展走上信息化轨道。

文化是时代的精髓，是特定的人群在一定的历史时期、一定的地域范围对其生产和生活模式、思维和行为方式的觉悟和理性化，它伴随着人类创造和使用工具能力的提高而不断发展。文者，经天纬地也；化者，变化、改变、造化、习俗、风气也。也可以说，文化作为社会的人们在生产和生活中思维和行为方式的理性化，是文治和教化的结果。因此，文化具有区域性、群体性和时代性。在信息时代的帷幕刚刚拉开、新时代的气息开始弥漫社会各个角落的20世纪70年代，先觉们就已开始创办以加速信息化的进程为宗旨、以培养信息资源开发人才为目标的信息管理与信息系统专业。

从与信息有关的学科纵向来看，信息管理与信息系统专业处于信息学、信息技术、信息管理、信息经济、信息社会学这个层次结构的中间，它下以信息学和信息技术为基础，上与信息经济和信息社会学相联系。从其涉及的学科横向来看，它处在管理学、信息科学与技术和有关专业领域的交叉点上。它对技术有极高的要求，又要求对组织有深刻的理解，对行为有合理的组织，反映了科学与人文融合的特点。这种交叉与融合正是信息管理与信息系统专业最重要的特征，是其他的学科或专业难以取代和涵盖的。

我国的信息管理与信息系统专业创建于20世纪70年代末。在近20年的时间里，已发展到151个点，成为培养信息化人才的重要领域。其发展速度之快、影响之深远已令世人和学术界刮目相看。然而作为一个新的、特别是与各行各业关系极为密切的专业，其课程体系、教学内容以及教学方法、手段，都要经历一个逐步完善、逐步成熟的过程，其教材体系的建设更需要较长期的实践和探索。没有这样一个过程，具有专业特点、符合中国实际的教材体系是不会建立的。近20年来，大家一直在课程体系的完善和建设有自己专业特点的教材方面不断进行探讨。1991年，全国10所财经类院校的经济信息管理专业的负责人在太原召开第一次研讨会。以后，1993年在大连、1995年在武汉、1997年在烟台，又有更多的院校参加到了这一研讨之中。这些研讨活动得到了国家教委有关部门的赞许和支持。通过研讨，大家在建设具有专业特点的教材体系、改变简单照搬其他专业教材上取得了共识。在武汉会议之后，即着手进行系列教材的编写工作。经协商，由张基温教授担任主编，由魏晴宇教授、陈禹教授担任顾问。

这套教材是我国信息管理与信息系统专业的第一套教材。尽管编写者为它付出了巨大

的辛劳，但在实践中我们也深深地感到了时代的鞭策和工作的难度。一方面，席卷全球的信息化大潮已经使信息、信息管理、信息系统成为全社会关注的热点，人们对其期望和要求越来越高；另一方面，在世纪之交的今天，作为现代社会先导技术的信息技术和相关学科的更新速度在不断加快，多种社会因素相互渗透、相互影响，新情况、新问题给专业的建设带来很多的困难。当然，这些对我们专业的发展和建设也是一种动力和机遇。为此，在这套教材问世之际，我们再一次表示一个心愿：希望与全国的同行共勉，在教材和专业建设上齐心协力，做出更大贡献。也由于如上种种原因，这套教材不会是完整的，也不会是完美的，一定存在这样那样的不足或错误，我们将会不断补充，不断修改，不断完善。任何建设性意见都是我们非常期盼的。为此，这一套教材将具有充分的开放性：每一本教材都是一个原型，每一条建设性意见都将会被采纳，并享有自己的知识产权。

全国高等院校计算机基础教育研究会
财经信息管理专业委员会
信息管理与信息系统专业系列教材编委会

1997 年 8 月

前　言

数据仓库(data warehouse,DW)是利用数据资源提供决策支持。它比利用模型资源辅助决策更有效,而且辅助决策的范围更宽。由于在现实中,数据大量存在,而且在迅速地增长,只要将面向应用(事务驱动)的数据库重新组织转变为面向决策分析的数据仓库,就可以帮助决策者从不同的视角,通过综合数据分析掌握现状;通过多维数据分析发现各种存在的问题;通过对数据层次的钻取找出问题产生的原因;通过历史数据预测未来。由于数据仓库辅助决策效果明显,数据仓库已经从20世纪90年代中期兴起,经过几年的发展,迅速形成了潮流。

数据挖掘(data mining,DM)是从数据中挖掘出信息和知识,是从人工智能的机器学习(machine learning,ML)中发展起来的。机器学习是让计算机模拟人的学习方法获取知识。机器学习中的大量学习方法已经引入到数据挖掘中。数据挖掘也是20世纪90年代中期兴起的。正是由于数据挖掘具有获取知识的能力,目前各数据仓库均将数据挖掘作为数据仓库的前端分析工具,用于提高数据仓库的决策支持能力。

数据仓库、数据挖掘和联机分析处理(on line analytical processing,OLAP)结合起来的新决策支持系统是以数据驱动的决策支持系统。而传统决策支持系统(decision support system,DSS)是以模型和知识驱动的决策支持系统,是由模型库系统、知识库系统、数据库系统和人机交互系统组成的。新决策支持系统利用的是数据资源,而传统决策支持系统利用的是模型资源和知识资源,它们两者辅助决策的方式和效果均不相同。新决策支持系统并不能代替传统决策支持系统,它们是相互补充的。新决策支持系统与传统决策支持系统结合起来形成的综合决策支持系统将是决策支持系统发展的新方向。

数据仓库、数据挖掘、联机分析处理等结合起来也称为商业智能(business intelligence,BI)。商业智能是一种新的智能技术,区别于人工智能(artificial intelligence,AI)和计算智能(computational intelligence,CI)。人工智能采用的技术是符号推理,符号推理过程形成了概念的推理链。计算智能采用的技术是计算推理,模拟人和生物的模糊推理、神经网络计算和遗传进化过程。商业智能是从数据仓库和数据挖掘中获取信息和知识,对变化的商业环境提供决策支持。商业智能是目前企业界正在大力推广的知识管理(knowledge management,KM)的基础。

作者于1997年6月30日在《计算机世界》报上发表了一组关于数据开采(数据挖掘)的文章,最早向国内学者介绍了数据挖掘概念和技术。作者又于1998年6月15日在《计算机世界》报上发表了一组关于数据仓库与决策支持系统的文章,在介绍基于数据仓库的决策支持系统上,提出了将基于数据仓库的决策支持系统和传统决策支持系统结合的综合决策支持系统,在国内产生了一定的影响。

本书的特点是从数据仓库和数据挖掘的兴起与演变来说明它们的本质,通过例子来解释它们的原理,既系统地介绍了数据仓库和数据挖掘的概念和技术,又介绍了它们之间的关

系，以及今后的发展。

在数据仓库的章节中，重点介绍数据仓库原理、联机分析处理、数据仓库设计与开发、数据仓库的决策支持应用。在数据挖掘的章节中重点介绍信息论方法、集合论方法、公式发现、神经网络和遗传算法，这些数据挖掘方法在现实中应用较广泛。由于数据挖掘的基础理论涉及面较宽，建议在本科生教学中对信息论原理和集合论方法只讲定义和例子，对神经网络和遗传算法只讲公式和应用，省略原理的深层内容和公式的推导。这些省略的内容适合研究生教学。

由于作者从事数据仓库与数据挖掘工作多年，并得到过国家自然科学基金项目的资助。在书中还介绍了作者领导的课题组完成的 IBLE 决策规则树方法、FDD 公式发现系统、遗传分类学习系统 GCLS 等。本书也包含了作者提出的综合决策支持系统概念和可拓数据挖掘概念及理论，这些内容适合研究生学习和参考。

欢迎和广大读者进行交流，共同为促进我国数据仓库和数据挖掘的发展而努力。

参加本书录入的有毕季明、廖建文、赵健、徐怡峰、田昊等同志，在此表示感谢！

陈文伟

2006 年 5 月 29 日于广州

目　录

第1章 数据仓库与数据挖掘概述

1.1 数据仓库的兴起

1.1.1 从数据库到数据仓库

由数据库(DB)发展到数据仓库(DW)主要在于如下几点。

- 数据太多,信息贫乏(data rich, information poor):随着数据库技术的发展,企事业单位建立了大量的数据库,数据越来越多,而辅助决策信息却很贫乏,如何将大量的数据转化为辅助决策信息成了研究的热点。
- 异构环境数据的转换和共享:由于各类数据库产品的增加,异构环境的数据也随之增加,如何实现这些异构环境数据的转换和共享也成了研究的热点。
- 利用数据进行事务处理转变为利用数据支持决策:数据库用于事务处理,若要达到辅助决策,则需要更多的数据。例如,如何利用历史数据的分析来进行预测。对大量数据的综合得到宏观信息等均需要大量的数据。

数据仓库概念提出后,在不到几年的时间内就得到了迅速的发展。数据仓库产品也不断出现并陆续进入市场。

1. 数据库用于事务处理

数据库存储大量的共享数据,作为数据资源用于管理业务中的事务处理,已经成为了成熟的信息基础设施。

数据库中存放的数据基本上是保存当前数据,随着业务的变化随时更新数据库中的数据。例如,学生数据库,随着新生的入校,数据库中要增加新学员的数据记录;随着毕业学生的离校,数据库中要删除这些学员的数据记录。数据库总是保持当前的数据记录。

不同的管理业务需要建立不同的数据库。例如,银行中储蓄业务要建立储蓄数据库,记录所有储蓄用户的存款及使用信息;信用卡业务要建立信用卡数据库,记录所有用户信用卡的存款及使用信息;贷款业务要建立贷款数据库,记录贷款用户的贷款及使用信息。

数据库是为事务处理需求设计和建立的,从而使计算机在事务处理上发挥极大的效果。但是,数据库在帮助人们进行决策分析时就显得不适应了。例如,银行想了解用户的经济状态(收入与支出情况)以及信誉如何(是否超支,还贷情况等)?是否继续贷款给他?单靠一个数据库是无法完成这种决策分析的。必须将储蓄数据库、信用卡数据库、贷款数据库集中起来,对某一个人进行全面分析,才能准确了解他的存款及收支情况、信用卡使用情况以及贷款和还贷情况。这样,银行才能有效地决定是否给此人继续贷款。

同时使用3个数据库进行操作并非是一件简单的事,由于3个管理业务各自独立,在建

立数据库时对同一个人可能使用了不同的编码，对于他的姓名可能有的用汉字，有的用汉语拼音，有的用英文。这为使用 3 个数据库地共同进行决策分析带来了困难。

2. 数据仓库用于决策分析

随着决策分析的需求扩大，兴起了支持决策的数据仓库。它是以决策主题需求集成多个数据库，重新组织数据结构，统一规范编码，使其有效地完成各种决策分析。

从数据库到数据仓库的演变，体现了以下几点。

(1) 数据库用于事务处理，数据仓库用于决策分析

事务处理功能单一，数据库完成事务处理的增加、删除、修改、查询等操作。决策分析要求数据较多。数据仓库需要存储更多的数据，不需要修改数据，主要提取综合数据的信息，以及分析预测数据的信息。

(2) 数据库保持事务处理的当前状态，数据仓库既保存过去的数据又保存当前的数据

数据库中数据随业务的变化一直在更新，总保存当前的数据，如学生数据库。数据仓库中数据不随时间变化而变化，但保留大量不同时间的数据，即保留历史数据和当前数据。

(3) 数据仓库的数据是大量数据库的集成

数据仓库的数据不是数据库的简单集成，而是按决策主题，将大量数据库中数据进行重新组织，统一编码进行集成。如银行数据仓库数据是由储蓄数据库、信用卡数据库、贷款数据库等多个数据库按“用户”主题进行重新组织、编码和集成而建立的。可见，数据仓库的数据量比数据库的数据量大得多。

(4) 对数据库的操作比较明确，操作数据量少。对数据仓库操作不明确，操作数据量大

一般对数据库的操作都是事先知道的事务处理工作，每次操作(增加、删除、修改、查询)涉及的数据量也小，如一个或几个记录数据。

对数据仓库的操作都是根据当时决策需求临时决定而进行的。如比较两个地区某个商品销售的情况。该操作所涉及的数据量很大，不是几个记录数据，而是两个地区多个商店的某商品的所有销售记录。

3. 数据库与数据仓库对比

数据库与数据仓库的对比如表 1.1 所示。

表 1.1　数据库与数据仓库对比

数据库	数据仓库
面向应用	面向主题
数据是详细的	数据是综合的或提炼的
保持当前数据	保存过去和现在的数据
数据是可更新的	数据不更新
对数据操作是重复的	对数据的操作是启发式的
操作需求是事先可知	操作需求是临时决定的

续表

数据库	数据仓库
一个操作存取一个记录	一个操作存取一个集合
数据非冗余	数据时常冗余
操作比较频繁	操作相对不频繁
查询的是原始数据	查询的是经过加工的数据
事务处理需要的是当前数据	决策分析需要过去、现在的数据
很少有复杂的计算	很多复杂的计算
支持事务处理	支持决策分析

1.1.2 从 OLTP 到 OLAP

1. 联机事物处理(on line transaction processing,OLTP)

联机事物处理是在网络环境下的事务处理工作,利用计算机网络技术,以快速的事务响应和频繁的数据修改为特征,使用户利用数据库能够快速地处理具体的业务。OLTP 是事务处理从单机到网络环境发展的新阶段。OLTP 应用要求多个查询并行,以便将每个查询分布到一个处理器上。

OLTP 的特点在于事务处理量大,但事务处理内容比较简单且重复率高。大量的数据操作主要涉及的是一些增加、删除、修改、查询等操作。每次操作的数据量不大且多为当前的数据,OLTP 的数据组织的数据模型采用实体-关系(E-R)模型。

OLTP 处理的数据是高度结构化的,涉及的事务比较简单,数据访问路径是已知的,至少是固定的。事务处理应用程序可以直接使用具体的数据结构,如表、索引等。

OLTP 面对的是事务处理操作人员和低层管理人员。

在过去三十多年中,OLTP 系统发展的目标就是能够处理大量的数据。每时间单位能够处理更多的事务,能支持更多的并发用户,且有更好的系统健壮性。大型的系统每秒能够处理 1000 个以上的事务。有些系统,像机票预订系统,每秒能够处理的事务峰值可以达到 2 万个。

数据库存储的数据量很大,经常每天要处理成千上万的事务,OLTP 在查找业务数据时是非常有效的。但是为高层领导者提供决策分析时,则显得力不从心。

2. 联机分析处理(on line analytical processing,OLAP)

关系数据库之父 E. F. Codd 在 1993 年认为,联机事务处理已经不能满足终端用户对数据库决策分析的需要,决策分析需要对多个关系数据库共同进行大量的综合计算才能得到结果。为此,他提出了多维数据库和多维分析的概念,即联机分析处理概念。关系数据库是二维数据(平面),多维数据库是空间立体数据。

近年来,人们利用信息技术生产和搜集数据的能力大幅度提高,大量的数据库被用于商业管理、政府办公、科学研究和工程开发等,这一势头仍将持续发展下去。于是,一个新的挑战被提出来:在信息爆炸的时代,信息过量几乎成为人人需要面对的问题。如何才能不被

信息的汪洋大海所淹没，从中及时发现有用的知识或者规律，提高信息利用率呢？要想使数据真正成为一个决策资源，只有充分利用它为一个组织的业务决策和战略发展服务才行，否则大量的数据可能成为包袱，甚至成为垃圾。OLAP 是解决这类问题的最有力的工具之一。

OLAP 专门用于支持复杂的分析操作，侧重对分析人员和高层管理人员的决策支持，可以应分析人员的要求快速、灵活地进行大数据量的复杂处理，并且以一种直观易懂的形式将查询结果提供给决策制定人，以便他们准确掌握企业（公司）的经营情况，了解市场需求，制定正确方案，以增加效益。OLAP 软件以它先进的分析功能和以多维形式提供数据的能力，正作为一种支持企业关键商业决策的解决方案而迅速崛起。

OLAP 的基本思想是决策者从多方面和多角度以多维的形式来观察企业的状态和了解企业的变化。

3. OLTP 与 OLAP 的对比

OLAP 是以数据仓库为基础，其最终数据来源与 OLTP 一样均来自底层的数据库系统，但由于二者面对的用户不同，OLTP 面对的是操作人员和低层管理人员，OLAP 面对的是决策人员和高层管理人员，因而数据的特点与处理也明显不同。

OLTP 和 OLAP 是两类不同的应用，它们的各自特点见表 1.2 所示。

表 1.2　OLTP 与 OLAP 对比表

OLTP	OLAP
数据库数据	数据库或数据仓库数据
细节性数据	综合性数据
当前数据	历史数据
经常更新	不更新，但周期性刷新
一次性处理的数据量小	一次性处理的数据量大
对响应时间要求高	响应时间合理
用户数量大	用户数量相对较少
面向操作人员，支持日常操作	面向决策人员，支持决策需要
面向应用，事务驱动	面向分析，分析驱动

1.1.3　数据字典与元数据

1. 数据库的数据字典

数据字典是数据库中各类数据描述的集合，在数据库设计中占有很重要的地位。数据字典通常包括数据项、数据结构、数据流、数据存储和处理过程 5 个部分，其中数据项是数据的最小组成单位。若干个数据项可以组成一个数据结构。数据字典通过对数据项和数据结构的定义来描述数据流、数据存储的逻辑内容。

（1）数据项

数据项是不可再分的数据单位。对数据项的描述通常包括数据项名、数据项含义说明、数据类型、长度、取值范围和取值含义等。