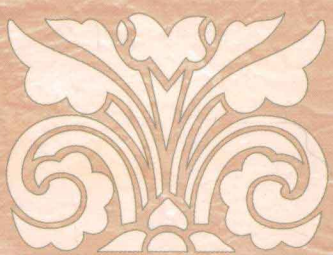
高等院校信息管理与信息系统专业系列教材

数据仓库与数据挖掘教程
（第2版）

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
(第2版)

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是决策支持的两项重要技术，它们共同的特点是都需要利用大量的数据资源，并从数据资源中提取信息和知识。由于数据资源丰富，因此数据仓库与数据挖掘的决策支持效果显著。

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，数据挖掘原理，基于信息论的决策树方法，基于集合论的粗糙集方法、K-均值聚类、关联规则挖掘，仿生物技术的神经网络，遗传算法，公式发现，知识挖掘，文本挖掘与 Web 挖掘。

本书从数据仓库的兴起来说明决策支持的特点，从数据挖掘的理论基础来说明数据挖掘的方法，并通过实例来详细讲解。希望读者在学习之后，亲自在计算机上去实践，这样才能更有效地掌握数据挖掘的方法。

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．—2 版．—北京：清华大学出版社，2011.11
（高等院校信息管理与信息系统专业系列教材）
ISBN 978-7-302-25913-8

Ⅰ．①数…　Ⅱ．①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(2011)第 115742 号

责任编辑：白立军　李玮琪
责任校对：李建庄
责任印制：何　芊

出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
社　总　机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62795954，jsjc@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：三河市君旺印装厂
装 订 者：三河市新茂装订有限公司
经　　销：全国新华书店
开　　本：185×260　　**印　　张**：19.75　　**字　　数**：496 千字
版　　次：2011 年 11 月第 2 版　　**印　　次**：2011 年 11 月第 1 次印刷
印　　数：1～3000
定　　价：29.5 元

产品编号：040341-01

第2版前言

数据仓库(Data Warehouse,DW)和数据挖掘(Data Mining,DM)是决策支持的两项重要技术。在数据仓库中利用多维数据分析来发现问题,并找出产生的原因,能从大量历史数据中预测未来;利用数据挖掘方法能从大量数据中获取知识。两项技术的共同特点是都需要利用大量的数据资源。

数据仓库和数据挖掘是在20世纪90年代中期兴起的,经过十多年的发展,在技术和应用两个方面都得到了很大的提高。为了提高数据仓库的决策支持效果,近年来开展了对综合数据的数据立方体的压缩技术研究,以及对多维数据分析的MDX语言的推广。本书第2版增加了这两项内容。为了强化数据挖掘中神经网络与遗传算法两项实用技术,在第2版中把它们独立列为两章。在神经网络中,按从易到难的顺序将内容重新安排了一下,并增加了径向基函数网络RBF的内容。在遗传算法中增加了进化计算的内容,以便扩大读者的视野。

本书仍保留了按数据仓库的形成过程来讲述其内容的方式,即从数据库到数据仓库以及对比,从联机事务处理OLTP到联机分析处理OLAP以及对比,用它们的对比来突出数据仓库决策支持的作用。按形成过程来讲述,既有利于掌握它们的连贯性,又有利于掌握数据仓库的新特点。

本书保留了依照数据挖掘的理论基础来讲述数据挖掘的方法:大家熟悉的决策树方法实质上是利用信息论中计算信息量的公式来选择属性构造决策树的结点;影响较大的粗糙集方法是典型的利用集合的覆盖原理;关联规则挖掘方法是对相关事务(项)的子集占整个集合的比例,大于阈值时建立关联规则的;在集合论方法中增加了影响最大的K-均值聚类方法。读者在懂得数据挖掘的方法的理论基础后,能够更好地掌握和使用这些方法。

本书第12章由原来的第12章的"数据仓库与数据挖掘的发展"变为"知识挖掘",这一章是全新的内容。第13章做了部分修改,增加了"Web日志分析与实例"一节。

作者从事数据仓库与数据挖掘研究工作多年,在本书第12章中介绍了作者完成的项目——"软件进化规律的知识挖掘",相信能对本科生有启发作用。掌握这些软件进化规律,一来能够帮助学员提高软件使用能力;二来能够引起他们的兴趣,再进一步去挖掘软件进化规律,促进软件进化。本书中也介绍了作者领导的团队完成的项目:IBLE决策规则树方法、FDD公式发现系统、遗传分类学习系统GCLS、变换规则的知识挖掘等。这些内容并不要求本科生掌握,关键在于启发他们如何去创新。这些内容更适合研究生学习和相关行业的工作人员参考。

建议在本科教学中,对信息论原理、集合论方法、神经网络和遗传算法,只讲公式和应用,概略地说明原理的深层内容和公式的推导。这些知识的详细内容适合于研究生教学。

王珊教授曾说过:我觉得数据仓库或者数据挖掘,有时候挖掘出来的东西并不是很有用的,可能要经过很长时间,也许在某些情况下得到一个非常好的结果,能够给领导者一个启

示。但是不会像宣传的那样，我们今天建立了数据仓库系统，明天就能够解决商业竞争中的很多问题，就能取得很大的效益。而且，领导者的素质也是一个重要因素。领导者能不能发现这些问题，技术人员给他的新提示他能不能接受，数据挖掘对他是否有效，等等。这些问题都影响了数据仓库和数据挖掘的效果。

这段话说明了一个问题，数据仓库和数据挖掘的应用比技术有时显得更重要。作者也希望学员在学习这门课程时，除学习原理与技术外，还要加强应用能力的锻炼，即通过计算机去亲自实现它，体会它的真正价值。

欢迎广大读者与作者进行交流，为促进我国数据仓库和数据挖掘的发展而共同努力。

陈文伟

2011 年 9 月于广州

第1版前言

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

由数据库发展到数据仓库，主要特征有如下几点。

- 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为了研究热点。
- 异构环境数据的转换和共享。随之各类数据库产品的增加，异构环境的数据也逐渐增加，如何实现这些异构环境数据的转换和共享也成为了研究热点。
- 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，若要达到辅助决策的目的，则需要更多的数据。例如，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，都需要大量的数据。

数据仓库概念提出后，在短短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。它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的数据，随着业务的变化再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。数据库总是保存当前的数据记录。

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，记录所有贷款用户的贷款及使用信息。

数据库是为满足事务处理需求而设计和建立的，从而使计算机在事务处理上发挥了极大的效果。但是，数据库在帮助人们进行决策分析时就显得不适用了。例如，银行想了解用户的经济状态(收入与支出情况)以及信誉情况(是否超支，还贷情况等)，决定是否继续贷款给他，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及还贷情况。这样，银行才能有效地决定是否给此人继续贷款。

同时使用三个数据库进行操作并非是一件简单的事，由于三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。这为使用三个数据库共同进行决策分析带来了困难。

2. 数据仓库用于决策分析

随着决策分析需求的扩大，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，它不需要修改数据，它主要从大量数据中提取综合信息以及利用历史数据的规律得到预测信息。

(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。

数据库中的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、财务数据库等。数据仓库中的数据不随时间变化而变化，但它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，将大量数据库中的数据进行重新组织，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. 数据库与数据仓库的对比

数据库与数据仓库的对比如表 1.1 所示。

表 1.1 数据库(DB)与数据仓库(DW)对比

数据库(DB)	数据仓库(DW)
面向应用	面向主题
数据是详细的	数据是综合的和历史的
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

续表

数据库(DB)	数据仓库(DW)
操作需求是事先可知的	操作需求是临时决定的
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
查询基本是原始数据	查询基本是经过加工的数据
事务处理需要的是当前数据	决策分析需要过去和现在的数据
很少有复杂的计算	有很多复杂的计算
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

联机事物处理(On Line Transaction Processing,OLTP)是在网络环境下面向交易的事务处理,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。其基本特征是用户的数据可以立即传送到计算中心进行处理,并在很短的时间内给出处理结果。这样做的最大优点是可以实时地处理用户的输入的数据,及时地回答。这样的系统也称为实时系统(Real time System)。

OLTP 主要用于银行业、航空、邮购订单、超级市场和制造业等的输入数据和取回交易数据。例如,银行为分布在各地的自动取款机(ATM)完成即时取款交易;机票预定系统每秒能处理的订票事务峰值可以达到 20 000 个。

OLTP 是事务处理从单机到网络环境的发展新阶段。OLTP 的特点在于事务处理量大,应用要求多个并行处理,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 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 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,在处理业务数据时是非常有效的。

OLTP 面对的是事务处理操作人员和低层管理人员。但是,在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

关系数据库之父 E. F. Codd 在 1993 年提出,联机事务处理(OLTP)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(On Line Analytical Processing,OLAP)概念。关系数据库是二维(平面)数据,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，必须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，增加效益。OLAP 软件以它先进的分析功能和用多维形式提供数据的能力，正作为一种支持企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，它们各自的特点如表 1.2 所示。

表 1.2　OLTP 与 OLAP 对比表

OLTP	OLAP
数据库数据	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
一次处理的数据量小	一次处理的数据量大
对响应时间要求高	响应时间合理
用户数量大	用户数量相对较小
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，它在数据库设计中具有很重要的地位。数