21世纪高等学校计算机专业
核心课程规划教材

数据仓库与数据挖掘

（第二版）

◎ 陈志泊　主编
韩　慧　王建新　孙　俏　聂耿青　编著

清華大學出版社

21世纪高等学校计算机专业
核心课程规划教材

数据仓库与数据挖掘

（第二版）

◎ 陈志泊　主编
韩　慧　王建新　孙　俏　聂耿青　编著

清華大學出版社
北京

内 容 简 介

本书主要介绍数据仓库和数据挖掘技术的基本原理和应用方法。全书共分为12章，主要内容包括数据仓库的概念和体系结构、数据仓库的数据存储和处理、数据仓库系统的设计与开发、关联规则、数据分类、数据聚类、贝叶斯网络、粗糙集、神经网络、遗传算法、统计分析、文本和Web挖掘。

本书既重视理论知识的讲解，又强调应用技能的培养。每章首先介绍算法的主要思想和理论基础，之后利用算法去解决实例中给出的任务，而且对于数据仓库的组建方法和多数章节中的数据挖掘算法，书中都使用Microsoft SQL Server 2005进行了操作实现。通过对具体实例的学习和实践，使读者掌握数据仓库和数据挖掘中必要的知识点，达到学以致用的目的。

本书每章均配有习题，习题形式为选择题、简答题和操作题，可以帮助读者进一步掌握和巩固所学知识。此外，本书提供多媒体教学课件和习题参考答案，读者可到清华大学出版社网站http://www.tup.com.cn/下载。

本书可以作为高等学校计算机及相关专业本科、研究生的数据仓库和数据挖掘教材，也可供相关领域的广大科技工作人员和高校师生参考。

图书在版编目(CIP)数据

数据仓库与数据挖掘/陈志泊主编. —2版. —北京：清华大学出版社，2017
(21世纪高等学校计算机专业核心课程规划教材)
ISBN 978-7-302-48399-1

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(2017)第218791号

责任编辑：刘向威
封面设计：刘　键
责任校对：李建庄
责任印制：李红英

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课 件 下 载：http://www.tup.com.cn，010-62795954

印 装 者：三河市金元印装有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：16.5　　**字　　数**：403千字
版　　次：2009年5月第1版　2017年11月第2版　　**印　　次**：2017年11月第1次印刷
印　　数：18501～20500
定　　价：39.00元

产品编号：076794-01

FOREWORD

前言

随着计算机和信息时代的迅猛发展，人类收集、存储和访问数据的能力大大增强，快速增长的海量数据集已经远远超出了人类的理解能力，传统的数据分析工具也显得力不从心。如何才能不被这些海量数据淹没，而是有效地组织这些数据，并且从中找出有价值的知识，帮助人类制定正确的决策？针对这一问题，数据仓库和数据挖掘技术应运而生，并且显示出强大的生命力。要将海量数据转换成为有用的信息和知识，首先要有效地收集和组织数据。数据仓库是良好的数据收集和组织工具，它的任务是搜集来自各个业务系统的有用数据，存放在一个集成的储存区内。在数据仓库丰富完整的数据基础上，数据挖掘技术可以从中挖掘出有价值的知识，从而帮助决策者正确决策。

本书主要介绍数据仓库和数据挖掘技术的基本原理和应用方法，全书共分为 12 章，主要内容包括数据仓库的概念和体系结构、数据仓库的数据存储和处理、数据仓库系统的设计与开发、关联规则、数据分类、数据聚类、贝叶斯网络、粗糙集、神经网络、遗传算法、统计分析、文本和 Web 挖掘。其中，前 3 章主要介绍数据仓库的基本原理和数据仓库系统的组建方法，后面的章节介绍当前流行的数据挖掘算法的主要思想和理论基础，并且给出丰富的应用实例。

本书紧跟数据仓库和数据挖掘技术的发展和人才培养的目标，有以下几个特点。

（1）可读性强，文字叙述深入浅出，易读易用，即使是初学者，阅读起来也比较容易。

（2）概念清晰，条理清楚，内容取舍合理。

（3）本书强调基础，重视实例。各章节都以经典算法为主，介绍其主要思想和基本原理，并且给出恰当和丰富的实例。

（4）书中实例和课后习题实用、丰富，通过练习，读者可以对各个知识点从不同角度得到训练，掌握和巩固所学知识。

（5）教学资源丰富，本书提供多媒体教学课件和习题参考答案，方便教学。对于上述资源，读者可到清华大学出版社的网站 http://www.tup.com.cn/

下载。

(6) 对于数据仓库的组建方法和多数章节中的数据挖掘算法,本书都使用 Microsoft SQL Server 2005 进行了操作实现,这种做法与市场主流开发工具和技术同步,有利于读者走向社会。

本书各章节之间衔接自然,同时各章节又有一定的独立性,读者可按教材的自然顺序学习,也可以根据实际情况挑选需要的章节学习。

本书可以作为高等学校计算机及相关专业本科、研究生学习数据仓库和数据挖掘的教材,也可供相关领域的广大科技工作人员和高校师生参考。

本书由陈志泊担任主编,第 1～3 章由聂耿青编写,第 5 章、第 6 章和第 11 章由韩慧编写,第 4 章和第 10 章由孙俏编写,第 7～9 章和第 12 章由王建新编写。

由于时间仓促,加之编者水平有限,对于书中不足之处敬请读者批评指正。

编　者

2017 年 8 月

CONTENTS

目录

CHAPTER 1

数据仓库的概念与体系结构 第1章

随着企事业单位信息化建设的逐步完善，各单位信息系统将产生越来越多的历史数据信息。如何处理这些历史数据呢？现各单位至少有如下三种做法。

(1) 将已经失效的历史数据简单地删除，以便减少磁盘的占用空间并提高系统性能。这种方法最简单。

(2) 先对历史数据作介质备份，然后删除，以防万一需要查看。

(3) 建立一个数据仓库系统，将各业务系统及其他档案数据中有分析价值的数据及需要存档的数据保存到数据仓库中，进而可以综合利用这些数据，建立分析模型，从中挖掘出符合规律的知识并用于未来的预测与决策中。

一方面，各信息化单位正逐步认识到这些历史业务数据就是金矿石，可以从中炼出金子来，因此越来越多的单位开始建立自己的数据仓库与数据挖掘系统，以从中淘出“金子”来。事实上，业务数据的积累年限越长，越容易发现规律，形成知识。

另一方面，基于 Web 的商务应用越来越普及，客户和供应商在商务网站上的活动提供了大量的点击流数据，通过分析可以进一步了解访问者的行为偏好，发现带普遍性的消费行为规律。同时，通过网站日志还可进一步获得访问者的活动细节，如时间、IP 地址、经常访问的页面和内容、在网页上的停留时间等。如果将这些数据连同客户的交易、付款、产品利润、业务查询等历史记录都从各业务系统中合并到数据仓库中，将可以进一步改进网站页面内容和风格，让客户和业务伙伴更加满意，甚至带来利润更高的相关业务。

1.1 数据仓库的概念、特点与组成

数据仓库(data warehouse)通常指一个数据库环境，而不是指一件产品，它提供用户用于决策支持的当前和历史数据，这些数据在传统的数据库中通常不方便得到。

简单地说,数据仓库就是一个面向主题(subject oriented)的、集成(integrate)的、相对稳定(non-volatile)的、反映历史变化(time variant)的数据集合,通常用于辅助决策支持。

1.1.1 数据仓库的特点

1. 面向主题

操作型数据库中的数据针对事务处理任务,各个业务系统之间各自分离;而数据仓库中的数据是按照一定的主题域进行组织的。主题是一个抽象的概念,是指用户使用数据仓库进行决策时所关心的重点领域,例如顾客、供应商和产品等。一个主题通常与多个操作型数据库相关。

2. 集成

操作型数据库通常与某些特定的应用相关,数据库之间相互独立,并且往往是异构的;而数据仓库中的数据是在对原有分散的数据库数据作抽取、清理的基础上经过系统加工、汇总和整理得到的。所以,必须消除源数据中的不一致性,以保证数据仓库内的信息是关于整个企事业单位一致的全局信息。也就是说,存放在数据仓库中的数据应使用一致的命名规则、格式、编码结构和相关特性来定义。

3. 相对稳定

操作型数据库中的数据通常实时更新,数据根据需要及时发生变化。数据仓库的数据主要用于决策分析,其所涉及的数据操作主要是数据查询和定期更新,一旦某个数据加载到数据仓库以后,一般情况下将作为数据档案长期保存,几乎不再做修改和删除操作。也就是说,针对数据仓库,通常有大量的查询操作及少量定期的更新操作。

4. 反映历史变化

操作型数据库主要关心当前某一个时间段内的数据,而数据仓库中的数据通常包含较久远的历史数据,因此总是包括一个时间维,以便可以研究趋势和变化。数据仓库系统通常记录了一个单位从过去某一时期到目前的所有时期的信息,通过这些信息,可以对单位的发展历程和未来趋势作出定量分析和预测。

1.1.2 数据仓库的组成

1. 数据仓库数据库

数据仓库数据库是整个数据仓库环境的核心,是数据信息存放的地方,对数据提供存取和检索支持。相对于传统数据库来说,其突出的特点是对海量数据的支持和快速的检索技术。

2. 数据抽取工具

数据抽取工具把数据从各种各样的存储环境中提取出来,进行必要的转化、整理,再存放到数据仓库内。对各种不同数据存储方式的访问能力是数据抽取工具的关键,可以运用

高级语言编写的程序、操作系统脚本、批命令脚本或 SQL 脚本等方式访问不同的数据环境。数据转换通常包括如下内容。

(1) 删除对决策分析没有意义的数据。

(2) 转换到统一的数据名称和定义。

(3) 计算统计和衍生数据。

(4) 填补缺失数据。

(5) 统一不同的数据定义方式。

3. 元数据

元数据是描述数据仓库内数据的结构和建立方法的数据。元数据为访问数据仓库提供了一个信息目录,这个目录全面描述了数据仓库中有什么数据、这些数据是怎么得到的、怎么访问这些数据。元数据是数据仓库运行和维护的中心内容,数据仓库系统对数据的存取和更新都需要元数据信息。根据元数据用途的不同可将元数据分为技术元数据和业务元数据两类。

(1) 技术元数据是数据仓库的设计和管理人员在开发和管理数据仓库时使用的元数据,包括数据源信息、数据转换的描述、数据仓库内对象和数据结构的定义、数据清理和数据更新时用的规则、源数据到目的数据的映射表,以及用户访问权限、数据备份历史记录、数据导入历史记录和信息发布历史记录等。

(2) 业务元数据是从单位业务的角度描述数据仓库的元数据,例如业务主题的描述,即业务主题包含的数据、查询及报表等信息。

4. 访问工具

访问工具是为用户访问数据仓库提供的手段,如数据查询和报表工具、应用开发工具、数据挖掘工具和数据分析工具等。

5. 数据集市(Data Mart)

数据集市是为了特定的应用目的,从数据仓库中独立出来的一部分数据,也称为部门数据或主题数据。在数据仓库的实施过程中往往可以从一个部门的数据集市着手,再逐渐用几个数据集市组成一个完整的数据仓库。需要注意的是,在实施不同的数据集市时,相同含义字段的定义一定要相容,以免未来实施数据仓库时出现问题。

6. 数据仓库管理

数据仓库管理包括安全与权限的管理、数据更新的跟踪、数据质量的检查、元数据的管理与更新、数据仓库使用状态的检测与审计、数据复制与删除、数据分割与分发、数据备份与恢复、数据存储管理等。

7. 信息发布系统

信息发布系统用于把数据仓库中的数据或其他相关的数据发送给不同的地点或用户。基于 Web 的信息发布系统是当前流行的多用户访问的最有效方法。

1.2 数据挖掘的概念与方法

数据挖掘(Data Mining)就是从大量数据中获取有效的、新颖的、潜在有用的、最终可理解的模式的过程。简单地说,数据挖掘就是从大量数据中提取或"挖掘"知识,又被称为数据库中的知识发现(Knowledge Discovery in Database, KDD)。

1.2.1 数据挖掘的分析方法

数据挖掘的分析方法可以分为直接数据挖掘与间接数据挖掘两类。

直接数据挖掘的目标是利用可用的数据建立一个模型,这个模型对剩余的数据(例如对一个特定的变量)进行描述,包括分类(classification)、估值(estimation)和预测(prediction)等分析方法。

在间接数据挖掘的目标中,没有选出某一具体的变量并用模型进行描述,而是在所有的变量中建立起某种关系,如相关性分组(affinity grouping)或关联规则(association rules)、聚类(clustering)、描述和可视化(description and visualization)及复杂数据类型的挖掘,如文本、网页、图形图像、音视频和空间数据等。

在后续的章节会详细介绍有关的数据挖掘分析方法。

1.2.2 数据仓库与数据挖掘的关系

若将数据仓库比作矿井,那么数据挖掘就是深入矿井采矿的工作。数据挖掘不是一种无中生有的魔术,也不是点石成金的炼金术,若没有足够丰富完整的数据,将很难期待数据挖掘能挖掘出什么有意义的信息。

要将庞大的数据转换成为有用的信息和知识,必须要先有效地收集数据。功能完善的数据库管理系统事实上是最好的数据收集工具,数据仓库的一个重要任务就是搜集来自其他业务系统的有用数据,存放在一个集成的储存区内。

决策者利用这些数据作决策,即从数据仓库中挖掘出对决策有用的信息与知识,是建立数据仓库与进行数据挖掘的最大目的。只有数据仓库先行建立完成,且数据仓库所含数据是干净(不会有虚假错误的数据掺杂其中)、完备和经过整合的,数据挖掘才能有效地进行,因此,从一定意义上可将两者的关系解读为数据挖掘是从数据仓库中找出有用信息的一种过程与技术。

1.3 数据仓库的技术、方法与产品

数据仓库技术是为了有效地把操作型数据集成到统一的环境中以提供决策型数据访问的各种技术和模型的总称。

1.3.1 OLAP 技术

1. 联机事务处理与联机分析处理的比较

数据处理通常分成两大类:联机事务处理(On-Line Transaction Processing,OLTP)和

联机分析处理(On-Line Analytical Processing,OLAP)。

OLTP是传统的操作型数据库系统的主要应用,主要是一些基本的日常事务处理,如银行柜台存取款、股票交易和商场POS系统等。OLAP是数据仓库系统的主要应用,支持复杂的分析操作,侧重决策支持,并且提供直观易懂的查询结果。表1.1列出了OLTP与OLAP之间的区别。

表1.1 OLTP与OLAP的比较

	OLTP	OLAP
用户	操作人员、低层管理人员	决策人员、高级管理人员
功能	日常操作型事务处理	分析决策
数据库设计目标	面向应用	面向主题
数据特点	当前的、最新的、细节的、二维的与分立的	历史的、聚集的、多维的、集成的与统一的
存取规模	通常一次读或写数十条记录	可能读取百万条以上记录
工作单元	一个事务	一个复杂查询
用户数	通常是成千上万个用户	可能只有几十个或上百个用户
数据库大小	通常在GB级(100MB~1GB)	通常在TB级(100GB~1TB及以上)

2. OLAP技术的有关概念

(1) 多维数据集。多维数据集是联机分析处理的主要对象,它是一个数据集合,通常从数据仓库的子集构造,并组织汇总成一个由一组维度和度量值定义的多维结构。

(2) 维度。维度是OLAP技术的核心,即人们观察客观世界的角度,通过把一个实体的一些重要属性定义为维(dimension),使用户能对不同维属性上的数据进行比较研究。因此,"维"是一种高层次的类型划分,一般都包含层次关系,甚至相当复杂的层次关系。例如,一个企业在考虑产品的销售情况时,通常从时间、销售地区和产品等不同角度来深入观察产品的销售情况。这里的时间、地区和产品就是维度。而这些维的不同组合和所考查的度量值(如销售额)共同构成的多维数据集则是OLAP分析的基础。

(3) 度量值。度量值也叫度量指标,是多维数据集中的一组数值,这些值基于多维数据集的事实数据表中的一列,是最终用户浏览多维数据集时重点查看的数值数据,也是所分析的多维数据集的中心值。如销售量、成本值和费用支出等都可能成为度量值。

(4) 多维分析。多维分析是指对以"维"形式组织起来的数据(多维数据集)采取切片(slice)、切块(dice)、钻取(drill down和roll up等)和旋转(pivot)等各种分析动作,以求剖析数据,使用户能从不同角度、不同侧面观察数据仓库中的数据,从而深入理解多维数据集中的信息。多维分析操作通常包括如下内容。

① 钻取可以改变维的层次、变换分析的粒度,包括向上钻取(roll up)、向下钻取(drill down)、交叉钻取(drill across)和钻透(drill through)等。向上钻取即减少维数,是在某一维上将低层次的细节数据概括到高层次的汇总数据;而向下钻取则正好相反,它从汇总数据深入到细节数据进行观察,增加了维数。

② 切片和切块是在一部分维上选定值后,度量值在剩余维上的分布。如果剩余维有两个则是切片,如果有三个则是切块。

③ 旋转是变换维的方向,即在表格中重新安排维的放置,例如行列互换。