21世纪高等学校规划教材 | 计算机科学与技术

数据仓库与数据挖掘教程

黄德才 编著

清华大学出版社

数据仓库与数据挖掘教程

黄德才 编著

清华大学出版社
北京

内 容 简 介

本书较详细地介绍了数据仓库和数据挖掘的原理、方法及应用技术。全书共有14章，分为4篇。第1章为绪论篇，介绍数据仓库与数据挖掘的基本概念及其相互关系；第2～6章为数据仓库原理及应用篇，主要介绍数据仓库的概念模型、逻辑模型和物理模型，以及数据仓库的规划、设计、实施和OLAP应用等；第7～10章为传统数据挖掘原理及算法篇，介绍数据的属性类型与相似性度量、关联规则挖掘、分类规则挖掘、聚类分析和离群点挖掘算法等；第11～14章为数据挖掘创新篇，主要内容取自编者近年指导研究生发表的学术论文，并根据教学需要进行适当补充修改而成，包括混合属性数据、数据流和不确定数据的聚类分析，以及量子遗传聚类算法等。

本书可作为普通高等院校计算机专业与IT相关专业高年级本科生和研究生的教材，也可作为经济管理类专业同名课程的教材和参考书，还可作为电子商务、金融保险等行业数据管理与数据分析人员的培训教材或自学参考书。

图书在版编目(CIP)数据

数据仓库与数据挖掘教程/黄德才编著. —北京：清华大学出版社，2016
21世纪高等学校规划教材·计算机科学与技术
ISBN 978-7-302-43412-2

Ⅰ. ①数… Ⅱ. ①黄… 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(2016)第075232号

责任编辑：郑寅堃　李　晔
封面设计：傅瑞学
责任校对：梁　毅
责任印制：宋　林

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：北京鑫海金澳胶印有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：26.5　　**字　　数**：657千字
版　　次：2016年8月第1版　　**印　　次**：2016年8月第1次印刷
印　　数：1～1500
定　　价：54.90元

产品编号：066534-01

出版说明

随着我国改革开放的进一步深化，高等教育也得到了快速发展，各地高校紧密结合地方经济建设发展需要，科学运用市场调节机制，加大了使用信息科学等现代科学技术提升、改造传统学科专业的投入力度，通过教育改革合理调整和配置了教育资源，优化了传统学科专业，积极为地方经济建设输送人才，为我国经济社会的快速、健康和可持续发展以及高等教育自身的改革发展做出了巨大贡献。但是，高等教育质量还需要进一步提高以适应经济社会发展的需要，不少高校的专业设置和结构不尽合理，教师队伍整体素质亟待提高，人才培养模式、教学内容和方法需要进一步转变，学生的实践能力和创新精神亟待加强。

教育部一直十分重视高等教育质量工作。2007 年 1 月，教育部下发了《关于实施高等学校本科教学质量与教学改革工程的意见》，计划实施“高等学校本科教学质量与教学改革工程”（简称“质量工程”），通过专业结构调整、课程教材建设、实践教学改革、教学团队建设等多项内容，进一步深化高等学校教学改革，提高人才培养的能力和水平，更好地满足经济社会发展对高素质人才的需要。在贯彻和落实教育部“质量工程”的过程中，各地高校发挥师资力量强、办学经验丰富、教学资源充裕等优势，对其特色专业及特色课程（群）加以规划、整理和总结，更新教学内容、改革课程体系，建设了一大批内容新、体系新、方法新、手段新的特色课程。在此基础上，经教育部相关教学指导委员会专家的指导和建议，清华大学出版社在多个领域精选各高校的特色课程，分别规划出版系列教材，以配合“质量工程”的实施，满足各高校教学质量和教学改革的需要。

为了深入贯彻落实教育部《关于加强高等学校本科教学工作，提高教学质量的若干意见》精神，紧密配合教育部已经启动的“高等学校教学质量与教学改革工程精品课程建设工作”，在有关专家、教授的倡议和有关部门的大力支持下，我们组织并成立了“清华大学出版社教材编审委员会”（以下简称“编委会”），旨在配合教育部制定精品课程教材的出版规划，讨论并实施精品课程教材的编写与出版工作。“编委会”成员皆来自全国各类高等学校教学与科研第一线的骨干教师，其中许多教师为各校相关院、系主管教学的院长或系主任。

按照教育部的要求，“编委会”一致认为，精品课程的建设工作从开始就要坚持高标准、严要求，处于一个比较高的起点上。精品课程教材应该能够反映各高校教学改革与课程建设的需要，要有特色风格、有创新性（新体系、新内容、新手段、新思路，教材的内容体系有较高的科学创新、技术创新和理念创新的含量）、先进性（对原有的学科体系有实质性的改革和发展，顺应并符合 21 世纪教学发展的规律，代表并引领课程发展的趋势和方向）、示范性（教材所体现的课程体系具有较广泛的辐射性和示范性）和一定的前瞻性。教材由个人申报或各校推荐（通过所在高校的“编委会”成员推荐），经“编委会”认真评审，最后由清华大学出版

社审定出版。

目前，针对计算机类和电子信息类相关专业成立了两个“编委会”，即“清华大学出版社计算机教材编审委员会”和“清华大学出版社电子信息教材编审委员会”。推出的特色精品教材包括：

(1) 21 世纪高等学校规划教材·计算机应用——高等学校各类专业，特别是非计算机专业的计算机应用类教材。

(2) 21 世纪高等学校规划教材·计算机科学与技术——高等学校计算机相关专业的教材。

(3) 21 世纪高等学校规划教材·电子信息——高等学校电子信息相关专业的教材。

(4) 21 世纪高等学校规划教材·软件工程——高等学校软件工程相关专业的教材。

(5) 21 世纪高等学校规划教材·信息管理与信息系统。

(6) 21 世纪高等学校规划教材·财经管理与应用。

(7) 21 世纪高等学校规划教材·电子商务。

(8) 21 世纪高等学校规划教材·物联网。

清华大学出版社经过三十多年的努力，在教材尤其是计算机和电子信息类专业教材出版方面树立了权威品牌，为我国的高等教育事业做出了重要贡献。清华版教材形成了技术准确、内容严谨的独特风格，这种风格将延续并反映在特色精品教材的建设中。

清华大学出版社教材编审委员会

联系人：魏江江

E-mail：weijj@tup.tsinghua.edu.cn

前言

随着计算机、网络和通信等信息技术的发展，数据采集的方法越来越丰富，存储设备的容量不断提升而成本逐年下降，特别是数据库技术在各行各业的普及应用，使人类积累了海量的数据，步入了大数据时代却陷入了“数据丰富、知识贫乏”的困境。人们迫切希望从所拥有的数据中获取有用的知识，以帮助其更好地进行有效决策。数据仓库与数据挖掘就是20世纪90年代兴起的两项独立的决策支持新技术，并在商业零售、金融保险、银行、电信等行业得到成功应用，“数据仓库与数据挖掘”已成为普通高等院校计算机、经贸管理和信息类相关专业研究生和高年级本科生的学位课程或选修科目。

笔者所在学校计算机学院从2005年开始讲授“数据仓库与数据挖掘”课程，但每学期的教材选择一直都是一件困难的事情。因为这是一门涉及交叉学科的新课程，并且由两个相对独立的知识体系构成，加之当时国内只有很少从国外引进的数据仓库或数据挖掘专著译本，如William H. Inmon于1993年出版的《建立数据仓库》，Jiawei Han等于2001年出版的《数据挖掘——概念与技术》等，作为教学参考书。虽然此后又陆续有一些国外专著通过翻译或以影印的方式在国内出版，但也因知识内容过多，叙述过于简练，加之涉及过多研究前沿，原则上都不太适宜当作该领域的入门教材。此后国内许多学者也相继编写了数据仓库与数据挖掘方面的教材，但受到我国高等教育长期以学术理论培养为主的教学模式影响，其数据仓库部分大多偏重概念和原理的叙述，缺乏建立数据仓库的实例，而数据挖掘算法则偏重算法结构的描述，缺少计算实例演示，常常让初学者感到理解困难。

基于笔者十余年“数据仓库与数据挖掘”的教学实践，按照教育部关于高等学校本科教育以培养更多应用型人才为目标的教学改革方向，以及全日制研究生以学术型和专业型两大类进行有差别培养的要求，我们迫切需要一本在教学时数限制严格的条件下，理论叙述深入浅出、实际应用具体完整；算法描述自然易懂，计算实例详略得当的数据仓库与数据挖掘方面的教材。

本教程正是在这种社会需求背景和实际教学需要的情况下，在总结十余年教学改革与实践经验所编写的讲义基础上修改而成的。教程兼顾了应用型人才与学术型人才培养的需求，并以一个完整具体的警务数据仓库实例，介绍数据仓库原理、数据仓库设计和实现方法，为读者真正架起了理论与实践的桥梁；还以大量的计算实例来增加读者对数据挖掘原理及其各种挖掘算法的理解深度。教学实践表明，这种以实际应用需求和计算实例驱动的教学组织方法，可提高学生阅读的“幸福指数”，激发学生的学习兴趣，增强学生的实际应用能力，促进学生对理论知识的理解和掌握，并总体上提高教学质量和教学效果。如果把早期出版的数据仓库、数据挖掘专著和教材比作“鲜牛奶”的话，作者更希望本书呈现给读者的是一份营养丰富且易于消化吸收的“酸牛奶”。

全书共14章，分为4篇，其主要内容的教学时数可以控制在32～48学时之间，另有15%左右的篇幅作为学生课外阅读内容。第1章为绪论篇，内容包括数据仓库概述、数据挖

掘概述、数据仓库与数据挖掘的联系与区别等；第2～6章为数据仓库原理及应用篇，内容包括数据预处理技术、数据仓库的概念模型、逻辑模型、物理模型，数据仓库的规划、数据仓库设计、数据仓库实施、数据仓库系统开发和OLAP技术等，并特别介绍了一个警务数据仓库的实现过程与OLAP应用实例；第7～10章为传统数据挖掘原理及算法篇，内容包括数据的属性类型、相似性与相异性度量、关联规则的Apriori算法、FP-增长算法、关联规则的评价和序列模式发现算法；分类问题的k-最近邻算法、决策树方法和贝叶斯方法；聚类分析的划分聚类算法、层次聚类方法、密度聚类方法、聚类的质量评价和离群点挖掘算法等；第11～14章为数据挖掘创新篇，内容包括混合属性数据集的聚类算法、数据流挖掘的聚类算法、不确定数据的聚类算法、量子计算与量子遗传聚类算法等。它们都取材于作者近年指导研究生发表的学术论文，因此在保留一些学术论文写作风格的同时，还根据教学需要进行了较多的补充和修改。

本书的编写得到了清华大学出版社和作者的同事及研究生的大力支持；杨良怀教授、陆亿红副教授和范玉雷博士对本书的出版给予了很大的帮助，他们分别参与了部分章节的讨论，杨教授还提供了一些有价值的参考资料；温州大学城市学院沈良忠副教授和笔者指导的博士生刘世华讲师为教程提供了警务数据仓库和OLAP应用的实例；笔者指导的博士生郑祺讲师、已经毕业的钱国红、钱潮恺和潘冬明硕士，以及在读硕士生夏聪、翁纯佳、周海松、张振宁、王骏、谷宗昌、吕存伟、任胜亮和魏方圆等都分别参与了部分章节的校对工作，在此一并向他们表示衷心感谢。此外，教程中还参考或引用了许多文献资料的部分内容，谨向这些文献的作者表示衷心的感谢和深深的敬意，对于那些没能在此一一提到名字的同事和研究生给予的支持，也表示深深的谢意。

限于作者水平，加之数据仓库与数据挖掘理论技术的内容十分丰富，且发展非常迅速，疏漏和不当之处在所难免，殷切希望广大师生和读者批评指正。

作者的电子邮件地址是：hdc@zjut.edu.cn。

黄德才

2015年12月于杭州

目 录

第1章 绪 论

本章简述数据仓库与数据挖掘相关的基本概念和引导性知识，其目的是为后续章节的学习做好基础知识的储备，并起到穿针引线的作用。

1.1节介绍传统数据库与事务处理、决策支持与分析处理之间的关系，特别是当事务处理与分析处理共享一个数据库系统时带来的读写冲突等问题，从而导出数据仓库的概念与特征。随后介绍数据仓库系统及其基本体系结构，以及数据仓库数据的逻辑组织方式。

1.2节介绍数据挖掘技术产生的背景，数据挖掘与知识发现概念的同一性、数据挖掘的数据来源、数据挖掘的任务和挖掘步骤，然后介绍数据挖掘技术的常见应用领域。

1.3节不仅分析了决策支持是数据仓库与数据挖掘的共同任务，而且分析比较了它们之间的不同点。虽然说数据仓库不是为数据挖掘而生的，数据挖掘也不是为数据仓库而活的，即它们是两个相对独立的知识体系，但两者都是为决策支持这一中心服务的，并具有“富矿金山”与“开采工具”之间的互补关系，即它们是决策支持这个中心的两个基本点。这就使我们能够充分利用数据仓库与数据挖掘技术的互补性，再加上其他数据分析技术来构造一个完美的决策支持系统环境。

1.4节简要介绍教程后续章节所涉及的主要知识点，并给出了教学时数分配的建议。

1.1 数据仓库概述

一般来说，计算机数据处理主要有两种方式：操作型处理和分析型处理。

1.1.1 从传统数据库到数据仓库

1. 传统数据库与操作型处理

数据库(DataBase，DB)是长期存储在计算机内的、有组织的、可共享的数据集合。其理论产生于20世纪60年代。在20世纪70年代之前的数据库技术称为第一代，支持层次数据模型和网状数据模型。20世纪70年代开始出现了关系数据库理论和关系数据库管理系统[1]。由于有严格的数学理论支持，关系数据库管理系统迅速取代了层次和网状数据库管理系统，并在商业领域得到普及应用，长盛不衰，至今枝繁叶茂。为了与数据仓库相区别，人们把现在普遍使用的关系数据库称为传统数据库，或操作型数据库。

从20世纪80年代开始，随着信息技术的发展，特别是数据库技术的成熟和个人计算机(PC)的诞生，基于数据库技术的数据收集、存储管理和检索利用的数据库应用系统，比如财

务管理系统、超市管理系统、户籍管理系统、宾馆住宿管理系统等，在企事业单位得到了广泛应用，20 世纪 90 年代时就取得了巨大的成功。

按照现在计算机数据处理方式的分类，像财务管理和超市管理这类数据库应用系统都称为联机事务处理(On-Line Transaction Processing)系统，简称 OLTP 系统，其数据存储在传统数据库中。它的核心任务是对传统数据库(也称为事务处理数据库或 OLTP 数据库)进行联机的日常操作，因此称为操作型处理，它们通常是对一个或一组记录的查询和修改操作，主要为企事业单位的特定数据管理和应用服务。用户希望在保证数据安全性和完整性的前提下，每次操作能够实时响应。

2. 传统决策支持与分析处理

随着时间的推移，事务处理数据库中的数据不断地积累增长，其数据量也从 20 世纪 80 年代的兆(M)字节或千兆字节(GB)级别跃升到兆兆字节(TB)级别，并且分布在不同的系统平台上，还具有多种存储形式，而这些数据却承载着企业生产和经营管理的大量信息。由于经济全球化导致市场竞争更加激烈，用户(即管理决策人员)已经不满足于企业仅仅用计算机去处理每天所产生的事务数据，而是希望通过对自身已经积累的大量数据进行分析，提取能够支持决策的关键信息。因此，能否从企事业单位这些纷繁复杂、大量历史数据环境中得到有用的决策支持信息，已成为企业生存、发展和壮大的重要任务，也成为社会管理和公共服务的必备手段。

比如，为了确保社会公共安全，特别是重大节假日期间，公安局、派出所等公共安全部门就需要知道历史上，特别是最近一段时间内，辖区内所有宾馆登记入住人员的情况，比如入住人次、住宿天数，哪些地方来的人多，哪个宾馆有多少前科人员入住等，以提前做好公共安全防护预案和警力部署。

这种对当前和大量历史数据的统计分析，并从中提取管理决策所需重要信息的数据处理方法，称为数据的分析处理或分析型处理，并将能够完成这种分析处理任务的计算机系统称为决策支持系统(Decision Support System，DSS)，而将决策支持系统分析所得到数据信息，提供给企事业单位董事会或主管领导决策参考的过程称为决策支持。鉴于分析处理的结果就是用于决策支持，因此，决策支持系统有时也称为分析型处理系统或分析处理系统，它通常需要对大量历史数据进行长时间的分析处理。用户对分析处理的时间长短并不十分在意，而对数据分析的深度和广度，以及分析结果的使用价值非常重视。

早期的分析处理系统都是在联机事务处理系统中，直接增加一些统计分析软件或决策支持程序来实现的(见图 1-1)。

从图 1-1 可以看出，分析型处理的传统方法，就是在事务处理(OLTP)的数据库系统环境中，直接增加分析型处理软件或程序来构成的分析型处理系统，它是一种“事务处理”+“分析处理”的“2 合 1(two in one)”系统，其明显的优点是投资少，见效快。但随着企业的快速发展，特别是那些全球化跨国经营的企业，不仅事务处理的数据量增长迅速，而且对决策分析处理的方法以及分析的深度都提出了更高的要求，常常导致两种不同的处理方式产生读写等冲突，比如分析处理锁住了一张表且正在进行统计分析，而事务处理希望立即对表中的部分数据进行修改，严重时根本无法满足企业管理和决策分析的实际需要。