21世纪高等学校电子商务专业规划教材

李春葆 蒋林 陈良臣 喻丹丹 曾平 编著

数据仓库与数据挖掘应用教程

清华大学出版社

21世纪高等学校电子商务专业规划教材

李春葆 蒋林 陈良臣 喻丹丹 曾平 编著

数据仓库与数据挖掘应用教程

清华大学出版社
北京

内 容 简 介

本书以 SQL Server 分析服务为环境介绍数据仓库和数据挖掘应用技术，包括数据仓库和数据挖掘概述、OLAP 和多维数据模型、数据仓库设计和 SQL Server 数据仓库开发实例、关联分析算法、决策树分类算法、贝叶斯分类算法、神经网络算法、回归分析算法、时间序列分析和聚类算法。

本书内容翔实，循序渐进地介绍各个知识点，并提供全面而丰富的教学资源，可作为各类高等院校计算机及相关专业“数据仓库和数据挖掘应用技术”和“SQL Server 高级应用”课程的教学用书，也适合计算机应用人员和计算机爱好者参考。

图书在版编目(CIP)数据

数据仓库与数据挖掘应用教程/李春葆等编著. --北京：清华大学出版社，2016
21 世纪高等学校电子商务专业规划教材
ISBN 978-7-302-43077-3

Ⅰ. ①数…　Ⅱ. ①李…　Ⅲ. ①数据库系统－教材 ②数据采集－教材　Ⅳ. ①TP311.13 ②TP274

中国版本图书馆 CIP 数据核字(2016)第 034105 号

责任编辑：魏江江　薛　阳
封面设计：常雪影
责任校对：时翠兰
责任印制：沈　露

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 刷 者：北京季蜂印刷有限公司
装 订 者：三河市溧源装订厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：20　　**字　　数**：509 千字
版　　次：2016 年 10 月第 1 版　　**印　　次**：2016 年 10 月第1 次印刷
印　　数：1～2000
定　　价：39.50 元

产品编号：067892-01

出版说明

电子商务是以信息网络技术为手段，以商品交换为中心的商务活动，是“互联网＋”的杰作之一。特别是在2015年年初的政府工作报告中，李克强总理首次提出“制定‘互联网＋’行动计划”，极大地推进了电子商务在我国的蓬勃发展，改造和影响着众多传统行业。电子商务系统是保证以电子商务为基础的网上交易实现的体系。

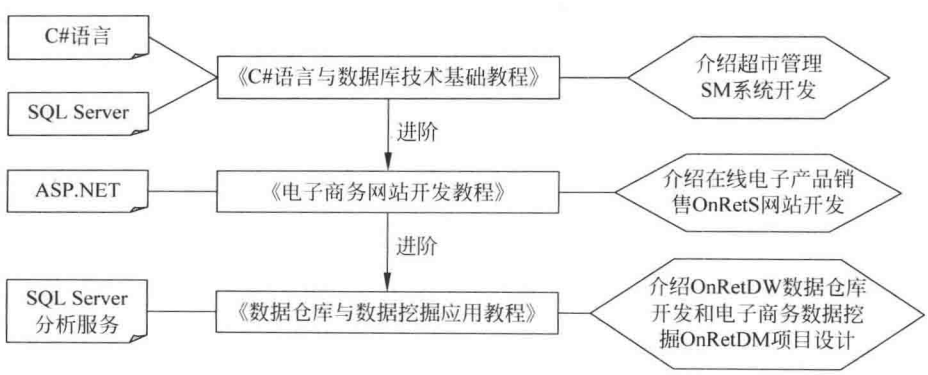
电子商务应用的快速发展，需要大量的专业技术人员，据专家测算，未来10年我国电子商务人才缺口将达到200万。为了加快电子商务系统人才培养，我们以国家卓越工程师计划为契机规划并出版本系列教材。

电子商务网站是电子商务系统的核心，电子商务网站开发涉及多方面的技术。本系列教材以Visual Studio为开发环境，以案例为向导，全面介绍电子商务网站的开发技术，涵盖的教程如下：

- C＃语言与数据库技术基础教程
- 电子商务网站开发教程
- 数据仓库与数据挖掘应用教程

本系列教材具有专业培养定位清晰、可操作性强的特点。《C＃语言与数据库技术基础》从零基础开始，循序渐进地介绍C＃语言的基本语法、面向对象编程、Windows窗体应用程序设计、SQL Server数据库操作、C＃访问数据库方法以及Windows界面的电子商务系统开发技术。《电子商务网站开发教程》以ASP. NET为背景，介绍动态网站的开发技术。《数据仓库与数据挖掘应用教程》介绍数据仓库数据和电子商务数据分析技术。

教程中涉及的相关案例如下。



本系列教材是武汉大学计算机学院和解放军理工大学在探索电子商务人才培养并结合国家卓越工程师计划的教学实践，总结和提炼的教学成果。教学改革是教育工作永恒不变的主题，也是需要不断探索的课题，需要不断地努力实践和完善。本系列教材虽然经过细致的编写与校订，仍然难免有疏漏和不足之处，需要不断地补充、修订和完善，我们热情欢迎使用本系列教材的教师、学生和读者朋友提出宝贵意见和建议，使之更臻成熟。

“21世纪高等学校电子商务专业规划教材”编委会

前　言

数据仓库是企业决策支持系统和联机分析处理(OLAP)的结构化数据环境,具有面向主题、集成性、稳定性和随时间变化的(时变性)的特征。数据挖掘(Data Mining)是从大量的、有噪声的、不完全的数据中提取隐含的、人们事先未知的有用知识和信息的过程。数据仓库和数据挖掘是电子商务数据分析的有效手段。本书讨论数据仓库和数据挖掘应用的相关技术,其内容组织如下。

第 1 章为数据仓库和数据挖掘概述,介绍数据仓库的特征、数据仓库系统及开发工具、商业智能和数据仓库的关系、数据挖掘的定义和数据挖掘过程。

第 2 章为 OLAP 和多维数据模型,介绍 OLAP 定义和特性、多维数据模型和数据仓库的维度建模。

第 3 章为数据仓库设计,介绍数据仓库规划与需求分析、数据仓库建模、数据仓库物理模型设计和数据仓库部署与维护。

第 4 章为 SQL Server 数据仓库开发实例,介绍一个基于在线电子产品销售数据的 OnRetDW 数据仓库的设计过程,包括需求分析、建模、数据抽取工具设计等。

第 5 章为关联分析算法,介绍关联分析的相关概念、Apriori 算法、SQL Server 挖掘关联规则方法和电子商务数据的关联规则挖掘过程。

第 6 章为决策树分类算法,介绍基本分类步骤、决策树分类、SQL Server 决策树分类方法和电子商务数据的决策树分类过程。

第 7 章为贝叶斯分类算法,介绍贝叶斯公式、朴素贝叶斯分类原理、SQL Server 朴素贝叶斯分类方法和电子商务数据的贝叶斯分类过程。

第 8 章为神经网络算法,介绍人工神经网络相关概念、用于分类的前馈神经网络、SQL Server 神经网络分类方法和电子商务数据的神经网络分类过程。

第 9 章为回归分析算法,介绍回归分析相关概念、线性回归分析、非线性回归分析、逻辑回归分析方法和电子商务数据的逻辑回归分析过程。

第 10 章为时间序列分析,介绍时间序列分析相关概念、确定性时间序列分析、随机时间序列模型、SQL Server 时间序列分析方法和电子商务数据的时间序列分析过程。

第 11 章为聚类算法,介绍聚类相关概念、k-均值算法及其应用、EM 算法及其应用、电子商务数据的聚类分析过程以及 Microsoft 顺序分析和聚类分析算法。

书中提供了大量的练习题和上机实验题供读者选用,附录 A 给出了部分练习题参考答案,附录 B 给出了所有上机实验题参考答案,附录 C 给出了书中数据库和包含的数据表。其中带“*”部分为选修内容。

本书紧扣数据仓库和数据挖掘开发所需要的知识、技能和素质要求，以技术应用能力培养为主线构建教材内容，具有以下特色：

☑ 内容全面、知识点翔实：在内容讲授上力求翔实和全面，细致解析每个知识点和各知识点的联系。

☑ 条理清晰、讲解透彻：从介绍数据仓库和数据挖掘的基本概念出发，由简单到复杂，循序渐进介绍数据仓库和数据挖掘系统的开发过程。

☑ 精选实例、实用性强：列举了大量的应用示例，读者通过上机模仿可以大大提高使用应用系统开发能力。

☑ 配套教学资源丰富：提供了教学 PPT、书中所有示例代码、相关数据库文件和 ETL 源程序，便于读者打开和调试。配套的教学资源可以从清华大学出版社网站下载。

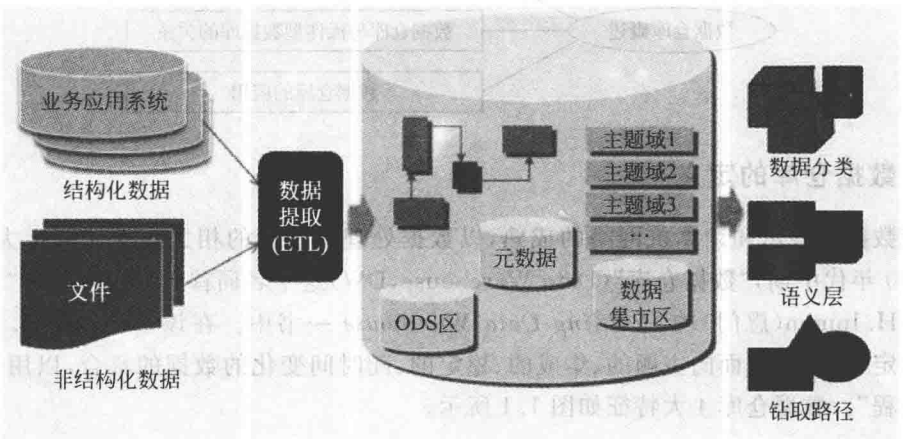
本教材的编写工作得到武汉大学教务部教改项目的资助，解放军理工大学和清华大学出版社给予了大力支持，连续多届选课的同学提出了许多宝贵的建议，编者在此表示衷心感谢。

编　者

2016 年 4 月

目　录

第 1 章　数据仓库和数据挖掘概述

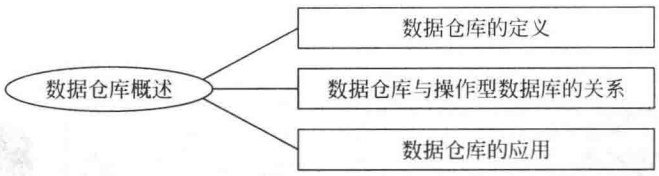


数据仓库体系结构

本章指南

1.1 数据仓库概述

知识梳理



1.1.1 数据仓库的定义

随着数据库技术和计算机网络的成熟，以数据处理为基础的相关技术得到巨大的发展。20 世纪 80 年代中期，“数据仓库”(Data Warehouse，DW)这个名词首次出现在号称“数据仓库之父”W. H. Inmon(恩门)的 *Building Data Warehouse* 一书中。在该书中，W. H. Inmon 把数据仓库定义为“一个面向主题的、集成的、稳定的、随时间变化的数据的集合，以用于支持管理决策过程”。数据仓库 4 大特征如图 1.1 所示。

1. 面向主题

主题是指用户使用数据仓库进行决策时所关心的重点领域，也就是在一个较高的管理层次上对信息系统的数据按照某一具体的管理对象进行综合、归类所形成的分析对象。例如，某保险公司有人寿保险和财产保险两类业务，构建有人寿保险和财产保险两个管理信息系统，如果要对所有顾客进行分析，需要构建面向顾客主题的数据仓库；如果要对所有保单进行分析，需要构建面向保单主题的数据仓库；如果要对所有保费进行分析，需要构建面向保费主题的数据仓库，如图 1.2 所示。

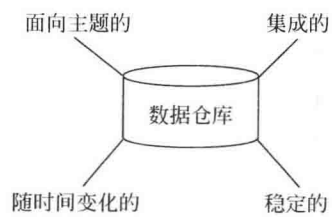


图 1.1 数据仓库的特征

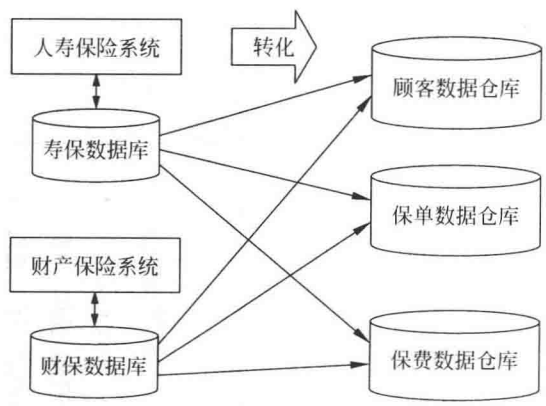


图 1.2 面向主题的示例

从数据组织的角度看，主题是一些数据集合，这些数据集合对分析对象做了比较完整的、一致的描述，这种描述不仅涉及数据自身，而且涉及数据之间的关系。面向主题的数据组织方式，就是在较高层次上对分析对象的数据的一个完整、一致的描述，能完整、统一地刻画各个分

析对象所涉及的企业的各项数据，以及数据之间的联系。

操作型数据库（如人寿保险数据管理系统）中的数据针对事务处理任务（如处理某顾客的人寿保险），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进行组织的。

面向主题组织的数据具有以下特点：

（1）各个主题有完整、一致的内容以便在此基础上做分析处理。

（2）主题之间有重叠的内容，反映主题间的联系。重叠是逻辑上的，不是物理上的。

（3）各主题的综合方式存在不同。

（4）主题域应该具有独立性（数据是否属于该主题有明确的界限）和完备性（对该主题进行分析所涉及的内容均要在主题域内）。

2. 集成性

数据仓库中存储的数据一般从企业原来已建立的数据库系统中提取出来，但并不是原有数据的简单拷贝，而是经过了抽取、筛选、清理、转换、综合等工作得到的数据。例如，某顾客数据仓库中的数据是从应用 A、B、C 中集成的，则需要将性别数据统一转换成 m、f，如图 1.3 所示。

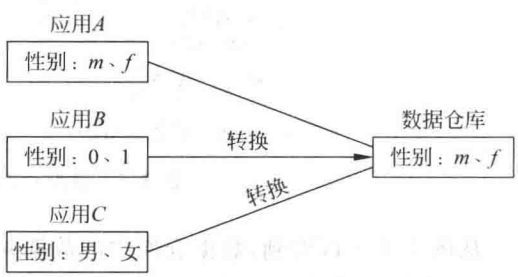


图 1.3　性别的集成

原有数据库系统记录的是每一项业务处理的流水账，这些数据不适合于分析处理。在进入数据仓库之前必须经过综合、计算，同时抛弃一些分析处理不需要的数据项，必要时还要增加一些可能涉及的外部数据。

数据仓库每一个主题所对应的源数据在源分散数据库中有许多重复或不一致之处，必须将这些数据转换成全局统一的定义，消除不一致和错误之处，以保证数据的质量。显然，对不准确，甚至不正确的数据分析得出的结果将不能用于指导企业做出科学的决策。

源数据加载到数据仓库后，还要根据决策分析的需要对这些数据进行概括、聚集处理。

3. 稳定性

数据仓库在某个时间段内来看是保持不变的。

操作型数据库系统中一般只存储短期数据，因此其数据是不稳定的，它记录的是系统中数据变化的瞬态。但对于决策分析而言，历史数据是相当重要的，许多分析方法必须以大量的历史数据为依托。没有大量历史数据的支持是难以进行企业的决策分析的，因此数据仓库中的数据大多表示过去某一时刻的数据，主要用于查询、分析，不像业务系统中的数据库那样，要经常进行修改、添加，除非数据仓库中的数据是错误的。

例如，操作型应用数据库中的数据可以随时被插入、更新、删除和访问（查询），可以从中抽取 10 年的数据构建数据仓库，用于对这 10 年的数据进行分析，一旦数据仓库构建完成，它主要用于访问，一般不会被修改，具有相对的稳定性，如图 1.4 所示。

4. 随时间而变化

数据仓库大多关注的是历史数据，其中数据是批量载入的，即定期从操作型应用系统中接收新的数据内容，这使得数据仓库中的数据总是拥有时间维度。从这个角度看，数据仓库实际是记录了系统的各个瞬态（快照），并通过将各个瞬态连接起来形成动画（即数据仓库的快照集