普通高等教育“十一五”国家级规划教材

教育部“高等学校教学质量与教学改革工程”立项项目

张兴会 等编著

数据仓库与数据挖掘技术

计算机科学与技术专业实践系列教材

清华大学出版社

普通高等教育“十一五”国家级规划教材

计算机科学与技术专业实践系列教材

教育部“高等学校教学质量与教学改革工程”立项项目

数据仓库与数据挖掘技术

张兴会 等编著

清华大学出版社
北京

内 容 简 介

数据仓库与数据挖掘是计算机专业和其他一些与计算机技术关系密切专业必修的核心课程。本书系统地介绍了数据仓库和数据挖掘的基本概念、相关知识和基本方法，每种数据挖掘方法都有详尽的实例描述和具体实现步骤。

本书结构严谨，条理清晰，语言浅显易懂，循序渐进地表达了知识内容；本书坚持理论与实际相结合，概念和具体方法相结合，使知识具体化，生动化；实例实现的过程建立在 SQL 2005 数据挖掘软件的基础上，以帮助读者在学习后达到学以致用的目的。

本书可以作为计算机类、信息类等相关专业本科生数据挖掘课程的教材，也可以作为其他专业技术人员的自学参考书。

图书在版编目（CIP）数据

数据仓库与数据挖掘技术/张兴会等编著．—北京：清华大学出版社，2011.6
（计算机科学与技术专业实践系列教材）
ISBN 978-7-302-24701-2

Ⅰ．①数…　Ⅱ．①张…　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 018775 号

责任编辑：汪汉友
责任校对：时翠兰
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，jsjjc@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：北京富博印刷有限公司
装 订 者：北京市密云县京文制本装订厂
经　　销：全国新华书店
开　　本：185×260　　**印　　张**：14.25　　**字　　数**：344 千字
版　　次：2011 年 6 月第 1 版　　**印　　次**：2011 年 6 月第 1 次印刷
印　　数：1～4000
定　　价：24.00 元

产品编号：034551-01

普通高等教育“十一五”国家级规划教材
计算机科学与技术专业实践系列教材

编 委 会

前　言

随着信息技术的高速发展，数据量的积累急剧增长，数据挖掘是为顺应这种需要而发展起来的数据处理技术，是知识发现(Knowledge Discovery in Database)的关键步骤。数据挖掘涉及比较多的数学基础知识，如何深入浅出地将这些知识及其应用方法介绍给学生是编写数据挖掘教材的关键所在。

为此，本书在编写时力求突出以下特征：

(1) 采用尽可能浅显易懂的语言表达知识内容；

(2) 理论与实际相结合，概念和方法相结合，使知识具体化，实用化；

(3) 实例是通过数据挖掘软件 SQL Server 2005 完成的；

(4) 每章最后结合实例，理论联系实际，帮助学生达到学以致用的效果。

本书共 11 章，包括 4 个主要部分，具体内容如下。

第一部分：第 1 章为数据挖掘和数据仓库概述，简要介绍了数据挖掘和数据仓库的发展趋势、基本概念等相关知识。

第二部分：第 2 章和第 3 章详细介绍了数据仓库的基本概念、相关知识，以及联机分析处理技术的基本方法和实例的具体实现。

第三部分：第 4 章～第 10 章详细介绍了关联规则方法、决策树方法、统计学习方法、神经网络方法、聚类分析、粗糙集方法等方法的相关知识和实例的具体实现。

第四部分：第 11 章介绍了一些复杂结构的数据挖掘以及数据挖掘的发展。

本书的亮点为，每章的最后一节都是本章理论方法的一个具体实现，便于读者深入掌握。读者可以根据自己的需要选择学习相关内容。本书可以作为计算机类、信息类等相关专业本科生数据挖掘课程的教材，也可以作为其他专业技术人员的自学参考书。

信息处理技术是信息科学、应用数学发展的一个重要分支，在教学中，主要通过理论教学、实验教学、课程设计等教学环节来提高学生的实践技能和应用水平，这样的教学方法也是天津职业技术师范大学长期为社会培养高素质职教师资和应用型高级专门人才过程中总结出来的一种行之有效的教学方法。为了使理论和实际相结合，使基本概念和知识与具体的方法、工具相结合，达到学以致用的效果，体现应用型大学手脑并用的办学理念，作者还特别编写了一本与本教材相配合使用的《数据仓库与数据挖掘工程实例》辅助教材，该教材共包括 10 个工程案例，介绍了利用数据挖掘与数据仓库工具如何建立数据仓库、如何进行数据预处理和进行数据挖掘等，目的就是希望通过通俗易懂的语言和详细的工程实例分析，使学生能够较好地掌握数据挖掘与数据仓库的理论知识和构建模型的操作过程，进一步提高学生对信息进行管理和利用的能力。

本书由张兴会统稿，王明春、郑晓艳、刘玲、刘新钰、童勇木参加了本书的编写、图表绘

制、模型构建、软件调试等工作。在本书编写过程中，安淑芝教授提出了宝贵的修改意见。另外，本书还参阅和引用了许多专家和学者的文献资料，在此表示衷心的感谢。

由于作者水平和能力有限，新技术的发展和更新较快，书中的不妥之处，欢迎读者批评指正。作者邮箱：xhzhang@tute.edu.cn。

作者

2011年4月于天津

目　录

第 1 章　数据挖掘和数据仓库概述

随着计算机技术和网络技术的发展，数据量急剧增长。人类处于信息爆炸的时代，被淹没在数据海洋之中。如何有效地组织和存储数据，如何从数据海洋中及时发现有用的知识、提高信息利用率，成为人们亟待解决的问题。但是，仅以目前数据库系统的录入、查询、统计等功能，无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势。正是在这样的背景下，数据挖掘(data mining，DM)技术应运而生，并越来越显示出强大的生命力。

数据挖掘技术的发展催生决策分析数据环境的改变，而传统的数据库管理系统因自身的局限性无法满足决策支持系统的要求，具体表现为：不能满足数据成几何级数增长的需要，不同部分的数据难以集成，访问数据的响应性能不断降低。要想使数据能够发挥其最佳效用，更好地为用户服务，数据必须经过严格的准备、组织和显示等步骤。因此，一种适用于决策支持系统的数据组织与管理技术——数据仓库(data warehouse，DW)技术应运而生，并逐渐成为支持分析与决策的重要技术。

1.1　数据挖掘引论

1.1.1　数据挖掘的由来

数据挖掘经历了逐渐演变的过程。在电子化数据处理的初期，人们就试图通过某些方法实现自动决策支持，于是机器学习成为关注的焦点。机器学习的过程就是将一些已知的并已被成功解决的问题作为范例输入计算机，机器通过学习这些范例，总结并生成相应的规则，这些规则具有通用性，使用它们可以解决某一类问题。机器学习的研究最早始于 20 世纪 60 年代，比较典型的结果有 Rosenblate 的感知机、Sammel 的西洋跳棋程序。

随着神经网络等技术的形成和发展，人们的注意力逐渐转向知识工程。知识工程不同于机器学习，不是为计算机输入范例，由其生成出规则，而是直接为计算机输入已被代码化的规则，计算机通过使用这些规则来解决某些问题，如专家系统就是这种方法所得到的成果。

20 世纪 80 年代，在新的神经网络等理论的指导下，重新回到机器学习的方法上，并将其成果应用于处理大型商业数据库，如 Michelski 等人的 AQ11 系统(1980 年)、Quiulan 的 ID3(1983 年)决策树方法、Rumelhart 等人研制的反向传播神经网络 BP 模型(1985 年)、Langley 等人的 BACON 系统(1987 年)等，这些显著成果的出现，使机器学习逐渐成为人工智能的主要学科方向之一。

1989 年，在美国底特律召开的第十一届国际联合人工智能学术会议上首次提到知识发现(knowledge discovery in database，KDD)这一概念；1993 年，美国电气电子工程师学会(IEEE)的知识与数据工程(knowledge and data engineering)会刊出版 TKDD 技术专刊，发

表的论文和摘要体现了当时知识发现的最新研究成果和动态。

随着来自各个领域的研究和应用开发不断增多，1995 年，在加拿大蒙特利尔召开了首届 KDD 国际学术年会，数据挖掘技术被分为工程领域的数据挖掘与科研领域的知识发现。由于把数据库中的“数据”形象地比喻为矿床，“数据挖掘”一词很快流传开来。此后，此类会议每年召开一次，数量和规模逐渐扩大，从专题研讨会一直发展到国际学术大会，并成为当前计算机领域的研究热点。目前，对 KDD 的研究主要围绕理论、技术和应用这三个方面展开。

1.1.2 数据挖掘的定义

数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。数据挖掘应该更正确地命名为“从数据中挖掘知识”。还有很多和这一术语相似的术语，如知识发现、数据分析、数据融合以及决策支持等。人工智能领域习惯称之为知识发现，而数据库领域习惯称之为数据挖掘。

用于数据挖掘的原始数据可以是结构化的，如关系数据库中的数据；也可以是半结构化的，如文本、图形、图像数据等。数据挖掘的方法可以是数学的，也可以是非数学的；可以是演绎的，也可以是归纳的。挖掘出的知识可以被用于信息管理、查询优化、决策支持、过程控制等；还可以用于数据自身的维护。

数据挖掘是一个完整的过程，其一般步骤如图 1-1 所示。

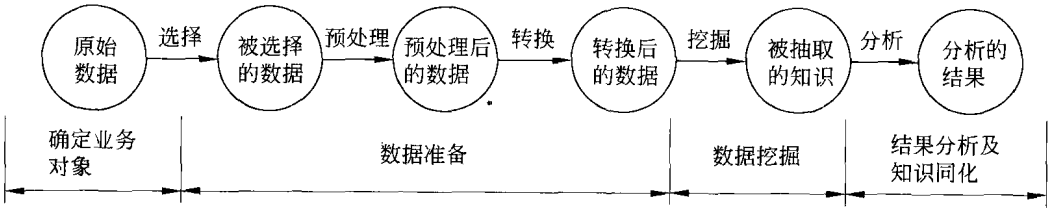

图 1-1 数据挖掘的过程

1. 确定业务对象

在开始数据挖掘之前，最基础的就是理解数据和实际的业务，在这个基础上提出问题，对目标有明确的定义。认清数据挖掘的目的是数据挖掘的重要一步，因此必须清晰地定义出业务对象。

2. 数据准备

数据准备是保证数据挖掘得以成功的先决条件，数据准备在整个数据挖掘过程中占有很大的工作量，大约是整个数据挖掘工作量的 60%。数据准备包括数据选择、数据预处理和数据转换。

(1) 数据选择。就是搜索所有与业务对象有关的内部和外部数据信息，获取原始的数据，从中选择出适用于数据挖掘应用的数据，建立数据挖掘库。

(2) 数据预处理。由于数据可能是不完全的、有噪声的、随机的、复杂的，数据预处理就要对数据进行初步的整理，清洗不完全的数据，为进一步的分析作准备。

(3) 数据转换。数据转换是构建面向分析的数据存储模式的关键，在转换过程中数据

会被格式化，并加载到适合分析的存储环境中。常见的数据转换问题包括：数据类型转换、对象名转换、数据编码转换、表结构转换。

3. 数据挖掘

数据挖掘就是对所得到的经过转换的数据进行挖掘，除了选择合适的挖掘算法外，其余工作应能自动地完成。

4. 结果分析与知识同化

结果分析就是对挖掘结果进行解释并评估，其使用的分析方法一般应根据数据挖掘操作而定，目前通常应用可视化技术。知识同化就是将分析所得到的知识集成到业务信息系统的组织结构中去。

1.1.3 数据挖掘的功能

数据挖掘具体功能主要有以下几个方面。

1. 概念描述

概念描述(concept description)，就是对某类对象的内涵进行描述，并概括这类对象的有关特征。具体的描述分为特征性(characterization)描述和区别性(discrimination)描述。前者用于描述某类对象的共同特征，后者用于描述不同类对象之间的区别。

描述数据允许数据在多个抽象层概化，便于用户考察数据的一般行为。

2. 关联分析

数据关联是数据中存在的一类重要的可被发现的知识，若两个或多个变量间存在着某种规律性，就称为关联。关联可分为简单关联、时序关联、因果关联。关联分析(association analysis)是从大量的数据中发现项集之间有趣的联系、相关关系或因果结构，以及项集的频繁模式。

3. 分类与预测

(1) 分类(classification)。分类是数据挖掘中的一项非常重要的任务。分类的目的是提出一个分类函数或者分类模型，该模型能把数据库中的数据项映射到给定类别中的一个。构造分类器，需要有一个训练样本数据集作为输入。

(2) 预测(prediction)。预测是利用历史数据建立模型，再运用最新数据作为输入值，获得未来变化的趋势或者评估给定样本可能具有的属性值或值的范围。

4. 聚类分析

(1) 聚类(clustering)。聚类是根据数据的不同特征，将其划分为不同的数据类。其目的是使得属于同一类别的个体之间的距离尽可能小，而不同类别的个体间的距离尽可能大。

(2) 聚类与分类的区别如下：分类需要预先定义类别和训练样本；而聚类分析直接面向源数据，没有预先定义好的类别和训练样本，所有记录都根据彼此相似程度加以归类。

5. 偏差分析

偏差分析(deviation analysis)又称为比较分析，是对差异和极端特例的描述，揭示事物偏离常规的异常现象，其基本思想是寻找观测结果与参照值之间有意义的差别。偏差包括分类中的反常实例、不满足规则的特例、观测结果对模型预测的偏差、量值随时间的变化等。

1.1.4 数据挖掘的常用方法

1. 聚类分析

聚类分析(clustering analysis)是一个比较活跃的数据挖掘研究领域,源于统计学、生物学以及机器学习等。聚类生成的组叫簇,簇是数据对象的集合。聚类分析的过程就是使同一个簇内的任意两个对象之间具有较高的相似性,不同簇的两个对象之间具有较高的相异性。

用于数据挖掘的聚类分析有划分的方法、层次的方法、基于密度的方法、基于网格的方法和基于模型的方法等。

2. 决策树

决策树(decision tree)主要应用于分类和预测,提供了一种展示类似在什么条件下会得到什么值这类规则的方法。决策树分为分类树和回归树两种,分类树对离散变量做决策,回归树对连续变量做决策。

决策树是一个类似于流程图的树结构,树的最顶层结点是根结点,中间的结点是内部结点,末梢的结点是叶结点,其中根结点是整个数据集合空间,每个内部结点表示在一个属性上的测试,每个分支代表一个测试输出,每个叶结点代表类或类分布。

建立决策树的过程,即树的生长过程是不断地把数据进行切分的过程,每次切分对应一个问题,也对应着一个结点。对每个切分都要求分成的组之间的"差异"最大。各种决策树算法之间的主要区别是"差异"衡量方式的区别。数据挖掘中决策树是一种经常用到的技术,常用的算法有 CHAID、CART、Quest、ID3 和 C4.5 等。

3. 人工神经网络

人工神经网络(artificial neural network,ANN)是一类比较新的计算模型,它是模仿人脑神经网络的结构和某些工作机制而建立的一种计算模型。这种计算模型的特点是利用大量的简单计算单元(即神经元)连成网络,来实现大规模并行计算。神经网络的工作机理是通过学习,来改变神经元之间的连接强度。由于人工神经网络具有自我组织和自我学习等特点,能解决许多其他方法难以解决的问题,因此得到较普遍的应用。

人工神经网络主要有前馈式网络、反馈式网络和自组织网络。

4. 粗糙集

粗糙集(rough set)是一种处理不确定、不完备数据和不精确问题的新的数学理论。粗糙集理论建立在分类机制的基础上,将知识理解为对数据的划分,并引入上近似(upper approximation)和下近似(lower approximation)等概念来刻画知识的不确定性和模糊性。模糊集和概率统计方法是处理不确定信息的常用方法,但这些方法需要一些数据的附加信息或先验知识,如模糊隶属函数和概率分布等,这些信息有时并不容易得到。粗糙集分析方法仅利用数据本身提供的信息,无须任何先验知识。

5. 关联规则挖掘

关联规则挖掘(association rule mining)是数据挖掘中最活跃的研究方法之一,最早由 Agrawal 等人提出(1993 年)。最初的动机是针对购物篮分析问题提出的,其目的是发现交易数据库中不同商品(项)之间的联系,由这些规则找出顾客购买行为模式,如购买了某一商品对购买其他商品的影响。发现这样的规则可以应用于商品货架设计、库存安排以及根据

购买模式对用户进行分类。

关联规则的基本思想：一是找到所有支持度大于最小支持度的频繁项集，即频集；二是使用第一步找到的频集产生期望的规则。其核心方法是基于频集理论的递推方法。关联规则挖掘的主要算法包含关联发现、序列模式发现、时序发现等。

6. 统计分析

统计分析(statistics analysis)是从事物的外在数量上的表现去推断该事物可能的规律。科学的规律性一般总是隐藏得比较深，最初总是从其数量表现上通过统计分析看出一些线索，然后提出一定的假说或学说，做进一步深入的理论研究。当理论研究提出一定的结论时，往往还需要在实践中加以验证，即观测一些自然现象或专门安排的实验所得资料是否与理论相符，在多大程度上相符，偏离可能是朝哪个方向，等等。

常见的统计分析有回归分析(多元回归、自回归)、判别分析(贝叶斯判别、费歇尔判别、非参数判别)以及探索性分析(主元分析、相关分析)等。

1.2 数据仓库引论

1.2.1 数据仓库的产生与发展

随着市场竞争的日趋激烈，信息对企业的生存、发展、壮大起着越来越重要的作用。由于计算机技术的普遍应用，承载信息的数据随着时间的推移而不断增长，并且分布在不同的系统平台上，具有多种存储形式。能否从纷繁复杂、大量沉淀的数据环境中得到有用的决策信息，已成为企业生存、发展、壮大的重要环节。

基于上述的需求，在20世纪80年代出现了数据仓库的思想。1988年，为解决全企业集成问题，IBM爱尔兰公司的Barry Devlin和Paul Murphy第一次提出了“信息仓库”的概念，其定义为：“一个结构化的环境，能支持最终用户管理其全部的业务，并支持信息技术部门保证数据质量”。在20世纪90年代初期，数据仓库的基本原理、框架架构，以及分析系统的主要原则都已经确定，主要技术包括关系型数据存取、网络、C/S架构和图形化界面。一些前沿的公司已经开始建立数据仓库。

1992年，美国著名的信息工程学家William H. Inmon在《建立数据仓库》(Building the Data Warehouse)一书中首先系统地阐述了关于数据仓库的思想、理论。该书不仅说明为什么要建数据仓库、数据仓库能带来什么，更重要的是，Inmon第一次提供了如何建设数据仓库的指导性意见。该书定义了数据仓库非常具体的原则，即：数据仓库是面向主题的、集成的、包含历史的、不可更新的、面向决策支持的、面向全企业的、最明细的数据存储、数据快照式的数据获取等。这些原则到现在仍然是指导数据仓库建设的最基本原则，因此，William H. Inmon被人们尊称为“数据仓库之父”。

数据仓库的盛行始于1995年，而且其作为数据库的高端扩展技术一直是一大热点。IBM所推崇的商业智能(BI)，其核心就是数据仓库；微软的SQL Server 7.0已经绑定了OLAP服务器，将数据仓库功能集成到数据库中，并建立了数据仓库联盟；Oracle公司也有自己的Oracle Express系列OLAP产品用来提供决策支持。

从目前形势看，数据仓库已成为继因特网之后，信息社会中获得企业竞争优势的关键。