大学本科计算机专业应用型规划教材

丛书主编：高林

# 数据仓库与数据挖掘

安淑芝 等 编著

清华大学出版社

大学本科计算机专业应用型规划教材

丛书主编：高林

# 数据仓库与数据挖掘

安淑芝 等 编著

清华大学出版社
北京

## 内容简介

本书是一本介绍数据仓库和数据挖掘的图书。全书力求深入浅出、通过浅显易懂的语言及实例介绍数据仓库与数据挖掘的基本概念及相关理论。从数据仓库的定义、结构、设计、数据访问方法及应用等方面对数据仓库做了较详细的介绍。从数据挖掘的定义、数据预处理方法、数据挖掘发现知识的类型及数据挖掘常用算法等几方面对数据挖掘的基本知识和算法等理论做了介绍。本书特别介绍了 SQL Server 2000 数据挖掘工具应用和 SPSS 数据挖掘工具应用。最后，给出了一个数据挖掘的应用实例。本书总的指导思想是在掌握基本知识和基本理论的基础上，更强调实际应用能力的培养。

本书可作为普通高等院校计算机科学与技术专业、软件工程专业或信息类等其他相关专业的教材，也可作为有关数据仓库与数据挖掘方面的培训教材，以及所有想学习数据仓库与数据挖掘知识的人的自学用书。



**图书在版编目(CIP)数据**

数据仓库与数据挖掘/安淑芝等编著. —北京：清华大学出版社，2005.6
（大学本科计算机专业应用型规划教材/高林主编）
ISBN 7-302-10688-6

Ⅰ. 数… Ⅱ. 安… 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(2005)第 022900 号

**出 版 者**：清华大学出版社　　**地　　址**：北京清华大学学研大厦
http://www.tup.com.cn　　**邮　　编**：100084
**社 总 机**：010-62770175　　**客户服务**：010-62776969
**组稿编辑**：谢　琛
**文稿编辑**：汪汉友
**印 刷 者**：国防工业出版社印刷厂
**装 订 者**：三河市化甲屯小学装订二厂
**发 行 者**：新华书店总店北京发行所
**开　　本**：185×260　**印张**：14　**字数**：323 千字
**版　　次**：2005 年 6 月第 1 版　2005 年 6 月第 1 次印刷
**书　　号**：ISBN 7-302-10688-6/TP・7230
**印　　数**：1～3000
**定　　价**：19.00 元

大学本科 计算机专业应用型规划教材

# 编 委 会

# 丛书序

大学本科计算机专业应用型规划教材

为适应我国“以信息业带动工业化，发挥后发优势，实现社会生产力的跨越式发展”以及大力发展制造业和优化产业结构的要求，应用型人才培养已成为高等学校人才培养的重要任务。

以微电子技术为基础、计算机技术为主体的信息技术，是当前人类社会中发展最快、渗透性最强、应用面最广的先导技术。信息技术的广泛应用推动着以信息产品制造业、软件业、信息系统集成业和信息咨询服务业为主体的信息产业的发展。在新的世纪里，信息已成为重要的生产要素和战略资源，信息技术成为先进生产力的代表，信息产业将发展成为现代产业的带头产业，人类即将跨越工业时代进入信息时代。因此，信息化成为当今世界经济和社会的发展趋势，大力推进社会和国民经济信息化是推进我国社会主义现代化建设的重要任务。计算机和信息技术的发展不仅需要大批专业技术人才，而且还产生了一批新的职业岗位，毋庸置疑，信息及其相关职业将成为未来最紧缺的职业。

计算机和信息技术与应用的人才需求将呈多元化、多层次趋势，表现在科学、技术、产业、应用、服务诸多方面。不仅需要从事科学、技术研发的人才，而且更需要把研发成果转变为现实产品的技术和管理人才；不仅要有能从事计算机和信息科学、技术工作的人才，而且更需要能从事计算机和信息产业、应用、服务工作的人才，以及在各类人才中的精英人才、领军人物。这实际是对我国计算机和信息类高等教育改革提出了新的要求和新的课题，要求我国高等教育进行结构调整，满足人才培养的多元化，大力培养具有计算机和信息技术专长的应用型人才——他们是这些领域的技术专家和管理专家，可以在相应的行业、企业担任各种技术工作。

目前，我国高等教育中应用型人才培养模式相对落后，如何发展应用型教育已成为课程改革的主要任务。本套教材是以培养计算机和信息类专业本科应用型人才为目的进行的课程与教材改革尝试。在本套教材的策划过程中，清华大学出版社多次组织了由行业企业专家和有丰富教学经验的一线教师参加的研讨会，对应用型高等教育的规律和在计算机教学中的体现进行了深入的研讨。在此基础上我们力求能从整体上把握计算机和信息类应用型人才培养的特征，并体现在这套教材的编写过程中：在教材编写的指导思想上，力求在保持学科科学性的同时，体现工程和技术学科的系统性；在教材

的内容组织上，尽量采用以问题为中心的写作方法，加强案例性教学；在理论联系实际和加强能力培养方面，增加方案性设计习题和实际训练性题目，以培养学生的专门技术能力和完成实际工作任务的能力。

计算机和信息类应用型教材编写还处于改革的初步尝试阶段，希望使用这套教材的教师也能够参与到教材建设工作中来，并提出宝贵意见，以便推动课程改革并提高教材质量。

高　林

2004 年 5 月

# 前 言



数据挖掘是信息和数据深度处理的必然需要，也是体现信息价值的重要工具。现在数据挖掘已经成为计算机、信息系统等很多专业本科生的必修教学内容，可见数据挖掘技术在当今科学中的重要性以及应用的广泛。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介绍数据挖掘的教材的关键。

为此，本书在写作上力求体现如下特点，其一是采用尽可能浅显易懂的语言，循序渐进地表达知识内容；其二是概念和具体的方法、工具相结合，使知识具体化，不枯燥；其三是尽可能结合应用的实例，使理论和实际相结合，达到学以致用的效果。

本书共 7 章，可分为三个主要部分。

第一部分：第 1～4 章为数据挖掘的基础知识，包括数据仓库和数据挖掘的基本概念和相关知识介绍；

第二部分：第 5、6 章介绍了数据挖掘的算法和工具；

第三部分：第 7 章是数据挖掘的实际应用。通过一个用决策树算法应用于人力资源管理系统的实例说明数据挖掘技术的具体应用。

读者可以根据自己的需要选择学习相关的内容。本书可以作为计算机、信息类等专业本科生数据挖掘课程的教材，也可以作为其他专业学生的参考书。

参加本书的编写人员还有张兴会、郑晓艳和刘玲。

由于作者水平有限，欢迎读者对于书中的不足给予指正。

清华大学出版社的编辑在本书的编写和出版过程中给予了大力支持和帮助。

本书作者对参考文献中列出的以及未列出的所有文献作者表示由衷的感谢。

作 者

2005 年 3 月

# 目 录

数据仓库与数据挖掘

# 第1章 数据仓库与数据挖掘

# 绪　论

随着计算机技术和计算机网络技术的发展，信息化程度快速增长，人们利用信息技术生产和搜集数据的能力大幅度提高。于是，信息过量几乎成为人人需要面对的问题。有人称现在是信息爆炸的时代，人们面对着“被数据淹没，却饥饿于知识”的挑战。“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、提高信息利用率”是人们迫切需要解决的问题。数据挖掘技术就是在这样的背景下应运而生和蓬勃发展，并越来越显示出强大的生命力。

数据挖掘技术是一门综合性的技术领域，主要涉及数据库、人工智能和数理统计3个技术领域。

## 1.1 初识数据挖掘

### 1.1.1 数据挖掘的产生

当今，数据容量规模已经达到万亿字节(TB)的水平。过量的数据被人们称为信息爆炸，带来的挑战是一方面规模庞大、纷繁复杂的数据体系让使用者漫无头绪、无从下手；另一方面在这些大量数据的背后却隐藏着很多具有决策意义的有价值的信息。那么，如何发现这些有用的知识，使之为管理决策和经营战略发展服务？计算机科学给出的最新回答是数据挖掘(data mining，DM)。

数据挖掘产生的前提是需要从多年积累的大量数据中找出隐藏在其中的、有用的信息和规律；计算机技术和信息技术的发展使其有能力处理这样大量的数据。

### 1.1.2 数据挖掘的应用价值

应用数据挖掘从大量数据中所发现的规律并不是“放置四海而皆准”的规律，而是面向某一应用的规律，具有具体的指导意义。

早期，数据挖掘主要应用于商业领域，如许多读者都熟知的“啤酒和尿布”的故事，就是零售业巨头“沃尔玛”从大量销售数据中分析出来的规律：美国的男士在下班后要去超市买婴儿尿布，他们在购买尿布的同时会买啤酒。“沃尔玛”因此将这两种“毫不相干”的商品摆放在靠近的货架上，并在其间摆放一些下酒小菜，使这些商品销售量大增。

随着人们对数据挖掘了解的逐步深入，其应用领域也逐步扩大，如科学研究、市场营

销、金融分析、体育比赛等。

### 1.1.3 数据挖掘的发展过程

数据挖掘是20世纪80年代，人工智能(aritificial intelligence，AI)研究项目失败后，AI转入实际应用时提出的。它是一个新兴的，面向商业应用的AI研究。

知识发现(knowledge discovery in database，KDD)和数据挖掘是数据库领域中最重要的课题之一，国际上第一次关于数据挖掘与知识发现的研讨会于1989年在美国的底特律召开，在此次会议上第一次提出了知识发现一词。

1995年，在加拿大召开了第一届KDD和DM国际学术会议。会议对KDD做了确切的定义，未对DM做确切定义。

目前KDD和DM已成为研究的热点和焦点，一批DM系统开发出来，在商业、经济、金融和管理等领域都取得了应用性的成果。

### 1.1.4 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近似的术语，如知识发现、数据分析、数据融合(data fusion)以及决策支持等。人工智能领域习惯称知识发现，而数据库领域习惯称数据挖掘。

人们把原始数据看作是形成知识的源泉，就像从矿石中采矿一样。

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。发现知识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。发现了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。

一般来说，数据挖掘是一个利用各种分析方法和分析工具在大规模海量数据中建立模型和发现数据间关系的过程，这些模型和关系可以用来作出决策和预测。支持大规模数据分析的方法和过程，选择或者建立一种适合数据挖掘应用的数据环境是数据挖掘研究的重要课题。

## 1.2 初识数据仓库

### 1.2.1 数据仓库的产生

随着市场经济竞争的加剧和企业数据量的积累，人们不满足于利用数据库对数据的处理和管理了，更希望能够将多方面、多渠道的数据综合处理和管理，从而更好地发挥这些信息资源的作用，帮助管理者进行决策。

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，Devlin 和 Murphy 发表了一篇关于数据仓库论述的最早文章，将其定义为："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 W H. Inmon 在《Building the Data Warehouse》(《建立数据仓库》)一书中首先系统地阐述了关于数据仓库的思想、理论。他在这本书中不仅仅说明为什么要建数据仓库以及数据仓库能给你带来什么，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建设的非常具体的原则，包括：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，从此数据仓库的研究和应用得到了广泛的关注，因而 Inmon 被人们尊称为"数据仓库之父"。

### 1.2.2 数据仓库的应用价值

人们为什么不能在原数据库上作决策，而一定要建造数据仓库呢？

传统数据库对日常事务处理(联机事务处理)(on line transaction process，OLTP)十分理想，但是要给予事务处理的数据库帮助决策分析就产生了很大困难，其原因主要是传统数据库的处理方式和决策分析中的数据需求不相称，主要表现在以下方面。

#### 1. 决策处理的系统响应问题

在日常事务处理中，用户对系统和数据库的要求是数据存取频率要高，操作时间要短；而在决策分析中，有的决策问题请求可能导致系统长达数小时的运行，有的决策分析问题的解决需要遍历数据库中大部分数据，这些是日常事务处理系统所无法承担的。因此操作型数据和决策分析型数据应该分离。

#### 2. 决策数据需求的问题

在进行决策分析时，需要有全面的、正确的集成数据。如果将数据集成问题交给决策分析程序解决，将大大增加决策分析系统的负担，并且没有必要在每次进行决策分析时都进行数据集成。

对在不同的应用系统中，存在的同一实体属性具有不同数据类型、不同字段名称以及不同格式等，在决策数据集成时需要进行转换。

决策数据需要动态更新，并且往往需要一些经过汇总、概括的数据。

#### 3. 决策数据操作的问题

从对数据的操作方式上讲，日常事务处理系统远远不能满足决策人员的需要，决策分析人员希望以专业用户的身份使用各种工具对数据进行多种形式的操作，对数据操作的结果以商业智能的方式表达出来，现有系统很难达到此要求。

由于系统响应问题、决策数据问题和决策数据操作问题的存在，导致企业无法使用现有的业务处理来满足决策分析的需要，因此决策分析需要一个能够不受传统事务处理的约束、高效率处理决策分析数据的支持环境，这就是数据仓库存在的价值。

### 1.2.3 数据仓库的发展过程

数据仓库是一种新的数据处理体系结构，它是企业内部各部门业务数据和各种外部数据进行统一和综合的中央数据仓库，它为企业决策支持系统提供所需的信息，它是一种信息管理技术。数据量越大，数据仓库的作用就越大。

从目前的形势看，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。据美国 Meta Group 市场调查机构的资料表明，《幸福》杂志所列的全球 2000 家大公司中已有 99%将因特网和数据仓库这两项技术都列入企业计划。

数据仓库是 1995 年开始盛行起来的，数据仓库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；微软公司的 SQL Server 7.0 就已经绑定了 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。

目前世界上最大的数据仓库是 NCR 公司建立的基于其 Tera data 数据库拥有 24TB 数据量的 Wal-Mart(沃尔玛)数据仓库系统，由此产生了“啤酒与尿布”的故事。

### 1.2.4 数据仓库的定义

W. H. Inmon 提出的数据仓库的定义：数据仓库是一个面向主题的、集成的、非易失的且随时间变化的数据集合，用来支持管理人员的决策。

其他定义：“数据仓库是一种体系结构，一种独立存在的不影响其他已经运行的业务系统的语义一致的数据仓储，可以满足不同的数据存取，文档报告的需要。”

公认的数据仓库概念基本上采用了 W. H. Inmon 的定义：数据仓库是面向主题的、集成的、不可更新的(稳定性)随时间不断变化(不同时间)的数据集合，用以支持经营管理中的决策制定过程。

### 1.2.5 数据仓库与数据挖掘的关系

#### 1. 数据仓库系统的数据可以作为数据挖掘的数据源

因为数据仓库系统已经按照主题将数据进行了集成、转换，因此数据仓库系统能够满足数据挖掘技术对数据环境的要求，可以直接作为数据挖掘的数据源。如果将数据仓库和数据挖掘紧密联系在一起，将获得更好的结果。

#### 2. 数据挖掘的数据源不一定必须是数据仓库系统

作为数据挖掘的数据源不一定必须是数据仓库。它可以是任何数据文件或格式，但必须事先进行数据预处理，处理成适合数据挖掘的数据。数据预处理是数据挖掘的关键步骤，并占有数据挖掘全过程工作量的很大比重。

因此，数据仓库与数据挖掘没有必然的关系，有些人简单地认为数据仓库是数据挖掘的前期准备，这种认识是不全面的。

## 1.3 进一步理解数据挖掘

### 1.3.1 数据挖掘的功能

数据挖掘通过预测未来趋势及行为，作出前瞻的、基于知识的决策。数据挖掘的目标是从数据中发现隐含的、有意义的知识。具体的功能有以下7个方面。

#### 1. 概念描述

概念描述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描述和区别性描述。

(1) 特征性描述。特征性描述用于描述某类对象的共同特征。

(2) 区别性描述。区别性描述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。

例如对超市的销售数据，销售经理并不想了解每个客户的事务，而愿意观察到高层的数据，譬如按地区对顾客分组，观察每组顾客购买频率和顾客的收入等。

#### 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。关联分析的目的就是找出数据中隐藏的关联网。

关联分析发现关联规则，这些规则展示属性值频繁地在给定数据集中一起出现的条件。

“啤酒和尿布”就是从大型超市的购物篮当中分析出的关联规则。

#### 3. 分类与预测

(1) 分类。所谓分类，就是依照所分析对象的属性分门别类、加以定义、建立类组。比如，将信用卡申请人分为低、中、高风险群，或是将顾客分到事先定义好的族群。分类的关键是确定对数据按照什么标准或什么规则进行分类。因此，分类时首先根据属性特征，为每一种类别找到一个合理的描述或模型，即确定分类规则；再根据规则对数据进行分类。

(2) 预测。所谓预测，就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。比如，预测哪些顾客会在未来的半年内取消该公司的服务，或是预测哪些电话用户会申请增值服务等。

#### 4. 聚类分析

聚类分析又称为无指导的学习，其目的在于客观地按被处理对象的特征分类，将有相