清华大学 计算机系列教材

徐华 编著

数据挖掘：方法与应用

清华大学出版社

清華大學 计算机系列教材

徐华 编著

数据挖掘：方法与应用

清华大学出版社
北京

内 容 简 介

本书主要根据作者近几年在清华大学面向研究生和本科生开设的“数据挖掘：方法与应用”课程的教学实践与积累，参考近几年国外著名大学相关课程的教学体系，系统的介绍数据挖掘的基本概念和基本原理方法；结合一些典型的应用实例展示用数据挖掘的思维方法求解问题的一般性模式与思路。

本书可作为有一定数据结构、数据库和程序设计基础的研究生或本科生开展数据挖掘知识学习和研究的入门性教材与参考读物。

图书在版编目(CIP)数据

数据挖掘：方法与应用/徐华编著．--北京：清华大学出版社，2014
清华大学计算机系列教材
ISBN 978-7-302-36901-1

Ⅰ．①数…　Ⅱ．①徐…　Ⅲ．①数据采集－高等学校－教材　Ⅳ．①TP274

中国版本图书馆 CIP 数据核字(2014)第 131376 号

责任编辑：白立军　顾　冰
封面设计：常雪影
责任校对：焦丽丽
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：北京密云胶印厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：12　　**字　　数**：286 千字
版　　次：2014 年 10 月第 1 版　　**印　　次**：2014 年 10 月第 1 次印刷
印　　数：1～2000
定　　价：25.00 元

产品编号：044864-01

序

“清华大学计算机系列教材”已经出版发行了30余种，包括计算机科学与技术专业的基础数学、专业技术基础和专业等课程的教材，覆盖了计算机科学与技术专业本科生和研究生的主要教学内容。这是一批至今发行数量很大并赢得广大读者赞誉的书籍，是近年来出版的大学计算机专业教材中影响比较大的一批精品。

本系列教材的作者都是我熟悉的教授与同事，他们长期在第一线担任相关课程的教学工作，是一批很受本科生和研究生欢迎的任课教师。编写高质量的计算机专业本科生（和研究生）教材，不仅需要作者具备丰富的教学经验和科研实践，还需要对相关领域科技发展前沿的正确把握和了解。正因为本系列教材的作者们具备了这些条件，才有了这批高质量优秀教材的产生。可以说，教材是他们长期辛勤工作的结晶。本系列教材出版发行以来，从其发行的数量、读者的反映、已经获得的国家级与省部级的奖励，以及在各个高等院校教学中所发挥的作用上，都可以看出本系列教材所产生的社会影响与效益。

计算机学科发展异常迅速，内容更新很快。作为教材，一方面要反映本领域基础性、普遍性的知识，保持内容的相对稳定性；另一方面，又需要紧跟科技的发展，及时地调整和更新内容。本系列教材都能按照自身的需要及时地做到这一点。如王爱英教授等编著的《计算机组成与结构》、戴梅萼教授等编著的《微型计算机技术及应用》都已经出版了第四版，严蔚敏教授的《数据结构》也出版了三版，使教材既保持了稳定性，又达到了先进性的要求。

本系列教材内容丰富，体系结构严谨，概念清晰，易学易懂，符合学生的认知规律，适合教学与自学，深受广大读者的欢迎。系列教材中多数配有丰富的习题集、习题解答、上机及实验指导和电子教案，便于学生理论联系实际地学习相关课程。

随着我国进一步的开放，我们需要扩大国际交流，加强学习国外的先进经验。在大学教材建设上，我们也应该注意学习和引进国外的先进教材。但是，“清华大学计算机系列教材”的出版发行实践以及它所取得的效果告诉我们，在当前形势下，编写符合国情的具有自主版权的高质量教材仍具有重大意义和价值。它与国外原版教材不仅不矛盾，而且是相辅相成的。本系列教材的出版还表明，针对某一学科培养的要求，在教育部等上级部门的指导下，有计划地组织任课教师编写系列教材，还能促进对该学科科学、合理的教学体系和内容的研究。

我希望今后有更多、更好的我国优秀教材出版。

清华大学计算机系教授，中国科学院院士

张钹

前　言

近年来，随着计算机硬件资源成本的持续下降，软件开发技术的不断进步，基于不同领域的大数据(Big Data)研究与应用性研发工作正在如火如荼地开展起来。作为大数据挖掘、分析与处理的关键方法与技术之一，“数据挖掘”正在被不同的专业领域所关注。“数据挖掘”也逐渐演变成一门具有通用性和基础性的数据处理方法与技术。正是在这样的大环境背景之下，作者于2011年春季学期开始开设了面向清华大学非计算机专业学生的专业课程“数据挖掘：方法与应用”。开设这门课程的主要目的是为了让不同专业领域的学生能够掌握数据挖掘的基本概念、基本方法和基本算法实现技术，能够针对不同专业领域的数据挖掘与分析问题，开展相应的数据挖掘与分析工作。

参照国外相关大学的教材、课件和应用实例，本书内容的编排顺序主体上是按照一个典型的知识发现过程进行编排的，分别是基本概念、数据预处理、数据仓库构建、关联规则挖掘与相关性分析、聚类分析(无监督的学习分类)、分类方法(有监督的学习分类)。在相关方法与算法讲解的基础之上，进一步展示用本书所介绍的数据挖掘与相关知识开展的一个快速消费品领域消费者调查问卷的挖掘与分析实例，以及在此基础上所构建的一个消费者皮肤状况预测模型。

作为面向非计算机专业学生的课程，本书以介绍概念和讲解方法的主要思想为主。对于有进一步深入学习需求的学生，建议进一步研读高级机器学习、高级数据挖掘等知识内容相关的书籍。在课程教学计划安排上，建议理论方法讲解安排32学时，同时安排16学时的课程实践与讨论环节，以进一步增强学生在数据挖掘与分析方面的应用实战能力，提升未来对于本专业领域数据挖掘与分析的能力。

由于作者水平所限，本书在编写过程中纰漏和疏忽之处在所难免，望读者不吝指正。

徐　华

2014年初春于清华园

关于教学计划编排的建议

采用本书作为教材时，视学生具体情况、教学目标及课时总量的不同，授课教师可从以下两种典型的学时分配方案中选择其一。

教学内容 \ 教学方案与学时分配			方案 A	方案 B
部分	章	节		
一、引言	第 1 章　绪论	1.1～1.8	2	2
二、基本方法	第 2 章　数据预处理	2.1～2.7	4	4
	第 3 章　数据仓库	3.1～3.8	2	2
	第 4 章　相关性与关联规则	4.1～4.6	4	4
	第 5 章　分类和预测	5.1～5.10	6	6
	第 6 章　聚类分析	6.1～6.9	6	6
	第 7 章　数据挖掘应用	7.1～7.6	2	2
三、应用与讨论	讨论课 1	文献调研讨论课	2	3
	讨论课 2	课程设计方案讨论课		3
	讨论课 3	课程成果展示讨论课	2	3

本书所有相关教学资料均向公众开放，包括勘误表、插图和讲义等。

目　录

第1章 绪 论

1.1 应用背景

自从20世纪80年代以来，随着信息技术的高速发展，特别是大型商业数据库的普及应用，各个单位、各个行业都积累了一定规模或超大规模(海量)的数据信息。这些数据信息往往以一定的形式存储在各种类型的商业数据库或者文件系统中。近年来，随着社会生活与商业应用的发展，很多公司和个人都迫切希望能从所拥有的海量数据集中发现对其生活、工作等有帮助的潜在信息或者规律，即希望能够从已有的数据中发现一些“知识”。

另一方面，互联网的普及与发展使得互联网也成为当今社会一个重要的数据源。我国的网页数量已经从2006年的45亿网页规模迅速膨胀为2013年的1500亿网页规模。这些页面中包含了丰富的信息内容，从这些海量数据中发现目前工作所需的有用信息或者知识已经成为人们的普遍需求。尽管在互联网上有像Google、百度、搜狗等这样一些搜索引擎工具，但这些网络信息的搜索工具主要在信息的物理层面上辅助人们做好相关的信息检索工作，并不能根据用户的需求从被检索信息中发现或者获取潜在的知识。中国互联网网页数量与增长率变化如图1.1所示。

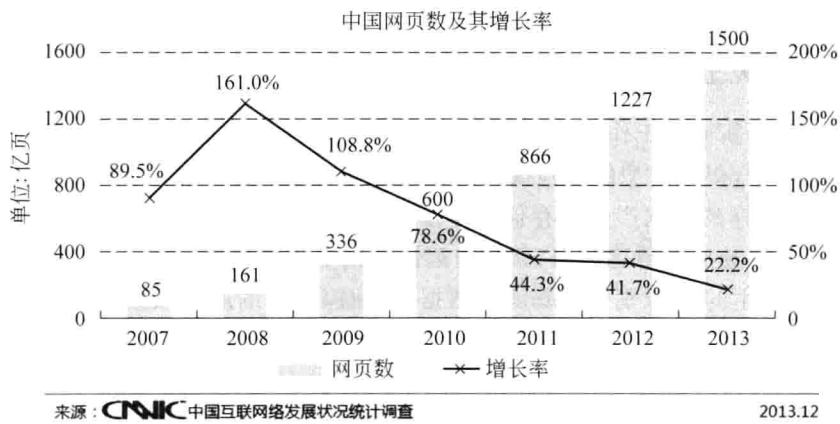

图1.1 中国互联网网页数量与增长率变化图

面向以上对于数据分析的需求，一门跨越数据库技术、信息检索技术、算法、统计学和机器学习等领域的新兴研究领域——“数据挖掘”应运而生。数据挖掘就是指从当前数据集中发现并获取有用信息的过程。数据挖掘是伴随着数据库技术的出现而出现的，同时它的发展为商业应用和科学研究所驱动。

下面探讨推动数据挖掘方法与技术发展的几个关键因素。首先介绍商业上驱动数据挖掘技术发展的原动力；其次介绍科学研究上驱动数据挖掘技术发展的潜在需求；最后介绍和

讨论数据挖掘技术与数据库技术的并行发展历程。

1.1.1 商业上的驱动

当前，商业领域是数据挖掘技术重要的应用领域之一。其数据主要来自于电子商务数据、Web 数据、商场或者零售连锁店的销售数据、金融与信用卡数据、各类交易数据等。电子商务数据主要包括发生电子商务行为相关的所有数据，例如网络书店上的电子商务数据，包括图书的销售数据、图书的库存数据、图书的进货数据、购书人的相关信息、购书人对于图书的浏览日志信息以及图书检索信息等。Web 数据主要是指从各大网站页面上所获得的新闻、评论等相关数据信息。金融数据主要是指银行的所有交易数据。信用卡数据主要包括持卡人持卡消费的所有交易数据。交易数据主要包括商业领域的各类交易数据，如股票交易、期货交易等相关的数据。

推动数据挖掘技术在商业领域研发与应用的一个直接原因是商用计算机等基础硬件设施的价格越来越便宜，同时计算机的运算能力越来越强，即性价比越来越高。由于计算机等基础设备性能价格比的提升，使得对于商业领域的海量信息处理成本大幅度降低，数据挖掘成为可能。

推动数据挖掘技术在商业领域研发与应用的另外一个原因是来自于商业领域强大的竞争压力。例如，在金融业务领域，当前国外各大金融机构为了给客户提供完善优质的金融业务服务，往往会采用数据挖掘方法进行客户关系的管理，即首先对客户类型进行区分，其次对客户的消费习惯进行分析，最后根据用户的消费习惯为客户推荐相应的金融服务。由于采用数据分析与挖掘技术给各大金融机构带来巨大的效益，在美国各大商业银行均构建了面向自身业务内容的数据挖掘与分析系统。

1.1.2 科学研究上的驱动

数据挖掘与分析研究工作发展的另外一个直接驱动力来自于科学研究工作的需求。在实际的科学实验中，很多大型的实验仪器设备或者实验系统会以很高的生成速度产生并存储大量的数据，这样的数据产生与存储的速度往往是每小时 GB 量级的数据。典型的科学实验系统包括卫星的远程传感数据、天文望远镜的太空扫描数据、微阵列产生的基因表达式数据、科学仿真产生的 T 级别的仿真实验数据、石油探测上的地质数据、气象卫星的云图数据等。针对上述规模的数据，传统的技术很难实现对于此类源数据的分析与挖掘工作。海量信息所研究的数据挖掘(即大数据挖掘与分析方法)方法与技术能够适用于超大规模数据的应用挖掘与分析工作。数据挖掘技术可以帮助不同领域的科学家实现对于数据的分类与划分、完成科学的假设性验证等方面的工作。

1.1.3 数据挖掘伴随着数据库技术而出现

数据挖掘技术是伴随着数据库技术的发展而兴起的。回顾数据库技术的发展历程，其发展主要可以分为如下几个阶段。

(1) 在 20 世纪 60 年代，随着电子计算机的出现，应用的发展需要使用计算机对不同业务领域的数据进行收集，因此“数据库(DataBase)”这一特殊的文件系统应运而生。数据库是“按照数据结构来组织、存储和管理数据的仓库”。管理数据库的系统通常称为数据库管

理系统(DataBase Management System,DBMS)。随着数据库应用的普及,基于计算机和数据库技术的信息管理系统也应运而生,它主要面向不同业务领域对于数据管理的需要,利用数据库技术实现对于信息的管理。伴随着网络技术的兴起,网络化的数据库管理系统(Network DBMS)也得到了深入研究与发展。

(2) 随着数据库技术的发展,1970 年美国 IBM 公司圣何塞(San Jose)研究室的研究员 E. F. Codd 首次提出了数据库系统的关系模型,开创了数据库的关系方法和关系数据理论的研究,为数据库技术奠定了理论基础。由于 E. F. Codd 的杰出工作,他于 1981 年获得 ACM 图灵奖(国际计算机领域的最高奖)。关系模型由关系数据结构、关系操作集合和关系完整性约束三部分组成。关系数据库是支持关系模型的数据库系统。

(3) 20 世纪 80 年代以来,各大数据库系统软件商新推出的数据库管理系统(DBMS)几乎都支持关系模型,非关系型系统的产品也大都加上了关系接口。数据库领域当前的研究工作也都是以关系方法为基础。在此期间,一些更高级的先进数据模型被提出,例如扩展的关系数据模型(Extended Relational Data Model)、面向对象模型(Object Oriented Model)和规约模型(Reduction Model)等被提出,相应地数据库技术也获得了持续发展。在这一发展阶段,数据库技术另一个重要的进步是出现了面向应用的数据管理系统,例如面向空间探索、科学研究和工程等应用的数据库管理系统。

(4) 20 世纪 90 年代以来,随着数据库技术的应用普及,人们开始关注从数据库中发现和获取隐含的知识,于是对于数据挖掘方法的研究取得了较深入的进展。特别是数据仓库(Data Warehouse)概念的提出,相关领域专家又深入研究了多媒体数据库(Multimedia Databases)和基于 Web 信息的数据库(Web Databases)。

(5) 进入 21 世纪以来,数据挖掘技术又在应用的深度和广度上获得了进一步的拓展。特别是在数据流的管理与挖掘上取得了重要的研究进展。在包括金融、生物、医药、产品研发等多个领域,数据挖掘技术获得了广泛的应用,在取得一定经济效益的同时,使相关企业的核心竞争力获得了显著提升。近年来,随着 Web 技术的广泛应用,针对 Web 内容的数据挖掘研究也获得了快速进展。

1.2 什么是数据挖掘

自 20 世纪 90 年代以来,随着数据库技术应用的普及,数据挖掘(Data Mining)技术已经引起了学术界、产业界的极大关注,其主要原因是当前各个单位已经存储了超大规模,即海量规模的数据,未来能够真正发挥这些数据的实际价值。由于数据分析和管理工作的应用需要,需将这些数据转换成有用的信息和知识,即从传统的数据统计向数据挖掘与分析进行转换。另外,通过数据挖掘技术获取的信息和知识还可以广泛应用于各个行业领域,包括市场开拓与分析、商务管理、生产控制、工程设计和科学探索等方面。

1.2.1 基本描述

"数据挖掘"也称为从数据中发现知识,具体来讲就是从大规模海量数据中抽取人们所感兴趣的非平凡的、隐含的、事先未知的和具有潜在用途的模式或者知识。回顾数据挖掘研究的历程,不同的名称都被赋予了数据挖掘的含义,包括从数据库中发现知识(Knowledge

Discovery in databases (KDD))、知识抽取(Knowledge Extraction)、数据/模式分析(Data/pattern Analysis)、数据考古(Data Archeology)、数据捕捞(Data Dredging)、信息收获(Information Harvesting)和商业智能(Business Intelligence)等概念都被赋予了数据挖掘的含义。

在解释数据挖掘的概念时,有一点需要特别强调,并非所有与数据库相关的操作与分析都属于数据挖掘研究的范畴。例如,对于数据库简单的搜索与查询处理操作并不属于数据挖掘研究的内容;而对于基于数据库已有的数据所构建的规约式专家系统也不属于数据挖掘的范畴。

1.2.2 关于知识发现

从一组大规模或者海量数据中发现和挖掘新的具有潜在用途的模式或者知识的过程也被称为知识发现 。如图 1.2 所示,一个典型的知识发现过程包括如下几个主要步骤:首先将存放在数据库中的数据经过数据清洗、数据抽取、数据转换、数据集成等预处理过程存入数据仓库中;其次,将清洗过的数据再次经数据抽取或者集成等过程,获得任务相关数据;第三,在此基础上进一步进行数据挖掘过程,获得潜在的有价值的模式或者规律;最后进行模式评估,评估所获得知识的有效性,以此最终获得相关知识。

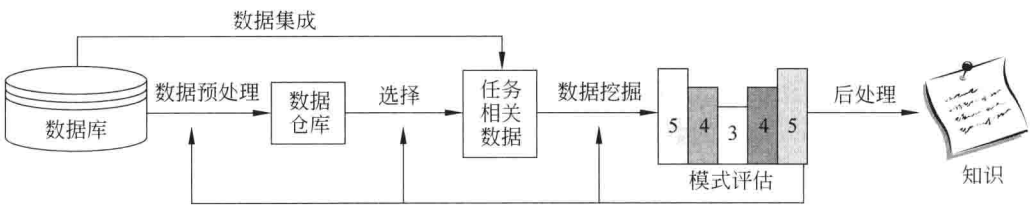

图 1.2 一个典型的知识发现过程

1. 数据挖掘与知识发现

从严格意义上讲,数据挖掘与知识发现是有区别的。对于数据库中的知识发现,主要是指发现数据中有用的信息和模式的过程。而数据挖掘是指在知识发现过程中使用相关的算法抽取有用的信息或者模式的过程。

2. 数据挖掘与商业智能

在商业领域,往往将对于商业数据的智能分析与挖掘的过程称为商业智能。图 1.3 中,一个典型的商业智能过程自底向上分别在 5 个层次开展相关的商业智能分析工作。在最底层是数据源,主要包括论文、文件、网络文档、科学实验、数据库系统等来自不同源头的数据信息,这一层次的工作主要面向数据库分析师;第二层次为数据预处理、数据集成,并形成相应的数据仓库;第三层次对经过预处理的数据进行统计汇总、综合查询和生成报告等工作;第四层次对有用的信息进行数据挖掘工作,第三和第四层次的工作主要面向数据分析师;第五层次将数据挖掘的结果以一定的形式展现出来,用到了数据的科学计算可视化技术,这一层次的工作主要是面向商业分析师;第六层次是决策层,主要是根据发现的知识进行商业上的决策,这一层次的工作主要是面向商业领域的决策者。

图 1.3　数据挖掘在商业智能实现过程中的关系图

1.3　数据挖掘的主要技术

1. 数据挖掘融合了多学科领域的知识

数据挖掘技术利用了来自如下一些领域的方法和技术：

(1) 来自于数据库技术的关系数据模型、结构化查询语言(SQL)、关联规则算法、数据仓库、扩展性技术等；

(2) 计算机算法相关的数据结构、算法分析与设计的理论方法；

(3) 信息检索相关的相似度度量、分层聚类、信息检索系统、近似检索、Web 搜索引擎等；

(4) 来自统计学的贝叶斯理论、回归分析、最大期望估计算法、K 均值算法、时间序列分析等；

(5) 来自机器学习的神经网络、决策树、支持向量机等算法。

近年来，数据挖掘也吸纳了来自其他研究领域的思想方法，这些领域包括最优化、进化计算、信息论、信号处理和科学计算可视化。相关领域的研究工作对数据挖掘应用的实施也起到了重要的支撑作用。

2. 传统的数据统计分析方法与数据挖掘

在谈到数据挖掘方法与技术的时候，很多研究者会问为何不采用传统数据统计的分析方法来获得相关的知识。我们知道，数据挖掘技术是伴随着数据库技术的发展而出现的，数据库中的数据，即数据挖掘分析的对象具有如下几个方面的特征：

(1) 海量数据。

数据挖掘所处理的数据规模往往要求能够扩展到处理以 TB 为计数单位的数据，数据规模是传统数据统计分析方法所面临的一大挑战。

(2) 高维数据。

存储在数据库中的数据往往是具有成千上万维度规模的数据，传统的数据分析方法处理如此高维度的信息将面临很大的困难。

(3) 高复杂性的数据。

当前数据库中所存储的数据往往是具有高复杂度的数据，这些数据具有如下的特点：规模巨大，随着时间而不断的累积增长。如下是在日常工作中几类典型的高复杂度数据。

① 数据流与传感数据。

② 时间序列数据、随时间而变化的数据序列。

③ 结构化数据、图、社会关系网络、多链接关系数据。

④ 异构数据库、法律数据。

⑤ 空间数据、时空描述数据、多媒体数据、Web 数据。

⑥ 软件程序、科学仿真数据等。

(4) 新的复杂数据应用。

近年来，随着计算机技术和网络技术的发展，新的数据挖掘的应用需求不断涌现。例如对于人口调查问卷的分析、日用化工产品性能的分析等。随着应用的发展，新的应用需求不断涌现，这些崭新的应用需求往往是传统数据统计分析方法所不能处理的。

根据当前在现实工作中数据挖掘所解决的问题，利用数据挖掘技术可以实现如下几个方面的功能。

1) 多维概念的描述：特征抽取与识别

在现实生活中描述或者陈述一个事物或者人物时，常常会用这类事物或者人物的某个特征来对其进行描述，以区别于其他被描述的对象或者特征。例如描述一个人时，常常用这个人物的姓名、性别、年龄、身高、体重等特征来描述。特征的识别与抽取就是通过规范化、总结和对比的方式抽取被分析对象的特征。

2) 频繁模式、相关性、关联规则与随机性

与随机性出现的事物和现象相比，数据挖掘就是从大量随机的被分析对象数据中获取规律性的频繁发生的关联模式与规律信息。经典的数据挖掘分析案例——啤酒与尿布案例就说明了这一点。20 世纪，国际上一些大型的超市利用数据挖掘技术分析了客户购买商品的搭配情况，发现了一个很有意思的现象，就是购买啤酒的男士往往也会同时购买小孩的纸尿布。针对这一有意思的现象，超市随即在商品摆放上将啤酒与小孩的纸尿布放在一起，从而明显提升了两种商品的销售数量。从上述利用数据挖掘方法开展商业数据的分析过程中可以看出，数据挖掘就是要从大量随机发生的事件中抽取频繁的具有相关性的规律使之服务于商业决策和日常生活。

3) 分类与预测

数据挖掘相关的研究工作中常常还力图构建一个模型或者描述函数来刻画或者区分不同的类型与概念，以实现对于未来潜在的预测需求。例如在实际工作中，往往会根据气候的类型来对相关国家进行分类，分为热带国家、温带国家和寒带国家。实际生活中，会根据小汽车的排量对小汽车进行分类，分为小排量汽车、大排量汽车等类型。

在实际应用数据挖掘技术解决相关问题的过程中，常常会采用分类技术与方法解决对未知的结果或者未知量化特征的预测。