清华大学 计算机系列教材

徐华 编著

数据挖掘

方法与应用——应用案例

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-应用案例

清华大学出版社
北京

内容简介

本书主要以作者近五年在清华大学开展数据挖掘应用研究和教学工作为基础，从所指导的多个数据挖掘与分析的应用案例中精选出包括交通、体育、金融、生物信息、社交网络、电力等领域代表性的数据挖掘与分析案例，结合基本的数据挖掘应用实施思路，展示了在不同行业领域开展数据挖掘与分析工作的实际过程。

本书可作为高等院校学生学习数据挖掘的参考读物，同时可供工程技术人员开展数据挖掘与分析工作时参考。

图书在版编目(CIP)数据

数据挖掘：方法与应用——应用案例/徐华编著. —北京：清华大学出版社，2017
（清华大学计算机系列教材）
ISBN 978-7-302-47211-7

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(2017)第 113144 号

责任编辑：白立军　张爱华
封面设计：常雪影
责任校对：李建庄
责任印制：刘海龙

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：北京泽宇印刷有限公司
经　　销：全国新华书店
开　　本：140mm×210mm　**印张**：5.875　**彩插**：4　**字　　数**：154 千字
版　　次：2017 年 8 月第 1 版　**印　　次**：2017 年 8 月第1次印刷
印　　数：1～2000
定　　价：19.00 元

产品编号：062379-01

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，基于移动互联网的数据采集能力不断提升，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。作为大数据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”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”，并于2014年10月出版了《数据挖掘：方法与应用》教材。在实际教学和应用研发过程中，我们深感数据挖掘工作与专业背景知识相结合的重要性，为了能让不同专业领域的同学和工程技术人员更加深入地理解如何开展一个高质量的挖掘和分析工作，我们从所指导的不同专业背景团队应用实施案例中精选出多个有代表性的实施案例进行介绍与点评。

本书所讨论的案例数据均来自于国内外相关开放数据，精选了交通、体育、金融、生物信息、社交网络和电力等领域代表性的案例，分别从问题描述、挖掘与分析过程和案例点评三大方面对上述领域的案例进行介绍与讨论。

作为《数据挖掘：方法与应用》一书配套的案例教材，本书在内容编排上以应用思路的讲解为主，特别强调将数据挖

掘方法与专业领域的背景知识相结合，挖掘与分析出高质量的结果。本书作为相关课程的配套实验教材，可作为高等院校学生学习数据挖掘的参考读物，同时也可为工程技术人员开展数据挖掘与分析工作提供实施思路的指导。

由于作者水平所限，疏漏之处在所难免，望读者不吝指正。

最后，感谢“清华大学 2015 年秋季学期本科教改立项项目”对本教材的立项支持。

徐　华

2017 年初春于清华园

目　录

第1章 绪 论

1.1 本书背景

在信息技术领域，衡量计算机数据存储规模的单位从小到大依次包括B、KB、MB、GB、TB、PB、EB、ZB、YB、DB、NB等。根据对IDC行业的统计，2004年全球的数据量为30EB，2005年达到50EB，2006年达到161EB，目前全球范围内数据存储的规模大约为ZB量级，现在每两天产生的数据量相当于有史以来到2003年的数据量总和。

自从有了一定规模的数据之后，各行各业都在想办法从已有的数据中发现有一定价值或者规律性的东西，即开展数据挖掘的应用性研究工作。利用已有数据，不同的机构和个人曾试图开展一些有价值的工作，然而同样的工作往往会得到不同的挖掘和分析结果。例如，针对1936年的美国总统选举，美国《文学文摘》(*Literary Digest*)杂志根据当时的电话号码簿及该杂志订户俱乐部会员名单，邮寄了1000万份调查问卷，回收了约240万份。工作人员对此回收的问卷数据进行了精确的计算，他们断言：在1936年美国总统选举中兰登以57%比43%，领先14个百分点击败罗斯福。而与之相反，一个名叫乔治·盖洛普的人，对《文学文摘》的调查结果的可信度提出质疑。他也组织了抽样调查，按照他的设计抽样方法仅仅调查了3000人。他的预测与《文学文摘》截

然相反，认为罗斯福必胜无疑。事实结果是罗斯福以62%比38%压制性地大胜兰登。这一结果使《文学文摘》销声匿迹，而盖洛普则名声大振。事实上，今天很多大数据挖掘与分析的工作都在重复《文学文摘》的例子。从《文学文摘》的例子中可以看出，对于同样一项调查分析工作，由于采用的数据采集、处理、挖掘和分析方法不同，往往会得到完全不同的分析结果。

所以从应用的角度来说，若要保证高质量的挖掘结果，除了要有高质量的数据之外，还要采用最适合的数据挖掘与分析方法。要选取最适合的挖掘与分析方法，除了要有数据挖掘与分析的基本知识外，还需要从相关专业领域的角度对行业数据有透彻和深入的理解，选取有效适合的挖掘分析策略与方法。

本书正是在这样的一个应用背景之下，精选了不同行业领域的真实数据，探讨和点评不同行业领域数据挖掘与应用的案例，并希望从不同行业领域的挖掘案例中得到相关的启发和指导。

1.2 数据挖掘应用概述

数据挖掘就是从大量的数据中挖掘出有价值的信息或者知识。目前随着各行业领域数据积累的不断增多，不同单位和个人纷纷采用不同的数据挖掘理论方法和工具开展相关行业领域的数据挖掘与分析工作。从挖掘的目标角度来看，主要完成两个方面的挖掘工作：一方面是发现规律模式的描述性挖掘；另一方面是对于未来趋势性挖掘。

1. 规律模式的描述性挖掘

传统的“啤酒—小孩纸尿布”这一经典的购物篮问题就是一类典型的描述性挖掘问题。近年来，结合公共卫生、消费心理、市场营销、自然语言处理等领域的专业知识，对消费者的行为特征进行分析，根据此行为特征与受众客户的心理，合理布局相关的商品位置，并推送相关商品的促销信息。这是一类崭新的购物篮分析与商业决策应用的经典案例。随着技术的发展，数据挖掘领域还实现了跨越多模态信息形式的综合性挖掘任务。例如，基于传统的消费者调查问卷数据与消费者洗衣视频的综合挖掘，通过消费者洗衣视频挖掘消费者的行为规律特征，并与调查问卷的基本数据进行跨维度分析，综合挖掘和分析消费者的行为特征。

2. 未来趋势预测性挖掘

传统的股票预测软件根据股票交易市场的历史数据，预测股票价格的趋势，这是一类典型的预测性挖掘任务。近年来，结合金融、自然语言处理、心理学等领域的专业知识，实现基于网络评论信息和社交媒体数据分析的量化投资分析与投资预测。这是在量化投资分析领域所实现的对于股票价格趋势分析的一种新的尝试，也是数据挖掘技术在金融领域的一类创新性应用。

1.3 本书的主要内容安排

本书共精选了交通、体育、金融、生物信息、社交网络和电力等领域的真实需求和应用问题，基于不同行业领域所开放的真实数据，开展了相应的数据挖掘与分析工作。本书所讨论和介绍的所有应用实例均是从上述领域精选出的数据挖掘与分析的典型实例。希望能够通过对相关案例的讨论、分析和点评，加深对于数据挖掘与分析方法在不同专业领域应用实施的理解，体会进行数据挖掘与分析工作的思路与实施策略。

本书首先对数据挖掘应用案例的背景情况进行介绍。第 2 章基于中国南京市所公开的出租车 GPS 数据，研究南京市出租车行驶轨迹规律的挖掘与分析方法。第 3 章针对美国 NBA 篮球赛的历史数据，研究对于未来比赛结果的预测方法。第 4 章针对大型商业银行系统的后台日志信息，研究商业银行系统的故障模式挖掘方法。第 5 章针对生物信息学领域的 RNA 测序，研究 RNA 排序的预测方法。第 6 章针对社交网络中的微信公众号的统计数据，研究其变化趋势的分析方法。第 7 章针对股票行业的多空状况，探讨多空的预测方法。第 8 章针对电力领域监测数据，研究如何采用数据挖掘和分析方法自动辨识电力系统中的不良数据。最后，第 9 章总结本书所讨论的内容。

1.4 小　结

当前，随着各行各业数据存储规模的扩大，海量数据或者大数据已经成为当前我们所面临的一个普遍问题。如何从海量的或超大规模数据中发现有价值的信息或知识，已经成为摆在各个行业的决策者、分析师和知识管理人员面前的一个普遍问题。本书立足于数据挖掘的基础理论方法，从六大行业领域精选了七个典型的挖掘与分析案例进行深入的讲解和点评。希望通过应用实例的介绍和点评，加深不同行业领域的读者对于数据挖掘应用的感性体验，深刻理解数据挖掘的实施思路。

第2章　基于GPS信息的出租车行车轨迹数据挖掘

2.1　概　　述

公共交通领域的数据挖掘与分析，对于改善民生、提升政府对于城市道路的管理与决策能力具有重要的意义。目前我国许多大中城市为便于实现对于出租车的管理，通过城市出租车相关的轨迹数据分析，初步实现了对于城市交通状况的分析和预测，为政府有关主管部门提供了管理与决策的依据。

本章数据主要来源于国内知名的数据共享网站“数据堂”(datatang.com)，数据的主要内容是南京市出租车GPS位置数据(2010年9月1日到9月2日)。具体数据记录了该城市7000辆出租车及每部出租车的GPS记录数据(采样频率：2条/分钟)，总量为4000万条。

2.2　出租车GPS数据挖掘问题描述

1. 挖掘与分析目标

在获得海量规模的出租车GPS数据之后，需要先深入分析一下目前可做的挖掘与分析目标。基于出租车GPS位