大数据分析与数据挖掘

简祯富　许嘉裕 / 编著

DATA

清华大学出版社

大数据分析与数据挖掘

简祯富　许嘉裕 / 编著

清华大学出版社
北京

内容简介

随着移动通信和行动装置普及、物联网和网络发展，以及云端技术的不断进步，现今数据产生、搜集和储存方式比以往更为方便。数据挖掘与大数据分析可以从海量数据中，找到值得参考的样型或规则，转换成有价值的信息、洞察或知识，创造更多新价值。

本书主要介绍数据挖掘与大数据分析的理论方法与实践应用，并加入丰富的实务案例介绍，具体说明如何应用数据挖掘与大数据分析技术以解决真实问题，深入浅出地剖析从数据中掏金的秘诀。全书共分为13章，内容涵盖数据挖掘基本概念与数据准备、数据挖掘的方法与实证、数据挖掘的进阶运用；书中也提供R语言与编程实例辅以说明，使读者更能融会贯通地应用数据挖掘方法，进而提升大数据分析和数字决策能力。

图书在版编目(CIP)数据

大数据分析与数据挖掘/简祯富，许嘉裕编著. —北京：清华大学出版社，2016

ISBN 978-7-302-42425-3

Ⅰ. ①大…　Ⅱ. ①简… ②许…　Ⅲ. ① 统计数据—统计分析 ②数据采集　Ⅳ. ①O212.1 ②TP274

中国版本图书馆CIP数据核字(2015)第306758号

责任编辑：冯　昕
封面设计：张京京
责任校对：王淑云
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
印 装 者：北京鑫海金澳胶印有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：23　　**字　　数**：560千字
版　　次：2016年3月第1版　　**印　　次**：2016年3月第1次印刷
印　　数：1～2000
定　　价：49.00元

产品编号：065663-01

前言 Foreword

“为大于其细，行远必自迩！”

1992年我到美国威斯康星大学麦迪逊分校(UW-Madison)攻读决策科学与作业研究博士时，发现我在新竹“清华大学”念的概率、统计、实验设计和统计方法等课程的教科书作者竟然都是麦迪逊的教授，所以选择统计作为副修；另一方面，我又在麦迪逊的医疗系统研究分析中心(Center for Health Systems Research and Analysis，CHSRA)担任研究助理，参与由Gustafson教授领导的大型研究团队发展的“综合医疗促进支持系统”(Comprehensive Health Enhancement Support System，CHESS)，计划的目的是借着提供信息(information)、转介服务(referral to service providers)、决策支持(decision support)和社会援助(social support)等方式，帮助面对疾病和健康危机的人(如癌症和艾滋病患者)及其亲友取得相关信息、寻求可利用的资源、分析决策，以及社群服务和互相扶持等。我的主要工作是分析系统所搜集的使用数据和用户填写的问卷调查数据等，并在每周研究团队的定期会议上进行汇报，通过各种可能的分析和数据探索，以证明CHESS的效益。因为我的指导教授当时只是团队中的助理教授，所以我特别卖力分析，生怕工作不保就没有奖学金了。有一天，研究团队的一位成员在会议后告诉我说，我做的工作好像“数据挖掘”(data mining)，他认为数据挖掘的方法将来可能会超越统计，虽然当时我觉得怎么可能有一种最近才发展的方法，可以超越已有几百年根基的统计学，但也让我注意到数据挖掘这个研究领域。

1996年我回到新竹“清华大学”任教，即成立“决策分析研究室”(Decision Analysis Laboratory，DALab)，和研究伙伴与学生们包括本书共同作者许嘉裕博士一起投入决策分析、数据挖掘和优化的研究和实践工作，并通过产学合作计划作研究，然而却苦无合适的教材训练学生，特别是结合实际案例的课本，因此就持续借着整理产学合作研究成果、撰写期刊论文和指导学生论文之机，准备撰写教科书的基础材料。数据挖掘和大数据分析是方法论，也是实证推导模式(empirically derived model)，因此必须结合方法发展与实证研究以检验研究效度。决策分析研究室研究团队与台积电、旺宏、台达电、联发科、广达电脑、创意电子、晶元光电、采钰、关东鑫林、茂迪、普生、力晶、世界先进等公司建立双赢的产学合作机制，做到学术研究贡献能够接连获奖，而实际效益能够达到合作厂商产业化的要求，作为更深一层理论研究的基础；更有幸从2005年借调台积电三年，实际应用所发展的分析方法在企业营运中，领导研究室的学生们和工业工程处同仁们一起推动台积电“IE十大建设”并发展相关的分析技术和数字决策系统，提供数字化系统化之决策依据，而从中得到产业导师宝贵的指导和回馈，也累积实战的经验和心得；进而执行台湾“科技部”“IC产业同盟”(Semiconductor Technologies Empowerment Partners Consortium，STEP Consortium)暨

深耕工业基础技术计划，并成立“清华-台积电卓越制造中心”(NTHU-TSMC Center for Manufacturing Excellence)，把累积多年的实证及大数据分析技术，推广到半导体供应链上、下游和其他高科技产业，借此提升产业的决策分析和智能制造能力；并通过主办“清华IC学堂”“半导体大数据分析竞赛”及产学合作成果发表研讨会等活动，培养具备跨界创新、团队合作能力的“资料科学家”。因此，本书在编撰过程中一再修改更新，希望一方面能深入介绍数据挖掘与大数据分析的基础方法和工具，另一方面则通过跨领域的实际案例和范例程序，以具体培养结合理论与实务的决策科学家。

非常感谢新竹“清华大学”和元智大学的良好学术研究环境和科学园区的地利人和，使我们可以结合理论与实务，从产业大数据和具体问题的实证中发展适用的方法、检验所学，再进而导向更深一层的研究。随着问题的广度和复杂度以及合作伙伴的阶层和领域而不断成长，这一路走来，虽然整个研究团队一直秉持自强不息、行胜于言的精神努力提升，但也得力于产业先进和合作伙伴们的提携协助和计划执行过程中的指导，因此要感谢的人非常多，希望借着本书的出版能使更多读者从中得到启发和实际的帮助，以造福社会和产业，也算是间接回报所有关心和帮助我们的人。尽管本书经过长期的准备，但完稿阶段所花费的心力远远超过预期，特别感谢专任助理梁婉玲编辑汇总的工作和与出版社的联络，减少本书错误的可能，以及决策分析研究室同学们一起打拼完成各项研究计划，这也是本书各案例的论文均引用完整作者名单的原因；也感谢在“数据挖掘”课程教学中每位互动的学生，让我们得到教学相长和调整教材的回馈建议。本书自2014年在台湾出版以来，引发学术界和产业界的广泛回响，成为多所大学和各大企业的指定教材。感谢北京清华大学出版社理工分社张秋玲社长和冯昕主任的支持，将全书重新编辑改版，去芜存菁，并增添一章全新章节，使内容更加丰富完整。然而，本书疏漏之处在所难免，盼诸位领导和前辈，不吝赐教，以提升大数据分析和数字决策能力。

简祯富　许嘉裕　谨识

IC产业同盟，2015冬

目录 Contents

第1篇 大数据分析与数据挖掘导论

第3篇 数据挖掘进阶运用

第1篇

大数据分析与数据挖掘导论

第1章

大数据分析与数据挖掘概论

1.1 前言

随着信息科技的进步和网络的发达、计算机运算能力的增强以及数据搜集与储存技术持续改进的影响，大幅改变数据的分析和应用方式，**“大数据分析”(big data analytics)**和**数据挖掘(data mining)**可以发掘先前未知且潜在有用的信息样型(patterns)或规则(rules)，进而转化为有价值的信息或知识，帮助决策者迅速做出适当的决策，是现代企业重要的竞争优势。

由于自动化的生产环境、智能手机的普及、电子商务的发展、物联网的建立以及社交网络的发达，现在多数人都可以不受时空地点限制地上网，浏览社交网络，在网络上聊天、购物，以及实时收看与查询最新的新闻报道与文章等，也可以用来管理远程的生产和服务系统。当你在微博上打卡点赞、收发电子邮件、到便利商店购买零食、搭乘大众交通工具、经过停车场利用信用卡缴费时，这些日常生活中的习惯与动作，随时随地正透过网络记录，快速累积成巨量数据或大数据。过去对商品的评价主要是通过口口相传，而现在则是借由在线文章发表，由社交网络快速扩散，这意味着网络经营的重要性已开始逐渐大过实体经营，大数据分析正引领着数字决策并带来新商机。

“数据”在经济学中属于非竞争性的商品，其与物质性的东西(例如食物、车等)不同，并不会因为使用次数增加而降低价值或造成耗损。因此，零售业者累积的事务数据可以一再使用，根据不同目的提取不同的数据，或运用于不同的目标对象上(Mayer-Schonberger & Cukier，2013)。除了传统的统计分析和数据挖掘外，大数据分析技术和应用正改变我们的生产方式、服务系统和生活形态。

每一秒，一间大型医院会增加 12 万笔健康相关的生理数据；每一分钟，YouTube 网站会接收到民众上传总长达 72 小时的视频；每一天，一家银行的信用卡交易次数达 500 万笔。时间分秒走过的同时，大量数据也随时都在快速累积，如图 1.1 所示。而在全世界数兆个传感器、超过五亿部智能手机、十亿台计算机上，每一天不断运作所产生的数据量估计高达 25 亿 GB(胡世忠，2013)。科技研究公司 IDC 更预估，到 2020 年全球数据量将累积达 40 000 ZB(Gantz & Reinsel，2012)，数据储存单位如表 1.1。

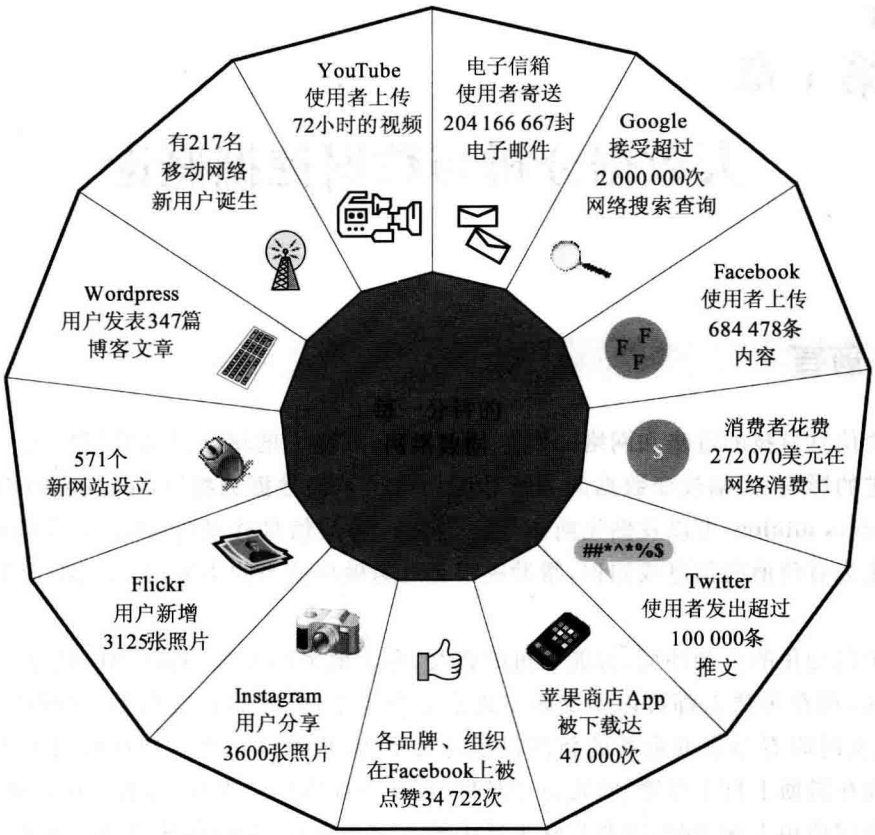

图 1.1 持续增加的大数据(胡世忠,2013)

表 1.1 数据的储存单位

储存单位/B	文件储存单位
Kilobyte (KB)	1 KB=1024 B=2^{10} B
Megabyte (MB)	1 MB=1024 KB=2^{20} B
Gigabyte (GB)	1 GB=1024 MB=2^{30} B
Terabyte (TB)	1 TB=1024 GB=2^{40} B
Perabyte (PB)	1 PB=1024 TB=2^{50} B
Exabyte (EB)	1 EB=1024 PB=2^{60} B
Zettebyte (ZB)	1 ZB=1024 EB=2^{70} B
Yottabyte (YB)	1 YB=1024 ZB=2^{80} B

大量的传感器与电子卷标置入到日常生活的电子设备中,例如手机、监控摄影机、环境温度传感器、水电天然气表等,随时感测人们的生活动态。例如,电力公司为了节省能源,开发的智能电表和智能电网即装置了大量的传感器,24 小时不间断地测量与传输终端顾客的电力使用信息。对终端顾客而言,智能电表能实时显示家中的用电量,协助用户调整用电习惯。对电力公司而言,则可透过实时用电量的监控,掌握电网供电状态,当耗电量可能超过