21世纪高等学校规划教材 | 计算机应用

互联网数据分析与应用

赵守香　姜同强　编著

清华大学出版社

互联网数据分析与应用

赵守香　姜同强　编著

清华大学出版社
北京

内 容 简 介

本书从“挖掘数据的潜在价值和应用”的视角出发，针对互联网企业、政府、传统企业、个人等不同的社会角色对数据挖掘和分析的应用需求，系统地介绍了大数据时代数据分析的作用、技术和具体应用。

本书共分为8章，内容包括大数据与数据分析、互联网数据存储、互联网数据分析工具、商务网站数据分析与应用、政府网站数据分析及应用、物联网数据分析与应用、移动商务数据分析与应用、微博数据分析与应用等，全面地介绍了数据分析的相关内容。

本书可以作为高校“数据分析与应用”、“商务智能”等本科、研究生课程的教材，也可以作为从事互联网数据管理和数据分析的专业人员的参考书。

图书在版编目(CIP)数据

互联网数据分析与应用/赵守香，姜同强编著. --北京：清华大学出版社，2015
21世纪高等学校规划教材·计算机应用
ISBN 978-7-302-37974-4

Ⅰ. ①互… Ⅱ. ①赵… ②姜… Ⅲ. ①互联网络－数据管理－高等学校－教材 Ⅳ. ①TP393.4

中国版本图书馆CIP数据核字(2014)第295238号

责任编辑：闫红梅　薛　阳
封面设计：傅瑞学
责任校对：李建庄
责任印制：李红英

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 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：185mm×260mm　　印　张：19.5　　字　数：484千字
版　　次：2015年9月第1版　　印　次：2015年9月第1次印刷
印　　数：1～2000
定　　价：34.50元

产品编号：054098-01

出版说明

随着我国改革开放的进一步深化，高等教育也得到了快速发展，各地高校紧密结合地方经济建设发展需要，科学运用市场调节机制，加大了使用信息科学等现代科学技术提升、改造传统学科专业的投入力度，通过教育改革合理调整和配置了教育资源，优化了传统学科专业，积极为地方经济建设输送人才，为我国经济社会的快速、健康和可持续发展以及高等教育自身的改革发展做出了巨大贡献。但是，高等教育质量还需要进一步提高以适应经济社会发展的需要，不少高校的专业设置和结构不尽合理，教师队伍整体素质亟待提高，人才培养模式、教学内容和方法需要进一步转变，学生的实践能力和创新精神亟待加强。

教育部一直十分重视高等教育质量工作。2007 年 1 月，教育部下发了《关于实施高等学校本科教学质量与教学改革工程的意见》，计划实施“高等学校本科教学质量与教学改革工程（简称‘质量工程’）”，通过专业结构调整、课程教材建设、实践教学改革、教学团队建设等多项内容，进一步深化高等学校教学改革，提高人才培养的能力和水平，更好地满足经济社会发展对高素质人才的需要。在贯彻和落实教育部“质量工程”的过程中，各地高校发挥师资力量强、办学经验丰富、教学资源充裕等优势，对其特色专业及特色课程（群）加以规划、整理和总结，更新教学内容、改革课程体系，建设了一大批内容新、体系新、方法新、手段新的特色课程。在此基础上，经教育部相关教学指导委员会专家的指导和建议，清华大学出版社在多个领域精选各高校的特色课程，分别规划出版系列教材，以配合“质量工程”的实施，满足各高校教学质量和教学改革的需要。

为了深入贯彻落实教育部《关于加强高等学校本科教学工作，提高教学质量的若干意见》精神，紧密配合教育部已经启动的“高等学校教学质量与教学改革工程精品课程建设工作”，在有关专家、教授的倡议和有关部门的大力支持下，我们组织并成立了“清华大学出版社教材编审委员会”（以下简称“编委会”），旨在配合教育部制定精品课程教材的出版规划，讨论并实施精品课程教材的编写与出版工作。“编委会”成员皆来自全国各类高等学校教学与科研第一线的骨干教师，其中许多教师为各校相关院、系主管教学的院长或系主任。

按照教育部的要求，“编委会”一致认为，精品课程的建设工作从开始就要坚持高标准、严要求，处于一个比较高的起点上；精品课程教材应该能够反映各高校教学改革与课程建设的需要，要有特色风格、有创新性（新体系、新内容、新手段、新思路，教材的内容体系有较高的科学创新、技术创新和理念创新的含量）、先进性（对原有的学科体系有实质性的改革和发展，顺应并符合 21 世纪教学发展的规律，代表并引领课程发展的趋势和方向）、示范性（教材所体现的课程体系具有较广泛的辐射性和示范性）和一定的前瞻性。教材由个人申报或各校推荐（通过所在高校的“编委会”成员推荐），经“编委会”认真评审，最后由清华大学出版

社审定出版。

目前，针对计算机类和电子信息类相关专业成立了两个“编委会”，即“清华大学出版社计算机教材编审委员会”和“清华大学出版社电子信息教材编审委员会”。推出的特色精品教材包括：

(1) 21 世纪高等学校规划教材·计算机应用——高等学校各类专业，特别是非计算机专业的计算机应用类教材。

(2) 21 世纪高等学校规划教材·计算机科学与技术——高等学校计算机相关专业的教材。

(3) 21 世纪高等学校规划教材·电子信息——高等学校电子信息相关专业的教材。

(4) 21 世纪高等学校规划教材·软件工程——高等学校软件工程相关专业的教材。

(5) 21 世纪高等学校规划教材·信息管理与信息系统。

(6) 21 世纪高等学校规划教材·财经管理与应用。

(7) 21 世纪高等学校规划教材·电子商务。

(8) 21 世纪高等学校规划教材·物联网。

清华大学出版社经过三十多年的努力，在教材尤其是计算机和电子信息类专业教材出版方面树立了权威品牌，为我国的高等教育事业做出了重要贡献。清华版教材形成了技术准确、内容严谨的独特风格，这种风格将延续并反映在特色精品教材的建设中。

清华大学出版社教材编审委员会

联系人：魏江江

E-mail：weijj@tup.tsinghua.edu.cn

前言

移动互联时代，数以百亿计的机器、企业、个人随时随地都会获取和产生新的数据。1分钟之内，新浪微博发送数万条微博，苹果应用商店下载次数以万计，淘宝卖出了几万件商品，百度产生了百万次搜索查询……所有这些行为都由海量的数据来呈现。

据CNNIC第35次调查报告显示：截止到2014年12月，我国的网民数量达到6.49亿，手机网民数达到5.57亿，互联网普及率为47.9%。这些网民每天产生海量的浏览数据、交易数据和原创数据，这些数据都存储在经营者的数据库里，占用了大量的存储资源。如何发现这些数据背后所隐藏的有价值的信息？如何利用这些数据来发现机会？

目前，很多高校在本科高年级和研究生阶段都开设了“数据挖掘”、“商务智能”、“客户关系管理”、“企业信息管理”、“物联网技术与应用”等课程，这些课程从不同侧面介绍了数据分析的技术、数学模型、行业应用等。随着大数据时代的到来，越来越多的高校已经意识到数据分析的重要性，培养学生，尤其是高年级本科生和研究生数据分析和应用的能力已经被列入了新的教学计划。

本书从"挖掘数据的潜在价值和应用"的视角出发，从互联网企业、政府、传统企业、个人等不同的社会角色对数据挖掘和分析的应用需求，系统地介绍了大数据时代数据分析的作用、技术和具体应用。

本书共分为8章：第1章 大数据与数据分析，第2章 互联网数据存储，第3章 互联网数据分析工具，第4章 商务网站数据分析与应用，第5章 政府网站数据分析及应用，第6章 物联网数据分析与应用，第7章 移动商务数据分析与应用，第8章 微博数据分析与应用，分别从不同侧面介绍了数据分析的相关内容。

在本书的写作过程中，参考了很多互联网上的相关资料，由于数量众多，不可能在参考文献里一一列出，在此一并向无私奉献自己的智慧和经验的同行们表达衷心的感谢！王晨、杨致远、郜庆月、张学伟为本书搜集了一手资料，对本书的体系架构提出了很好的建议，在此表示感谢！

赵守香
北京工商大学计算机与信息工程学院
2015年5月

目录

第1章 大数据与数据分析

【内容提要】

本章从大数据的出现、大数据的影响及大数据对数据处理的要求出发，分析了大数据环境下对数据分析与利用的重要性，主要内容包括：

(1) 大数据产生的背景和特征；
(2) 电子商务的快速发展与大数据；
(3) 物联网的兴起与大数据；
(4) 移动商务的快速渗透与大数据；
(5) 数据分析给企业带来的机会与好处；
(6) 大数据环境下数据分析的需求；
(7) 数据分析的含义；
(8) 数据分析的作用；
(9) 大数据与数据分析。

1.1 概述

互联网、移动互联网、物联网、云计算的快速兴起以及移动智能终端的快速普及，使得当前人类社会的数据增长比以往任何一个时期都要快。数据的爆炸式增长正在出乎人们的想象。据预计，2020 年，全球以电子形式存储的数据量将达 35ZB，是 2009 年全球存储量的 40 倍。而在 2010 年年底，根据 IDC 的统计，全球数据量已经达到了 1.2ZB。如果将这些数据的 50%刻录在 DVD 上，那么把这些 DVD 盘片堆起来就可以从地球垒到月球一个来回。

与此同时，伴随着物联网、移动智能终端以及移动互联网的快速发展，移动网络中数据流量的增长速度也非常迅猛。从 2011 年开始，全球移动数据流量年增长率将保持在 50%以上，并处于一个稳定增长的态势。到 2016 年，全球移动数据流量将达到 2011 年全球移动数据流量的 18 倍，达到 129.6EB。

数据的疯狂增长，使得适应和应对数据增长成为整个社会关注的焦点。“大数据”的概念正是在这一背景下应运而生。

1.1.1 大数据的含义

我们先来看一组数据：1分钟之内，新浪微博发送数万条微博，苹果应用商店下载次数以万计，淘宝卖出了几万件商品，百度产生了百万次搜索查询……所有这些行为都由海量的数据来呈现。

在2013年12月12日电商的促销期，淘宝网推出“时光机”——一个根据淘宝买家几年来的购买商品记录、浏览点击次数、收货地址等数据编辑制作的“个人网购志”，从而记录和勾勒出让人感怀的生活记忆。背后是基于对4.7亿淘宝注册用户网购数据的分析处理，这正是大数据的典型应用。

随着传统互联网向移动互联发展，全球范围内，除了个人电脑、平板电脑、智能手机、游戏主机等常见的计算终端之外，更广阔的、泛在互联的智能设备，如智能汽车、智能电视、工业设备和手持设备等都连接到网络之中。基于社会化网络的平台和应用，让数以百亿计的机器、企业、个人随时随地都会获取和产生新的数据。

互联网搜索引擎是大数据最为典型的应用之一。百度日处理数据量达到数十拍字节，并呈现高速增长的态势。如果一张光盘容量为1GB，这相当于垒在一起的几千万张光盘。微软Bing(在中国为必应)搜索引擎，一周需要响应100亿次量级的搜索请求。通过和Facebook的合作，每天有超过10亿次的社交网络搜索请求通过Bing来处理。

短短的18个月，中国移动互联网流量增加了10倍。随着社交网络的逐渐成熟、移动带宽迅速提升，更多的传感设备、移动终端接入网络，产生的数据及其增长速度比历史上任何时期都要多，互联网上的数据流量正在迅猛增长。在云计算、物联网等技术的带动下，中国的移动互联网已经步入“大数据”时代了。

而根据市场调研公司IDC的报告，全球信息总量每过两年就会增长一倍，2011年全球产生的数据总量为1.8ZB(1ZB为百万拍字节)，相当于全球历史数据总和。

继云计算后，大数据(Big Data)成为信息技术领域最为热门的概念之一。

简单说，大数据是指那些超过传统数据库系统处理能力的数据。但是，大数据的问题并不仅仅是规模，数据产生的速度以及数据的多样性同样是大数据不可忽略的两个基本特性。根据摩尔定律，计算能力每一年半到两年的时间将增加一倍。可是，现有的网络带宽并没有以同样的速度在增加。因此，如此迅猛的数据洪流的产生，正在给电信运营商的网络运营带来极大的挑战。

在IT业界，有人把大数据产业定义为建立在对互联网、物联网等渠道广泛大量数据资源收集基础上的数据存储、价值提炼、智能处理和分发的信息服务业；也有人概括大数据战略为：致力于让所有用户能够从几乎任何数据中获得可转换为业务执行的洞察力，包括之前隐藏在非结构化数据中的洞察力。

总之，大数据是对大量、动态、能持续的数据，通过运用新系统、新工具、新模型的挖掘，从而获得具有洞察力和新价值的东西。

1.1.2 大数据的定义

根据维基百科的定义，大数据指难以用常用的软件工具在可容忍的时间内抓取、管理以

及处理的数据集。大数据的显著特征有：数据体量巨大；数据类型繁多，包括结构化数据以及非结构化数据如网页、日志、视频、图片等；要求的处理速度快。

大数据技术涵盖了从数据的海量存储、处理到应用多方面的技术，包括海量分布式文件系统、并行计算框架、NoSQL 数据库、实时流数据处理以及智能分析技术如模式识别、自然语言理解、应用知识库等。

1.1.3 大数据的特征

虽然有多种解读，但业界一般认为，大数据有 4 个 V 字开头的特征：Volume(容量)、Variety(种类)、Velocity(速度)和最重要的 Value(价值)。

1. Volume

是指大数据巨大的数据量与数据完整性。IT 业界所指的数据，诞生不过六十多年。而一直到个人电脑普及前，由于存储、计算和分析工具的技术和成本限制，许多自然界和人类社会值得记录的信号，并未形成数据。几十年前，气象、地质、石油物探、出版业、媒体业和影视业是大量、持续产出信号的行业，但那时 90%以上采用的是存储模拟信号，难以通过计算设备和软件进行直接分析。拥有大量资金和人才的政府和企业，也只能把少量最关键的信号，进行抽取、转换、装载到数据库中。

尽管业界对达到怎样的数量级才算是大数据并无定论，但在很多行业的应用场景里，数据集本身的大小并不是最重要的，是否完整才最重要。

2. Variety

Variety 则意味着要在海量、种类繁多的数据间发现其内在关联。互联网时代，各种设备通过网络连成了一个整体。进入以互动为特征的 Web 2.0 时代，个人计算机用户不仅可以通过网络获取信息，还成为信息的制造者和传播者。这个阶段，不仅数据量开始了爆炸式增长，数据种类也开始变得繁多。

这必然促使我们对海量数据进行分析、处理和集成，找出原本看来毫无关系的那些数据的“关联性”，把似乎没有用的数据变成有用的信息，以支持我们做出的判断。

3. Velocity

Velocity 可以理解为更快地满足实时性需求。数据的实时化需求正越来越清晰。对普通人而言，开车去吃饭，会先用移动终端中的地图查询餐厅的位置，预计行车路线的拥堵情况，了解停车场信息甚至是其他用户对餐厅的评论。吃饭时，会用手机拍摄食物的照片，编辑简短评论发布到微博或者微信上，还可以用 LBS(基于位置的服务)应用查找在同一间餐厅吃饭的人，看有没有好友在附近……

如今，通过各种有线和无线网络，人和人、人和各种机器、机器和机器之间产生无处不在的连接，这些连接不可避免地带来了数据交换。而数据交换的关键是降低延迟，以近乎实时——这意味着小于 250ms——的方式呈献给用户。

4. Value

但比前面三个 V 更重要的，就是 Value，它是大数据的最终意义——获得洞察力和价值。大数据的崛起，正是在人工智能、机器学习和数据挖掘等技术的迅速发展驱动下，呈现了这么一个过程：将信号转化为数据，将数据分析为信息，将信息提炼为知识，以知识促成决策和行动。

就大数据的价值而言，就像沙子淘金，大数据规模越大，真正有价值的数据相对越少。

所以真正好的大数据系统，重要的不是越多越好，其实越少越好。开始数据要多，最后还是要少，把 ZB、PB 最终变成一个比特，也就是最后的决策，这才是最关键的。

正如我们常说的：书刚开始越读越厚，到最后就越读越薄了。

数据挖掘和应用可以多方位创造价值，如表 1-1 和图 1-1 所示。

表 1-1 数据有效性的 5 个方面

数据有效性	内 容
数据质量	准确性：数据中没有错误的程度
	范围：数据覆盖的深度和广度
	合时性：及时获取数据以采取行动和决策的程度
	有效：相关数据及时更新程度
可用性	数据简明简要呈现的程度
	数据易操作和易处理
	多种数据来源中数据的一致性
智能性	数据预测准确性
	相关数据的趋势性
	需求模式分析
	精确营销推荐性
远程访问性	数据可被管理者远程获得的程度
	管理者可远程应用的程度
支持移动销售	利用便携系统与消费者交换价格信息
	利用便携系统与消费者交换订单信息
	利用便携系统与消费者交换运输信息

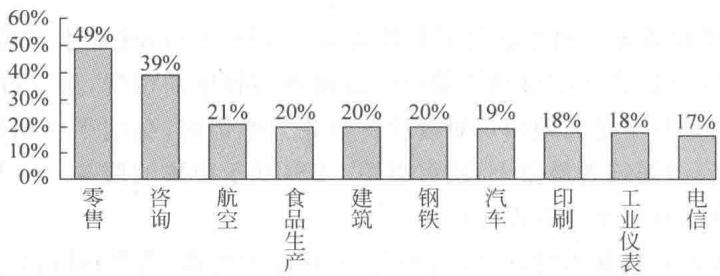

图 1-1 数据可用性提高 10%，各行业员工销售额提高百分比统计

1.1.4　大数据与云计算

云计算和大数据是一个硬币的两面，大数据正在引发全球范围内深刻的技术和商业变革。

云计算是大数据的IT基础，而大数据是云计算的一个杀手级应用。云计算是大数据成长的驱动力，而另一方面，由于数据越来越多、越来越复杂、越来越实时，这就更加需要云计算去处理，所以二者之间是相辅相成的。

30年前，存储1TB(1000GB)数据的成本大约是16亿美元，如今存储到云上只需不到100美元；但存储下来的数据，如果不以云计算进行挖掘和分析，就只是僵死的数据，没有太大的价值。

目前，云计算已经普及并成为IT行业的主流技术，其实质是在计算量越来越大、数据越来越多、越来越动态、越来越实时的需求背景下被催生出来的一种基础架构和商业模式。个人用户将文档、照片、视频、游戏存档记录上传至“云”中永久保存，企业客户根据自身需求，可以搭建自己的“私有云”，或托管或租用“公有云”上的IT资源与服务，这些都已不是新鲜事。可以说，云是一棵挂满了大数据的苹果树。

表1-2说明了大数据对数据全周期的要求都大大提高了。数据周期为相应产业带来的发展契机见图1-2。

表1-2　数据资产化对数据全周期提出更高要求

数据周期	供应商	要　求
数据存储	存储器供应商	需要更大、更快、更准地存储各项数据，同时对所存储数据有安全性，具有良好的灾备功能
数据处理	服务器供应商	需要更快地计算处理海量数据
数据管理	操作系统与数据库软件商	需要高效地查询和操作海量数据
数据应用	商业智能软件商	对海量数据进行挖掘、分析和优化
数据安全	信息安全提供商	对数据提供安全保护

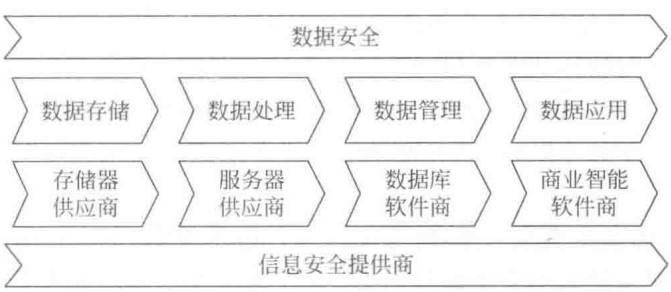

图1-2　数据周期为相应行业带来发展契机

1.1.5　大数据与商业模式变革

大数据的出现，正在引发全球范围内深刻的技术与商业变革。在技术上，大数据使从数