商业分析

Business Analytics

商业大数据分析

许　鑫◎编著

华东师范大学出版社

商业分析

Business Analytics

商业大数据分析

许　鑫◎编著

华东师范大学出版社

图书在版编目（CIP）数据

商业大数据分析/许鑫编著. —上海：华东师范大学出版社，2015.9
（商业分析丛书）
ISBN 978-7-5675-4012-5

Ⅰ.①商… Ⅱ.①许… Ⅲ.①商业信息—数据管理 Ⅳ.①F713.51

中国版本图书馆 CIP 数据核字(2015)第 210082 号

商业大数据分析

编　　著　许　鑫
策划组稿　孙小帆
项目编辑　孙小帆
特约审读　金　天
装帧设计　卢晓红

出版发行　华东师范大学出版社
社　　址　上海市中山北路 3663 号　邮编 200062
网　　址　www.ecnupress.com.cn
电　　话　021-60821666　行政传真 021-62572105
客服电话　021-62865537　门市(邮购)电话 021-62869887
地　　址　上海市中山北路 3663 号华东师范大学校内先锋路口
网　　店　http://hdsdcbs.tmall.com/

印 刷 者　苏州工业园区美柯乐制版印务有限责任公司
开　　本　787×1092　16 开
印　　张　20.25
字　　数　397 千字
版　　次　2015 年 10 月第 1 版
印　　次　2015 年 10 月第 1 次
书　　号　ISBN 978-7-5675-4012-5/F·339
定　　价　42.00 元

出 版 人　王　焰

本书简介

目标：

聚焦于商业大数据分析，通过对大数据概念和主题的梳理帮助读者理解商业大数据，介绍大数据处理、大数据分析、大数据应用、大数据管理等多个方面的知识要点和具体案例，帮助阅读者跟上大数据时代发展步伐，理解和应用大数据分析创造商业价值。

内容组织：

本书分为基础编、技术编、分析编和管理编四部分。

基础编包括什么是大数据（第 1 章）和大数据领域应用（第 2 章）两章，涉及大数据的起源、概念、本质探讨、主题分析以及在各个领域的应用，帮助阅读者深入理解大数据和商业大数据。

技术编包括大数据处理概述（第 3 章）、云计算（第 4 章）、大数据平台工具（第 5 章）三章，涉及大数据处理需求、处理模式、基本流程、关键技术以及云计算服务模式、虚拟化技术、云计算安全、移动云计算等知识点，介绍了以 Hadoop 为代表的一系列平台工具，帮助阅读者提升技术实务上的见解。

分析编首先介绍大数据分析概述（第 6 章），然后在统计分析方法和数据挖掘方法方面重点介绍非结构化数据与文本挖掘（第 7 章）、社交媒体与社会网络分析（第 8 章）、多维异构数据的分析方法（第 9 章）。

管理编包括商业环境中的大数据分析（第 10 章）、大数据商业分析中的人（第 11 章）、大数据政策框架与隐私问题（第 12 章）三章。第 10 章不但介绍了大数据背景下的商务管理研究，还面向业务问题和商业决策进行了实务探讨；第 11 章对人的关注不仅重点探讨了数据科学家及数据分析师，还对与人有关的企业组织结构和文化氛围进行了分析；第 12 章首先探讨了大数据政策框架，然后结合数据隐私保护重点探讨了大数据的隐私问题。

体例特点：

本书在介绍和论述的过程中结合了有关大数据的最新资料，包括新的技术应用、领域专家最新访谈等，同时给出来源方便拓展阅读。每章结尾部分均有本章小结，有助于帮助阅读者进一步总结和思考。

目录

基础编

1 什么是大数据

大数据本身是一个比较抽象的概念，单从字面来看，它表示具有庞大规模的数据。但是仅仅数量上的庞大显然无法看出大数据这一概念和以往的“海量数据”(massive data)、“超大规模数据”(very large data)等概念之间有何区别，鉴于大数据尚未有一个公认的定义，还是让我们从大数据的起源开始谈起。

1.1 大数据的起源

自古代有了第一次计数和农作物产量记录以来，数据收集和分析便成为改进社会生产和管理的重要手段。17、18 世纪的微积分、概率论和统计学所提供的基础性工作，为科学家提供了一系列新工具，用来准确预测星辰运动、确定公众犯罪率、离婚率和自杀率。这些工具常常带来惊人的进步。在 19 世纪，约翰·斯诺(John Snow)运用近代早期的数据科学绘制了伦敦霍乱爆发的“群聚”地图，霍乱在过去被普遍认为是由“有害”空气导致的，斯诺通过调查被污染的公共水井进而确定了“霍乱”的元凶，并同时奠定了疾病细菌理论的基础①。到了 20 世纪，从数据中撷取洞见以提振经济行为成为工业界的惯常做法，弗雷德里克·温斯洛·泰勒(Frederick Winslow Taylor)在宾夕法尼亚州的米德瓦尔钢铁厂采用秒表和笔记本来分析生产力，这大大增加了车间产量，也铸就了他的信念，即数据科学可以为生活中每一个方面都带来革命性影响②。进入 21 世纪，数据比以往任何时候都更加深入地与我们的生活交织在一起，我们期待着用数据解决各种问题、改善福利，以及推动经济繁荣。数据的收集、存储与分析技术不断提升，这种提升看上去正处于一种无限向上的轨迹之中。它们的加速是因为处理器能力的增强、计算与存储成本的降低，以及在各

① Scott Crosier, John Snow: The London Cholera Epidemic of 1854, Center for Spatially Integrated Social Science, University of California, Santa Barbara, 2007, http://www.csiss.org/classics/content/8.

② Simon Head, The New Ruthless Economy: Work and Power in the Digital Age, (Oxford University Press, 2005).

类设备中嵌入传感器技术的增长。2011 年，新生成的和复制的信息量估计超过了 1.8 ZB(泽字节)[①]，而 2013 年这一数字是 4 ZB[②]。

小贴士

泽字节(ZB 或 Zettabyte)

1 泽字节等于 10^{21} 字节，或相应的信息单元。想想看，一个字节可表示文本中的一个字符。1 ZB 相当于存储 323 兆份列夫·托尔斯泰所著的 1 250 页的《战争与和平》所需的容量[③]。或者想象一下，假定每一个美国人每秒钟拍一张照片并连续拍 1 个月，所有这些照片存储起来容量就相当于 1 ZB。而根据 IDC 的数据，全球的数据产生量在 2011 年达到的 1.8 ZB(或者说 1.8 万亿 GB)相当于每个美国人每分钟写 3 条 Twitter 信息，总共写 2.697 6万年。

IDC(International Documentation Centre，互联网数据中心)估计全球数据总量到 2020 年将增长 50 倍，主要源于嵌入服装、媒体设备和建筑物内的传感器逐渐增多，而文件、电子邮件和视频等非结构化信息约占未来十年数据产生量的 90%[④]。与此同时，Gartner(高德纳，又译顾能公司，全球最具权威的 IT 研究与顾问咨询公司)针对 IT 机构和用户发布的 2012 年及未来重大预测显示，到 2015 年，超过 85%的财富 500 强企业将无法有效利用数据为企业带来竞争优势[⑤]。就在我们编写此书的时候，世界上每天大约有 5 亿张照片上传或分享，另外每分钟还有超过 200 小时的视频上传分享。但是这些人们自己产生的信息(即从语音通话、电子邮件、文本到上传的图片、视频、音乐等全方位交流产生的信息)与每天产生的其他相关电子记录等数字化信息相比，在数量上也还是相形见绌的，因为我们已经处在一个所谓的“物联网”(Internet of Things)初级阶段，各种各样的应用设备、运输工具以及持续增长的“可穿戴”技术产品已可以彼此交换信息。在这样的背景下，“大数据”的概念受到学术界、企业界越来越多的关注。

---

① John Gantz and David Reinsel, Extracting Value from Chaos, IDC, 2012, http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.

② Mary Meeker and Liang Yu, Internet Trends, Kleiner Perkins Caulfield Byers, 2013, http://www.slideshare.net/kleinerperkins/kpcb-internet-trends-2013.

③ “2016: The Year of the Zettabyte,” Daily Infographic, March 23, 2013, http://dailyinfographic.com/2016-the-year-of-the-zettabyte-infographic.

④ http://tech.xinmin.cn/it/2011/06/29/11249210.html.

⑤ http://network.chinabyte.com/299/12220299_2.shtml.

小贴士

物 联 网

“物联网”这个术语用来描述具有可交换信息能力的设备网络，这些设备通常嵌入了传感器，并通过有线或无线网络连接后进行彼此间的信息交换。它们可能包括你的温控器、家电、汽车，甚至是你咽下去的“小药片”，医生可以用它来监控你的肠胃以及消化道的健康状况。这些连接的设备通过互联网传输、编制和分析数据。

物联网作为“物物相连的互联网”，通过智能感知、识别技术与普适计算广泛应用于网络的融合中，其有两层含义：一是物联网的核心和基础仍然是互联网，是在互联网基础上的延伸和扩展的网络；二是其用户端延伸和扩展到了任何物品与物品之间，进行信息交换和通信。

拓展阅读

三大风暴带来大数据分析

大数据的到来不是一夜之间悄然而至的，它的出现经过了相当长一段时间的酝酿。事实上，大数据的萌芽已经经历数十年，伴随企业处理大量的交易数据——由此甚至可以追溯到主机时代。如果有人问你大数据是如何产生的？你仅仅需要根据现实中真正发生了什么而迅速作出你的回答：

(1) 计算机应用爆发。大数据分析是四大全球化趋势的必然结果：摩尔定律(技术的获取成本越来越低)、移动计算(智能手机和平板电脑广泛使用)、社交网络(Facebook、Twitter 等)、云计算(甚至不必拥有硬件或软件，租用即可)。

(2) 数据爆发性增长。大公司数十年形成大量的交易数据，以及物联网带来的海量数据如洪水般涌入，原有的数据处理和数据分析方式和工具已经不能适应现实需要。

(3) 汇聚完美风暴。在传统的数据管理、分析软件、商用硬件的汇聚中，给信息技术和商业主管人员创造了新的选择——大数据分析。

其实对于一些行业资深人士而言，大数据并不新奇。有相当多的企业早已需要在数日里处理数以亿万计的交易数据，例如像万事达国际的数据仓库，常需要在期末处理十亿笔以上的交易数据。实际上信息技术行业的资深人士一直致力于处理更多的海量数据。在

过去二十年里，保罗·肯特(Paul Kent)作为一位研发专家兼SAS数据仓库副主管开发了许多处理大数据场景下的软件，在2012年的SAS全球论坛中，肯特指出：如果有足够的存储能力，完全能够改变用户的游戏规则。

“人们可以存储比以往更多的数据。我们已经到达了巅峰时刻，此时大家不再需要决定选择哪部分进行保存或历史上已经保存了多少。你可以经济可行地保存所有的历史数据以及任何相关的其他数据，当你遇到新问题的时候，还可以回头查阅历史信息并寻找新的答案。这是以前不可能实现的。”

米莎·戈什(Misha Ghosh)是一位发明家，他拥有好几项专利。在作为万事达国际的顾问之前，戈什就职美国银行长达11年，其间他进行数据分析以尝试解决业务难题，他指出：除了一些软硬件的改变以外，巨大的改变是在数据系统的建立上。他将数据系统分为三个阶段：(1) 早期的附属阶段(dependent)，数据仓库作为新面孔出现，但是用户并不知道数据仓库能够满足他们怎样的需求，IT行业坚信有必要建立数据系统，它的时代即将到来。(2) 紧接着的独立自主阶段(independent)，用户明白了信息技术与分析平台的结合不仅仅能够实现公司的业务需求，也是公司获得更多机会的方法。(3) 大数据时代的相互依赖阶段(interdependent)，这是一个相互协作的阶段，公司之间有了更多的社会合作，企业也突破了传统意义上的边界。

当前全球化经济形式中产生了前所未有的数据量，人们尝试通过每日产生的大量数据以获得庞大无比的力量，这些海量的数据是我们以前从未见过的，新鲜、强大，也很可怕，但令人兴奋。

编译自“Minelli, Michael, Michele Chambers, 和 Ambiga Dhiraj 的 *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses*. John Wiley & Sons, 2012.”

拓展阅读

大数据产生的三个阶段

人类历史上从未有哪个时代和今天一样产生如此海量的数据。数据的产生已经完全不受时间、地点的限制。从最初采用数据库作为数据管理的主要方式开始，人类社会的数据产生方式大致经历了三个阶段，而正是数据产生方式的巨大变化才最终导致大数据的产生。

(1) 运营式系统阶段。数据库的出现使得数据管理的复杂程度大大降低,实际中数据库大都为运营系统所采用,作为运营系统的数据管理子系统。比如超市的销售记录系统、银行的交易记录系统、医院病人的医疗记录系统等。人类社会数据量第一次大的飞跃正是建立在运营式系统开始广泛使用数据库。这个阶段最主要特点是数据往往伴随着一定的运营活动而产生并记录在数据库中,比如超市每销售出一件产品就会在数据库中产生相应的一条销售记录。这种数据的产生方式是被动的。

(2) 用户原创内容阶段。互联网的诞生促使人类社会数据量出现第二次大的飞跃。但是真正的数据爆发产生于 Web 2.0 时代,而 Web 2.0 的最重要标志就是用户原创内容(UGC, User Generated Content)。这类数据近几年一直呈现爆炸性的增长,主要有两个方面的原因。首先是以博客、微博为代表的新型社交网络的出现和快速发展,使得用户产生数据的意愿更加强烈。其次是以智能手机、平板电脑为代表的新型移动设备的出现,这些易携带、全天候接入网络的移动设备使得人们在网上发表自己意见的途径更为便捷。这个阶段数据的产生方式是主动的。

(3) 感知式系统阶段。人类社会数据量第三次大的飞跃最终导致了大数据的产生,今天我们正处于这个阶段。这次飞跃的根本原因在于感知式系统的广泛使用。随着技术的发展,人们已经有能力制造极其微小的带有处理功能的传感器,并开始将这些设备广泛地安置于社会的各个角落,通过这些设备来对整个社会的运转进行监控。这些设备会源源不断地产生新数据,这种数据的产生方式是自动的。

简单来说,数据产生经历了被动、主动和自动三个阶段。这些被动、主动和自动的数据共同构成了大数据的数据来源,但其中自动式的数据才是大数据产生的最根本原因。

摘自"孟小峰,慈祥. 大数据管理:概念,技术与挑战[J]. 计算机研究与发展,2013,50(1):146-169."

1.2 大数据的概念

"大数据"一词是从英语"Bigdata"一词直译而来。2008 年 9 月《科学》(*Science*)杂志发表了一篇文章"Bigdata: Science in the Petabyte Era","大数据"这个词开始走入人们的视野。2011 年 5 月,以倡导云计算而著称的 EMC 公司在"云计算相遇大数据"的年会上抛出了大数据的概念;同

年6月，IBM、麦肯锡等众多国外机构发布大数据相关研究报告予以积极跟进。麦肯锡在其研究报告中指出："数据已经渗透到每一个行业和业务职能领域，逐渐成为重要的生产要素，而人们对于海量数据的运用将预示着新一波生产率增长和消费者盈余浪潮的到来。"[①]至此，"大数据时代"作为一个正式的概念逐步进入公众的视野并引发了一系列后续社会影响。

关于"大数据"有多种定义，差别取决于你是一位计算机科学家，还是一位金融分析师，抑或是一位为风险投资人推销一个概念的企业家。多数定义都反映出了不断增长的捕捉、聚合与处理数据的技术能力，以及这个数据集在数量、速率与种类上的持续扩大。换言之，"现在，数据可以更快获取，有着更大的广度和深度，并且包含了以前做不到的新的观测和度量类型。"[②]更确切地说，大数据集是"庞大的、多样化的、复杂的、纵深的或分布式的，它由各类仪器设备、传感器、网上交易、电子邮件、视频、点击流，以及现在与未来所有可以利用的其他数字化信号源产生"。[③] 大数据的发展趋势可以通过表1-1的比较加以理解。

表1-1　大数据的趋势

	从　　前	现　　在
数　据	① 数据是资源 ② 数据是业务的副产品	① 数据是资产 ② 为了业务创新，需要收集数据、制造数据
分　析	① 大多数情况是事后分析 ② 寻找原因与规律	通常在事前进行分析，希望通过分析结果使某些事情发生
提供方式	信息共享与发布	将信息视为产品，运营信息
提供方	① 由IT部门提供 ② 业务部门有专门的分析团队	业务人员具有一定的自服务能力。简单的分析由业务部门解决，复杂的分析由IT部门专门的分析团队解决
分析团队	① 人数有限 ② 聚焦于分析技能提升	① 设置CDO、CAO、数据科学家等角色 ② 每个业务人员有一定的分析技能
数据管理	① 数据管理方法散落企业各处 ② 对数据标准、数据质量关注度不高	① 重视建设企业级数据管控方法 ② 设置数据管家等角色，保障数据质量

① Mckinsey Global Institute, Big Data: The Nest Frontier for Innovation, Competition and Productivity [EB/OL]. May 2011.

② Liran Einav and Jonathan Levin, "The Data Revolution and Economic Analysis," Working Paper, No. 19035, National Bureau of Economic Research, 2013, http://www.nber.org/papers/w19035; Viktor Mayer-Schonberger and Kenneth Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think, (Houghton Mifflin Harcourt, 2013).

③ National Science Foundation, Solicitation 12-499: Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA), 2012, http://www.nsf.gov/pubs/2012/nsf12499/nsf12499.pdf.