TURING

大数据的冲击

野村综研大数据专家权威解析

野村综合研究所（日本）城田真琴 著
野村综合研究所（上海）朱四明 审读
周自恒 译

风靡日本、韩国的超级畅销书

独家披露
野村综合研究所的
第一手资料

eBay、麦当劳等
美国、日本标杆企业的实践案例

人民邮电出版社
POSTS & TELECOM PRESS

TURING

大数据的冲击

野村综合研究所（日本）城田真琴 著
野村综合研究所（上海）朱四明 审读
周自恒 译

人民邮电出版社
北 京

图书在版编目 (CIP) 数据

大数据的冲击 / (日) 城田真琴著 ; 周自恒译 . --
北京 : 人民邮电出版社 , 2013.6
ISBN 978-7-115-31787-2

Ⅰ . ①大　Ⅱ . ①城…　②周…　Ⅲ . ①数据处理
Ⅳ . ① TP274

中国版本图书馆 CIP 数据核字（2013）第 093150 号

内 容 提 要

本书是日本最畅销的大数据商业应用指南。书中结合野村综合研究独家披露的调查数据，网罗了美国、日本标杆企业与政府的应用案例，总结了大数据的商业模式，以及在大数据应用中需要注意的隐私问题，并就如何为大数据时代做好准备展开了深入的探讨，提出了诸多有益的建议。

本书适合商业人士以及与大数据相关的 IT 从业者阅读。

◆ 著　　[日] 城田真琴
审　读　朱四明
译　　周自恒
责任编辑　乐　馨
责任印制　焦志炜

◆ 人民邮电出版社出版发行　　北京市崇文区夕照寺街 14 号
邮编　100061　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京铭成印刷有限公司印刷

◆ 开本：720 × 960　1/16
印张：16.75
字数：200 千字　　2013 年 6 月第 1 版
印数：1-5 000 册　　2013 年 6 月北京第 1 次印刷
著作权合同登记号　图字：01-2013-3031 号

定价：49.00 元

读者服务热线：(010)51095186 转 604　印装质量热线：(010)67129223
反盗版热线：(010)67171154
广告经营许可证：京崇工商广字第 0021 号

译 者 序

当我用 Gmail 阅读电子邮件时，页面上会显示 Google 提供的广告，这些广告往往和我正在阅读的这封邮件的内容密切相关；当我逛 Amazon 时，即便没有登录，Amazon 也能为我推荐我可能会感兴趣的商品，而且还真是相当地准，让我总能有意想不到的收获；当我带着我的 Android 手机上下班、出差、旅游时，谷歌纵横（Google Latitude）就会在后台默默地记下我所经过的地点，我可以随时查阅过去任意时间的位置记录，还可以和朋友分享。最近，很多网站都热衷于发布“年度盘点”，以信息图（Infographics）的形式对其掌握的数据进行汇总，并从中得出一些有趣的观点。例如支付宝的年度对账单，可以按性别、地域等维度分析不同群体的消费习惯。而迅雷的年度下载盘点，则可以在一定程度上反映出各地的网速水平。

上面所提到的这些，都是我们身边的大数据。在这样一个信息爆炸的时代，我们不得不感叹，大数据已经不再是一个虚无缥缈的概念，而

是与每个人的生活息息相关，实实在在且触手可及。大数据火了，它催生出无数新的服务和商业模式，也让一些传统行业找到了新的机会，同时产生了对“数据科学家”这种新兴复合型人才的迫切需求。而数据运用和隐私保护之间到底该如何权衡，也成了一个令各方势力争论不休的话题。大数据到底是什么？它为我们带来了什么？我们又该怎样去运用它？——这本书的目的，正是帮助大家思考上面这三个问题，迎接大数据所带来的机遇和挑战。

值得一提的是，这本书只用了短短一章的篇幅介绍关于大数据在技术层面上的内容，更多的则是围绕着大数据运用的成功案例、商业模式、隐私保护、法律框架、人才培养、经营战略等话题展开讨论，是一本无需具备技术背景也能够无障碍阅读的综述类著作。本书作者城田真琴先生，是野村综合研究所（NRI，简称野村综研）的高级研究员。野村综研是亚洲最大的咨询公司，堪称日本的麦肯锡，同时也是日本最大的系统集成商。作为本书译者的我，也曾有幸供职于野村综研的某合资子公司，可以说颇有一些缘分。这样的背景，让这本书的内容显得十分扎实，散发着咨询公司所特有的风范。在著书过程中，除了查阅文献和数据，作者还亲自采访了案例中相关公司的关键人物，掌握了大量的一手资料。相信大家通过这本书，一定能够对大数据有一个更加全面和深入的理解。

最后，感谢图灵公司各位编辑的辛勤工作，感谢作者城田真琴先生和野村综研（上海）咨询有限公司在本书翻译过程中所给予的帮助和支持。

周自恒

2013 年 2 月于上海

中文版序

作为本书的主题，“大数据”一词无论是在我的家乡日本，还是在欧美的IT业界，都已成为时下当仁不让的热点。但这个词对中国的各位读者来说也许并非如此耳熟能详。在新浪微博、人人网、QQ空间等社交网络中发表的文本数据，以及由物联网所产生的各种传感器网络数据，这些都是大数据的一部分。无论中国的各位读者是否听说过“大数据”这个词，大数据每天都正从大家身边不断地产生。

当然，仅仅看到每天产生出的大数据是没有意义的。我们还必须通过对数据进行适当的收集、存储和分析，将由此所获得的信息转化为具体的行为，并最终付诸实施。

例如，从事B2C业务的企业，通过对从社交网络中收集到的和自家产品相关的言论进行分析，就可以在新产品投入市场后的第一时间了解其评价。不过，仅仅做到这一步的话，还只能享受到大数据所带来的一半价值，因为我们还需要根据数据的分析结果，发现产品不足，并对

其进行改进。能够做到这一步，才可以说是真正享受到了大数据所带来的价值。

未来几年，大数据将对通信、金融、零售、制造、交通、物流、医疗、公共服务、农业等各个领域带来巨大的冲击。中国拥有世界上最多的人口，也必将成为全世界最大的数据生产国。另一方面，中国拥有清华大学、北京大学、浙江大学、上海交通大学等汇聚了众多优秀理工科人才的高等学府，有望培养出在欧美正十分紧俏的数据科学家。综上所述，我认为中国在成为世界最大的数据生产国的同时，还具备有效运用这些数据的潜力。

本书以“什么是大数据”为题介绍了大数据的基本知识、支撑大数据的技术、欧美及日本企业运用大数据的案例、大数据与个人信息保护及隐私保护之间的关系等。这些知识对于深入理解大数据是不可或缺的。本书自在日本上市以来，在大型书店取得了畅销书第一名的成绩，获得了极大的反响，在率先推出译本的韩国也备受好评。希望中国的各位读者能够从本书中获益。

城田真琴

2012 年 3 月

前　言

“Google、Amazon、Facebook、Twitter，这些称霸全球互联网的企业，它们的成功都具备一个共同的因素，你知道是什么吗？”

面对这样的问题，恐怕有些人会说：“是因为它们的商业模式非常创新。”而有些人则可能会说：“是因为它们的创业者非常优秀。”

然而，本书想要强调的，则是“数据分析”。看到这个词，可能你会说：“什么嘛，就这么简单？”虽然乍看之下会觉得很简单，但我们所列举的这些企业，它们每天不断存储和分析的数据量是十分庞大的，而这正是本书的主题——“大数据”。

充分运用大数据，并由此获得巨额的收益，Google 可以称得上是精通此道的鼻祖。据说，Google 每个月要处理 900 亿次的 Web 搜索，为此每月需要处理的数据量高达600PB[①]。使用Google各种服务的用户，

①1PB = 100 万 GB，这个信息量据说相当于 100 万年新闻早报的总和。

以及与之相关的各种数据，都是分析的对象。

在 Google 的搜索框中，只需要输入一部分关键字，就会显示出一些搜索关键字的建议，例如，只要输入“云”，系统就会自动提示“云免费”、“云是什么”、“云服务”等[①]。这样的搜索关键字建议，都是对用户庞大的搜索历史记录进行分析后得出的。此外，即便不以片假名的方式输入，而是直接输入罗马拼音“kuraudo”[②]，Google 也会给出正确的搜索建议。这种“输入修正功能”（或者叫“你要找的是不是……”功能），也是通过相同的原理实现的。

“购买了此商品的顾客还购买了这些商品”，这恐怕是世界上最广为人知的一种商品推荐系统了，而创造出这个系统的正是 Amazon。Amazon 通过分析商品的购买记录、浏览历史记录等庞大的用户行为历史数据，并与行为模式相似的其他用户的历史数据进行对照，提供出最适合的商品推荐信息。以这种数据分析为核心的服务设计发挥了巨大的作用，推动了 Amazon 成长为 2011 年销售额高达约 480 亿美元（约合人民币 3000 亿元）的巨型企业。

Twitter 拥有超过 1 亿的活跃用户，平均每天产生 2.5 亿条推文（根据 2011 年 10 月公布的数据）。每条推文最多 140 个字，数据量约为 200 个字节，这些推文平均每天相当于产生了约 48GB 的数据流量。而从 Twitter 整个生态圈来看，平均每天可产生约 8TB[③] 的数据。

---

①这些搜索关键字建议是根据日文翻译过来的，用中文搜索出现的搜索建议会有所不同。——译者注

②在日文中，“云”（クラウド）是外来语，即英文“cloud”的音译，而“kuraudo”则是其在日文中实际的读音，这里的例子类似于用中文搜索时直接输入汉语拼音“yun”。——译者注

③1TB 相当于 10^{12} 字节。

Facebook于2012年2月提出了IPO申请[①]。其公布的数据显示，每月活跃用户达到8.45亿，每日活跃用户达到4.83亿，着实令人惊叹。Facebook是世界最大的由用户产生内容的网站。

Facebook的所有用户平均每个月在Facebook上花费的时间高达7000亿小时，平均每个用户每个月会创建90条内容（包括新闻、博客等）。整体上来看，每个月产生的内容高达300亿条。根据公布的数据推测，Facebook所拥有的数据量超过30PB。

Facebook可以为用户提供类似“也许你还认识这些人”的提示，这种提示可以准确到令人恐怖的程度，而这正是对庞大的数据进行分析而得到的结果。

通过分析庞大的数据来获得有价值的信息或判断，这个被称为“大数据”的概念正受到越来越广泛的关注。它所掀起的巨大波澜早已经突破了IT业界的范畴，连报纸和电视新闻节目都对此制作了专题报道。

精通IT的读者在这里可能会有一点疑问：“通过对大量数据的分析来提升业绩，并不是这些新兴互联网企业的专利吧？对销售、库存等业务数据进行分析，帮助公司提升竞争优势，这种被称为‘商业智能’(BI)的方法已经由来已久，为什么现在却要特意翻出来大谈特谈一番呢？”

说起来，可能还真的是这么回事。例如，美国大型超市连锁集团沃尔玛，每小时就要处理约100万笔交易，在企业的数据仓库中产生和存储的数据量高达2.5PB。企业通过分析每天产生的大量数据，对商品的库存和定价做出极致的优化，这样的努力对于企业业绩的提升可以说功

①Facebook于2012年5月18日在纳斯达克正式上市。——译者注

不可没，这是不争的事实。

然而，在这里我们也要注意到两个重要的差异。

第一，同为海量数据，和传统意义上的销售额、库存量等数值数据相比，Google、Facebook 等互联网企业所处理的网站点击流（clickstream）数据和社交数据在管理和分析方法上是大相径庭的。目前大数据潮流的核心，并不是数值数据等结构化数据，而是网站点击流数据和社交数据，或者是传感器数据等这些无法存放在传统关系型数据库中的非结构化数据。

第二，从结果来看，掌握用于海量数据管理和处理新技术的，已不是沃尔玛、花旗银行这样的大企业，而是互联网企业和社交媒体企业。和 Facebook 的 30PB 相比，沃尔玛的 2.5PB 不仅在数据量上，而且在数据的多样性（网站点击流、社交媒体上的文字、人与人之间的联系等）和数据产生频率上都有很大差别。在这些方面，传统型大企业有很多东西需要向新兴互联网企业和社交媒体企业学习。

笔者有幸采访过的美国 B2B 企业中，经常能够听到这样的声音："Google、Amazon、Twitter、Facebook 等公司每天都产生、管理和分析大量的数据，传统型大企业需要将这些面向消费者的企业作为学习的榜样。"

实际上，现在用于大数据存储和处理的技术，如 Hadoop、NoSQL 数据库[①]等，大多数是从 Google、Amazon、Facebook 这样的互联网企业、社交媒体企业中诞生的。

①详见第 2 章。

在互联网世界之外，也有大数据的身影，其中由传感器网络所产生的传感器数据是最具代表性的一种。对各种机器的状态进行采集，并存储和分析这些数据，这样的尝试从很早就已经开始了，如自动贩卖机的管理系统、公交车和汽车的运行管理系统、重型机械的监控系统等。然而，随着技术的进步和通信成本的下降，能够对各种信息进行采集并对数据进行廉价存储的环境已经日趋成熟，今后应该会迎来进一步的普及。目前带有 GPS 功能的智能手机，以及 Suica、PASMO 等交通 IC 卡等，都已经显现出这样的趋势。

今后，随着智能电网、智能城市有望在全世界推广，传感器数据也必定会不断增加。而且，由于传感器是每秒都在进行测量和记录的，它们所产生的数据量，很可能会快速超过网站上由人类产生的信息、文本等数据量。

此外，各种设备和机器通过通信手段与互联网服务相结合所诞生的“M2M”（Machine to Machine）、“物联网”（Internet of Things）等词汇最近也受到了广泛的关注，这也将推动传感器数据的进一步增加。

将传感器所产生的庞大数据进行提取、分析，转化为有意义的信息并为商业服务，这样的尝试才刚刚崭露头角。这样一块蓝海市场[①]，必将带来巨大的商机。

综上所述，Google、Amazon 这样的互联网企业，及时发现了一般企业不重视的那些数据的价值，并独自开发出能够低成本存储和处理这些数据的技术，从数据中提取出有价值的信息，并将其整合到业务

①指尚未开拓的新兴市场，这一说法来自《蓝海战略》（*Blue Ocean Strategy*）一书，其中将现存的传统市场称为“红海市场”，将尚未开拓的新兴市场称为“蓝海市场”。——译者注

流程中，最终通过这样的方式发挥了自身的竞争优势。目前，跟随着Google 和 Amazon 的脚步，有越来越多的企业开始积极进行大数据的分析，通过提供新型服务和提高客户满意度来提升自身的竞争优势，这样的势头在各个行业中都愈发显著。

当然，原本通过对数值数据等结构化数据的深入分析建立起竞争优势的沃尔玛这样的大企业也不甘落后。沃尔玛于 2011 年 4 月收购了擅长社交媒体分析的创业型公司 Kosmix，在大数据的运用上迈出了重要的一步。沃尔玛通过对各卖场附近发布的推文和 Facebook 留言进行分析，掌握各卖场不同的需求，并由此制定商品种类和库存的调整策略。例如，从社交媒体的数据可以看出，在加州山景城有很多居民喜欢自行车，因此可以根据这一特点对卖场的商品种类进行调整。

除了社交媒体、非接触式 IC 卡这些 10 年前还不存在的新型数据，还有一些数据在过去产生时就被舍弃了，或者是保存下来也没有得到很好的运用，经过一段时间之后就被舍弃了，在这些数据中是不是也埋藏着一些“宝藏”呢？这也正是目前一些企业对大数据的运用跃跃欲试的一个重要的动机。

最近在美国经常听到“Data is the new oil”（数据就是石油）这样的说法。这句话的意思是，正如炼油所具备的巨大经济价值一样，数据只要进行适当的分析，也可以产生出巨大的价值。在这种思想的影响下，为了“最大限度地利用大数据所带来的机会”，美国政府于 2012 年 3 月宣布对大数据运用相关的研究开发投入 2 亿美元的巨额资金，展示了尽举国之力的积极态度。

本书涵盖了大数据在日本国内外企业中的应用事例，以及大数据在

商业领域中的运用要点、课题等内容，旨在尽量以通俗易懂的方式，介绍大数据的日本国内外的现状以及将来的发展趋势。

第 1 章对大数据作出了明确的定义，并讲解现在大数据为什么会如此受关注。

第 2 章讲解了支撑大数据存储、处理、分析的技术，以及其中主要领军者的动向。这一章会涉及很多技术性话题，对技术不感兴趣的读者可以跳过，如果在第 3 章之后遇到一些看不懂的术语，再回过头来参考这一章。

第 3 章介绍了一些欧美企业对大数据的运用事例，这些企业包括 eBay、Zynga、Centrica、Catalina Marketing 等。

第 4 章介绍了一些通过运用大数据带来大幅业绩增长的日本企业，这些企业包括小松、Recruit、GREE[①]、麦当劳等。

第 5 章介绍了笔者所总结的企业用户运用大数据的机会和模式。

第 6 章就大数据的商业应用中无法避免的隐私问题，介绍了国内外的指导意见以及法律法规方面的趋势。

第 7 章介绍了将位于封闭世界中的数据开放出来以促进创新的 Open Data 运动，以及数据交易市场 Data Marketplace。

第 8 章介绍了伴随着大数据时代的到来，企业需要如何应对，例如如何培养和吸引需求急剧高涨的“数据科学家”人才。

希望读者阅读本书后，能够对“大数据”这一企业在今后不得不面对的崭新世界加深一些理解。

①GREE 是日本的一家社交网站（http://gree.jp），与中国的格力电器无关。——译者注

目 录 CONTENTS

第 1 章　什么是大数据

第 2 章　支撑大数据的技术

第 3 章 以大数据为武器的企业 欧美企业篇

第 4 章 以大数据为武器的企业 日本企业篇