信息通信技术普及丛书

走近大数据

中国通信企业协会　组编
段云峰　张韬　编著

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

信息通信技术普及丛书

走近大数据

中国通信企业协会　组编
段云峰　张韬　编著

中国矿业大学（北京）藏书 图书馆

人民邮电出版社
北京

图书在版编目（CIP）数据

走近大数据 / 中国通信企业协会组编 ; 段云峰, 张韬编著. -- 北京 : 人民邮电出版社, 2018.12
（信息通信技术普及丛书）
ISBN 978-7-115-49281-4

Ⅰ. ①走… Ⅱ. ①中… ②段… ③张… Ⅲ. ①数据处理－普及读物 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2018)第203692号

内 容 提 要

本书首先介绍了大数据的一些基本概念，阐述了大数据发展的历史必然性；然后围绕大数据生态的各个技术和组件进行了基本的介绍；接着介绍了建设大数据系统要考虑的一些关键内容；最后以附件的形式给出了一些企业建设大数据系统的案例情况。

本书适合从事与大数据行业相关的人员，以及产业经营管理人员阅读参考，也可以作为高校信息技术、管理、电子商务类专业师生的参考书。

◆ 组　　编　中国通信企业协会
编　　著　段云峰　张　韬
责任编辑　李　强
责任印制　彭志环

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
三河市祥达印刷包装有限公司印刷

◆ 开本：700×1000
印张：16　　2018 年 12 月第 1 版
字数：287　千字　　2018 年 12 月河北第 1 次印刷

定价：78.00 元

读者服务热线：(010)81055488　印装质量热线：(010)81055316
反盗版热线：(010)81055315

《信息通信技术普及丛书》编委会

主　　编： 苗建华　中国通信企业协会会长

副 主 编： 刘桂清　中国电信集团有限公司副总经理

顾晓敏　中国铁塔股份有限公司副总经理

赵中新　中国通信企业协会副会长兼秘书长

张同须　中国移动通信有限公司研究院院长

张　涌　中国联通网络技术研究院院长

执行主编： 柏国林　中国通信企业协会副秘书长

编 辑 组

组　　长： 赵俊涅　中国通信企业协会综合业务发展部主任

副 组 长： 冯志宏　中国通信企业协会综合业务发展部副主任

刘　婷　中国通信企业协会综合业务发展部副主任

王建军　人民邮电出版社信通传媒图书出版中心主任

前 言

时至今日，大数据应用无所不在，但什么才是真正的大数据？怀着对大数据的迷茫，“摸着石头过河”也有些年头了。一路走来，我们不断地思考和摸索，遇到很多问题，也解决了很多问题。这也许都是发展的必然历程吧。

在人人都讲大数据，都说他们用了、在用或要用大数据的今天，我们结合自身的工作体验和实战经验，想给大家带来一些我们对大数据内涵与外延的认识，也希望能针对相关的技术和产品的应用做一些普及性解读，希望对感兴趣的读者有所助益。

在云计算、物联网和人工智能的协同下，大数据的发展空间被全方位拓展，人们的想象边界也被一下子打开了，仿若潘多拉盒子被打开似的，充斥着各种未知的因素，让人望而生畏。而另一方面，大数据也确实给食品溯源、药品管控、聚集引导和公共安全这些人们热切关心的问题带来各种可能的解决方案。但是在数据采集、使用、共享、流通以及隐私保护等方面的建章立制，才是大数据能真正为人们所用的基础，用好大数据其实是一件很美好的事情。

最后，我们也将此书献给支持我们的家人们，没有你们的默默支持、宽容、理解和爱，本书难以成文！

段云峰　张韬

2018 年 8 月

此为试读，需要完整PDF请访问：www.ertongbook.com

目 录

此为试读，需要完整PDF请访问：www.ertongbook.com

第1章

大数据，心中有数

1.1 从一场亲子讲座谈起

从我们2013年编撰第一个内部大数据讲座材料至今，已有5个年头了。这5年，大数据正如甚至超过人们预期的那样迅猛发展。

在一场亲子讲座上，资深的教育界老师针对孩子的启蒙规律和启蒙学习进行讨论。当然，讲座主题是偏离本书讨论范围的，但启发我们的是，老师说道，“我们身处大数据时代，我们讲的这些启蒙规律和学习方法都是基于大数据的。”是啊，大数据不用来为我们最热衷的孩子教育服务太可惜了！

诚然，绝不仅仅教育应讲大数据，我们传统的交通、医疗、环保、旅游、体育和政务等各方面也都在讲大数据，大家都说自己用了、在用或要用大数据。伴随移动互联网浪潮衍生发展的新兴业态都成了大数据的载体，并深深受益于大数据，不断有共享单车、共享汽车等共享经济的商业模式应运而生，这些新的商业模式也反哺了大数据相关产业的发展（如图1-1所示）。

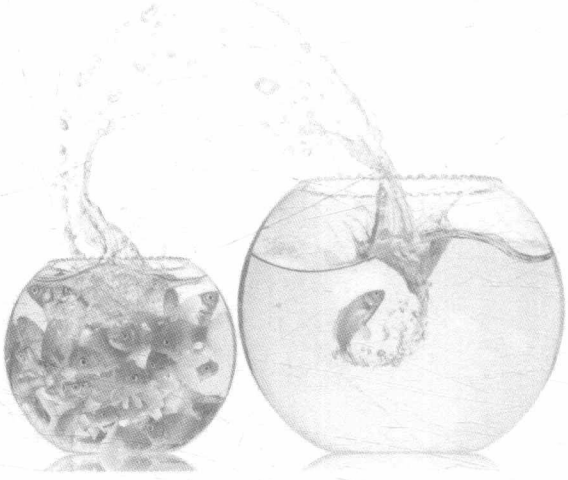

图1-1 捕获价值增长机会

人人讲大数据的今天，大数据究竟怎样改变了我们身处的世界，我们又如何认识、如何理解和如何应对这些改变呢？为了让我们能做到对大数据心中有数，更好地捕获大数据价值，增长企业、团队或个人的发展机会，我们在这里一起漫谈大数据。

1.2　数据非今日变大，为什么今天火了

人类文明的发展进程，伴随着信息的传播方式和记录方式的发展。最初的信息是人们对物品计数的记录和传播，这就是最初的数字。计数方式从连续到离散，传播方式从口耳相传到图形记录，直至文字记录。在这一系列过程中，数据一直就有，而绝非今天才产生。但为什么今天的数据突然被冠以“大”之名，火了呢？

1989 年，也许是埃里克·拉森第一次使用了“大数据”[1]，数据科学发展的未来大潮开始萌发。1994 年，比尔·盖茨拍了张照片（如图 1-2 所示），幽默地显示一张光盘能装下的数据比大量的纸张能记录下的数据都多，而他在 1981 年曾说“640KB 内存应该对任何人都够用了”，但是不久 DOS 编写人员就要着手编写内存管理程序，因为 640KB 实在太小了。随着数据存储技术的发展，伴随着各种商业生态链的极速发展，更伴随着互联网的诞生和发展、网络 2.0 时代以及物联网的推波助澜，大数据的发展经过了兴起，已经渐入佳境。

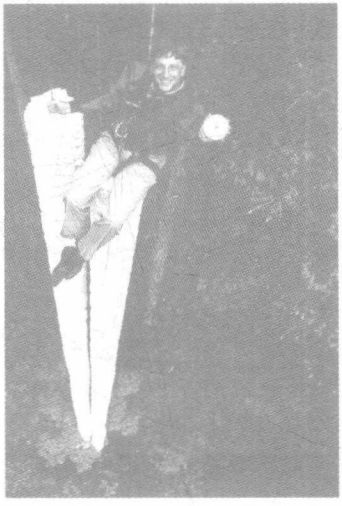

图 1-2　比尔·盖茨的光盘存储量类比

2008 年 9 月，美国《自然》（*Nature*）杂志在 Google 成立 10 周年之际，出版了 *The next google* 专刊，讨论未来 10 年大数据会带来的变化，并提出大数据真正重要的是新用途和新见解，而不是数据本身[2]。2010 年，Google 前执行主席埃里克·施密特说，现在两天所产生的数据量是人类文明开始到 2003 年的总和。2011 年 2 月，《科学》（*Science*）杂志刊登了名为 *Dealing with data* 的专辑，通过社会调查的方式，讨论数据对科学研究的重要性及大数据对人们的影响[3]。

2013 年 5 月，《外交》杂志上撰文[4]称：人们认为，公元前 3 世纪，埃及亚

---

[1] Bernard Marr, A Brief History of Big Data Everyone Should Read.

[2] Big bata: the next Google, Nature 455, 8-9 (2008).

[3] Dealing with data, 11 Feb. Issue of Science, 2011.

[4] Kennether Neil Cukier and Viktor Mayer-Schoenberger, The Rise of Big Data——How It's Changing the Way. We Think about the World, Foreign Affairs, May/June 2013 Issue.

历山大图书馆（如图 1-3 所示）收藏了人类所有的知识。而今天全世界有足够多的信息，预计所有信息量达到 1 200 艾字节[5]，以至于将这些信息分配给每个活着的人，每人获得的信息量将是整个亚历山大图书馆藏书的 320 倍之多。如果把所有这些信息存储在光盘上，这些光盘将会分别堆成 5 摞，每摞都可被从地球一直堆到月球。

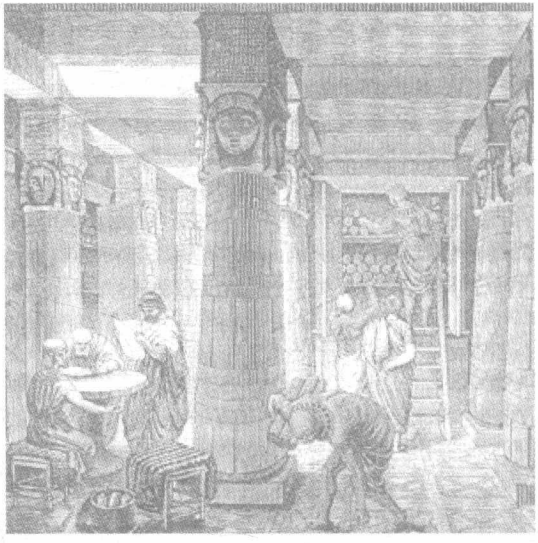

图 1-3　埃及亚历山大图书馆

暂且不论这些比喻是否恰当，互联网，特别是移动互联网的惊人发展使数据更以爆炸性势头得以增长。互联网上经常使用两张拍摄于不同时间、同一位置的照片的对比来说明当今数字信息化技术发展对人们生活的改变。

2015 年 6 月，爱立信发布《移动互联网报告》显示，2010 年手机数据流量才刚刚达到语音流量的 2 倍，而在 2014 年，手机数据流量已经是语音流量的 20 倍以上。仅在 2014—2015 年，数据流量增长了 55%。报告预测，到 2020 年全球数据流量较 2014 年又会增长 10 倍以上。

2016 年 2 月，业务管理软件平台 DOMO 发布了一系列数据，显示互联网每分钟运行着大量的在线数据（如图 1-4 所示）：每分钟有 4 310 人登录亚马逊网站，Netflix 用户每分钟会观看 77 160 小时的视频，苹果用户每分钟会下载 51 000 个应用，Instagram 用户每分钟发布 123 060 张照片，YouTube 用户每分钟会上传 300 小时的新视频，Twitter 用户每分钟发布 347 222 条推文，Facebook 用户每分钟点赞 4 166 667 次，Uber 每分钟获得 694 个订单；每分钟平均收发邮

[5] 艾字节（Exabytes），计算机存储单位，也常用 EB 来表示。1EB=1024PB=1024^{6}KB。

件达到 2.4 亿封；Google 的搜索量每分钟可达 278 万次[6]。2015 年世界互联网大会上，腾讯公司指出，其微信红包一天的收发量是 22 亿个，平均每分钟红包收发量是 1 527 777 个。根据支付宝官方大事记，2015 年“双十一”期间共完成 7.1 亿笔支付，平均每分钟完成 493 055 笔交易，当天淘宝活跃用户量超过一个亿，平均每分钟活跃用户超过 69 444 人。

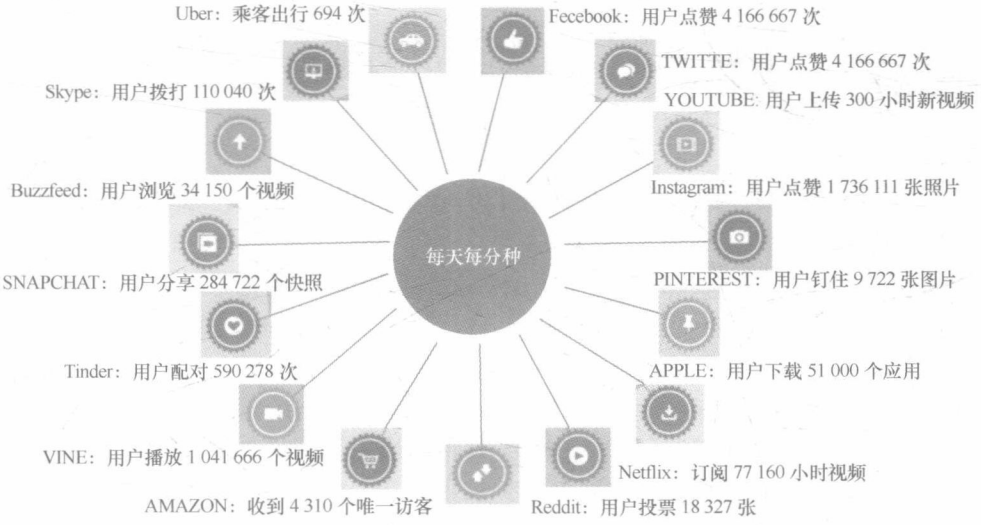

图 1-4　DOMO 发布的在线数据分析

2018 年 1 月 31 日，中国互联网信息中心（CNNIC，China Internet Network Information Center）发布的我国第 41 次《中国互联网络发展状况统计报告》[7]指出：截至 2017 年 12 月，我国网民规模达 7.72 亿人，手机网民规模达 7.53 亿人，网站总数为 533 万个，即时通信用户规模达 7.20 亿人，网络购物用户规模达 5.33 亿人，网上外卖用户规模达 3.43 亿人，在线旅行预订用户规模达 3.76 亿人，使用网上支付的用户规模达 5.31 亿人，网络音乐用户规模达 5.48 亿人，网络文学用户规模达 3.78 亿人，网络游戏用户规模达 4.42 亿人，网络视频用户规模达 5.79 亿人，网络直播用户规模达 4.22 亿人，在线政务用户规模达 4.85 亿人。巨量的在线和交易用户规模的背后是，各个网络交互环节中产生的海量数据。

回到本节我们提出的问题，数据并非今日变大，为什么现在火了呢？从上面一系列的数据来看，第一个原因是大数据有来源。现今互联网、移动设

[6] Brandt Ranj and Brandt Ranj, 15 things that happen on the internet every minute, Business Inside, Feb. 3, 2016.
[7]《中国互联网络发展状况统计报告》，中国互联网信息中心，2018 年 1 月 31 日。

备和物联网等的迅猛发展，使人们每分每秒都在产生着巨量数据，使大数据有了更广泛的来源。而这些逐渐产生的数据，极大地挑战了信息化技术的存储能力和处理能力。那么，随之而来的第二个原因是，信息技术的突破性发展，使大数据价值有可能在有限投入和有效时间内被发掘和发挥出来，得以绽放价值。

1.3 大数据带来的改变渐渐发生了

大数据逐渐深入的发展，改变了我们和我们所处的社会，改变了我们的生活和工作各个可触及的范围。

关于大数据，最经典和最易被提及的应用案例是美国塔吉特卖场对于17岁女孩怀孕的预测。该事件源自《纽约时报》的一篇报道，报道是关于一位怒气冲冲的父亲对塔吉特卖场将带有婴儿用品优惠券的广告邮件，寄送给他正在念高中女儿的质问。而事实是，这位父亲的女儿确实怀孕了。塔吉特卖场从这名女孩搜寻商品的关键词和在社交网站所显露的行为轨迹，成功预测其怀孕的信息。有数据显示，许多孕妇在第2个妊娠期开始，会购买许多大包装的无香味护手霜；在怀孕的最初20周大量购买补充钙、镁和锌的善存片之类的保健品。由此，塔吉特构建了“怀孕预测指数”，可在小误差范围内实现对顾客怀孕情况的预测。

FareCast是早期大数据创业公司的一个缩影，该公司通过“哈姆雷特”项目，从旅游网站上搜集41天的12 000个价格样本的分析基础上，开发了一个虚拟价格预测系统。获得风险投资后成立FareCast公司，分析了10万亿条北美70多个城市的机票价格记录，预测这些城市之间的机票最低价格，实现了75%的准确率[8]。

而伴随大数据成长的谷歌公司，初创于搜索引擎技术。该公司通过收集和分析人们输入的搜索关键词，实现特定区域的搜索关键字聚合，建立评估模型；再来建立搜索流感话题人数和真正流感患者人数之间的关系，将该模型应用到聚合后的搜索关键字后，可以一定程度地实现对流感在不同国家和地区中扩散情况的预测。

Gartner的分析师Doug Laney列举了55个大数据应用案例，其中有这样两个案例引人关注。2013年1月，PredPol公司与洛杉矶警方合作进行可精确到

[8] 维克托·迈尔-舍恩伯格，《大数据时代》。

此为试读，需要完整PDF请访问：www.ertongbook.com