大数据丛书系列之四

总主编◎曾　羽　龙奋杰

大数据科学

DASHUJU
KEXUE

主　编◎霍雨佳　周若平　钱晖中

电子科技大学出版社

大数据丛书系列之四

总主编◎曾　羽　龙奋杰

大数据科学

DASHUJU
KEXUE

主　编◎霍雨佳　周若平　钱晖中

电子科技大学出版社

图书在版编目(CIP)数据

大数据科学 / 霍雨佳，周若平，钱晖中主编. -- 成都：电子科技大学出版社，2017.7
ISBN 978-7-5647-4820-3

Ⅰ. ①大… Ⅱ. ①霍… ②周… ③钱… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2017）第 176277 号

大数据科学

霍雨佳　周若平　钱晖中　主编

策划编辑　杨仪玮　李燕芩
责任编辑　杨仪玮　李波翔

出版发行　电子科技大学出版社
　　成都市一环路东一段 159 号电子信息产业大厦　邮编 610051
主　　页　www.uestcp.com.cn
服务电话　028-83203399
邮购电话　028-83201495

印　　刷　成都市火炬印务有限公司
成品尺寸　165mm × 240mm
印　　张　18.5
字　　数　352 千字
版　　次　2017 年 7 月第一版
印　　次　2017 年 7 月第一次印刷
书　　号　ISBN 978-7-5647-4820-3
定　　价　65.00 元

目　录

第1篇

基础篇

第1章　大数据概述

“大数据”，或称之为海量数据，一般指所含的数据集规模巨大，现在大众的软件工具无法在合理的时间进行采集、存储、分析管理的数据信息。因其在各个行业的广泛应用，使之关注热度历年来居高不下。作为人们获得新的认知、理念和创造价值的源泉，大数据数据来源可以囊括我们从日常生活中可以普遍见到的上传到网页上的图像、视频、录音；高速公路上车辆与收费记录、日常监控录像、医院的治疗病例、高端的基因测序，天文学中通过望远镜收集的信息数据等。

根据国际数据公司（IDC）《数字世界》研究项目在2012年对全球数据量的统计结果可知：2011年，全球数据量的规模将达到里程碑式的1.8ZB，2010年全球产生的数据量为1.2ZB，2009年为0.8ZB，全球数字数据量每两年翻一番[①]。据预测，中国数据量规模在2012年至2020年间将从364EB增至8.6ZB，中国在全球数字世界中所占的份额将在2012年至2020年间从13%增至21%。如果你对这些数字仍然感到难以把控的话，接下来一组名为“互联网上一天”的数据可以清晰地告诉你，一天之内，互联网产生的全部内容可以刻满1.68亿张DVD；发出的邮件有2940亿封之多（相当于美国两年的纸质信件数量）；发出的社区帖子达200万个（相当于《时代》杂志770年的文字量）……[②]人类计量数据量的单位也从TB级别上升到PB、EB乃至ZB级别。毫无疑问，大数据时代已经来临。

1.1　产生背景

从2006到在2016年，大数据在IT行业、医疗、民生、金融、学术等多个领域中炙手可热，行业领导人也对其保持着高度的重视，关注其能够带来的科学价值和社会价值。

第三次科学技术革命的蓬勃发展，为大数据时代的到来奠定了良好的基础。互联网的普及、信息技术的发展、云计算的发展成熟、遍布的智能终端等，无时无刻不记录着人类产生的“数据足迹”，我们每个人都在毫无意识地成为数据的提供者。如今的数据已经不单单是信息技术及科学研究领域人

① http：//news.makepolo.com/1451080.html
② http：//www.360doc.com/content/15/0311/15/4786202_454311432.shtml

员的专有名词，冠有量词属性的“大数据”早已成为一种引人注目的新思潮，成为人们认识事物、分析事物、探索新发现及追求创新的新范畴。故在此，对大数据的产生背景进行一个梳理。

1.1.1 互联网与大数据

互联网技术的不断普及，其每天所催生的巨量数据使得世间万物不断走向数据化，数据量化的节奏不断加快。在由“万事皆数”过渡到“万物皆数”的过程中，互联网每天所产生的数据，对大数据时代的来临有着关键性作用。

（1）互联网的产生

互联网的发明最初是根据军事的需要而产生的。1969年，美军在ARPA（阿帕网，美国国防部研究计划署）制定的协定下，首先用于军事连接，后将美国西南部的加利福尼亚大学洛杉矶分校、斯坦福大学研究学院、UCSB（加利福尼亚大学）和犹他州大学的四台主要的计算机连接起来[①]。这个协定由剑桥大学的BBN和MA执行，在1969年12月开始联机执行。随着计算机在战争中军事上的广泛运用，计算机上保存的各自军事机密就越多，其安全性显得尤为重要，军事家们担心如果计算机上的重要军事机密数据泄露的话，会导致整个战争的失败。因此，如何在原有计算机上除能够进行储存数据外，能够通过某种渠道促使两台计算机，或是多台计算机之间能够进行数据之间的相互传递和备份，成为了军事家们的现实需要并极力促进其实现，这促进了早期互联网的形成。

20世纪七八十年代，我们对各类远程信息的接收，大多是依靠收音机或者电视这样的渠道。这种模式的信息数据传播主要是通过信息源利用信号塔进行信息四面八方传播，接收者利用收音机、电视进行信息接收。需要注意的是，这种“信号塔”与“接收”的模式，在整个过程中，接收者是被动的，而且发送信息方也对所传播的信息质量、受众数量、受众偏好等信息数据没有一个完全的记录。

互联网的诞生，使人类进行信息交流的方式发生了质的变化。经济的高速发展，电子产品更新换代进程的加快，使得互联网与大众的基础规模日益扩大。人类也越来越习惯通过互联网接收和传播信息数据。与传统的信息接收和传播模式相比，互联网模式只有当用户利用互联网进行访问，服务站在接收到请求后，会立即在万千信息中寻找到用户所需求的信息内容进行回馈。在这个过程中，我们的用户客户端早已记录下了用户的访问记录数据量、内容、停留的时间等多种信息数据。尽管用户会删除这些数据，但数据

① http：//www. jishi60. com/show/20/5455-0.html

后台已将这些记录保存下来，这些留下来的海量信息数据其背后所蕴含着难以估算的巨大价值。

在淘宝网上营业的一家食品销售店铺，为了扩大自身的竞争力和更准确地知道消费者的需求，需要对消费者购买意愿及购买能力进行掌握。为了解这些信息，需要对进入店铺的消费者进行一个调查研究，所需了解信息包括：长久时间以来顾客进入网上店铺的浏览量？每次浏览的食品有哪些？加入购物车的食品是什么品类？在店铺停留的时间有多长？从确定购买到支付完成的消费周期有多久？如果按照传统的销售策略，进行该项目就需要一系列的工作，或是进行大样本调研问卷进行调查获得数据资料，或是安装摄像头进行监控调查，或是通过选取忠诚的消费者样本，电话咨询，不管怎么样，都需要大量的人力、物力、财力。而如果想节约成本的话，就失去了消费者购买商品后所留下的大量消费信息。互联网大数据的出现会使这一系列的成本降得微乎其微。通过互联网上的用户每次操作访问记录，就可以很清楚地得到这一系列的数据。在这过程中，当消费者点击淘宝上店铺的页面，就相当于进入了店铺，对展示的食品进行点击，就类似于消费者拿起食品进行分析比较，与商家进行谈话交流，可以测算出消费者的关注问题，把食品放在购物车就相当于试穿，通过购物车进行结算就相当于购买，而从最开始进入店铺到支付结算，关闭网页的时间记录就相当于一个消费周期，消费者通过不断的对比所确认购买的服装价格相当于此类消费者的购买能力，同时对于消费者的类型、品味兴趣都有了一个精准的记录。并且获得这些大样本数据的成本为“零”。在实体店中的消费者行为和心理都如实地反映在了访问记录中，而且还可以大幅度降低商家的存储、操作成本！

（2）互联网催生大数据

从1969年互联网诞生到2016年，互联网技术已发展成熟。世界经济技术的高速发展，使互联网作为王榭的堂前燕飞入寻常百姓家的普及面日益扩大。

根据世界银行报告报告，截止到2016年初，七大人口大国之中，美国、日本、俄罗斯的互联网普及率最高，分别为87.4%、90.6%和70.5%，巴西、中国次之，分别为57.6%和49.3%，而印度和印尼最少，分别在18%和33%（见图1-1）。从中国的互联网普及率来看，中国互联网络信息中心（CNNIC）在京发布第38次《中国互联网络发展状况统计报告》显示，截止到2016年6月，中国网民规模达7.10亿，互联网普及率达到51.7%，超过全

球平均水平3.1个百分点（见图1-2）。移动互联网塑造的社会生活形态进一步加强，“互联网+”行动计划推动政企服务多元化、移动化发展。

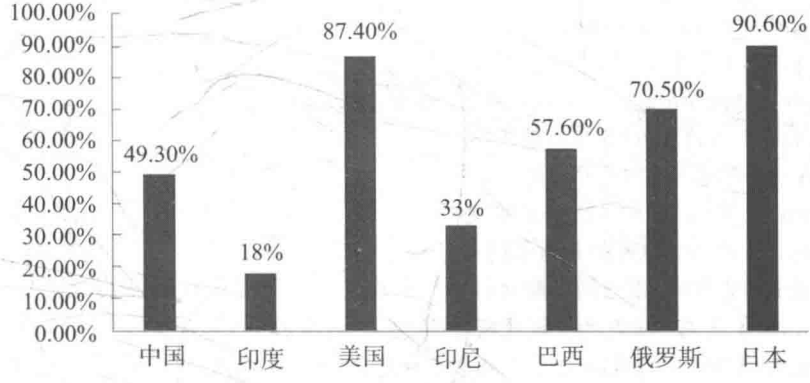

图1-1 全球人口大国互联网普及率

图1-2 中国网民规模和互联网普及率

互联网的迅猛发展和快速普及，使得大量的数据信息在采集、存储、传输、处理、管理等方面越来越便捷。同时，互联网的发展也使得其所产生的数据类型变得复杂多样，由最初的结构化数据，发展到非结构化数据、半结构化数据等。就大数据而言，在互联网上一天，都会潜在的拥有众多数据的“产生者”和“发送者”，这些“产生者”和“发送者”每时每刻都贡献出各种各样、难以计量的数据。这些数据可以是结构化数据，如数字、符号，也可以是非结构化数据，如文本、图象、声音、影视、超媒体等。这些接连不断出现的数据，催生着大数据浪潮的来临。根据Excelcom公司2016年发布的一份“互联网一分钟产生数据”的表1-1，可以看出互联网对于大数据的催生作用到底有多大。

表 1-1　互联网一分钟产生的数据量

① Netflix 共有 69 444 小时长的视频被观看；
② 1 亿 5 千万封电子邮件已发送；
③ Uber 产生了 1389 次驾驶；
④ Snapchat 上分享了 527 760 张照片；
⑤ Apple's AppStore 上已有 51 000 个 app 被下载；
⑥ Amazon 产生了$203 596 的销售额；
⑦ Linkedin 创建了 120 多个新账号；
⑧ Twitter 上发布了 347 222 条新推文；
⑨ Google 上产生 240 万的新搜索请求；
⑩ Tinder 上又有 972 222 的新配对；
⑪ Youtube 上已有 278 万的视频被观看；
⑫ Spotify 上的音乐播放时长已达 38 052 小时；
⑬ Facebook 共产生 701 389 个账号登陆；
⑭ Vine 上的小视频播放了 100 万次。

1.1.2　信息技术与大数据

信息技术指人们获取信息、传递信息、存储信息、处理信息、显示信息、分配信息的相关技术，它包括现代通信技术、电子计算机技术、微电子技术等[①]。在信息科学技术中，信息的收集能力、存储能力、对于信息数据的处理分析能力、以及信息之间远程传递能力是最为关键的，如果把人类历史上信息科技技术发展的五次信息技术革命看成是世界不断数字化的过程，就会发现一条信息技术进步与大数据时代来临的逻辑线[②]。

（1）信息采集技术

获得信息，是人认识世界的基础，因此也是改造世界和人类本身发展的基础。人类从客观世界获得的信息越多、内容越丰富，则人对客观世界的认识也越深刻，对客观世界的利用、改造及对人类本身发展也越有利。传统数据采集来源单一，且存储、管理和分析数据量也相对较小，大多采用关系型数据库和并行数据中心即可处理。大数据时代下，如何从数据量极大、增长速度极快、数据类型复杂、实用性极高的数据中采集出使用者所需要的信息，为传统的数据采集技术提出了难题。2006 年后出现的数据采集技术使得这一难题得到攻克。这些技术包括 Hadoop 的 Chukwa，Cloudera 的 Flume，

① 杨强. 信息技术的发展历程及其未来趋势[J].美丽中国，2009（3）：120-123.
② 赵国栋，易欢欢，糜万军等著. 大数据时代的历史机遇：产业变革与数据科学[M].北京：清华大学出版社.