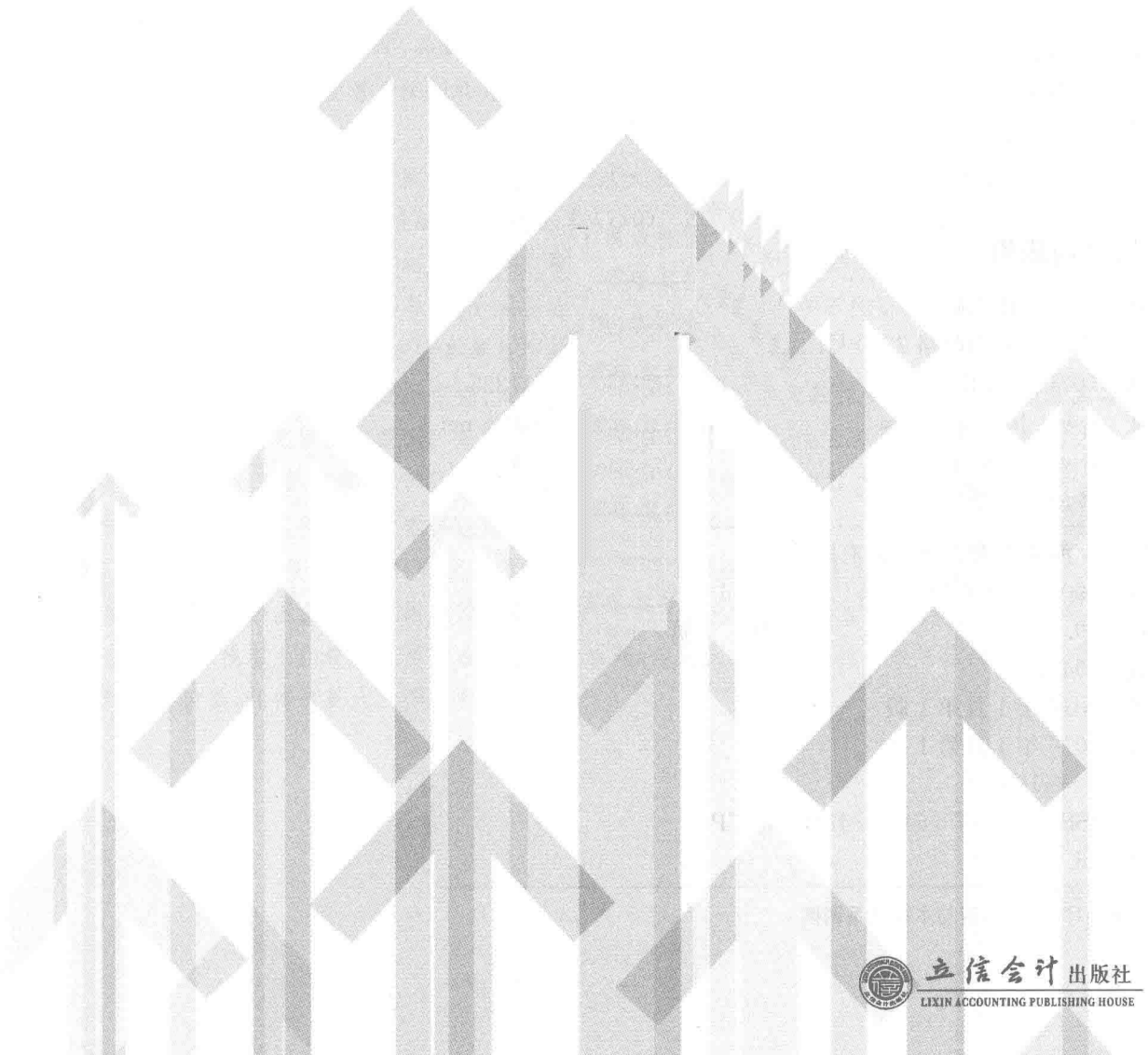
大数据分析与应用

樊重俊 刘臣 霍良安○编著

立信会计出版社
LIXIN ACCOUNTING PUBLISHING HOUSE

大数据分析与应用

樊重俊 刘臣 霍良安○编著



立信会计出版社
LIXIN ACCOUNTING PUBLISHING HOUSE

图书在版编目(CIP)数据

大数据分析与应用 / 樊重俊，刘臣，霍良安编著.
—上海：立信会计出版社，2016.1
ISBN 978-7-5429-4915-8

Ⅰ.①大… Ⅱ.①樊…②刘…③霍… Ⅲ.①数据库系统 Ⅳ.①TP311.13

中国版本图书馆 CIP 数据核字(2016)第 042550 号

策划编辑　黄成艮
责任编辑　黄成艮
封面设计　陈　楠

大数据分析与应用

出版发行	立信会计出版社		
地　　址	上海市中山西路 2230 号	邮政编码	200235
电　　话	(021)64411389	传　　真	(021)64411325
网　　址	www.lixinaph.com	电子邮箱	lxaph@sh163.net
网上书店	www.shlx.net	电　　话	(021)64411071
经　　销	各地新华书店		
印　　刷	上海天地海设计印刷有限公司		
开　　本	890 毫米×1 240 毫米	1/16	
印　　张	29.25	插　　页	1
字　　数	854 千字		
版　　次	2016 年 1 月第 1 版		
印　　次	2016 年 1 月第 1 次		
印　　数	1—2 100		
书　　号	ISBN 978-7-5429-4915-8/TP		
定　　价	56.00 元		

如有印订差错，请与本社联系调换

序

2015 年 3 月，李克强总理在政府工作报告中，首次提出“互联网＋”行动计划，推动移动互联网、云计算、大数据、物联网等与现代制造业结合，促进电子商务、工业互联网和互联网金融健康发展。目前，中国拥有庞大的互联网人群和巨大的移动互联网应用市场，因此，中国已经成为“数据矿藏量”最大的国家之一，通过大数据分析方法与技术探索大数据背后的价值并加以实践应用，是建设智慧城市、智慧交通、智慧医疗、智慧金融等的重要途径，同时也是各行业升级转型的重要措施。

随着新一代信息技术的发展和应用，尤其是物联网、移动互联网、社交网络等技术的发展，大数据时代已经到来，大数据的分析与应用已经并将持续成为当今信息处理的热点研究内容。大数据不仅数据规模大、数据类型复杂，更需要采取新的数据思维来进行分析，在数据获取方式、数据处理方式、数据分析方法上进行突破，这必然引领理论与技术的革新，颠覆传统数据管理模式。

大数据从数据挖掘、商业智能发展而来，是信息技术发展的必然产物，是新一代信息技术的重要领域。从大数据的理念到 Hadoop 开发技术，介绍大数据的书刊纷纷出现，但是每本大数据书籍的讲解的角度、深度和广度都有所不同，导致读者不能较为全面地认识大数据的理论方法与应用技术。因此，需要从读者角度出发，为读者系统地构建大数据的微观理论方法、中观技术实践、宏观项目实施与行业应用于一体的大数据思维体系。

由樊重俊教授编著的《大数据分析与应用》从大数据分析方法与技术应用的角度切入，建立大数据业务价值与技术架构之间的映射关系，内容丰富，深入浅出，繁简适度，使大众读者对大数据的基本概念、系统架构及其发展应用有相对全面的、较深层次的认识，能够系统地了解大数据的思维体系。这种结构化、系统化的思想贯穿全书，成为本书的一大特色。因此，本书不仅面向大众读者，对大数据相关的管理人员和技术人员都有一定的参考与帮助。

全书从大数据由来、大数据挖掘、大数据应用、大数据技术、大数据安全等不同的角度为读者展示了一个较为全面的、完整的大数据。分析了行业共性业务需求和个性业务需求，并且详细阐述了满足这些业务需求的大数据的技术，探讨了大数据下的商业智能技术和现有技术架构的整合，介绍大数据处理、存储的方法与技术，研究了大数据挖掘方法及实践案例，介绍了大数据可视化工具。

大数据在一些互联网公司有了很好的应用，各行业都在积极探索大数据的分析技术及其应用，本书列举相关实例，给出大数据应用的流程和方法论，强调了大数据对商业社会的巨大的变革，从侧面也反映出，即使在技术上和资金上都面临了很大的难题，但是大数据的分析与应用已经创造出重大的社会效益和经济效益，并将有更大的开拓空间。

于新南威尔士大学

2016 年 3 月

前　言

在“互联网+”的大环境下，大数据时代已经来临，信息的爆炸式增长使得大数据的分析方法及应用技术成为当务之急，未来国家治理、商业决策甚至个人生活都离不开大数据的分析与应用。大数据的特征不仅仅是规模大，而且具有多样性、复杂性、分布性、关联性等数据特征，这给传统的数据分析方法与应用技术带来了挑战。因此，必须加入新的数据思维，融合传统的数据分析与挖掘方法进行深化，以适应大数据的分析与管理。本书主要研究大数据分析与应用，在大量理论研究与实践的基础上，详细介绍了大数据分析方法，并结合具体应用领域对大数据分析方法进行具体阐述。本书强调理论联系实践，不仅仅介绍了基础的数据分析方法，并且结合各个行业的特征进行详细的实践介绍。全书共有19章，包括以下内容：

第1章主要介绍了大数据的基本概念与发展现状。第2章主要归纳了各行业中大数据的个性化应用，同时梳理了大数据应用的处理流程与企业中大数据应用的共性需求。第3章主要研究了大数据下的商业智能新概念与应用领域，同时介绍了大数据时代下新型的Hadoop与MPP结合的新架构、云平台以及大数据一体机。第4章主要对数据抽取和清洗原理与方法进行了详细的介绍，同时介绍了数据抽取和清洗的ETL工具。第5章主要介绍了大数据存储技术，包括大数据存储面临的挑战、数据存储方式、非关系型数据库、常见非关系型系统、分布式文件系统等。第6章主要介绍了云计算基本概念以及云服务的类型，同时分析了云计算技术以及云计算和大数据的异同。第7章主要介绍了数据挖掘基本概念以及新型的大数据挖掘分析技术，同时具体介绍了文本、语音、图像、空间以及Web数据挖掘的方法。第8章主要介绍了复杂数据的分析与建模过程，包括决策树、神经网络和隐马尔科夫模型等方法。第9章主要介绍了三类适合于大数据特征的预测方法，包括回归分析、时间序列分析以及深度学习方法。第10章主要从基本分词技术、索引以及文本信息检索模型等三个方面对信息挖掘作了介绍，同时对分词的方法以及评判标准、索引构建和更新、文本信息检索这四种基本模型展开了研究。第11章通过四个实例向读者介绍了如何利用MapReduce程序解决实际问题。第12章主要介绍了大数据时代下电子商务发展的新特点、新应用以及新态势。第13章重点介绍了移动互联网上的大数据及其分析等相关概念，同时介绍了大数据挖掘在移动互联网中的实际应用和未来发展情况。第14章主要介绍了社交网络中大数据及其分析的基本概念，同时重点阐述了以网络分析为主和以自然语言处理和情感分析为主的两类社交网络中大数据的相关分析技术与方法。第15章主要阐述了大数据在物流方面的应用原理以及具体应用。第16章介绍了大数据可视化的相关概念及部分大数据可视化工具。第17章主要阐述大数据实施项目的具体方法及常见的应用案例，同时详细介绍了两个大数据应用项目的实施过程。第18章主要指出大数据时代的信息安全面临的挑战及其特征并总结出了其应对策略，同时分析了大数据引起的个人隐私问题。第19章主要对IBM、Oracle、SAS和SAP等在内的部分主流厂商的大数据解决方案进行了介绍。

全书由樊重俊、刘臣、霍良安、杨云鹏、王雅琼、金阳统稿，并对各个章节的内容进行了调整、修改与补充。其中，第1章由王雅琼、樊重俊、霍良安撰写；第2章由王雅琼、樊重俊、刘臣撰写；第3章由金阳、

此为试读，需要完整PDF请访问：www.ertongbook.com

樊重俊、霍良安撰写；第 4 章由郭晓猛、刘臣、樊重俊撰写；第 5 章由郭晓猛、刘臣、杨云鹏、樊重俊撰写；第 6 章由何蒙蒙、樊重俊撰写；第 7 章由杨飞、樊重俊撰写；第 8 章由杨飞、樊重俊撰写；第 9 章由王雅琼、樊重俊撰写；第 10 章由刘臣、王育清撰写；第 11 章由刘臣、周立欣撰写；第 12 章由王宇莎、樊重俊、霍良安撰写；第 13 章由金阳、樊重俊、刘臣撰写；第 14 章由李佳婷、霍良安、樊重俊撰写；第 15 章由何蒙蒙、樊重俊撰写；第 16 章由霍良安、宋乃祥、蒋杰辉撰写；第 17 章由苏颖、樊重俊撰写；第 18 章由苏颖、樊重俊撰写；第 19 章由王宇莎、樊重俊撰写。参考文献由杨云鹏、王雅琼、樊重俊进行了汇总与整理。张惠珍、朱小栋、袁光辉、樊鸿飞对本书给出了一些修改建议。

新南威尔士大学计算机科学及工程学院数据库研究实验室主任林学民教授认真审阅了全书，并对本书给出了一些修改建议。

中国管理科学与工程学会副理事长、中国系统工程学会常务理事、上海市系统工程学会副理事长兼学术委员会主任、上海交通大学产业组织与技术创新研究中心主任陈宏民教授；英国雷丁大学商务信息、系统与会计学院院长、信息科学研究中心主任刘科成教授；全国高等学校计算机教育研究会常务理事、中国计算机学会教育专业委员会常委、复旦大学计算机科学技术学院赵一鸣副院长；上海机场(集团)有限公司冉祥来博士；华东理工大学李英教授、同济大学王洪伟教授、上海大学熊励教授也对本书给予了一些有益的建议。上海理工大学管理学院常务副院长高岩教授、上海理工大学信息管理系主任马良教授对本书的编写与出版给予了很多关心与支持。在此一并感谢。

大数据分析方法众多，且随着云计算、机器学习、智能化算法研究的持续深化，大数据分析方法与技术应用仍处于不断发展变化中。本书在编写的过程中，力求寻找适用于大数据分析与应用的理论研究、方法研究和技术水平，不断验证新方法与新技术在大数据中的应用，并参考了一些专家学者已公开发表的研究成果，多数研究成果均标注参考文献，但由于时间或疏漏未标明的研究成果，还请谅解。本书由于篇幅、时间以及环境等条件限制，书中疏漏之处在所难免，殷切希望同行专家和读者批评指正。

本书的研究与编著获得上海市教育委员会科研创新重点项目(No. 14ZZ131)、国家自然科学基金(71303157，71401107)的支持；获得沪江基金资助(A14006)，获得上海市一流学科项目资助(项目编号：S1201YLXK)，本书的出版获得立信会计出版社与黄成良老师的大力支持，特此感谢！

樊重俊

于上海理工大学

2016 年 3 月

目　录

此为试读，需要完整PDF请访问：www.ertongbook.com

第 1 章

大数据概述

早在20世纪80年代，著名的未来学家阿尔文·托夫勒在《第三次浪潮》一书中，将大数据热情地赞颂为“第三次浪潮的华彩乐章”。随着互联网的应用越来越广泛、新型社交网络的出现和快速发展、云计算技术的发展以及新型移动设备的出现，数据量呈现了爆炸性增长的趋势，大数据时代已经到来。数据从简单的“数字”概念逐渐成为“数字、文本、图片、视频”等的统称，数据结构也从结构化数据向非结构化数据扩展，数据作为一种基础性资源，如何更好地分析和应用数据已经成为大数据时代普遍关注的话题。大数据的规模效应给数据存储、管理以及数据分析带来了极大的挑战，也促使了数据分析方法的变革。

本章追溯了大数据的产生来源与发展历程。在对现有大数据研究资料进行全面归纳和总结的基础上，首先介绍了大数据的基本概念；其次，全面阐述大数据在学术界、产业界和政府机构的研究现状和大数据的应用现状；最后归纳总结大数据时代所面临的机遇与挑战。

1 大数据的产生与发展

1.1 大数据的产生原因

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，数据的总量在不断地增加，增加的速度也在不断地加快。而要掌握大数据的概念，首要任务就是从动态上了解大数据的成因。大数据的成因，不仅是人类信息技术的进步，而且是信息技术领域不同时期多个进步交互作用的结果。从开始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被动、主动和自动三个阶段，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。大数据产生的原因主要来自四大方面，一是数据存储成本的降低与存储硬件体积的减小；二是企业思维模式的转变；三是生活的数字化驱动；四是社交网络的飞速发展[1]。

1.1.1 数据存储成本的降低

大数据产生的重要前提是数据存储成本的大幅降低、存储硬件的体积日益减小。1965年，英特尔(Intel)创始人之一戈登·摩尔(Gordon Moore)提出著名的摩尔定律，即：当价格固定时，大约每隔18～24个月，集成电路上的元器件的数目便会增加1倍，其性能也将提升1倍，也就是说，每隔18～24个月，一美元所能买到的电脑性能将翻1倍以上。

摩尔定律所阐述的趋势已经持续了超过了半个世纪。半个多世纪以来，计算机硬件的发展规律基本符合摩尔定律，硬件的处理速度、存储能力不断提升，与此同时，硬件的价格却在持续降低。其中，商用硬盘存储器每兆价格从1995年的6 000多美元下降到2010年的0.005美分。

另外，随着计算机硬件价格的降低，其体积也在迅速变小。2014年，英特尔公司发布了14纳米(一纳米等于十亿分之一米)的晶体管，这比21纳米的晶体管缩小了1/3，而且更便宜、更节能。英特尔的

发明使得大部分科学家相信摩尔定律可以延续到 2020 年。预计到 2020 年，1TB 硬盘的价格将下降到 3 美元。而一所普通大学的图书馆，其馆藏量一般在 1TB～2TB 之间。可以很形象地理解为：只需要花一杯咖啡的价格就能够把一个图书馆的全部信息拷进一个小硬盘。

由于存储器的价格下降速度飞快，人们才得以廉价保存海量的数据；由于存储器的体积越来越小，人们才可以便捷携带海量的数据。这都在一定程度上促进了大数据时代的到来。

1.1.2 企业思维模式的转变

大数据真正存在的意义是在于其大价值，正是由于企业意识到大数据的价值所在，企业的商业思维发生了巨大的转变。企业开始注重对于企业内外部数据的挖掘，在海量的数据中搜索出隐藏的规律和价值，从而为决策者提供更好地参考。大数据时代的到来，人类对于数据的搜索和利用能力得到了巨大的提升，这种提升主要表现在企业大数据的挖掘上。数据挖掘一般是指，通过一定的算法从海量数据中分析出隐藏在数据背后价值的过程。

近几十年来，各大企业不断利用数据挖掘技术发掘出海量数据背后的商业价值。首先，最经典的是沃尔玛“啤酒和尿布”案例。沃尔玛通过精细的数据分析发现，年轻父亲在买尿布的同时喜欢买啤酒犒劳自己，于是沃尔玛采取捆绑销售策略，将“啤酒和尿布”这两种看似毫无关联的物品组合销售，大大提高了两者的销量。

其次是亚马逊的“预判发货”案例。2014 年 1 月，亚马逊宣布了一项新的专利“预判发货”，通过分析用户的行为数据，预测用户购买意向，在用户正式下单之前就寄出包裹，实现“先发货，后购买”。预判发货的核心就是通过算法预测并模拟整个发货过程，实现智能化发货。亚马逊有 1 亿多用户，这些用户的消费数据日积月累，可以说是海量数据。数据虽然多，但却不能直接表示出用户的收入、喜好等信息，所以整个预判过程都需要亚马逊依靠数据挖掘来完成。亚马逊“预判发货”的依据包括：用户之前的订单、搜索商品的痕迹、收藏夹、购物车内的商品，甚至还有用户的鼠标在某件商品上的停留时间。根据“预判发货”，用户从下单到收到快递的时间将被大幅缩短。为了降低预判发货的风险，亚马逊还会自动模糊填写用户的收货地址，让商品先接近潜在购买人群所在的区域，之后再在运输途中将确定的信息填写完整。同时，亚马逊还会向可能感兴趣的用户推荐一些“正在途中”的商品，从而提高预判发货的成功率。亚马逊称，这种预判式的发货比较适合畅销书或者一上市就吸引大批买家的商品。当然，预判发货不是完全准确的。当送货出现失误时，亚马逊会给予用户一定的折扣，或是将预测失误的已发货商品作为礼物赠送给用户，借此来提升公司的口碑。

另外，自 2012 年以来，全球的制造业正在变革，未来的工业制造将呈现数字化、智能化、定制化、互联化以及绿色化等特点。而这场变革的到来离不开 3D 打印机的应用，3D 打印将实现生产制造的数字化和定制化，将终结人类大规模工业生产的历史，引发商业组织和管理形态的重大变革。3D 打印是以数据包为基础的生产，只要这个数据包在打印机上运行，并且具备打印的原材料，生产就可以完成。也就是说，3D 在设计、生产、流通和消费各个环节上都离不开数据的驱动和协同。从这个角度上看，3D 打印为大数据时代贡献了一种新的数据种类：物理实体数据。随着可以打印的物品越来越多和 3D 打印机的价格不断下降，3D 打印机将快速走进普通家庭，数据包的数量将快速大量增加。

除了制造业的数字化生产将带来的数据激增以外，通用电气的“工业互联网”计划，也将带来数据的爆炸式增加。通用的工业互联网计划是指在其数万种产品上都安装传感器，通过网络将设备运行状态实时传至平台。让这些传感器实时监测生产过程还只是通用电气工业互联网计划的一部分，通用电气的目标是“让每件产品产生记忆”，未来产品在出厂前就被植入了传感器，记录它的生产过程，运送给顾客的过程，并且在顾客使用的每时每刻记录产品的运行情况，一旦出现故障，通用电气可以快速地整合生产记录，销售记录和运行记录这三种数据来进行分析。可以想象，全世界上百亿台带有微处理器的机器未来都装上传感器，日夜不停的自动产生数据，这将产生的数据量难以计量。通用电气公司估计，因为这种数据爆炸式增长，全世界数据中心将大量增加，数据中心的需求将每两年翻 1 倍。企业思维模式

的改变，对数据的创造、采集、挖掘、利用能力的日益增强，是大数据时代真正来临的最重要原因。

在信息时代，个人行为和社会状态的数据无处不在，这些数据是多源的、即时的、分散的、多形式的、碎片化的，同时又是海量的。通过采用自调适参数的算法，根据计算、挖掘次数的增多，不断调整自己算法的参数，使得对数据的挖掘和预测的结果更为精准，利用算法分析的结果，在这些海量的、零碎的数据中找到规律，从而发现大数据背后真正的价值动向，从而提供创造性、突破性的产品和服务。大数据时代下，企业对数据挖掘的利用还在不断的进步，有望在将来达到一个新的高度。

1.1.3 生活的数字化驱动

物联网的出现使得数据的产生从主动式产生变成自动式产生，而大数据真正产生的原因正是由于人们生活中自动式数据的产生。随着科学技术的进步发展，人们已经有能力制造极其微小的带有处理功能的传感器，感知式系统的广泛使用使得海量的数据自动生成。

近几年，越来越多的穿戴式设备进我们的生活，这些设备可以记录佩戴者的物理位置、热能消耗、体温、心跳、睡眠模式、步伐多少以及健身目标等数据。2014 年 2 月，日本东京大学的研究人员发明了一种比羽毛还轻的传感器，把它放在纸尿片内，纸尿片就会发出信号，看护就会知道并及时更换。谷歌眼镜被用于美国纽约市警察日常巡逻，以便他们快速记录事故现场的情形，并通过网络和同事快速分享。

智能家居通过物联网技术将与家居生活有关的各种设备（如音视频设备、安防系统、数字影院系统等）进行集成，构建了高效的住宅设施控制与家庭日程事务的管理系统。比如，你坐在办公室里，就可以调节家里冰箱的温度；在下班的路上就可以控制电饭煲的开关，并关上窗户，打开空调。2015 年，随着“互联网＋”模式的推动，国内互联网企业和家电公司纷纷推出智能系统和产品。如 360 公司 2015 年 4 月推出了智能安全门锁，为用户提供手机开锁、人脸开锁以及远程监控等功能。

在智能交通方面，智能导航服务利用出租车 GPS 的历史轨迹的分析结果为出行者设计个性化的路线，有效缓解交通拥堵问题；UPS 利用传感器等设备帮助调度中心监督并优化行车路线，根据过去积累的大数据制定最佳行车路线。2012 年 8 月，谷歌宣布无人驾驶汽车已经完成了 50 多万公里的安全行车测试，其本质就是把驾驶的任务“外包”给智能算法，对于无人驾驶汽车而言，最重要的组成部分就是全身上下的激光雷达、摄像头、红外相机、GPS（全球定位系统）和一系列传感器等感应设备，正是这些设备，无人驾驶汽车才能不断地收集路面的情况、汽车的地理位置、前后车辆精确的相对距离等数据。

穿戴式设备、智能家居以及智能交通的案例，都是以数据为载体而存在的。其内置的传感器最重要的任务就是大数据的采集。传感器等微小计算设备实现了无处不在的数据自动采集，这也意味着人们数据收集能力的提高，为大数据的产生提供了技术上的支持。

1.1.4 社交网络的飞速发展

自 2004 年起，以脸谱网（Facebook）、推特（Twitter）为代表的社交媒体相继问世，这拉开了一个互联网的崭新时代——Web2.0 时代。进入 Web2.0 时代之后，互联网开始成为人们实时互动、交流协同的载体。而真正的数据爆发就产生于 Web2.0 时代，Web2.0 的最重要标志就是用户原创内容。网络数据近几年一直呈现持续增长的势头，主要有两个方面的原因：一是首先是以博客、微博、微信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；二是以智能手机、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产生网络数据的途径更为便捷。

由于社交媒体的出现，全世界的网民都开始成为数据的生产者，每个网民犹如一个信息系统、一个传感器，不断地制造数据，这引发了人类历史上迄今为止最庞大的数据爆炸。比如，Facebook 用户每天共享的东西超过 40 亿，Twitter 每天处理的推特数量超过 3.4 亿；而每分钟 Tumblr 博客作者会发布 2.7万个新帖子，Instagram 用户会共享 3 600 张新照片。

除了数据总量极速增加，社交媒体还使数据的类型变得多元化：微博、微信中的信息大小、格式完全不一样，有文字、图片、音频、视频等。因为没有统一的结构，在社交媒体上产生的数据，也被称为非结构

化数据。这部分数据的处理，远远比结构化数据要困难的多。在这种前所未有的数据生产速度下，社交媒体的出现虽然才 10 年，但目前全世界大约超过 75%的数据都是非结构化数据。就目前来看，社交媒体的出现，是大数据爆发的直接原因。

1.2 大数据的发展历程

1.2.1 大数据应用的发展

正是由于大数据的广泛存在，才使得大数据问题的解决很具挑战性。而它的广泛应用，则促使越来越多的人开始关注和研究大数据问题。以下列举了若干个大数据发展中具有代表性的大事件[2]。

2005 年，Hadoop 项目诞生。Hadoop 原本来自于谷歌一款名为 MapReduce 的编程模型包，最初只与网页索引有关，被 Apache 软件基金会引入并成为分布式系统基础架构。Hadoop 可以帮助用户在不了解分布式底层细节的情况下，开发分布式程序，充分利用集群的威力进行高速运算和存储，从而以一种可靠、高效、可伸缩的方式进行数据处理。Hadoop 的框架最核心的设计就是 HDFS 和 MapReduce，HDFS 为海量的数据提供了存储，则 MapReduce 为海量的数据提供了计算。

2008 年末，“大数据”得到部分美国知名计算机科学研究人员的认可，业界组织“计算社区联盟”(ComputingCommunityConsortium)发表了一份有影响力的白皮书《大数据计算：在商务、科学和社会领域创建革命性突破》。这份白皮书指出大数据真正重要的是新用途和新见解，而非数据本身，这在一定程度上改变了人们固有的思维方式。计算社区联盟是最早提出大数据概念的机构。

2009 年中，美国政府通过启动 Data. gov 网站的方式向公众提供各种各样的政府数据。该网站的超过 4. 45 万量数据集被用于保证一些网站和智能手机应用程序来跟踪信息，包括航班信息，产品召回信息和特定区域内失业率信息等，这一行动激发了从肯尼亚到英国范围内的政府们相继推出类似举措。

2010 年 2 月，肯尼斯·库克尔在《经济学人》上发表了长达 14 页的大数据专题报告《数据，无所不在的数据》。库克尔在报告中提到：“世界上有着无法想象的巨量数字信息，并以极快的速度增长。从经济界到科学界，从政府部门到艺术领域，很多方面都已经感受到了这种巨量信息的影响。”科学家和计算机工程师已经为这个现象创造了一个新词汇：“大数据”。库克尔也因此成为最早洞见大数据时代趋势的数据科学家之一。

2011 年 2 月，IBM 的沃森超级计算机每秒可扫描并分析 4TB(约 2 亿页文字量)的数据量，并在美国著名智力竞赛电视节目《危险边缘》“Jeopardy”上击败两名人类选手而夺冠。后来纽约时报认为这一刻为一个“大数据计算的胜利”。

2011 年 5 月，全球知名咨询公司麦肯锡(McKinsey&Company)全球研究院(MGI)发布了一份报告——《大数据：创新、竞争和生产力的下一个新领域》，这是专业机构第一次全方位地介绍和展望大数据。报告指出，大数据已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。报告还提到，“大数据”源于数据生产和收集能力和速度的大幅提升——由于越来越多的人、设备和传感器通过数字网络连接起来，产生、传送、分享和访问数据的能力也得到彻底变革。

2011 年 12 月，工信部发布物联网十二五规划，提出将信息处理技术作为 4 项关键技术创新工程之一提出来，其中包括了海量数据存储、数据挖掘、图像视频智能分析，这都是大数据的重要组成部分。

2012 年 1 月份，瑞士达沃斯召开的世界经济论坛上，大数据是主题之一，会上发布的报告《大数据，大影响》(Big Data, Big Impact)宣称，数据已经成为一种新的经济资产类别，就像货币或黄金一样。

2012 年 3 月，美国奥巴马政府在白宫网站发布了《大数据研究和发展倡议》，这一倡议标志着大数据已经成为重要的时代特征。2012 年 3 月 22 日，奥巴马政府宣布 2 亿美元投资大数据领域，是大数据技术从商业行为上升到国家科技战略的分水岭，在次日的电话会议中，政府将数据定义成“未来的新石油”，大数据技术领域的竞争，事关国家安全和未来，并表示，国家层面的竞争力将部分体现为一国拥有

数据的规模、活性以及解释、运用的能力；国家数字主权体现对数据的占有和控制。数字主权将是继边防、海防、空防之后，另一个大国博弈的空间。

2012年4月，美国软件公司Splunk于19日在纳斯达克成功上市，成为第一家上市的大数据处理公司。鉴于美国经济持续低迷、股市持续震荡的大背景，Splunk首日的突出交易表现尤其令人们印象深刻，首日即暴涨了一倍多。Splunk是一家领先的提供大数据监测和分析服务的软件提供商，成立于2003年。Splunk成功上市促进了资本市场对大数据的关注，同时也促使IT厂商加快大数据布局。

2012年7月，联合国在纽约发布了一份关于大数据政务的白皮书，总结了各国政府如何利用大数据更好地服务和保护人民。这份白皮书举例说明在一个数据生态系统中，个人、公共部门和私人部门各自的角色、动机和需求。例如，为满足对价格关注和更好服务的需求，个人提供数据和众包信息，并对隐私和退出权力提出需求；公共部门出于改善服务，提升效益的目的，提供了诸如统计数据、设备信息、健康指标、及税务和消费信息等，并对隐私和退出权力提出需求。白皮书还指出，人们如今可以使用的丰富的数据资源，包括旧数据和新数据，来对社会人口进行前所未有的实时分析。

2014年4月，世界经济论坛以“大数据的回报与风险”主题发布了《全球信息技术报告(第13版)》。报告认为，在未来几年中针对各种信息通信技术的政策甚至会显得更加重要。全球大数据产业的日趋活跃，技术演进和应用创新的加速发展，使各国政府逐渐认识到大数据在推动经济发展、改善公共服务，增进人民福祉，乃至保障国家安全方面的重大意义。

2014年5月，美国白宫发布了2014年全球“大数据”白皮书的研究报告《大数据：抓住机遇、守护价值》。报告鼓励使用数据以推动社会进步，特别是在市场与现有的机构并未以其他方式来支持这种进步的领域；同时，也需要相应的框架、结构与研究，来帮助保护美国人对于保护个人隐私、确保公平或是防止歧视的坚定信仰。

大数据是一场革命，将改变我们的生活、工作和思维方式。庞大的新数据来源所带来的量化转变已经引起了学术界、企业界和政界的高度重视。

1.2.2 大数据技术的发展

大数据技术是一种新一代技术和构架，它以成本较低、以快速的采集、处理和分析技术，从各种超大规模的数据中提取价值。大数据技术不断涌现和发展，让我们处理海量数据更加容易、更加便宜和迅速，成为利用数据的好助手，甚至可以改变许多行业的商业模式，大数据技术的发展可以分为六大方向：

1) *在大数据采集与预处理方向*

这方向最常见的问题是数据的多源和多样性，导致数据的质量存在差异，严重影响到数据的可用性。针对这些问题，目前很多公司已经推出了多种数据清洗和质量控制工具(如IBM的Data Stage)。

2) *在大数据存储与管理方向*

这方向最常见的挑战是存储规模大，存储管理复杂，需要兼顾结构化、非结构化和半结构化的数据。分布式文件系统和分布式数据库相关技术的发展正在有效地解决这些方面的问题。在大数据存储和管理方向，尤其值得我们关注的是大数据索引和查询技术、实时及流式大数据存储与处理的发展。

3) *大数据计算模式方向*

由于大数据处理多样性的需求，目前出现了多种典型的计算模式，包括大数据查询分析计算(如Hive)、批处理计算(如Hadoop MapReduce)、流式计算(如Storm)、迭代计算(如HaLoop)、图计算(如Pregel)和内存计算(如Hana)，而这些计算模式的混合计算模式将成为满足多样性大数据处理和应用需求的有效手段。

4) *大数据分析与挖掘方向*

在数据量迅速膨胀的同时，还要进行深度的数据深度分析和挖掘，并且对自动化分析要求越来越高，越来越多的大数据数据分析工具和产品应运而生，如用于大数据挖掘的R Hadoop版、基于MapRe-

duce 开发的数据挖掘算法等。

5）大数据可视化分析方向

通过可视化方式来帮助人们探索和解释复杂的数据，有利于决策者挖掘数据的商业价值，进而有助于大数据的发展。很多公司也在开展相应的研究，试图把可视化引入其不同的数据分析和展示的产品中，各种可能相关的商品也将会不断出现。可视化工具 Tabealu 的成功上市反映了大数据可视化的需求。

6）大数据安全方向

当我们在用大数据分析和数据挖掘获取商业价值的时候，黑客很可能在向我们攻击，收集有用的信息。因此，大数据的安全一直是企业和学术界非常关注的研究方向。通过文件访问控制来限制呈现对数据的操作、基础设备加密、匿名化保护技术和加密保护等技术正在最大程度的保护数据安全。

2 大数据的概念

本节在对现有的大数据研究资料进行全面的归纳和总结的基础上，阐述了大数据的定义、特征、数据类型和重要作用。

2.1 大数据的定义

大数据本身是一个比较抽象的概念，单从字面来看，它表示数据规模的庞大。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。针对大数据，目前存在多种不同的理解和定义。

麦肯锡在其报告《Big data：The nextfrontier for innovation，competition and productivity》中给出的大数据定义是：大数据指的是大小超出常规的数据库工具获取、存储、管理和分析能力的数据集。但它同时强调，并不是说一定要超过特定 TB 值的数据集才能算是大数据。

维基百科对“大数据”的解读是：“大数据”（Big Data），或称巨量数据、海量数据、大资料，指的是所涉及的数据量规模巨大到无法通过人工在合理时间内达到截取、管理、处理、并整理成为人类所能解读的信息。

百度百科对“大数据”的定义为：“大数据”（Big Data），或称巨量资料，指的是所涉及的资料量规模巨大到无法透过目前主流软件工具，在合理时间内达到撷取、管理、处理、并整理成为帮助企业经营决策更积极目的的资讯。

研究机构 Gartner 认为，“大数据”是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。从数据的类别上看，“大数据”指的是无法使用传统流程或工具处理或分析的信息。它定义了那些超出正常处理范围和大小、迫使用户采用非传统处理方法的数据集。

按照美国国家标准与技术研究院（National Institute of Standards and Technology，NIST）发布的研究报告的定义，大数据是用来描述在我们网络的、数字的、遍布传感器的、信息驱动的世界中呈现出的数据泛滥的常用词语。大量数据资源为解决以前不可能解决的问题带来了可能性。

大数据是一个宽泛的概念，每个人的见解都不一样。笔者在综合各家观点的基础上，给出了自己的定义：“大数据”是在体量和类别特别大的杂乱数据集中，深度挖掘分析取得有价值信息的能力。大数据不仅仅在于数量的大，“大”只不过是信息技术不断发展所产生的海量数据的表象而已。我们更加关注“数据”的深度分析和应用，对于数据有价值的深度挖掘分析和在新形势下的数据应用是我们需要探讨的重点。

此为试读，需要完整PDF请访问：www.ertongbook.com