JISHU YU YINGYONG

大数据技术系列 DASHUJU JISHU XILIE

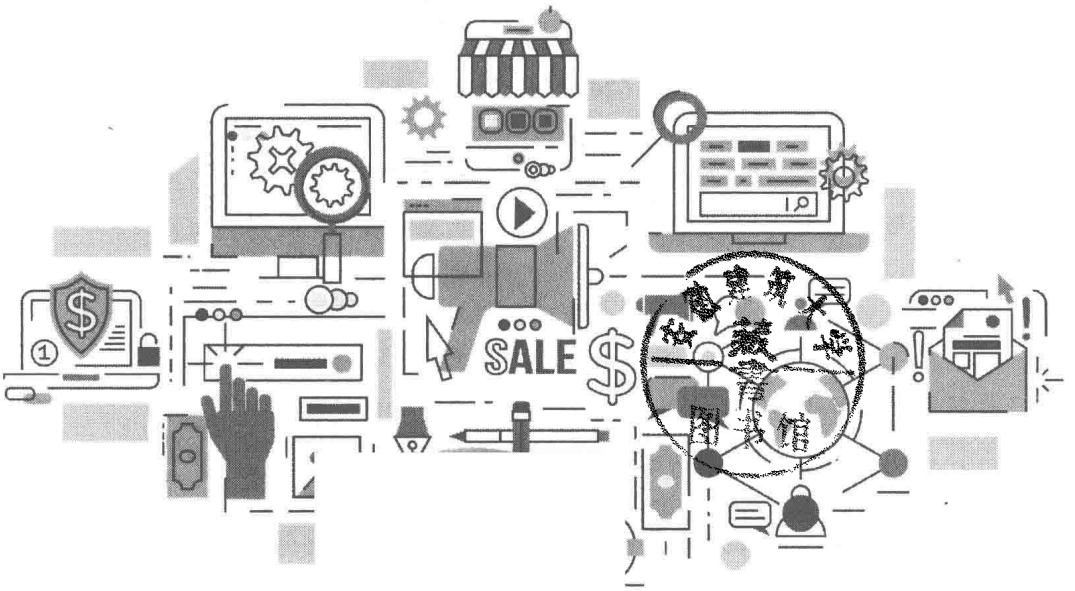
大数据技术与应用

主编 姚树春 周连生 张强 侯勇

西南交通大学出版社

大数据技术与应用

主编　姚树春　周连生　张强　侯勇



西南交通大学出版社
·成都·

内容简介

随着信息时代的信息爆炸式增长，大数据已经无处不在。数量、种类庞大和速度加快的数据浪潮是政府、企业面临的一个全新课题。本书首先讲述了什么是大数据、常见大数据源及其应用价值分析、大数据的商业应用；然后从具体实用的角度介绍了大数据应用的相关技术和工具、主要数据挖掘工具及平台；最后讲解了如何组建优秀的大数据分析专家团队，以及大数据的应用，包括几个经典的大数据应用案例、政府工作中的大数据应用、目前互联网中的大数据商机、大数据与未来之路。本书集知识性、实用性、可读性于一体，可以帮助读者了解大数据的概念、特点、重要性、价值；了解大数据处理的基础理念和常见工具；熟悉大数据的处理流程、方法和技术；结合实际案例构建大数据应用的战略蓝图、管理流程和实施策略。

图书在版编目（CIP）数据

大数据技术与应用 / 姚树春等主编. —成都：西南交通大学出版社，2018.6
（大数据技术系列）
ISBN 978-7-5643-6275-1

Ⅰ. ①大… Ⅱ. ①姚… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2018）第 147624 号

大数据技术系列
大数据技术与应用

主编 / 姚树春　周连生　张　强　侯　勇

责任编辑 / 黄庆斌
特邀编辑 / 刘姗姗
封面设计 / 何东琳设计工作室

西南交通大学出版社出版发行
（四川省成都市二环路北一段 111 号西南交通大学创新大厦 21 楼　610031）
发行部电话：028-87600564　028-87600533
网址：http://www.xnjdcbs.com
印刷：成都中永印务有限责任公司

成品尺寸　185 mm × 260 mm
印张　13.5　字数　321 千
版次　2018 年 6 月第 1 版　印次　2018 年 6 月第 1 次

书号　ISBN 978-7-5643-6275-1
定价　39.80 元

课件咨询电话：028-87600533
图书如有印装质量问题　本社负责退换

序　言

十年前，“数据”对于每个普通人来说，还是一个非常专业甚至陌生的词汇。随着科学技术的飞速发展，今天，“数据”已经深入到大家生活的方方面面，线上交流、网上购物、快递外卖、旅行记录等等。每个人，每天都会源源不断地产生大量的数据。据 IDC《数字宇宙》(Digital Universe) 的研究报告表明，2006 年，全世界产生数据量为 0.18 ZB。到 2020 年，全球新建和复制的信息量将超过 40 ZB，数据量呈现数百倍数量级的增长。

大数据 (Big Data)，或称巨量资料，是以容量大、类型多、存取速度快、价值密度低为主要特征的数据集合。大数据正快速发展为对数量巨大、来源分散、格式多样的数据进行采集、存储和关联分析，从中发现新知识、创造新价值、提升新能力的新一代信息技术和服务业态。

信息技术与经济社会的交汇融合引发了数据迅猛增长，数据已成为国家基础性战略资源，大数据正日益对全球生产、流通、分配、消费活动以及经济运行机制、社会生活方式和国家治理能力产生重要影响。

数据量的飞速增长也带来了大数据技术和服务市场的繁荣发展。大数据解决方案不断成熟，各领域大数据应用全面展开，为大数据发展带来了强劲动力。我国大数据仍处于起步发展阶段，各地发展大数据积极性较高，行业应用得到快速推广，市场规模增速明显。数据显示，2015 年我国大数据市场规模达到 115.9 亿元，增速达 53.1%。预计到 2021 年，我国大数据市场规模将突破 350 亿元。在大数据时代，各行各业对数据的分类检索和储存智能化要求越来越高，大数据对人们来说意味着宝藏，大数据技术就是打开这座宝藏的一把金钥匙。

由姚树春、周连生、张强、侯勇主编的《大数据技术与应用》，是基于大数据团队几年来的教学实践和科学研究成果，通过精心组织内容，并多次修改，用最新理论和数据，深入浅出地介绍了什么是大数据、大数据的价值、大数据的相关技术、大数据的案例应用等。本书为大家打开一扇了解“大数据”的窗户，无论是专业人士还是普通大众，都值得一读。

在《大数据技术与应用》即将出版之际，作者邀请我写该书的《序言》。

本书作者主要来自江苏汇誉通数据科技有限公司的周连生总经理、张强技术总监、苏州工业园区服务外包职业学院大数据教学团队的姚树春老师、蚌埠学院的侯勇老师。江苏汇誉通数据科技有限公司在给中科院相关部门进行大数据培训过程中，积累了大量的大数据学习资源，为编写本书提供了大量的资料。

当前，云计算、大数据、人工智能、区块链技术等层出不穷，技术革命带动产业变革，万物互联、人机交互、天地一体的网络空间正在逐步形成。2015 年 8 月国务院印发了《促进大数据发展行动纲要》。2017 年初，工业和信息化部印发了《国家大数据产业发展规划(2016—2020 年)》。可以说，数据科学的春天正朝我们姗姗走来，让我们一起张开双臂，学习新的技术，热烈地拥抱这个春天吧！

冯瑞教授

2018 年 6 月

目　录

第1篇　理念篇

第 2 篇 技术篇

第3篇 应用篇

第 1 篇　理念篇

1　信息时代背景及大数据基本介绍

1.1　信息时代的主要数据源

随着信息技术的迅速发展，管理信息系统、互联网、物联网、移动终端等新技术与设备正在不断改变现代企业的环境。互联网和其他全球性媒介已经初步消除了国界对信息的隔离。互联网上的公共网页和全球共享数据几乎对所有组织和个人都是公开的，大多可以被自由访问、下载。互联网网页资源、博客和论坛、企业公开报表、社会各组织的公共数据库、各类媒体资源均涉及有关政治、经济、管理、生活等各领域广泛的海量信息。大数据的主要来源分为如下几个方面。

1.1.1　互联网

互联网的出现，把每个人的计算机连接起来，改变了人们的生活，成为大家获取、分享各类数据的首要渠道。互联网成为大规模接近各类人群生活的工具和平台，人们在互联网上的一言一行都被忠实地记录下来。就像古代皇帝身边总有一位兢兢业业的史官，随身携带纸笔，记下皇帝的起居作息、金口玉言一样，互联网就像每个人的“史官”，它从不知疲倦，对事不分大小，都悉心而精准地记录着一切。互联网日志、博客、微博、论坛中就像无数的“史官”如实记录着大家的数字化生活。

1.1.2　社交网络

社交网络把真实的人际关系完美的映射到互联网空间，并借助互联网的特性而大大升华。广义上看，社交网络使得互联网甚至具备某些人类的特质，譬如“情绪”——人们分享各自的喜怒哀乐，并相互传染传播。社交网络为大数据带来一类最具活力的数据类型——人们的喜好和偏爱。

大型的社交网络平台事实上构成了以“个人”为枢纽的、不同的数据的集合。借助“分享”按钮，人们在不同网站上的购物信息、浏览的网页都可以“分享”到社交网络上。就像人们在雪地上留下脚印，社交网络把网民在不同网站上留下的“脚印”链接起来，形成完整的行为轨迹和“偏好”链。更重要的是，社交网络大数据中储存着网民的关系链，及其喜好和偏爱的传播路径，这些都具有极大的开发价值。

1.1.3 云计算

云计算改变了数据的存储和访问方式。在云计算出现之前，数据大多分散保存在每家企业的服务器中，或每个人的计算机中。云计算，尤其是公用云计算，把所有的数据集中存储到“数据中心”，也即所谓的“云端”，用户通过浏览器或者专用应用程序来访问。

一些大型的网站，通过提供基于“云”的服务，积累了大量的数据，成为事实上的“数据中心”。“数据”是最为核心的资产。云服务商往往不惜花费高昂的费用来保管这些数据。谷歌公司甚至购买了单独的水力发电站，为其庞大的数据中心提供充足的电力。根据一些公开资料显示，谷歌在全球分布着 36 个数据中心。谷歌公司数据中心一景如图 1.1 所示。

图 1.1 谷歌公司数据中心

这几年兴起的建设云计算基地的风潮，客观上为“大数据”的诞生提供了必备的储存空间和访问渠道。各大银行、电信运营商、大型互联网公司、政府各个部委都拥有各自的“数据中心”。银行、电信、互联网公司绝大部分已经实现了全国级的数据集中工作。云计算为大数据提供了存储空间和访问渠道。

1.1.4 物联网

物联网就是“物物相连的互联网”。由此可见，第一，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；第二，其用户端延伸和扩展到了任何物品与物品之间，都能进行信息交换和通信。物联网通过智能感知、识别技术与普适计算、泛在网络的融合应用，被称为继计算机、互联网之后世界信息产业发展的第三次浪潮。物联网是传感器技术进步的产物。传感器可以监测温度、压强、风力、桥梁、矿井的安全，还可以监测飞机、汽车的行驶状态。现在常用的智能手机，就包括重力感应器、加速度感应器、距离感应器、光线感应器、陀螺仪、电子罗盘、摄像头等各类传感器。这些不同类型的传感器，无时无刻不在产生大量的数据。这些大量的数据被持续地收集起来，成为大数据的重要来源之一。

1.1.5 智能终端

智能终端简称移动智能终端，由英文 Smart Phone 及 Smart Device 于 2000 年之后翻译而来。智能终端包括智能手机、便携式计算机、PDA、平板电脑等。智能终端的普及给大数据带来了丰富、鲜活的数据。2017 年，微信团队在微信公开课上发布的《2016 微信数据报告》显示：2016 年 9 月平均日登录用户 7.68 亿，较前一年增长 35%；典型用户日人均发送消息次数 74 次，比前一年增长 67%；典型用户月人均成功通话 8 次，月人均通话 65 分钟；微信音视频通话，日成功通话 1 亿次，较上一年增长了 180%；微信朋友圈中，典型用户发表的原创内容占 65%；典型用户月发送红包 28 次。微信连接每一个群体，微信中的应用越来越多，信息量也越来越大。

随着信息基础设施持续完善，网络带宽的持续增加，存储设备性价比不断提升，这些为大数据的存储和传播提供了物质基础；云计算为大量数据的集中管理和分布式访问提供了必要的场所和分享的渠道；物联网与智能终端持续不断的产生大量数据，其数据类型丰富，内容鲜活，是大数据重要的来源。现在，大数据正在更深层次地影响国家、企业的发展以及人们的生活。

1.1.6 信息时代数据增长的特点

1. 数据量呈现指数级增长

各项研究成果都表明，未来数年全球数据总量将会呈现指数级增长。据 IDC（互联网数据中心）公布的调查数据显示，未来全球数据增长率将维持 50%左右，到 2020 年，全球数据总量将达到 44ZB（十万亿亿字节，通常用 B、KB、MB、GB、TB、PB、EB、ZB、YB、BB 来表示数据量，它们之间的关系是 210 倍），中国将达到 8.6 ZB，占全球的 21%。中国信息产业研究院的数据显示，2015 年，我国大数据市场规模约为 116 亿元，同比增长 38%。

预计未来几年，随着应用效果的逐步显现，我国大数据市场规模还将维持40%左右的高速增长。

2. 不同行业的数据强度和内容差别很大

各个行业都呈现数据快速增长的现象，但不同行业数据存储量有所不同，数据产生和存储的类型也有所区别。证券、投资服务以及银行等金融服务领域拥有最高的平均数字化数据存储量，通信和媒体公司、公共事业公司以及政府等企业和组织也有规模显著的数字化数据存储。这些数据强度高的行业更加具有通过数据来创造价值的潜力。

3. 新技术应用将持续推动数据增长

在各部门和地区之间，企业正在加快收集数据的步伐，这推动了传统的事务数据库的增长；医疗卫生等面向消费者的行业中，多媒体的广泛使用刺激了数据的持续扩张；社交媒体的广泛普及以及物联网应用的不断创新也进一步推动了数据不断增长……这些相互交叉的动力刺激了数据的增长，并将继续推动数据池的迅速扩张。

1965年，戈登·摩尔（Gordon Moore）作为英特尔公司的创始人之一，发现“芯片上可容纳的晶体管数目，每隔18个月左右便会增加一倍，性能也将提升一倍。”后来人们发现这不仅适用于对存储器芯片的描述，也说明了计算能力和磁盘存储容量的发展。于是，摩尔定律成为许多工业对于性能预测的基础。

1977年，世界上第一条光纤通信系统在美国芝加哥投入商用，速率为45 Mb/s，自此，拉开了信息传输能力大幅跃升的序幕。有人甚至将光纤传输带宽的增长规律称为超摩尔定律，认为带宽的增长速度比芯片性能提升的速度还要快。

事实上，计算机存储的价格从20世纪60年代1万美元1 MB，降到现在的1美分1 GB的水平，其价差高达亿倍；在线实时观看高清电影，在几年前还是难以想象的，现在却变得习以为常了；网络的接入方式也从有线连接向高速无线连接的方式转变。毫无疑问，网络带宽和大规模存储技术的高速持续发展，为大数据时代提供了廉价的存储和传输服务。

4. 数据中隐藏着巨大的宝藏

当前大数据规模以及其存储容量正在迅速增长，大数据已经渗透到各个行业和业务职能领域，成为重要的信息资源。企业决策所需的信息、知识已经分散在各类大数据资源中，如何有效利用这类信息和知识成为组织成功的关键。“以天下之目视者，则无不见；以天下之耳听者，则无不闻；以天下之心思虑者，则无不知”，今天的互联网中已经包含了“天下之目”“天下之耳”“天下之心”。在全球化的今天，信息获取广度、深度、及时性上都有很大提高，综合利用这些资源为组织提供管理决策服务，从数据中挖掘宝藏的能力将成为各种组织的核心竞争力之一。

1.2 大数据及其特点

1.2.1 大数据的概念

随着人跟人、人与机器、机器与机器在交易、沟通、通信中产生的数据量越来越大，人

类开始走进大数据时代。早在1980年，著名未来学家阿尔文·托夫勒便在《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”。麦肯锡（美国著名的咨询公司）在其报告《Big data：The next frontier 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。

2013年5月，第462次香山科学会议在北京香山饭店召开，本次会议的主题是“数据科学与大数据的科学原理与发展前景”。来自中国科学院大学管理学院、中国科学院虚拟经济与数据科学研究中心、复旦大学、美国伊犁诺大学芝加哥分校、中科院科技政策与管理科学研究所的专家主持了学术讨论会，与会专家和学者给出了“大数据（BIG DATA）”概念的一个科学性描述，即大数据是来源多样、类型多样、大而复杂、具有潜在价值，但难以在期望时间内处理和分析的数据集。通俗地讲，大数据是数字化生存时代的新型战略资源，是驱动创新的重要因素，正在改变人类的生产和生活方式。

1.2.2 大数据的主要来源

人与人交易、沟通产生数据。移动通信、社交网络每时每刻都在大量的产生数据，传统的商业领域、电子商务和金融交易也同样如此，机器与机器、智能设备与网络中产生的数据，其数量更为巨大。随着时间的增长，物联网的发展也产生了更多的数据。在美国，评估净利润前15个行业中，每一家公司当年所产生的数据都大过美国国会图书馆所有的数据。

美国互联网数据中心指出：互联网上的数据每年将增长50%，每两年便将翻一番，目前世界上90%以上的数据是最近几年才产生的。此外，数据又并非单纯指人们在互联网上发布的信息，全世界的工业设备、汽车、电表上有着无数的数码传感器，随时测量和传递着有关位置、运动、震动、温度、湿度乃至空气中化学物质的变化，也产生了海量的数据信息。

伴随着多媒体、社会媒体以及物联网的发展，企业将收集更多的信息，从而带来数据呈现指数级的增长。全球可统计的数据存储量在2011年约为1.8ZB（1.8万亿GB），2013年达到4.4ZB，到2020年这一数值将增长到35ZB，2014年－2020年的预计年复合增长率达到84%。大数据已经成为当前人类最宝贵的财富。

1.2.3 大数据的特征

国际数据公司（IDC）从大数据的四个特征来对其进行定义：即海量的数据规模（Volume）、快速的数据流转（Velocity）、多样的数据类型（Variety）、巨大的数据价值（Value）。大数据的核心能力，是发现规律和预测未来。

我们认为，通过四个“V”，能够更好地把握大数据的特征。

1. 数据体量巨大（Volume）

截至目前，人类生产的所有印刷材料的数据量是200 PB（1 PB=210 TB），而历史上全人类说过所有话的数据量大约是5 EB（1 EB=210 PB）。当前，典型个人计算机硬盘的容量

为 TB 量级，而一些大企业的数据量已经接近 EB 量级。

2. 处理速度快（Velocity）

这是大数据区别于传统数据挖掘的最显著特征。根据 IDC 的“数字宇宙”的报告，预计到 2020 年，全球数据使用量将达到 35.2 ZB。在如此海量的数据面前，处理数据的效率就成为企业的生命。

3. 数据类型繁多（Variety）

这种类型的多样性也让数据被分为结构化数据和非结构化数据。相对于以往便于存储的以文本为主的结构化数据，非结构化数据越来越多，包括网络日志、音频、视频、图片、地理位置信息等，这些多类型的数据对数据的处理能力提出了更高要求。

4. 价值密度低（Value）

大数据具有巨大的商业价值，但不可否认的是，大数据价值密度的高低与数据总量的大小成反比。以视频为例，一部 1 小时的视频，在连续不间断的监控中，有用数据可能仅有一两秒。如何通过强大的机器算法更迅速地完成数据的价值“提纯”成为目前大数据背景下亟待解决的难题。

1.3 大数据的重要性及其价值

如今，数据已经成为可以与物质资产、人力资本相提并论的重要的生产要素。大数据的使用将成为未来提高竞争力、生产力、创新能力以及创造消费者价值的关键要素。

目前，大数据市场已经达到 700 亿美元规模并以每年 15%的速度增长，数据存储巨头 EMC 的 CEO Pat Gelsinger 透露，大数据处理目前的市场规模已达 700 亿美元并且正以每年 15%~20%的速度增长。几乎所有主要的大科技公司都对大数据感兴趣，对该领域的产品及服务进行了大量投入。其中包括 IBM、Oracle、EMC、HP、Dell、SGI、日立、Yahoo 等，而且这个列表还在继续加长。

近年来，IBM、甲骨文、EMC、SAP 等国际 IT 巨头掀起了“大数据”市场的收购热潮，共花费超过 15 亿美元用于收购相关数据管理和分析厂商，也使得“大数据（Big Data）”成为继“云计算”之后又一个在 IT 界炙手可热的名词，成为继传统 IT 之后下一个提高生产率的技术前沿。

对大数据的利用是成为企业提高核心竞争力并抢占市场先机的关键。在未来 3 到 5 年，我们将会看到那些真正理解大数据并能利用大数据进行挖掘分析的企业和不懂得大数据价值的企业之间的差距。真正能够利用好大数据并将其价值转化成生产力的企业必将形成有力的竞争优势，奠定行业领导者的地位。

在零售领域，对大数据的分析可以使零售商实时掌握市场动态并迅速做出应对。沃尔玛已经开始利用各个连锁店不断产生的海量销售数据，并结合天气数据、经济学、人口统计学进行分析，从而在特定的连锁店中选择合适的上架产品，并判定商品减价的时机。

在互联网领域，对大数据的分析可以为商家制定更加精准有效的营销策略提供决策支